NIH Public Access
Author Manuscript
AJR Am J Roentgenol. Author manuscript; available in PMC 2014 January 15.

Published in final edited form as:
AJR Am J Roentgenol. 2012 July ; 199(1): 224–235. doi:10.2214/AJR.11.7324.

Training the ACRIN 6666 Investigators and Effects of Feedback on Breast Ultrasound Interpretive Performance and Agreement in BI-RADS Ultrasound Feature Analysis

Wendie A. Berg[1,2], **Jeffrey D. Blume**[3,4], **Jean B. Cormack**[3], and **Ellen B. Mendelson**[5]
[1]American College of Radiology Imaging Network (ACRIN)

[3]Center for Statistical Sciences, Brown University, Providence, RI

[5]Feinberg School of Medicine, Northwestern University, Chicago, IL

Abstract

OBJECTIVE—Qualification tasks in mammography and breast ultrasound were developed for the American College of Radiology Imaging Network (ACRIN) 6666 Investigators. We sought to assess the effects of feedback on breast ultrasound interpretive performance and agreement in BI-RADS feature analysis among a subset of these experienced observers.

MATERIALS AND METHODS—After a 1-hour didactic session on *BI-RADS: Ultrasound*, an interpretive skills quiz set of 70 orthogonal sets of breast ultrasound images including 25 (36%) malignancies was presented to 100 experienced breast imaging observers. Thirty-five observers reviewed the quiz set twice: first without and then with immediate feedback of consensus feature analysis, management recommendations, and pathologic truth. Observer performance (sensitivity, specificity, area under the curve [AUC]) was calculated without feedback and with feedback. Kappas were determined for agreement on feature analysis and assessments.

RESULTS—For 35 observers without feedback, the mean sensitivity was 89% (range, 68–100%); specificity, 62% (range, 42–82%); and AUC, 82% (range, 73–89%). With feedback, the mean sensitivity was 93% (range, 80–100%; mean increase, 4%; range of increase, 0–12%; $p < 0.0001$), the mean specificity was 61% (range, 45–73%; mean decrease, 1%; range of change, −18% to 11%; $p = 0.19$), and the mean AUC was 84% (range, 78–90%; mean increase, 2%; range of change, −3% to 9%; $p < 0.0001$). Three breast imagers in the lowest quartile of initial performance showed the greatest improvement in sensitivity with no change or improvement in AUC. The kappa values for feature analysis did not change, but there was improved agreement about final assessments, with the kappa value increasing from 0.53 (SE, 0.02) without feedback to 0.59 (SE, 0.02) with feedback ($p < 0.0001$).

CONCLUSION—Most experienced breast imagers showed excellent breast ultrasound interpretive skills. Immediate feedback of consensus *BI-RADS: Ultrasound* features and histopathologic results improved performance in ultrasound interpretation across all experience variables.

Keywords

BI-RADS; breast ultrasound; feedback; observer agreement; observer performance; training

© American Roentgen Ray Society

[2]Present address: Department of Radiology, Magee-Womens Hospital, University of Pittsburgh School of Medicine, 300 Halket St, Pittsburgh, PA 15213. Address correspondence to W. A. Berg (wendieberg@gmail.com).
[4]Present address: Department of Biostatistics, Vanderbilt University, Nashville, TN.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Mammography remains the mainstay of breast cancer screening and is the only screening method proven to reduce breast cancer mortality [1]. Successful mammographic identification of early breast cancer requires that the cancer be visible, be detected, and be properly interpreted. In dense breast tissue, from 30% [2] to 70% [3] of cancers go undetected mammographically, with most of these obscured by overlying tissue.

To address reduced mammographic sensitivity in women with dense parenchyma, supplemental screening with ultrasound has been proposed. Results from more than 50,000 screening breast ultrasound examinations show a consistent rate of additional cancer detection of 2.7–4.6 per 1000 women screened [4–12]. More than 90% of cancers found only on screening ultrasound are invasive, with a mean size of 9–11 mm [4, 5, 13]; the cancers seen only sonographically are nearly all node-negative where reported [4, 5, 7, 8, 10, 11]. Such results prompted the multicenter American College of Radiology Imaging Network (ACRIN) protocol 6666 to evaluate supplemental screening breast ultrasound of women with dense breasts at elevated risk of breast cancer [4, 5]. As with mammography, sonographic recognition of early breast cancer requires both detection of the cancer as distinct from surrounding normal tissue and proper interpretation of the findings. As currently practiced, detecting a suspicious lesion on breast ultrasound requires real-time recognition of suspicious features by the operator, although various approaches to automated 3D breast sonography are currently undergoing validation [14, 15]. There is no computer-assisted detection for realtime breast ultrasound, but computer-aided diagnosis is in development [16]. Thus, interpretive skills are at least as important for successful performance of breast ultrasound as in mammography.

Concerns about the operator dependence of breast ultrasound and particularly of breast ultrasound interpretation prompted development of investigator qualification tasks that were required for the ACRIN 6666 protocol [17] before opening sites to participation. We sought to reduce sources of variability in breast ultrasound performance and interpretation and to develop generalizable techniques and criteria for interpretation. It was known from the work of Berg et al. [18] that training in the BI-RADS for mammography [19] improved both interpretive skills and agreement on mammographic feature analysis among radiologists. We hoped that training in *BI-RADS: Ultrasound* [20] would produce similar benefits. The purpose of this study was to assess the effects of feedback on breast ultrasound interpretive performance and agreement in BI-RADS feature analysis among experienced observers.

Materials and Methods

Investigator qualification tasks were approved by the ACRIN Institutional Review Board and the National Cancer Institute's National Cancer Therapy Experimental Protocols committees. Each potential investigator (i.e., observer) in this study agreed to participate and to have his or her results analyzed. By protocol, each of the observers stated that he or she met all the requirements of the Mammography Quality Standards Act for mammography-interpreting physicians and that he or she had a minimum experience in the previous 2 years scanning and interpreting at least 500 breast sonograms per year and interpreting at least 2500 mammograms per year. In addition to the interpretive skills tasks described herein, investigators also had to successfully scan and identify lesions in a breast ultrasound phantom [21].

Demographic variables were collected for the observers. Specifically, we collected information about the number of years practicing breast imaging, percentage of time spent in clinical breast imaging, who routinely performs breast imaging in their practice (technologist; attending radiologist [i.e., the observer in this protocol]; fellow; resident, then

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

attending radiologist; technologist, then attending radiologist; fellow, then attending radiologist), number of mammograms interpreted per week, and number of breast ultrasound examinations performed and interpreted per week. We also asked observers the indications for performing whole-breast ultrasound in their usual clinical practice (i.e., never, for newly diagnosed cancer in that breast, for most diagnostic ultrasound, for diagnostic and screening purposes).

In June 2003, we conducted a 1-hour didactic session in the *BI-RADS: Ultrasound* [20] before the interpretive skills task. Feature analysis was illustrated with examples, and BI-RADS final assessment categories were reviewed, together with their usual recommendations: category 1, negative; 2, benign; 3, probably benign; 4A, low suspicion of malignancy; 4B, intermediate suspicion of malignancy; 4C, moderate suspicion of malignancy; and 5, highly suggestive of malignancy. Observers were instructed that BI-RADS categories 1 and 2 implied routine follow-up, use of BI-RADS 3 implied a recommendation for 6-month follow-up ultrasound, and use of BI-RADS 4A or higher implied a recommendation for biopsy. Specific interpretive criteria were discussed and are detailed in the ACRIN 6666 protocol [17].

Case Set: Ultrasound

Two orthogonal B-mode ultrasound images of each of 70 lesions, including 25 (36%) malignancies, were prepared and embedded in a PowerPoint (Microsoft) presentation. No Doppler, elastographic, or mammographic images of these findings were supplied. All images had been acquired using a linear-array transducer with a maximum frequency of at least 12 MHz.

During development of the quiz, the ultrasound cases were first shown to three observers with 14, 21, and 25 years of experience in breast ultrasound, respectively. These “expert” observers were asked to describe the BI-RADS features [20] of each lesion and to provide a final assessment. Cases were selected so that all BI-RADS features [20] (special cases, mass shape, margins, echogenicity, and posterior features) were represented for which all three experts believed that the images were good examples of the features being tested. For each case, at least two experts agreed on the salient features and recommended management concordant with its malignant or benign cause (which would have included biopsy for a benign lesion with suspicious features), and these descriptions and assessments were used as the consensus for feedback.

Lesions had been proven by core biopsy or at least 4 years’ follow-up. Observers were asked to assume that the lesion was nonpalpable (although 14 of the 25 cancers and nine of 45 benign lesions were palpable) and that the patient was otherwise asymptomatic. Of the 25 cancers, 20 (80%) were invasive, with a median size of 12 mm (range, 5–17 mm), including one metastatic intramammary node replaced by tumor and five were ductal carcinoma in situ (DCIS), including one fibroadenoma involved by DCIS. One (excised) complex sclerosing lesion with associated atypical ductal hyperplasia and papilloma was included as were five negative cases and 39 benign lesions: eight fibroadenomata; five complicated cysts (including one with milk of calcium); three simple cysts; five fibrocystic changes; four (excised) papillomas; three fat necrosis; three benign lymph nodes; and one each ruptured cyst, fibrosis, galactocele, lactational changes, lipoma, granular cell tumor, epidermal inclusion cyst, and sebaceous cyst.

Mammographic Qualification Tasks

To remove potential bias of including investigators who might be expert in breast ultrasound but less skilled at mammography, mammographic interpretive skills tasks were also

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

required. A set of 23 cases of masses or asymmetries including nine invasive cancers (39%) with a median size of 9 mm, three normal variants, and 11 benign findings had been prepared and validated across multiple observers as previously detailed [18]. One or two ultrasound images were also shown for 13 of these cases but were not included in ultrasound interpretive skills analysis. A separate PowerPoint presentation of 32 cases of cropped orthogonal magnification views of mammographic calcifications (with no associated mass) had also been prepared and validated across multiple observers as previously detailed [18]. The calcification examples included magnification views of 17 malignancies (53%), of which 14 (82%) were pure DCIS, two were mixed infiltrating and intraductal carcinoma, and one was invasive ductal carcinoma. *BI-RADS: Mammography* [19] mass margins, calcification morphology and distribution, and final assessments were recorded. All mammographic findings were clearly visible; detection was not being tested.

Ultrasound Interpretive Skills Task

Each of the first 35 observers was shown the ultrasound quiz first without feedback projected from PowerPoint. Observers were first asked to describe whether the lesion was a special case as defined by *BI-RADS: Ultrasound* [20] (i.e., complicated cyst, clustered microcysts, intraductal mass, mass in or on skin, lymph node, or postsurgical scar, with addition of “cyst” and “no mass or lesion,” and excluding foreign body) and, if so, which one. If the lesion was not a special case, the observer was asked to describe the BI-RADS features [20] and to choose the most appropriate descriptor from each major category: mass shape (oval or gently lobulated, round, or irregular); margins (circumscribed or not); orientation (parallel or not); echo pattern (i.e., echogenicity: anechoic, hyperechoic, complex, hypoechoic, isoechoic, mixed hyper- and hypoechoic); posterior features (none, enhancement, shadowing, or combined); and calcifications (macrocalcifications, microcalcifications in a mass, or microcalcifications out of a mass). For each case, observers were asked to provide a BI-RADS [20] final assessment. For three cases (two malignant and one benign), the only sonographic finding was calcifications; although these cases were included for training in BI-RADS feature analysis because they are representative of findings expected at screening, these cases were excluded from scoring for qualifying as an investigator because calcification morphology is important in management and requires correlation with mammograms.

Responses were collected on paper for the first group of 35 observers. The same set of ultrasound cases was immediately shown a second time to the first 35 observers. The second time, after the observers had recorded their description and interpretation of each case, they were provided with feedback after each case. Feedback consisted of the expert consensus description of features and final assessment and the histopathologic truth or benign or negative status.

Because of time and logistic constraints, the remaining 65 investigators who performed qualification tasks over the ensuing 5 years were not included in the study of feedback per se. During a second training session in September 2003, 13 observers used laptops with a CD-ROM with the same cases displayed. Each observer had to lock in his or her feature analysis and assessments before proceeding to the next case, and feedback was given only at the end of the entire reading session. The final training session with 23 observers in January 2004 used response keypads provided by the American College of Radiology (ACR), and feedback after each case was projected in PowerPoint. The CD-ROM was also made available to individuals who did not attend central training sessions to complete independently after reviewing the *BI-RADS: Ultrasound* [20] lexicon; 29 observers completed the quiz in this manner without feedback. The last observer completed all tasks in February 2008. For observers with suboptimal performance on the task by CD-ROM, a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

version of the CD-ROM was also created to provide feedback after each case, and the observer was allowed to complete the task a second time; for this last group of observers, only the results from their first attempt are included.

Statistical Considerations

A malignancy coded as BI-RADS category 4A, 4B, 4C, or 5 was considered a true-positive and a benign or high-risk lesion coded as BI-RADS category 1, 2, or 3 was considered a true-negative. Sensitivity and specificity by observer and in aggregate were calculated for each task. Receiver operating characteristic (ROC) curve analysis was performed (Stata version 10.0, StataCorp), and area under the ROC curve (AUC) was calculated.

We compared observer performance (AUC) for the subset of 35 observers who completed the ultrasound task without feedback and then with feedback. Nonlinear multivariate ordinal regression models were used to assess predictive accuracy and model the AUC for each demographic variable both with feedback and without feedback (SAS, version 9.1 NL mixed, SAS Institute) for these 35 observers. Including a random effect for observer made the model unstable, so a model without random effects was used only for descriptive and exploratory purposes. The two models provided similar overall estimates, indicating that the observed trends were robust to modeling assumptions (i.e., a model for observers' scores vs a model for observers' AUCs). CIs and *p* values for changes in AUC after feedback were obtained from a global F test that the diagnostic accuracies were all equal under a random-effects model [22] and separate random-effects models were used to adjust this effect for observer demographics.

For the 35 observers who received training in *BI-RADS: Ultrasound* [20] and completed the ultrasound interpretive skills task both without feedback and with feedback after each case, we calculated kappa as a measure of agreement for each of the categories of feature analysis and final assessment. We also calculated kappa using grouped final assessments (category 1 or 2; category 3; and category 4A, 4B, 4C, or 5) as well as kappas for individual feature descriptors and assessments. According to the criteria of Landis and Koch [23], a kappa value of less than 0 indicates less agreement than expected by chance; 0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; 0.81–0.99, almost perfect agreement; and 1.0, perfect agreement.

To qualify as an investigator, individuals had to achieve at least 88% sensitivity and at least 40% specificity on each of the required tasks.

Results

Experience

The 100 observers who participated represented 26 institutions in the United States, Canada, and Argentina, of whom 82 were eventual investigators (from 21 institutions) in the ACRIN 6666 protocol. The majority had at least 10 years' experience in breast imaging (Table 1). Only 11% of observers interpreted an average of more than 300 mammograms per week, and most interpreted fewer than 40 breast ultrasound examinations per week. Fifty-seven observers (57%) had a technologist performing breast ultrasound examinations for them, although 53 of these 57 observers (93%) reported that they personally routinely scan the patient after their technologist has finished. Fifty-nine observers (59%) performed at least some whole-breast ultrasound examinations in their usual practice before ACRIN 6666 for both diagnostic and screening examinations; only 16 (16%) had never performed whole-breast ultrasound before the protocol. There were no important differences between the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

subset of 35 observers who completed the ultrasound task both without feedback and with feedback and the other 65 observers.

Performance and Effects of Feedback

Without feedback, a range of sensitivities was observed for the ultrasound task of from 0.68 to 1.00 (mean, 0.89 [Table 2]; median, 0.92 across 77 observers). At thresholds of 0.88 or greater sensitivity and at least 0.40 specificity, 17 of these 77 observers (22%) failed (with one failure due to low specificity). For the 23 observers who completed the task only with feedback, three (13%) failed due to low sensitivity ($p = 0.34$ vs independent group of observers' failure rate without feedback).

For the subset of observers performing the ultrasound task both without feedback and with feedback, eight of 35 observers (23%) failed the ultrasound interpretive skills task without feedback and only one of 35 (3%) failed with feedback given after each case ($p = 0.016$). With feedback for these 35 observers, there was significant improvement in mean sensitivity of 3.9% (95% CI, 2.5 to 5.3) (Table 3); for 20 observers, sensitivity increased, and for 15, there was no change. This improvement in sensitivity was achieved with an in-significant drop in specificity, averaging 1.4% (range of change, −18% to 11%) (Table 3); the decrease in specificity was significant for only one observer (loss of 18%; $p < 0.05$). Without feedback, AUCs ranged from 0.73 to 0.89 with a mean of 0.82 (median, 0.82). With feedback, AUCs ranged from 0.78 to 0.90 with a mean of 0.84 (median, 0.84). The mean increase was 0.022 (median, 0.018; 95% CI, 0.012 to 0.032) (Table 3). The change in AUC with feedback was significant for only four observers, each of whom showed improvement. Three breast imagers in the lowest quartile of initial sensitivity performance showed the greatest improvement in sensitivity (with no change or improvement in AUC). The Pearson correlation coefficient (r) of the AUC readings between before feedback and after feedback was 0.645.

Although a different set of cases was used from those in the ultrasound quiz, mammographic interpretation of masses (supplemented by relevant ultrasound images as needed) was uniformly excellent without feedback, with mean sensitivity of 0.99 (range, 0.89–1.0) and AUC of 0.93 across the first 71 observers (and therefore not required of the last 29 observers). The mean sensitivity of all 100 observers for the calcifications task was 0.93 (range, 0.71 to 1.0) (Table 2). Using a cutoff of 0.88 sensitivity, 10 of 100 observers (10%) failed the calcifications task.

Relationship of Observer Experience Variables to Breast Ultrasound Interpretive Performance

For the group of 35 observers who completed the breast ultrasound interpretive skills task both without feedback and then with feedback, the overall mean AUC improved from 0.817 without feedback to 0.840 with feedback ($p < 0.0001$). Initial performance improved with increasing years of experience in breast imaging, with peak performance at 6–10 years (Table 4) (AUC = 0.86 after feedback; 95% CI, 0.83 to 0.89) compared with 0.80 for less than 2 years' experience (95% CI, 0.74 to 0.85). Observers spending less than 25% of their time in breast imaging performed less well than those spending more time (Table 4).

Our model shows that observers for whom technologists perform breast ultrasound scanning showed worse interpretive skills than any other group with a mean AUC of 0.68 (95% CI, 0.61 to 0.75) without feedback and 0.74 (95% CI, 0.68 to 0.81) with feedback. Note that these estimates are based on the performance of two observers; Figure 1 and Table 4 display these trends.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The number of mammograms interpreted per week was an important potential confounder to include in the model but yielded only a weak association with performance. Our model suggested that there are trends in the degree of experience: Interpreting at least 40 breast ultrasound examinations per week and performing whole-breast ultrasound in women with newly diagnosed cancer were associated with improved interpretive performance, but the model was not definitive.

The model created from the first 35 observers validated well against the full set of 100 observers (Fig. 2). Figure 2 shows high calibration for these data. Immediate feedback significantly improved interpretive performance across all experience categories.

Specific cases with observer improvement after feedback on prior cases are shown in Figures 3–5.

Kappa Values

Table 5 summarizes the kappa values for BI-RADS features [20] and final assessments for the first 35 observers. Moderate agreement was seen for each of the major categories of ultrasound features except posterior features, for which there was good agreement. Only fair agreement was seen for characterization of a lesion as a simple cyst, with a kappa value of 0.25; however, there was moderate agreement for characterization of a lesion as a complicated cyst, with a kappa value of 0.51. There was essentially no agreement for classifying an unusual vertically oriented simple cyst, and a lobulated cyst was also problematic for some observers (Fig. 6). When these terms were grouped, the kappa value for simple cyst or complicated cyst was 0.58. Good agreement was seen for the common findings of clustered microcysts and lymph nodes.

With feedback, there was no change in kappa values for most features, but better agreement was seen for calcification descriptors (Table 5). With feedback, agreement on final assessments improved significantly for all assessments except BI-RADS category 1 (Table 5). Even with feedback, there remained only slight-to-fair agreement on the optional BI-RADS assessment subcategories of 4A, 4B, and 4C. The results for kappas with the subset of 35 observers were consistent with those obtained in the full dataset of 71 observers who were trained in BI-RADS.

Discussion

Proper interpretation of breast ultrasound requires both detection and feature analysis. To participate in performing ultrasound of patients in the ACRIN 6666 protocol, investigators had to successfully complete a separate task, wherein the ability to detect known 3- to 10-mm lesions was assessed by scanning a phantom [21]. Although perception of the abnormality was an issue for a few cases in this series, detection was not specifically tested because observers were shown two orthogonal images of each lesion: We primarily assessed feature analysis. Our results showed that feedback on feature analysis and assessments improved interpretive skills of experienced observers in breast ultrasound.

Typically breast ultrasound is integrated with mammography to achieve optimal interpretation. This task deliberately required observers to record their impression of ultra sound findings without the benefit of correlative mammography. Similarly, the ACRIN 6666 protocol required that the initial interpretations of breast ultrasound and mammographic examinations were independent in an effort to reduce potential bias from targeting vague, noncallable mammographic abnormalities when performing the ultrasound examination. Importantly, interpretive performance for mammographic masses (with correlative ultrasound when needed) was uniformly outstanding across these same observers,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

suggesting that the usual practice of integrating mammography and ultrasound is better than interpreting ultrasound in isolation, and the ACRIN 6666 protocol did require integration of mammographic and sonographic findings for clinical management. Further, it is important that investigators were successful at both mammographic and sonographic interpretive tasks and therefore the results of ACRIN 6666 should not be biased by lack of investigator skills in one modality compared with the other.

Reduced interpretive performance with mammographic calcifications compared with masses in this series was not surprising. In practice, positive predictive values for biopsy recommendations are typically lower for calcifications than for masses [24, 25]. This difference reflects greater diagnostic uncertainty for calcifications than for masses.

Immediate feedback was of value to all observers regardless of prior experience or practice patterns, but the greatest improvement in sensitivity was seen for three observers in the lowest quartile of initial performance. Feedback is an integral part of clinical medical and most other education and, from a systems standpoint, implies that there is the opportunity for change based on prior performance and, thereby, to learn how and why errors are made and avoid them in the future [26]. Breast imaging specialists have been shown to have higher rates of detection of early breast cancer, with fewer false-positives, than general radiologists [27]. One advantage common to many breast imaging specialists may be the performance and review of cases going to percutaneous biopsy and to regularly compare imaging and histopathologic findings—that is, to obtain frequent feedback on outcomes. Miglioretti et al. [28] showed that recent training in mammography and experience performing breast biopsies were associated with similar significant increases in both sensitivity and false-positive rates for diagnostic mammography.

Experience requirements for mammographic interpretation vary dramatically in countries with screening programs, with the requirement lowest in the United States at 960 mammograms interpreted every 2 years, and highest at 5000 mammograms per year for screeners in the United Kingdom. Barlow et al. [29] showed improved sensitivity and reduced specificity among radiologists interpreting at least 1000 mammograms per year compared with those interpreting fewer, and Haneuse et al [30] reported similar results for radiologists interpreting at least 1000 diagnostic mammograms per year. Increasing years of experience was associated with decreased sensitivity and increased specificity of mammographic interpretation [29]. Smith-Bindman et al. [31] reported that across 196 facilities, the mean number of mammograms interpreted per year per physician in the United States is currently 1777 and that 10% of the United States' capacity would be curtailed if the requirement were increased to 1000 mammograms per year.

Currently, for facility accreditation in breast ultrasound, the ACR has variable experience requirements for breast ultrasound–interpreting physicians, with an initial requirement of supervision or performance and interpretation of 300 (board-certified radiologists) to 500 (other specialties) breast ultrasound examinations in the prior 36 months, then 100 examinations per year thereafter [32]. The American Society of Breast Surgeons requires performance of 100 breast ultrasound examinations per year [33], and the International Breast Ultrasound School recommends performance of 500 examinations, of which 300 include cytology or histopathology correlation "to achieve accuracy and confidence" [34]. Both nationally and internationally, there is no requirement for the physician who interprets breast ultrasound to also meet experience requirements or certifications for mammographic interpretation. Indeed, in many countries, particularly in Asia, there are physician specialists in ultrasound who interpret only ultrasound and radiologists who interpret mammograms and no one individual who routinely does both.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

All of our observers self-reported a minimum experience of 500 breast ultrasound examinations and 2500 mammogram interpretations in the prior 2 years. Based on our results, experience performing breast ultrasound appears to be important to proper interpretation. This is supported by better performance in the task described herein among the observers who spend at least 25% of their time in clinical breast imaging, who interpret at least 40 breast ultrasound examinations per week, and who perform ultrasound themselves or at least rescan after their technologists rather than solely relying on their technologists.

Analysis of kappas yielded some interesting findings. Although “cyst” is not a formal term in the original *BI-RADS: Ultrasound* [20], observers were instructed that this term should describe a circumscribed mass that is anechoic with posterior enhancement and an imperceptible wall. The only fair agreement we observed in our series—for the three typical simple cysts—is a cause for concern because cysts are quite common. Cysts were present in more than 37% of participants in ACRIN 6666 in year 1 and more than 47% of participants over the 3 years of screening ultrasound [35], with the largest cyst 8 mm or smaller in 70% of the women with cysts. The operator dependence of breast ultrasound was not found to be problematic in prior work [36, 37], but characterization of simple cysts was unreliable for cysts smaller than 8 mm [36]. Some high-grade invasive carcinomas can appear mostly circumscribed and anechoic and can show posterior enhancement [38]: Overlap in the appearance of such invasive carcinomas with the appearance of cysts can be problematic. Indeed, in the series of Hong et al. [39], among anechoic masses going to biopsy, 16% were malignant. Elastography may improve the accuracy of breast ultrasound [40–44], in part by helping to distinguish simple cysts from complex cystic and solid or anechoic solid masses, but further validation is warranted and small masses deeper than 2 cm from the skin can remain problematic even on elastography [44].

Moderate agreement was observed for breast ultrasound margin descriptions in our study with no change with feedback. Reasoning that clinical management is largely derived from the determination of whether margins are circumscribed or not, we did not evaluate subdescriptors for margins that were not circumscribed (i.e., microlobulated, indistinct, angular, spiculated). Lee et al. [45] reported only slight-to-fair agreement for each of the terms “indistinct,” “angular,” and “microlobulated,” with kappas of 0.20, 0.21, and 0.25, respectively; however, they found good agreement on spiculated margins, with a kappa value of 0.66. In practical use, recognition that at least a portion of a mass’ margin is indistinct, angular, or microlobulated may be more important than distinguishing one of these features from the others.

Overall, our interobserver agreement results for *BI-RADS: Ultrasound* were similar to those of Lazarus et al. [46] and Lee et al. [45]. We did not specifically evaluate agreement on lesion boundary (abrupt interface or echogenic halo), but moderate agreement on this feature was seen in Lee et al. In practice, the lesion boundary is not ever “abrupt” for indistinctly marginated masses even when they lack an echogenic halo; this area of confusion will be addressed in the next edition of *BI-RADS: Ultrasound* (Mendelson EB, written communication, May 17, 2010).

At least in part because of the large number of choices, there was less agreement on echo pattern—that is, echogenicity—with a kappa value of 0.41. We had no cases that, by consensus, were anechoic in part because cysts were considered special cases; nearly anechoic solid masses were considered “hypoechoic,” as in “markedly hypoechoic (solid)” in the work of Stavros et al. [47] and validated by Baker et al. [48]. Lazarus et al. [46] found even less overall agreement among only five radiologists describing echo pattern, with a kappa value of 0.29. Isoechoic masses can be particularly subtle on ultrasound, and these

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

masses were overrepresented among missed cancers in our series. Importantly, we included negative cases in our series, which allowed outright lack of perception of isoechoic masses.

Microcalcifications can be extremely subtle on ultrasound because many calcifications are too small to be resolved even with high-frequency ultrasound transducers. Most DCIS lesions are diagnosed because of suspicious calcifications identified on screening mammography [49, 50]. It is not surprising therefore that, across six series encompassing 150 cancers, only 6% of cancers seen only sonographically were DCIS [13]. Despite these limitations, agreement on the presence of microcalcifications on ultrasound either in a mass or out of a mass was fair without feedback and was significantly improved to moderate after feedback in this series. Ultrasound can be used with success to guide biopsy of malignant calcifications [51, 52]; it remains difficult, however, to interpret microcalcifications outside a mass on ultrasound because their morphology and often their distribution cannot be discerned. Two of the six cancers most frequently dismissed as benign in our series were micro-calcifications due to DCIS, with the intraductal location of calcifications subtle, but recognizable, for one of these cases (Fig. 7).

Subdivision of BI-RADS category 4 is optional in the 4th edition *of BI-RADS: Mammography* [19] and was used in the ACRIN 6666 protocol for mammography, ultrasound, and MRI. Use of categories 4A, 4B, and 4C may facilitate communication with pathologists and referring physicians for lesions going to biopsy. Subdivision of BI-RADS 4 categories also facilitates ROC analysis. Although the overall kappa value for BI-RADS 4 was 0.52, only slight agreement was seen for the subcategories of 4A, 4B, and 4C despite preliminary instruction (during training) in specific lesions appropriate for each subcategory (e.g., for 4A: new or enlarging oval, circumscribed isoechoic mass compatible with fibroadenoma vs complicated cyst, or intraductal mass, or possible abscess; 4B, complex cystic and solid masses; and 4C, indistinctly marginated masses with or without microcalcifications). Lee et al. [45] found moderate agreement for subcategory 4A, with a kappa value of 0.57; slight agreement for subcategory 4B, with a kappa value of 0.09; and fair agreement for subcategory 4C, with a kappa value of 0.38.

Although feedback after each case improved agreement on final assessments, we could not attribute this improvement to a particular feature or group of features or lesion types. Indeed, except for improved recognition of the presence of calcifications, we did not observe improvement in most feature descriptors with feedback.

Our study has a few limitations. Ultrasound often depends on real-time scanning to distinguish artifactual refractive edge shadowing from true posterior shadowing due to a mass or to distinguish an isoechoic mass from a fat lobule. Observers were presented with only two orthogonal B-mode ultrasound images without Doppler imaging or elastography, both of which improve specificity of breast ultrasound interpretation [43, 44]. It has been shown that interpretive performance and agreement in clinical practice exceed that in observer studies [53]. We did not assess intraobserver variability, although intraobserver agreement has been substantial to almost perfect in prior evaluation of the *BI-RADS: Ultrasound* lexicon [45] and its precursors [48].

In summary, most experienced breast imaging observers performed well on a breast ultrasound interpretive skills task. The use of most *BI-RADS: Ultrasound* feature descriptors showed at least moderate agreement across experienced observers; simple cyst and calcification descriptors were the most inconsistently used terms. Feedback on feature analysis and diagnosis after each case improved breast ultrasound interpretive performance even among experienced observers, as did direct experience performing ultrasound scanning, including rescanning after the technologist. Results of supplemental screening

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

ultrasound in ACRIN 6666 [5] are expected to be generalizable to other observers who meet similar experience requirements, and the ultrasound CD-ROM used herein can be made available on request. Implementing consistent feedback in usual practice by reviewing biopsy-proven cases and follow-up may achieve similar or even greater gains in radiologist performance.

Acknowledgments

We thank Jose Cayere of the ACR for help developing the CD-ROM of the ultrasound qualification task and Christopher Merritt of Thomas Jefferson University School of Medicine for reviewing ultrasound quiz cases as part of the expert consensus. We are grateful to the dedicated staff of the Breast Center at the Feinberg School of Medicine, Northwestern University, and to Cynthia Olson of the ACRIN for facilitating the training sessions. We thank Michele Wittling of the ACR for assistance with audience response for one of the training sessions. We are especially appreciative of the many breast imaging radiologists who participated in these tasks and in the ACRIN 6666 protocol. Please note that the ultrasound CD-ROM can be made available on request by contacting the corresponding author.

This work was supported by grants from The Avon Foundation and the National Cancer Institute (CA89008). The mammographic interpretive skills tasks were developed through support from the Susan G. Komen Foundation. The clinical trials.gov identifier for this work is NCT00072501.

W. A. Berg has consulted for and received travel funds from SuperSonic Imagine to perform a reader study, present results, analyze data, and prepare manuscripts for publication. She is a consultant for Naviscan, Inc., and is on the medical advisory board for Philips Healthcare. E. B. Mendelson has received research support and travel funds from SuperSonic Imagine and Siemens, Inc.; is a consultant for and is on the medical advisory board of Quantason; and is on the scientific advisory boards of Hologic, Inc. (travel expenses only), and Toshiba (TAMS) Ultrasound.

References

1. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2002; 137:347–360. [PubMed: 12204020]
2. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005; 353:1773–1783. [PubMed: 16169887]
3. Mandelson MT, Oestreicher N, Porter PL, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000; 92:1081–1087. [PubMed: 10880551]
4. Berg WA, Blume JD, Cormack JB, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA. 2008; 299:2151–2163. [PubMed: 18477782]
5. Berg WA, Zhang Z, Lehrer D, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA. 2012; 307:1394–1404. [PubMed: 22474203]
6. Buchberger W, Niehoff A, Obrist P, DeKoekkoek-Doll P, Dunser M. Clinically and mammographically occult breast lesions: detection and classification with high-resolution sonography. Semin Ultrasound CT MR. 2000; 21:325–336. [PubMed: 11014255]
7. Corsetti V, Ferrari A, Ghirardi M, et al. Role of ultrasonography in detecting mammographically occult breast carcinoma in women with dense breasts. Radiol Med (Torino). 2006; 111:440–448. [PubMed: 16683089]
8. Crystal P, Strano SD, Shcharynski S, Koretz MJ. Using sonography to screen women with mammographically dense breasts. AJR. 2003; 181:177–182. [PubMed: 12818853]
9. Gordon PB, Goldenberg SL. Malignant breast masses detected only by ultrasound: a retrospective review. Cancer. 1995; 76:626–630. [PubMed: 8625156]
10. Kaplan SS. Clinical utility of bilateral whole-breast US in the evaluation of women with dense breast tissue. Radiology. 2001; 221:641–649. [PubMed: 11719658]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

11. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology. 2002; 225:165–175. [PubMed: 12355001]
12. Leconte I, Feger C, Galant C, et al. Mammography and subsequent whole-breast sonography of nonpalpable breast cancers: the importance of radiologic breast density. AJR. 2003; 180:1675–1679. [PubMed: 12760942]
13. Berg WA. Supplemental screening sonography in dense breasts. Radiol Clin North Am. 2004; 42:845–851. vi. [PubMed: 15337420]
14. Kelly KM, Dean J, Comulada WS, Lee SJ. Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. Eur Radiol. 2010; 20:734–742. [PubMed: 19727744]
15. Wenkel E, Heckmann M, Heinrich M, et al. Automated breast ultrasound: lesion detection and BI-RADS classification—a pilot study. Rofo. 2008; 180:804–808. [PubMed: 18704878]
16. Drukker K, Sennett CA, Giger ML. Automated method for improving system performance of computer-aided diagnosis in breast ultrasound. IEEE Trans Med Imaging. 2009; 28:122–128. [PubMed: 19116194]
17. Berg, WA.; Mendelson, EB.; Merritt, CRB.; Blume, J.; Schleinitz, M. ACRIN Website. [Accessed February 8, 2012] ACRIN 6666: screening breast ultrasound in high-risk women. acrin.org/Portals/0/Protocols/6666/Protocol-ACRIN%206666%20Admin%20Update%2011.30.07.pdf. Published November 9, 2007. Updated November 30, 2007.
18. Berg WA, D'Orsi CJ, Jackson VP, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology. 2002; 224:871–880. [PubMed: 12202727]
19. 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.
20. Mendelson, EB.; Baum, JK.; Berg, WA.; Merritt, CRB.; Rubin, E. Breast Imaging Reporting and Data System, BI-RADS: Ultrasound. Reston, VA: American College of Radiology; 2003.
21. Berg WA, Blume JD, Cormack JB, Mendelson EB, Madsen EL. Lesion detection and characterization in a breast US phantom: results of the ACRIN 6666 Investigators. Radiology. 2006; 239:693–702. [PubMed: 16641344]
22. Zhou, X-H.; Obuchowski, NA.; McClish, DK. Statistical methods in diagnostic medicine. New York, NY: Wiley; 2002.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
24. Burnside ES, Ochsner JE, Fowler KJ, et al. Use of microcalcification descriptors in *BI-RADS* 4th edition to stratify risk of malignancy. Radiology. 2007; 242:388–395. [PubMed: 17255409]
25. Liberman L, Abramson AF, Squires FB, Glassman JR, Morris EA, Dershaw DD. The Breast Imaging Reporting and Data System: positive predictive value of mammographic features and final assessment categories. AJR. 1998; 171:35–40. [PubMed: 9648759]
26. Ende J. Feedback in clinical medical education. JAMA. 1983; 250:777–781. [PubMed: 6876333]
27. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology. 2002; 224:861–869. [PubMed: 12202726]
28. Miglioretti DL, Smith-Bindman R, Abraham L, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. J Natl Cancer Inst. 2007; 99:1854–1863. [PubMed: 18073379]
29. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst. 2004; 96:1840–1850. [PubMed: 15601640]
30. Haneuse S, Buist DS, Miglioretti DL, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. Radiology. 2012; 262:69–79. [PubMed: 22106351]
31. Smith-Bindman R, Miglioretti DL, Rosenberg R, et al. Physician workload in mammography. AJR. 2008; 190:526–532. [PubMed: 18212242]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

32. American College of Radiology Website. [Accessed February 14, 2011] Breast ultrasound accreditation program requirements. www.acr.org/accreditation/breast/breast_ultrasound_reqs.aspx. Revised November 22, 2011
33. The American Society of Breast Surgeons Website. [Accessed February 14, 2011] Breast ultrasound certification. www.breast-surgeons.org/certification/breast_ultrasound_certification.php. Published 2008
34. Madjar, H.; Rickard, M.; Jellins, J.; Otto, R., editors. International Breast Ultrasound School Website. [Accessed February 14, 2011] IBUS guidelines for the ultrasonic examination of the breast. www.ibus.org/guidelines.html. Published May 24, 1998
35. Berg WA, Sechtin AG, Marques H, Zhang Z. Cystic breast lesions and the ACRIN 6666 experience. Radiol Clin North Am. 2010; 48:931–987. [PubMed: 20868895]
36. Berg WA, Blume JD, Cormack JB, Mendelson EB. Operator dependence of physician-performed whole-breast US: lesion detection and characterization. Radiology. 2006; 241:355–365. [PubMed: 17057064]
37. Bosch AM, Kessels AG, Beets GL, et al. Interexamination variation of whole breast ultrasound. Br J Radiol. 2003; 76:328–331. [PubMed: 12763948]
38. Lamb PM, Perry NM, Vinnicombe SJ, Wells CA. Correlation between ultrasound characteristics, mammographic findings and histological grade in patients with invasive ductal carcinoma of the breast. Clin Radiol. 2000; 55:40–44. [PubMed: 10650109]
39. Hong AS, Rosen EL, Soo MS, Baker JA. BI-RADS for sonography: positive and negative predictive values of sonographic features. AJR. 2005; 184:1260–1265. [PubMed: 15788607]
40. Burnside ES, Hall TJ, Sommer AM, et al. Differentiating benign from malignant solid breast masses with US strain imaging. Radiology. 2007; 245:401–410. [PubMed: 17940302]
41. Schaefer FK, Heer I, Schaefer PJ, et al. Breast ultrasound elastography: results of 193 breast lesions in a prospective study with histopathologic correlation. Eur J Radiol. 2011; 77:450–456. [PubMed: 19773141]
42. Wojcinski S, Farrokh A, Weber S, et al. Multi-center study of ultrasound real-time tissue elastography in 779 cases for the assessment of breast lesions: improved diagnostic performance by combining the BI-RADS(R)—US classification system with sonoelastography. Ultraschall Med. 2010; 31:484–491. [PubMed: 20408116]
43. Cho N, Jang M, Lyou CY, Park JS, Choi HY, Moon WK. Distinguishing benign from malignant masses at breast US: combined US elastography and color Doppler US—influence on radiologist accuracy. Radiology. 2012; 262:80–90. [PubMed: 22084209]
44. Berg WA, Cosgrove DO, Doré CJ, et al. Shear-wave elastography improves the specificity of breast US: the BE1 Multinational Study of 939 masses. Radiology. 2012; 262:435–449. [PubMed: 22282182]
45. Lee HJ, Kim EK, Kim MJ, et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. Eur J Radiol. 2008; 65:293–298. [PubMed: 17531417]
46. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology. 2006; 239:385–391. [PubMed: 16569780]
47. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology. 1995; 196:123–134. [PubMed: 7784555]
48. Baker JA, Kornguth PJ, Soo MS, Walsh R, Mengoni P. Sonography of solid breast lesions: observer variability of lesion description and assessment. AJR. 1999; 172:1621–1625. [PubMed: 10350302]
49. Dershaw DD, Abramson A, Kinne DW. Ductal carcinoma in situ: mammographic findings and clinical implications. Radiology. 1989; 170:411–415. [PubMed: 2536185]
50. Stomper PC, Connolly JL, Meyer JE, Harris JR. Clinically occult ductal carcinoma in situ detected with mammography: analysis of 100 cases with radiologic-pathologic correlation. Radiology. 1989; 172:235–241. [PubMed: 2544922]
51. Moon WK, Im JG, Koh YH, Noh DY, Park IA. US of mammographically detected clustered micro-calcifications. Radiology. 2000; 217:849–854. [PubMed: 11110953]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

52. Soo MS, Baker JA, Rosen EL. Sonographic detection and sonographically guided biopsy of breast microcalcifications. AJR. 2003; 180:941–948. [PubMed: 12646433]

53. Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. 2008; 249:47–53. [PubMed: 18682584]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

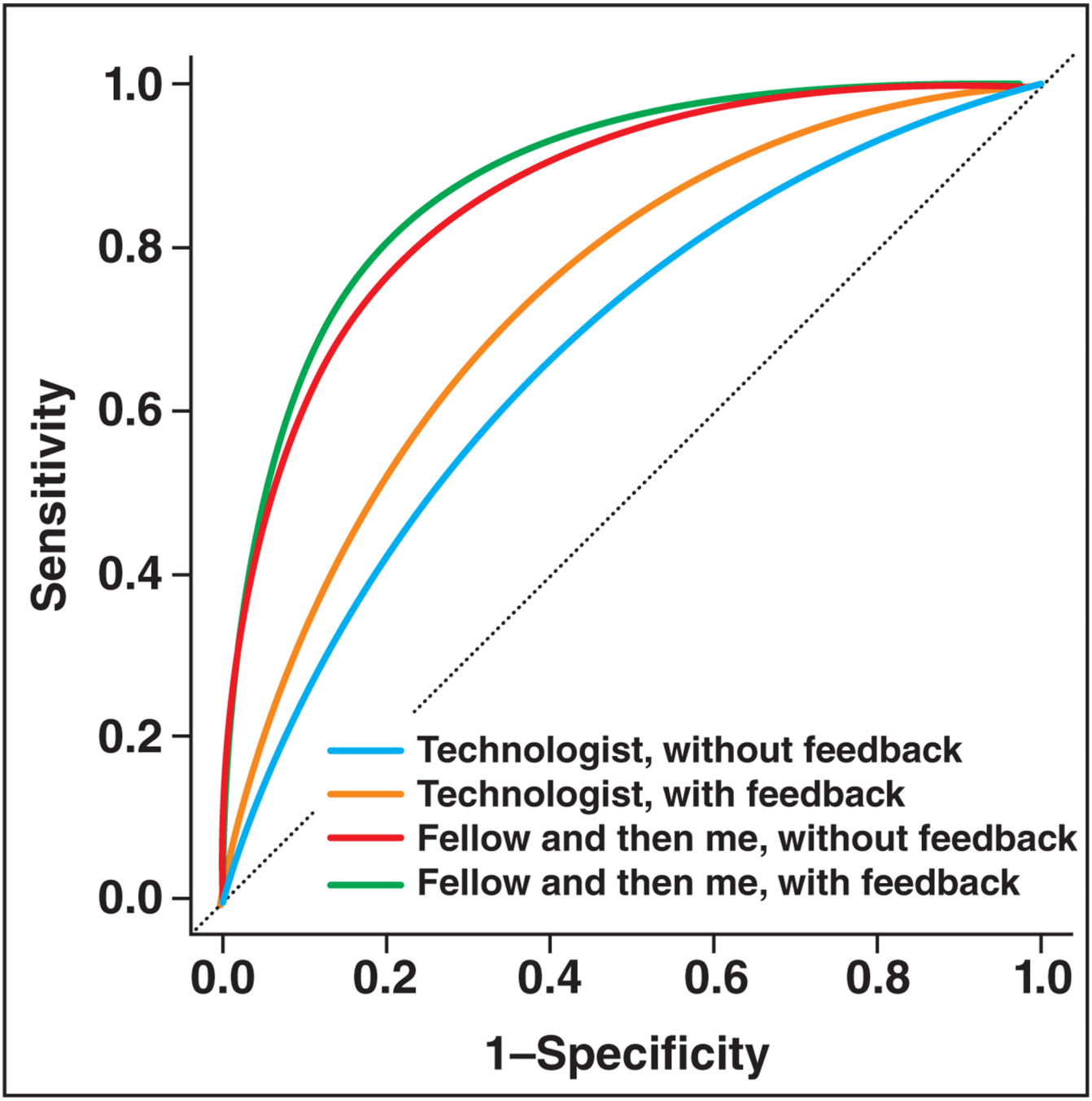

Fig. 1.
Plot of radiologists' interpretive skills in breast ultrasound without feedback and with feedback based on who routinely performs ultrasound examinations at their facility.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

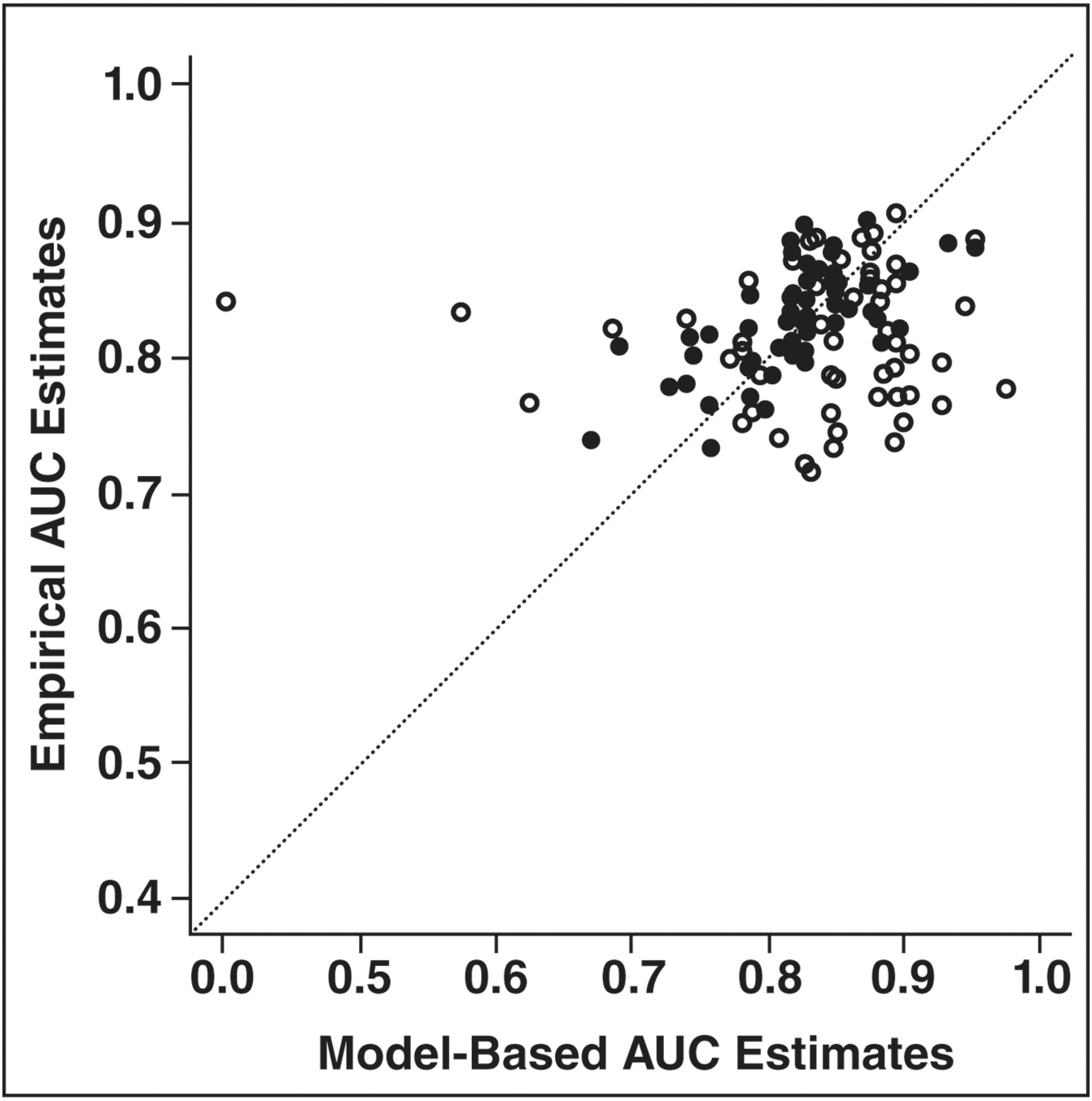

Fig. 2.
Calibration plot shows comparison of empirical areas under the curve (AUCs) versus those predicted by nonlinear ordinal model. Points lying on 45° line (*dotted*) are perfectly predicted by model. Figure shows good calibration and no evidence of bias or attenuation of AUC estimates. ○ = without feedback, ● = with feedback.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

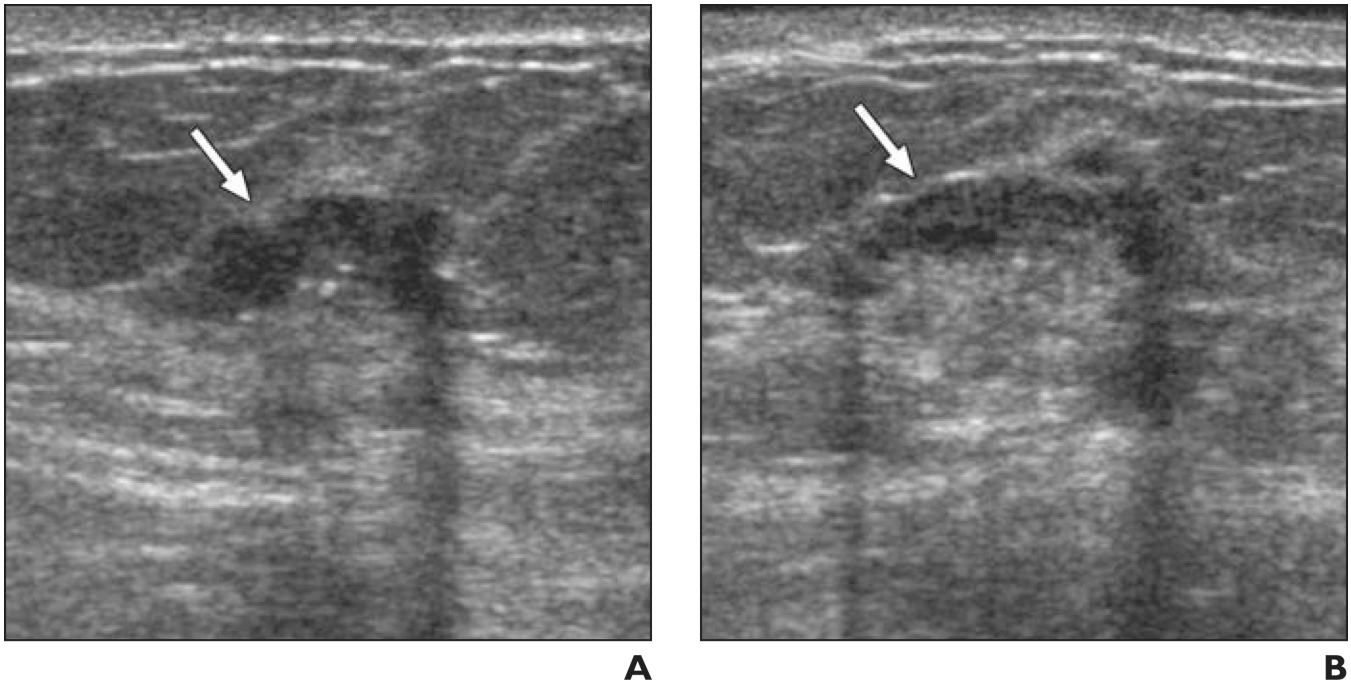

Fig. 3. 41-year-old woman

A and **B**, Radial (**A**) and antiradial (**B**) images of mass (*arrows*) and calcifications due to grade III invasive ductal crcinoma considered benign ($n = 5$) or probably benign ($n = 4$) by nine of 35 observers without feedback and suspicious by all 35 observers after feedback on prior cases.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

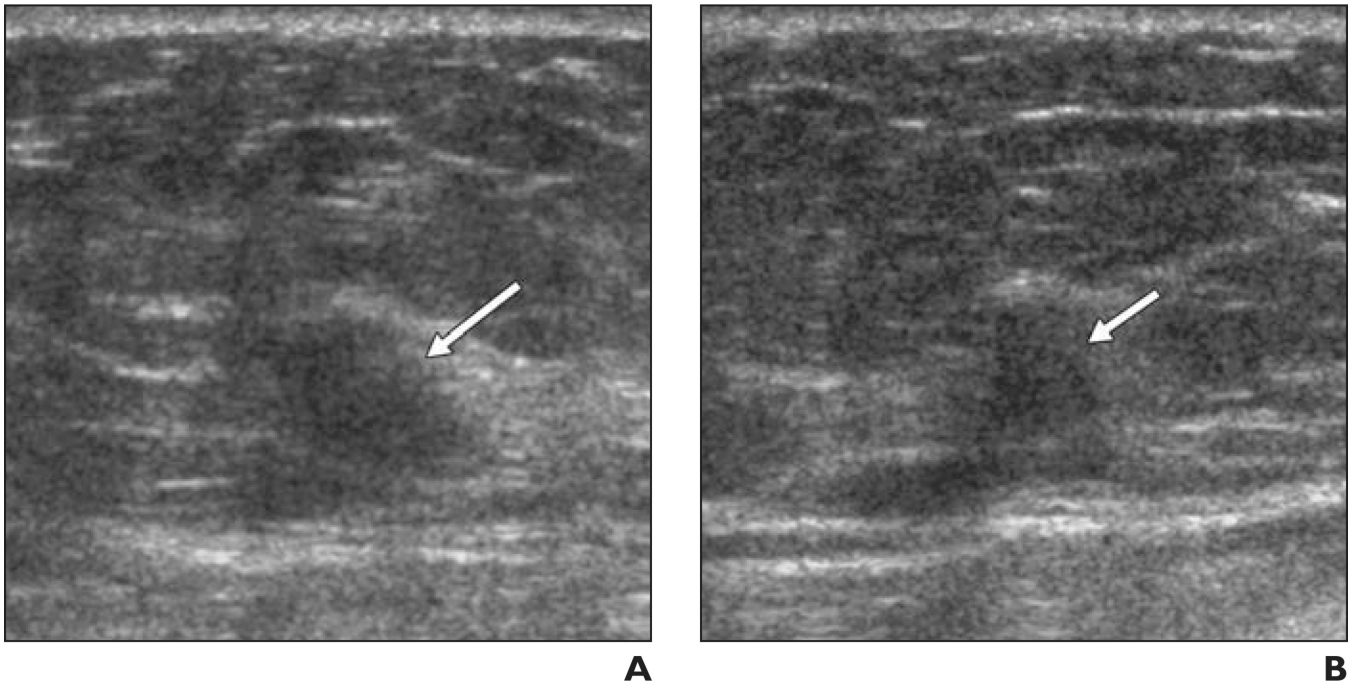

Fig. 4. 46-year-old woman

A and **B**, Radial (**A**) and antiradial (**B**) images of indistinctly marginated isoechoic mass (*arrows*) due to grade II invasive ductal carcinoma. Without feedback, six of 35 observers classified this case as negative and one classified this case as BI-RADS 3. Even with feedback, three observers classified it as negative.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

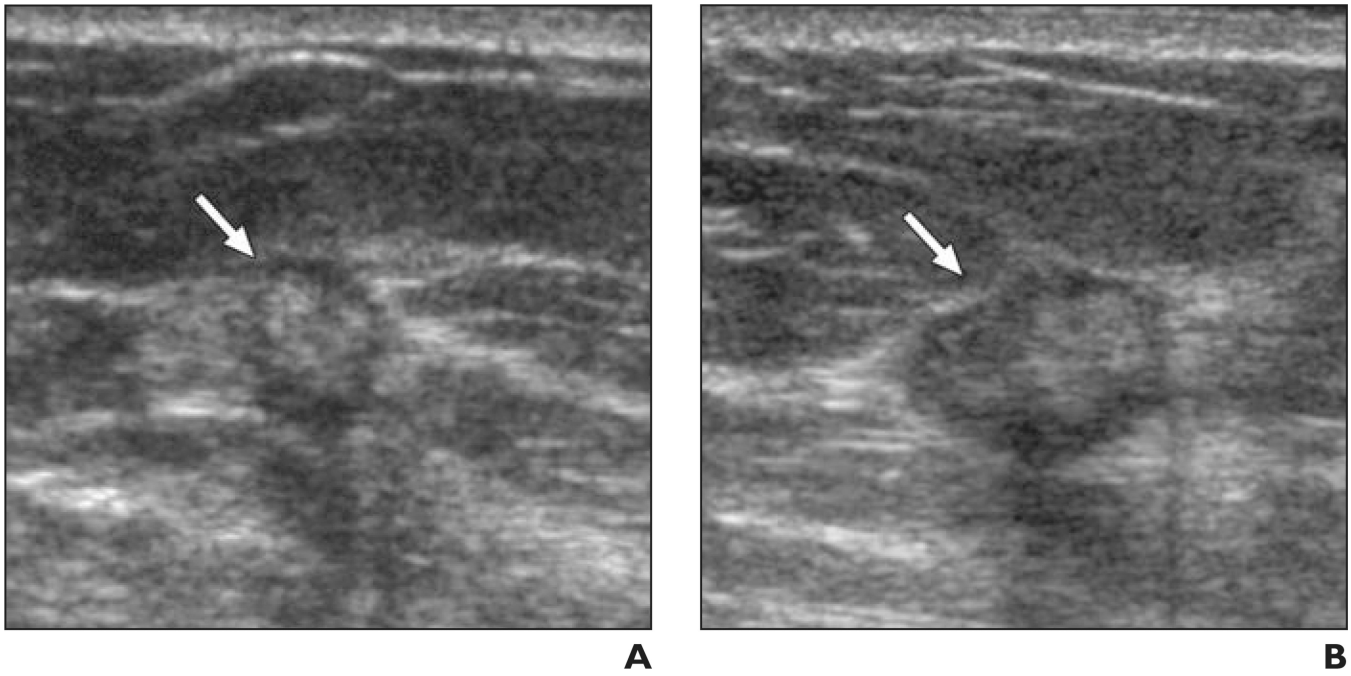

Fig. 5. 55-year-old woman

A and **B**, Radial (**A**) and antiradial (**B**) ultrasound images of irregular mass (*arrows*) with central hyperechogenicity due to grade I invasive ductal carcinoma. Without feedback, 21 of 35 observers (60%) considered this case suspicious; after feedback on prior cases, 26 of 35 observers (74%) recognized this case as suspicious. All misclassifications of this case were "benign lymph node."

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

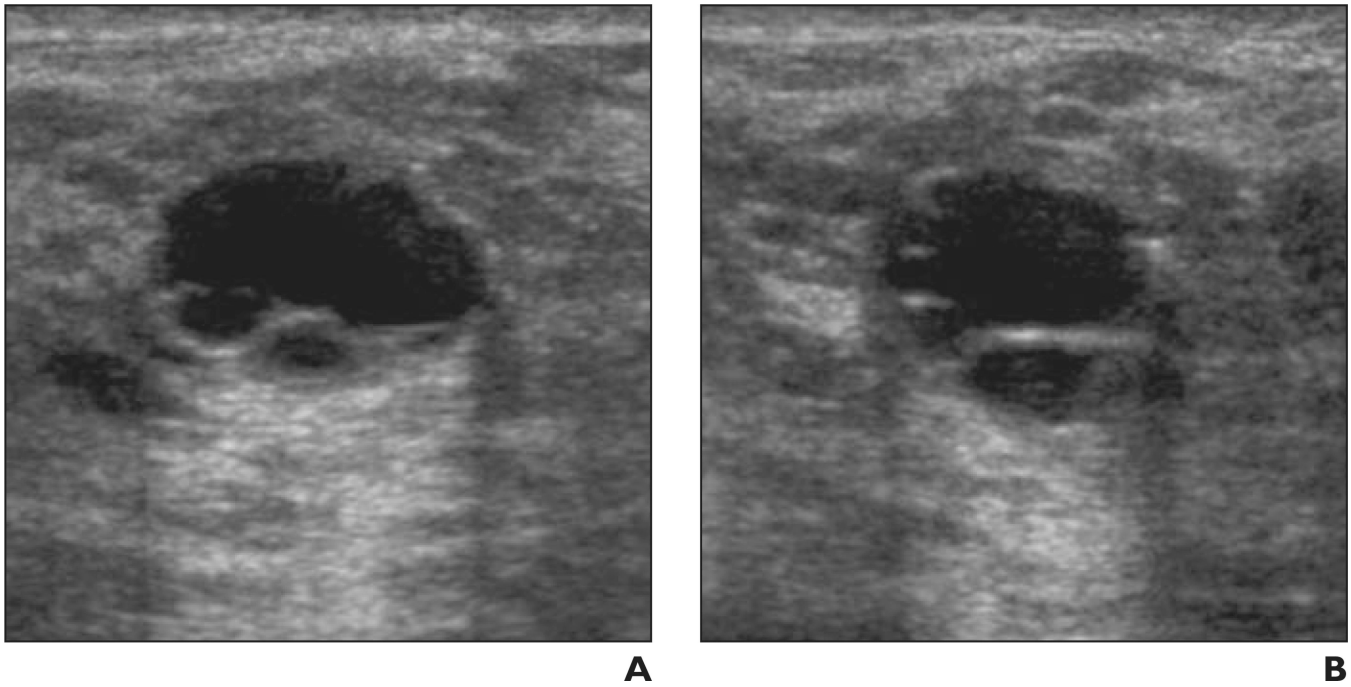

Fig. 6. 41-year-old woman

A and **B**, Radial (**A**) and antiradial (**B**) ultrasound images of lobulated simple cyst. Case was considered suspicious by five of 35 observers (14%) without feedback (three of whom described it as complicated cyst). After feedback on prior cases, three observers still considered mass suspicious, but two of these three classified mass as simple cyst.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

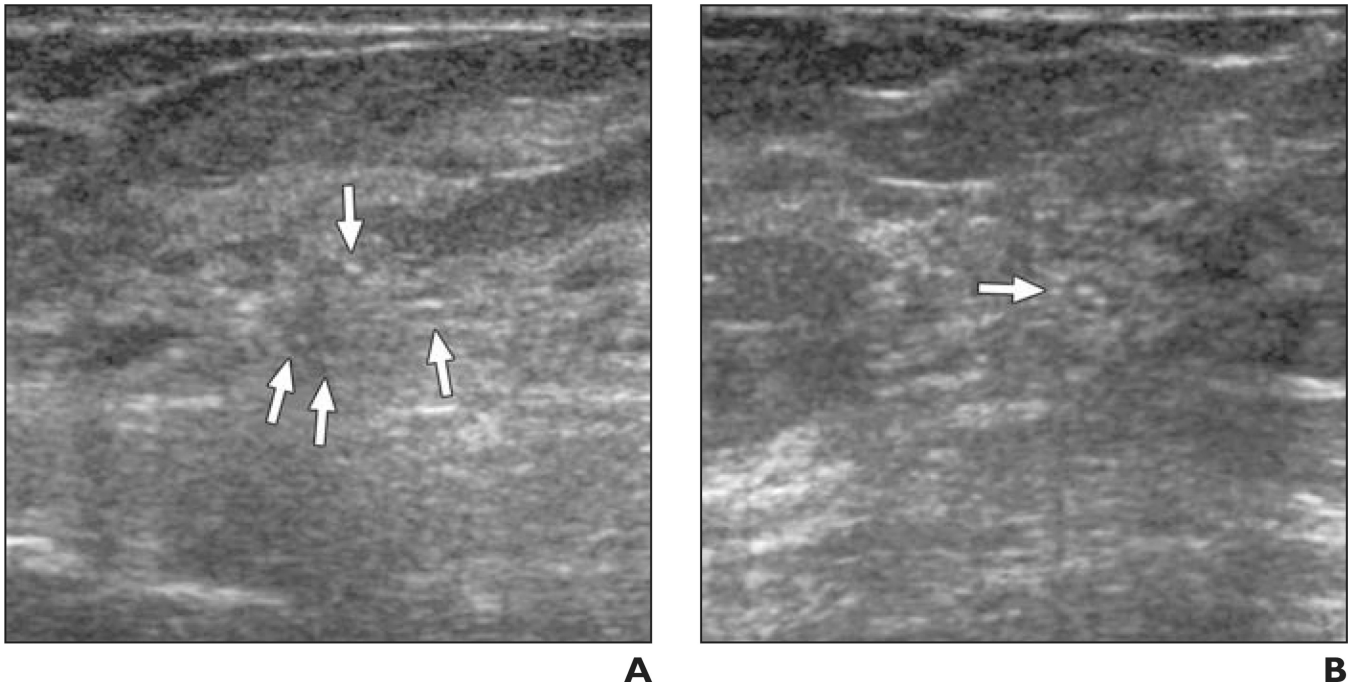

Fig. 7. 74-year-old woman

A and **B**, Radial (**A**) and antiradial (**B**) images show calcifications without mass (arrows) due to extensive low-nuclear-grade micropapillary ductal carcinoma in situ. Intraductal location of calcifications is subtle, but evident, particularly in **B**. This was considered negative (*n* = 9), benign (*n* = 4), or probably benign (*n* = 3) by 16 of 35 observers even after feedback on prior cases.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 1

Summary of Experience in Breast Imaging for 100 Observers Completing Interpretive Skills Tasks

Characteristic	No. of Observers	No. (%)	
		Subgroup With Feedback and Without Feedback	Remaining Observers
No. of observers	100	35	65
Ability to fill out forms			
Obsessive compulsive	61	23 (66)	38 (58)
My research associate does everything	29	9 (26)	20 (31)
Don't count on it	10	3 (8.6)	7 (11)
Years in breast imaging			
< 2	4	2 (5.7)	2 (3.1)
2–5	22	8 (23)	14 (22)
6–10	23	5 (14)	18 (28)
> 10	51	20 (57)	31 (48)
% Time spent in clinical breast imaging			
< 25	5	3 (8.6)	2 (3.1)
25–49	8	3 (8.6)	5 (7.7)
50–74	17	6 (17)	11 (17)
75–100	70	23 (66)	47 (72)
Who performs breast ultrasound of your patients?			
Technologist	4	2 (5.7)	2 (3.1)
You	27	11 (31)	16 (25)
Resident then you	7	3 (8.6)	4 (6.2)
Technologist then you	53	15 (43)	38 (58)
Fellow then you	9	4 (11)	5 (7.7)
How many mammograms do you interpret per week?			
0–99	9	5 (14)	4 (6.2)
100–149	25	7 (20)	18 (28)
150–199	28	11 (31)	17 (26)
200–299	27	10 (29)	17 (26)
300–399	4	1 (2.9)	3 (4.6)
400–499	6	1 (2.9)	5 (7.7)
500 or more	1	NA	1 (1.5)
How many breast ultrasound examinations do you interpret per week?			
< 20	15	4 (11)	11 (17)
21–39	46	15 (43)	31 (48)
40–69	34	16 (46)	18 (28)
70–99	4	NA	4 (6.2)
≥ 100 or more	1	NA	1 (1.5)
Do you currently perform whole-breast ultrasound?			

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Characteristic	No. of Observers	No. (%)	
		Subgroup With Feedback and Without Feedback	Remaining Observers
Never	16	8 (23)	8 (12)
In patients with new cancer in that breast	25	9 (26)	16 (25)
Diagnostic and screening	59	18 (51)	41 (63)

Note—NA = not applicable, no entries.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 2

Breast Imaging Interpretive Performance of 100 Experienced Observers

Examination Presented to Observer	Mean (95% CI)		
	Sensitivity	Specificity	Area Under the Curve
Ultrasound without feedback (n = 77)[a]	0.892 (0.876 to 0.908)	0.602 (0.581 to 0.623)	0.815 (0.805 to 0.826)
Ultrasound with feedback (n = 23)[a]	0.920 (0.896 to 0.945)	0.601 (0.582 to 0.619)	0.821 (0.803 to 0.840)
Difference[b]	0.028 (0.0007 to 0.057)	−0.0017 (−0.029 to 0.026)	0.006 (−0.015 to 0.027)
Mammographic masses and asymmetries[c] (n = 71)	0.994 (0.988 to 1.000)	0.717 (0.696 to 0.738)	0.930 (0.923 to 0.938)
Mammographic calcifications (n = 100)	0.930 (0.918 to 0.942)	0.794 (0.779 to 0.810)	0.924 (0.916 to 0.931)

[a] Seventy-one observers received 1 hour of didactic training in *BI-RADS: Ultrasound* [20]. Thirty-five observers performed the ultrasound task without feedback and then with feedback and only the without-feedback results are presented here; 13 had no concurrent feedback; 23 had concurrent feedback after each case. Another 29 observers completed the ultrasound task on CD-ROM with no feedback initially. Results are thus presented for 77 observers without feedback and 23 observers with immediate feedback.

[b] Difference between ultrasound performance characteristics for the group of 77 observers without feedback and the 23 with feedback.

[c] Only 71 observers completed the mammographic masses task because it was not required of the latter 29 observers. Sonographic images of 13 of these 23 cases were also provided to observers.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 3

Interpretive Performance of the Subset of 35 Experienced Breast Imagers Who Completed the Ultrasound Tasks Both Without Feedback and With Feedback After Each Case

Examination Presented to Observer	Mean (95% CI)		
	Sensitivity	Specificity	Area Under the Curve
Ultrasound without feedback	0.890 (0.865 to 0.915)	0.620 (0.591 to 0.650)	0.817 (0.805 to 0.830)
Ultrasound with feedback	0.929 (0.912 to 0.946)	0.607 (0.581 to 0.632)	0.840 (0.830 to 0.850)
Difference with feedback	0.039 (0.025 to 0.053)	−0.014 (−0.035 to 0.007)	0.022 (0.012 to 0.032)
p[a]	< 0.0001	0 .19	< 0.0001
Mammographic masses and asymmetries[b]	0.99 (0.98 to 1.00)	0.72 (0.69 to 0.76)	0.93 (0.92 to 0.94)
Mammographic calcifications	0.90 (0.88 to 0.93)	0.81 (0.79 to 0.84)	0.91 (0.90 to 0.93)

[a] *p* value (paired Student *t* test) that there is no difference in ultrasound with feedback compared with ultrasound without feedback.

[b] Sonographic images of 13 of these 23 cases were also provided to observers.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 4

Summary of Area Under the Curve (AUC) Values by Experience Variables for 35 Observers Without Feedback and With Feedback After Each of 70 Breast Ultrasound Cases

	Without Feedback		With Feedback	
Characteristic	**AUC**	**CI**	**AUC**	**CI**
Across all demographics	0.813	0.797 to 0.829	0.840	0.825 to 0.854
Ability to fill out forms				
Obsessive compulsive	0.817	0.799 to 0.835	0.844	0.828 to 0.861
My research associate does everything	0.803	0.775 to 0.832	0.832	0.807 to 0.857
Don't count on it	0.810	0.764 to 0.856	0.831	0.789 to 0.872
Years in breast imaging				
< 2 years	0.763	0.701 to 0.826	0.797	0.741 to 0.853
2–5 years	0.799	0.769 to 0.829	0.834	0.808 to 0.860
6–10 years	0.840	0.807 to 0.872	0.862	0.833 to 0.891
> 10 years	0.816	0.797 to 0.836	0.841	0.823 to 0.859
% Time in clinical breast imaging				
< 25	0.759	0.706 to 0.811	0.798	0.752 to 0.844
25–49	0. 811	0.765 to 0.857	0.828	0.786 to 0.869
50–74	0.809	0.776 to 0.842	0.848	0.819 to 0.876
75–100	0.821	0.803 to 0.839	0.845	0.829 to 0.861
Who routinely performs breast ultrasound?				
Technologist	0.680	0.608 to 0.752	0.743	0.680 to 0.806
Myself	0.792	0.765 to 0.818	0.827	0.804 to 0.850
Resident then myself	0.793	0.745 to 0.840	0.808	0.764 to 0.852
Technologist then myself	0.836	0.816 to 0.856	0.857	0.839 to 0.875
Fellow then myself	0.867	0.835 to 0.899	0.885	0.856 to 0.913
No. of mammograms per week				
0–99	0.772	0.732 to 0.812	0.822	0.788 to 0.855
100–149	0.805	0.774 to 0.836	0.840	0.813 to 0.867
150–199	0.819	0.794 to 0.843	0.843	0.821 to 0.865
200–299	0.841	0.817 to 0.864	0.853	0.830 to 0.875
300–399	0.727	0.632 to 0.821	0.739	0.650 to 0.828
400–499	0.812	0.734 to 0.890	0.873	0.815 to 0.932
No. of breast ultrasound per week				
< 20	0.796	0.755 to 0.838	0.828	0.791 to 0.864
21–39	0.798	0.775 to 0.821	0.828	0.808 to 0.849
40–69	0.830	0.810 to 0.851	0.854	0.836 to 0.872
Whole-breast ultrasound				
Never	0.804	0.775 to 0.834	0.829	0.803 to 0.856
In patients with new cancer	0.850	0.826 to 0.874	0.860	0.838 to 0.883
Diagnostic and screening	0.798	0.777 to 0.819	0.834	0.816 to 0.853

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE 5

Kappa Values for BI-RADS Ultrasound Features and Assessments for 35 Trained Observers Without Feedback and Then With Feedback for 70 Proven Cases

Descriptor	No. of Cases[a]	κ (SE)[b]	κ (SE) for Descriptors Without Feedback	Change in κ With Feedback	*p*
Special case	70	0.579 (0.026)		0.016	0.10
No mass	7		0.577 (0.035)	0.035	0.07
Simple cyst	3		0.247 (0.053)	0.044	0.13
Complicated cyst	7		0.511 (0.034)	−0.031	0.20
Clustered microcysts	3		0.765 (0.037)	0.041	0.14
Intraductal mass	3		0.526 (0.029)	0.008	0.60
Mass in or on skin	2		0.578 (0.050)	0.002	0.95
Lymph node	3		0.856 (0.040)	0.023	0.25
Not a special case	42		0.565 (0.028)	0.008	0.52
Shape	42	0.585 (0.020)		−0.027	0.06
Oval or gently lobulated	19		0.590 (0.021)	−0.033	0.05
Round	2		0.307 (0.041)	−0.030	0.36
Irregular	21		0.672 (0.022)	−0.022	0.15
Margins	42	0.513 (0.022)		0.003	0.89
Circumscribed	14		0.543 (0.022)	0.001	0.97
Not circumscribed	28		0.509 (0.023)	0.002	0.94
Orientation	42	0.461 (0.021)		−0.036	0.03[c]
Parallel	33		0.457 (0.021)	−0.033	0.03[c]
Not parallel	9		0.483 (0.022)	−0.042	0.02[c]
Echogenicity	42	0.412 (0.021)		0.007	0.43
Anechoic	0		NA	NA	
Hyperechoic	5		0.774 (0.039)	−0.001	0.97
Complex	4		0.446 (0.059)	0	> 0.99
Hypoechoic	23		0.448 (0.026)	−0.005	0.75
Isoechoic	6		0.279 (0.038)	0.058	0 .10
Mixed hyper- and hypoechoic	4		0.210 (0.028)	−0.001	0.95

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Descriptor	No. of Cases[a]	κ (SE)[b]	κ (SE) for Descriptors Without Feedback	Change in κ With Feedback	*p*
Posterior features	42	0.637 (0.020)		0.016	0.32
None	17		0.638 (0.025)	0.016	0.36
Enhancement	16		0.609 (0.029)	0.011	0.57
Shadowing	9		0.843 (0.019)	0.022	0.17
Calcifications	70	0.515 (0.018)		0.031	0.01[c]
None	55		0.640 (0.017)	0.029	0.03[c]
Macrocalcifications (≥ 0.5 mm)	4		0.428 (0.040)	0.030	0.34
Microcalcifications in a mass	7		0.377 (0.027)	0.044	0.04[c]
Microcalcifications out of a mass	4		0.384 (0.044)	0.017	0.43
Final assessment	70	0.462 (0.019)[d]		0.078	< 0.0001[c]
1	4		0.591 (0.030)	0.053	0.06
2	10		0.374 (0.026)	0.045	0.02[c]
3	15		0.320 (0.030)	0.069	0.01[c]
4	32		0.518 (0.022)	0.106	< 0.0001[c]
4A	9		0.197(0.020)	0.054	0.03[c]
4B	16		0.170 (0.021)	0.069	0.01[c]
4C	7		0.123 (0.028)	0.031	0.27
5	9		0.524 (0.048)	0.100	< 0.0001[c]
Grouped final assessments	70	0.525 (0.017)		0.068	< 0.0001[c]
1 or 2	14		0.489 (0.021)	0.058	< 0.01[c]
3	15		0.320 (0.030)	0.069	0.01[c]
4A, 4B, 4C, or 5	41		0.676 (0.016)	0.076	< 0.0001[c]

Note—NA = not applicable, no entries.

[a] Number of cases so described by expert consensus.

[b] Values presented are without feedback.

[c] Indicates $p < 0.05$.

AJR Am J Roentgenol. Author manuscript; available in PMC 2014 January 15.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

[d] Kappa value for final assessments with category 4 grouped was 0.462 before feedback and 0.541 after feedback. With category 4 subdivided into 4A, 4B, and 4C, overall kappa value for final assessments before feedback was 0 314, increasing to 0.370 after feedback ($p = 0.0001$).