Published in final edited form as:
Acad Radiol. 2013 July ; 20(7): 847–853. doi:10.1016/j.acra.2013.02.007.

Interreader Scoring Variability in an Observer Study Using Dual-Modality Imaging for Breast Cancer Detection in Women with Dense Breasts

Karen Drukker, PhD, **Karla J. Horsch, PhD**, **Lorenzo L. Pesce, PhD**, and **Maryellen L. Giger, PhD**
Department of Radiology, MC2026, University of Chicago, 5841 S Maryland Ave, Chicago, IL 60637 (K.D., K.J.H., M.L.G.); and Computation Institute, University of Chicago, Searle Chemistry Laboratory, 5735 South Ellis Avenue, Chicago, IL 60637 (L.L.P.)

Abstract

Rationale and Objectives—To evaluate variability in the clinical assessment of breast images, we evaluated scoring behavior of radiologists in a retrospective reader study combining x-ray mammography (XRM) and three-dimensional automated breast ultrasound (ABUS) for breast cancer detection in women with dense breasts.

Methods—The study involved 17 breast radiologists in a sequential study design with readers first interpreting XRM-alone followed by an interpretation of combined XRM + ABUS. Each interpretation included a forced Breast Imaging Reporting and Data System scale and a likelihood that the woman had breast cancer. The analysis included 164 asymptomatic patients, including 31 breast cancer patients, with dense breasts and a negative screening XRM. Of interest were interreader scoring variability for XRM-alone, XRM + ABUS, and the sequential effect. In addition, a simulated double reading by pairs of readers of XRM + ABUS was investigated. Performance analysis included receiver operating characteristic analysis, percentile analysis, and κ statistics. Bootstrapping was used to determine statistical significance.

Results—The median change in area under the receiver operating characteristic curve after ABUS interpretation was 0.12 (range 0.04–0.19). Reader agreement was fair with the median interreader κ being 0.26 (0.05–0.48) for XRM-alone and 0.34 (0.11–0.55) for XRM + ABUS (95% confidence interval for the difference in κ, 0.06–0.11). Simulated double reading of XRM + ABUS demonstrated tradeoffs in sensitivity and specificity, but conservative simulated double reading resulted in a significant improvement in both sensitivity (16.7%) and specificity (7.6%) with respect to XRM-alone.

Conclusion—A modest, but statistically significant, increase in interreader agreement was observed after interpretation of ABUS.

Keywords

Reader variability; observer study; 3D ultrasound; mammography

Address correspondence to: K.D. kdrukker@uchicago.edu.

Conflict of interest statement: M.L.G. is a stockholder in R2 Technology/Hologic, co-founder and equity holder of Quantitative Insights, Inc., and receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi and Toshiba. K.D. received royalties from Hologic. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

Breast imaging methods for the early detection and diagnosis of cancer continue to evolve. Mammography, as the primary screening modality, allows for the early detection of nonpalpable breast cancers and has been shown to reduce breast cancer mortality (1,2). Although the overall sensitivity of mammography is 70% to 90%, the sensitivity can range from 30% to 98% depending on whether the breast consists mostly of extremely dense glandular tissue or contains mostly fat (3). Tumors diagnosed in women with dense breast tissue are currently usually larger and of higher histological grade with a greater likelihood of lymph node metastases, resulting in poorer prognosis (4,5). Moreover, the presence of dense breast tissue is associated with an elevated risk for breast cancer with the relative risk more than 5 times greater for women with the most dense breast tissue than for women without dense breast tissue (6–8). Nearly 40% of women in the United States have dense breasts and the poor sensitivity of mammography in women with Breast Imaging Reporting and Data System (BI-RADS) composition/density 3 or 4 has resulted in several states passing legislation requiring women be informed of the breast density and the possible need for additional screening with modalities other than mammography (9).

Based on initial clinical studies using conventional ultrasound (10–14), the addition of automated breast ultrasound (ABUS) to screening x-ray mammography (XRM) is expected to yield a benefit to patients with dense breast tissue by providing earlier detection of breast cancers that might be missed by mammography. Hence, a multireader multicase (MRMC) clinical reader study was conducted evaluating the use of ABUS in conjunction with XRM in the breast cancer screening of women with dense breasts and a negative screening XRM (tumor BI-RADS assessment category 1 or 2) (15). That study involved both semicontinuous reader scoring data (the likelihood of malignancy) and two-category data (cancer versus non-cancer) (16). The reader-assigned likelihoods of malignancy served as the decision variables in an MRMC receiver operating characteristic (ROC) analysis (17–19). The BI-RADS assessment categories were used to determine sensitivity and specificity given a predetermined cutoff for the distinction between patients with and without cancer. A statistically significant increase in the overall area under the ROC curve was obtained as well as a statistically significant increase in sensitivity, while a slight decline in specificity failed to reach statistical significance (Table 1) (15) (and Giger et al, manuscript in preparation). In contrast, the work presented here focused more on individual readers and cases and analyzed (1) the reader scoring behavior of the participating radiologists, (2) agreement (or lack thereof) between readers, (3) the impact of the consecutive reading with two modalities (XRM and ABUS in this instance), and (4) the potential of improvement from double reading by pairs of readers. The latter was done through simulations using the reader data. It is important to note that it was not our intent to critique individual radiologists or to determine which radiologist was "better."

MATERIALS AND METHODS

Study Design

The study mimicked the clinical use of ABUS as a potential adjunct screening modality and involved an institutional review board (IRB)-approved, sequential-design, MRMC ROC reader study (15,19), which included a cancer-enriched set of screening XRMs and ABUS from asymptomatic women with breast density BI-RADS 3 or 4. For each patient ("case"), each reader interpreted and initially scored the screening XRM-alone, the "XRM-alone" condition. Immediately after viewing and assessing the XRM-alone, each reader's ratings were locked and they then interpreted the XRM and ABUS images combined (the "XRM + ABUS" condition). Each reader indicated any potential lesion, gave it a description and indicated a likelihood of it being cancerous, and provided an initial BI-RADS assessment of 0 (recall), 1 (negative), or 2 (benign). In the event of an initial BI-RADS score of 0, a forced BI-RADS 3, 4a, 4b, 4c, and 5 rating was given. Finally, each reader gave the case a likelihood that the woman had cancer (likelihood of malignancy) using a 0% to 100% scale. The forced BI-RADS assessment ratings and likelihood of malignancy ratings were used in the work presented here.

Screening full-field digital mammograms were displayed on a Hologic SecurView DX FFDM viewer (Hologic Inc, Bedford, MA) and ABUS images were displayed on a U-Systems Somo•VIEWer Workstation (U-Systems Inc, a GE Healthcare Company, Sunnyvale, CA).

Data Set

The data set was collected in an HIPAA-compliant manner at 13 clinical sites across the United States for a reader study (15) overseen by our IRB. The data collection was performed under a separate multicenter protocol, and institution-specific IRBs governed patient enrollment and informed consent. The study discussed here had access to deidentified data only. The data set consisted of digital screening mammograms and three-dimensional ABUS images obtained on the Somo•v® system (U-Systems, Inc, a GE Healthcare Company) for asymptomatic women with a mammogram-assigned BI-RADS composition/density category 3 or 4. The original data set contained images from 200 patients. The reader study analysis (15) and the study discussed here, however, focused on those patients with a clinical screening XRM-assigned BI-RADS assessment category 1 (negative) or 2 (normal with benign findings), limiting the number of patients (N_P) to 164 (31 patients with breast cancer and 133 healthy patients [ie, “normals”]).

The cancer cases were biopsy-proved cancers that were diagnosed as a result of any workup of the patient at 365 days or earlier after the original digital screening XRM. Two board-certified gold-standard radiologists, each reading more than 2,000 mammograms per year in their practice and with experience in ABUS, independently reviewed all cancer cases (patients). The median tumor size was 12 mm.

Readers

A total (N_R) of 17 breast radiologists participated in the reader study, which was conducted over several days, allowing for ample training of the radiologists in the interpretation of ABUS images (15). Readers were Mammography Quality Standards Act (MQSA)-qualified radiologists who were either fellowship trained in breast imaging and/or had 10 years of experience in breast imaging in a practice that was at least 70% breast imaging. The participating radiologists ranged in experience from 2 to 18 years in breast imaging (median 12 years), from 1,850 to 14,600 mammograms reviewed annually (median 4,180) and from 603 to 5,000 hand-held ultrasound examinations reviewed annually (median 825). Nine of the 17 readers were breast imaging fellowship trained. Six radiologists were in private practice, four practiced at a community hospital, and seven practiced at an academic teaching hospital.

Analyses

The current study was only concerned with the readers’ overall by-patient assessment, using the reader-assigned likelihoods of malignancy and BI-RADS categories regardless of correct identification of tumor location.

ROC analysis—The area under the ROC curve (AUC) was the figure of merit in each reader’s performance assessment. The proper binormal model (20) was used to determine the AUC_{XRM} and the $AUC_{XRM+ABUS}$ for each reader. The decision variable in this analysis was the reader-assigned likelihood of malignancy (LOM). Please note again that in the current study we were interested in individual readers and that an MRMC ROC analysis (19,20) of this reader study has been presented elsewhere (15).

Likelihood of malignancy—Percentile analysis and box plots were used to assess the impact of the XRM + ABUS interpretation on the reader-assigned LOM. For each reader, the change in LOM, $\Delta(\text{LOM}) = \text{LOM}_{\text{XRM + ABUS}} - \text{LOM}_{\text{XRM}}$, for all cases was determined and presented separately for actually normal patients and patients with breast cancer. The median, 25th and 75th percentiles, and outliers of the change in LOM were calculated.

Forced BI-RADS assessment—The reader-assigned (forced) BI-RADS score was used in the determination of whether a patient was deemed to have breast cancer. A BI-RADS cutoff of 4a was used for this purpose (ie, a score of 4a or higher indicated breast cancer). The number of patients diagnosed by each reader as having breast cancer was recorded for XRM-alone and for XRM + ABUS. Of interest was the impact of the ABUS interpretation on the number of (unnecessary) biopsies and (additional) cancers found (if this were a real-life situation): How many additional cancers does an “average reader” find after interpreting ABUS, and at what cost of additional unnecessary biopsies? How many additional readers might interpret an “average case” as being cancerous after interpreting ABUS? Note that the BI-RADS scores were not used in ROC analysis. Reader variability was explored visually by displaying the reader-assigned forced BI-RADS scores for all readers and all cases in a color-coded fashion.

Cohen κ—Reader agreement was assessed quantitatively through the calculation of κ statistics (21,22). A BI-RADS cutoff of 4a was once again used to be indicative of cancer, and κ was calculated for each pair of readers. Percentile analysis and box plots were used to assess κ for all cases, for normal cases, for cancer cases, for both XRM-alone and XRM + ABUS conditions. The overall interreader κ as a measure for agreement for the $N_R = 17$ readers was the median of the $N_{RP} = N_R(N_R - 1)/2 = 136$ pairwise values for κ. The statistical significance of the change in overall κ for the given set of readers was assessed through a bootstrap analysis (23,24) with N = 10,000 iterations. In each bootstrap iteration, $N_{RP} = 136$ reader *pairs* and $N_p = 164$ cases were randomly selected with replacement. The pairwise interreader κ values were calculated for the XRM-alone and XRM + ABUS conditions. The change in median interreader κ, $\Delta(\kappa) = \kappa_{\text{XRM + ABUS}} - \kappa_{\text{XRM}}$, was recorded each iteration, and the 95% confidence interval (CI) for $\Delta(\kappa)$ was calculated after the N = 10,000 bootstrap iterations were completed. In this report, all 95% CIs were calculated from the 2.5th–97.5th percentiles for the entity of interest.

Double reading—“Double reading” by two readers is known for its potential to improve diagnostic accuracy (eg, in the interpretation of mammograms) (25,26). Here, the impact on diagnostic accuracy through double reading was investigated in simulations by a posteriori pairing the $N_R = 17$ readers into $N_{RP} = 136$ pairs for the XRM + ABUS condition. For this purpose, two approaches were considered: An aggressive approach aimed at improving sensitivity and a conservative approach aimed at improving specificity. In the former, a case was considered cancerous if one or both readers diagnosed it as such (BI-RADS 4a or higher) and considered to be normal only if both readers diagnosed it as normal. In the latter conservative approach, a case was considered to be cancerous only if both readers diagnosed it as such and considered to be normal if one or both readers diagnosed it as normal. Sensitivity and specificity were calculated for each reader pair, and N = 10,000 bootstrap iterations were again used to assess statistical significance with respect to single reading conditions.

We also a posteriori constructed a consensus diagnosis of the cohort of readers for both XRM-alone and XRM + ABUS conditions. The consensus diagnosis was the majority opinion (ie, nine or more radiologists) on whether a given case was BI-RADS 4a or higher.

In all plots, readers are ordered by their AUC value for XRM-alone and all cases ordered by the median Δ(LOM) after separation into actually positive ("cancers") and actually negative cases ("normals"). Ranges are indicated as (x–y) and 95% CIs as [x; y].

RESULTS

ROC Analysis

All readers' AUC values were higher for the XRM + ABUS condition than for the XRM-alone condition (Fig 1) as shown by the histogram of the AUC values and that of the change in AUC, Δ(AUC). Remember that when referred to by number, readers are ordered throughout this report by increasing $\mathrm{AUC}_{\mathrm{XRM}}$ value. The error bars are ± standard error as given by the proper binormal ROC model (20).

Likelihood of Malignancy

The change in reader-assigned LOM was much larger for the cancer cases than for the normal cases (Fig 2). Each reader obtained a median change in LOM of 0.10 (range 0.00–0.47) for the cancer cases. Changes in reader scoring tended to have smaller magnitudes for the normal cases, but both positive and negative changes in individual case LOMs were observed, with all readers obtaining either zero or small negative change, median 0.00 (range −0.01 to 0.00), in the median normal case LOM.

Forced BI-RADS Assessment

The sequential effect on cancer detection (based on a BI-RADS cutoff of 4a) of interpreting XRM + ABUS after XRM-alone was heterogeneous with most, but not all, readers correctly identifying more cancer cases and most, but not all, readers incorrectly recommending additional work-up for normal cases (Fig 3). A typical reader erroneously identified seven more normal cases as cancer under the XRM + ABUS condition, while finding an additional 10 cancers (Fig 3a). Reader assessment of a typical normal case remained unchanged, while a typical cancer case was identified by five more readers (Fig 3c).

Importantly, the readers differed in which cases they misdiagnosed (Fig 4). Under the XRM-alone condition, there was only a single cancer case of the 31 (3%) correctly identified by all readers as such and 42 of the 133 normal cases (32%) were interpreted by all readers as normal. Under the XRM + ABUS condition, these numbers were 5 of the 31 cancer cases (16%) and 23 of the 133 normal cases (20%), respectively.

Cohen κ

The κ statistics indicated fair reader agreement (21) for this specific set of cases and readers with a median interreader κ of 0.26 (range 0.05–0.48 for individual radiologist pairs) for XRM-alone and of 0.34 (range 0.11–0.55) for XRM + ABUS (Fig 5). Generalizing by bootstrapping reader *pairs* and cases, the difference in the median interreader κ, $\Delta\kappa = 0.09$, was statistically significant with a 95% CI for $\Delta\kappa$ of [0.06; 0.11]. On dividing cases into normal cases and cancer cases in the calculation of κ (Fig 5b and c), it appeared that κ was higher for the cancer cases. A size-effect (31 cancer cases versus 133 normal cases) may have been largely responsible for this, however, with the 95% CI for the median interreader κ for 1,000 sets of 31 randomly selected normal cases being [−0.04; 0.36] for the XRM-alone condition (which includes the median value of $\kappa = 0.36$ observed for the cancer cases for the XRM-alone condition).

Double Reading

For the aggressive double reading, pairwise readers approach, the median sensitivity for the reader pairs was 71.0% and the median specificity was 79.0%. For the conservative approach, those values were 48.4% and 97.0%, respectively (Table 2). The 95% CIs for the changes in sensitivity and specificity with respect to single reading conditions all excluded zero, indicating statistical significance.

The consensus diagnosis from the 17 radiologists (agreement of 9 or more radiologists) correctly diagnosed 7 cancer cases (23%) and 129 normal cases (97%) for the XRM-alone condition. For XRM + ABUS, 2 additional recommendations for unnecessary work-up ensued, reducing the consensus specificity to 95% (127/133), while 11 more cancers were found, increasing the consensus sensitivity to 55% (18/31).

DISCUSSION

This report presents an analysis of reader scoring and inter-reader agreement/variability using rating data from a realistic clinical reader study involving XRM and ABUS images. It is important to note that the data set was considered to be difficult because it contained images only of women with dense breasts (BI-RADS density 3 or 4) and all cancer cases were missed on XRM at the clinical collection site. It was therefore not surprising that the reader performance for the XRM-alone condition was low. As previously noted, an MRMC ROC analysis of this reader study has been presented elsewhere (15), which found a statistically significant increase in AUC and sensitivity, while the slight decrease in specificity failed to reach statistical significance. Here, we focused on individual reader scoring behavior, reader agreement, and simulated the potential impact of double reading by pairs of readers.

One might have expected a more substantial improvement in reader agreement for the XRM + ABUS condition with respect to the XRM-alone condition, since overall the readers demonstrated an improvement in diagnosis, but it is important to note that reader agreement does not necessarily imply a high performance and vice versa. The sequential effect of interpreting XRM + ABUS after XRM-alone varied by reader, and although the increase in interreader κ was statistically significant, the overall agreement remained fair.

The apparent persistence of reader variability, however, is not necessarily detrimental since disagreement facilitates opportunities for diagnostic improvement through joint decision making, such as double reading by pairs of readers (27), which is impossible if readers also agree on misdiagnosed cases. For the aggressive double reading approach, the median sensitivity was equal to the highest value obtained by an individual reader, while the median specificity decreased with respect to the single reading approach but was still within the range of values obtained for individual readers. For the conservative double reading approach, the median specificity was equal to the highest value obtained by an individual reader, but at a considerable loss in sensitivity. These results illustrate that even in dual-modality imaging studies, double reading has potential to improve diagnostic accuracy (ie, sensitivity or specificity) depending on what is clinically of interest. We demonstrated that a fair interreader agreement may allow for substantial improvements in sensitivity or specificity under double reading conditions. Perhaps most importantly, the simulated conservative approach to double reading for XRM + ABUS simultaneously improved sensitivity *and* specificity (boldface in Table 2) over single reading of XRM-alone, indicating that potential loss in specificity upon introduction of a new adjunct imaging modality (5) may be counterbalanced by double reading. While it is usually not clinically feasible to obtain a consensus opinion from 17 radiologists, the consensus specificity reported here (95%) was substantially higher than the overall specificity of 84% reported in the MRMC analysis of variance (18) of this reader study (5), while the sensitivities were more comparable at 55% versus 58%, respectively. A potential limitation to our double reading and consensus simulations is that the paired and consensus scores were constructed a posteriori with the readers' scores completely independent from each other, which may not reflect clinical practice.

In conclusion, results from a retrospective clinical reader study involving the interpretation of XRM and ABUS for breast cancer detection in women with dense breasts have demonstrated both statistically significant increases in detection performance and statistically significant increases in interreader agreement after interpretation of ABUS.

Acknowledgments

The authors are thankful to Charles E. Metz, PhD, for helpful discussions in the early stages of this project.

Funding: This work was supported in part by DOE grant DE-FG02-08ER6478, NIH P50-CA125183, NIH S10 RR021039 and was supported in part by research funding from U-Systems Inc, a GE Healthcare company.

References

1. Bock K, Borisch B, Cawson J, et al. Effect of population-based screening on breast cancer mortality. Lancet. 2011; 378(9805):1775–1776. [PubMed: 22098846]
2. Tabar 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. [PubMed: 2858707]
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(13): 1081–1087. [PubMed: 10880551]
4. Ghosh K, Brandt KR, Sellers TA, et al. Association of mammographic density with the pathology of subsequent breast cancer among post-menopausal women. Cancer Epidemiol Biomarkers Prev. 2008; 17(4):872–879. [PubMed: 18398028]
5. Chiu SY, Duffy S, Yen AM, et al. Effect of baseline breast density on breast cancer incidence, stage, mortality, and screening parameters: 25-year follow-up of a Swedish mammographic screening. Cancer Epidemiol Biomarkers Prev. 2010; 19(5):1219–1228. [PubMed: 20406961]
6. Boyd NF, Martin LJ, Yaffe MJ, et al. Mammographic density and breast cancer risk: current understanding and future prospects. Breast Cancer Res. 2011; 13(6):223. [PubMed: 22114898]
7. Martin LJ, Melnichouk O, Guo H, et al. Family history, mammographic density, and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2010; 19(2):456–463. [PubMed: 20142244]
8. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007; 356(3):227–236. [PubMed: 17229950]
9. Are You Dense Advocacy. Are You Dense Advocacy, Inc; http://www.areyoudenseadvocacy.org/dense/ [Accessed 4/1/2013]
10. Hooley RJ, Greenberg KL, Stackhouse RM, et al. Screening US in patients with mammographically dense breasts: initial experience with Connecticut Public Act 09-41. Radiology. 2012; 265(1):59–69. [PubMed: 22723501]
11. Corsetti V, Houssami N, Ghirardi M, et al. Evidence of the effect of adjunct ultrasound screening in women with mammography-negative dense breasts: interval breast cancers at 1 year follow-up. Eur J Cancer. 2011; 47(7):1021–1026. [PubMed: 21211962]
12. Corsetti V, Houssami N, Ferrari A, et al. Breast screening with ultrasound in women with mammography-negative dense breasts: evidence on incremental cancer detection and false positives, and associated cost. Eur J Cancer. 2008; 44(4):539–544. [PubMed: 18267357]
13. Corsetti V, Ferrari A, Ghirardi M, et al. Role of ultrasonography in detecting mammographically occult breast carcinoma in women with dense breasts. Radiol Med. 2006; 111(3):440–448. [PubMed: 16683089]

14. 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(18):2151–2163. [PubMed: 18477782]

15. Giger, ML.; Miller, DP.; Bancroft Brown, J., et al. Clinical reader study examining the performance of mammography and automated breast ultrasound in breast cancer screening. 98th Assembly and Annual Meeting of Radiological Society of North America; 2012.

16. Samuelson F, Gallas BD, Myers KJ, et al. The importance of ROC data. Acad Radiol. 2011; 18(2):257–258. author reply 9–61. [PubMed: 21232688]

17. 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. [PubMed: 18423323]

18. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. Acad Radiol. 1997; 4(4):298–303. [PubMed: 9110028]

19. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol. 1992; 27(9):723–731. [PubMed: 1399456]

20. Pesce LL, Metz CE. Reliable and computationally efficient maximum-likelihood estimation of "proper" binormal ROC curves. Acad Radiol. 2007; 14(7):814–829. [PubMed: 17574132]

21. Kundel HL, Polansky M. Measurement of observer agreement. Radiology. 2003; 228(2):303–308. [PubMed: 12819342]

22. Berry CC. The kappa statistic. JAMA. 1992; 268(18):2513–2514. [PubMed: 1404812]

23. Efron, B.; Tibshirani, R. An Introduction to the Bootstrap. Chapman & Hall; 1993.

24. Gruszauskas NP, Drukker K, Giger ML, et al. Performance of breast ultrasound computer-aided diagnosis: dependence on image selection. Acad Radiol. 2008; 15(10):1234–1245. [PubMed: 18790394]

25. Waldmann A, Kapsimalakou S, Katalinic A, et al. Benefits of the quality assured double and arbitration reading of mammograms in the early diagnosis of breast cancer in symptomatic women. Eur Radiol. 2012; 22(5):1014–1022. [PubMed: 22095439]

26. Duijm LE, Groenewoud JH, Hendriks JH, et al. Independent double reading of screening mammograms in The Netherlands: effect of arbitration following reader disagreements. Radiology. 2004; 231(2):564–570. [PubMed: 15044742]

27. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol. 1996; 3(11):891–897. [PubMed: 8959178]

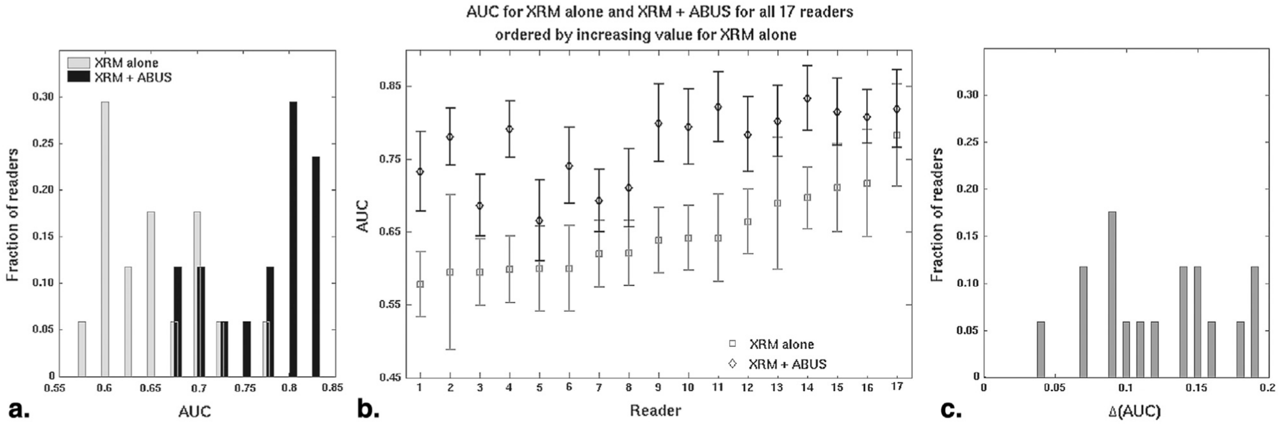

Figure 1.

Receiver operating characteristic (ROC) performance assessment of the 17 readers for the conditions x-ray mammography (XRM)-alone and XRM + three-dimensional automated breast ultrasound (ABUS), as **(a)** histogram of the area under the ROC curve (AUC) values, **(b)** the AUC values per reader, and **(c)** histogram of the change in AUC, Δ(AUC). When referred to by number, readers are ordered throughout this report by increasing AUC_{XRM} value. The error bars are ± standard error as given by the proper binormal ROC model.

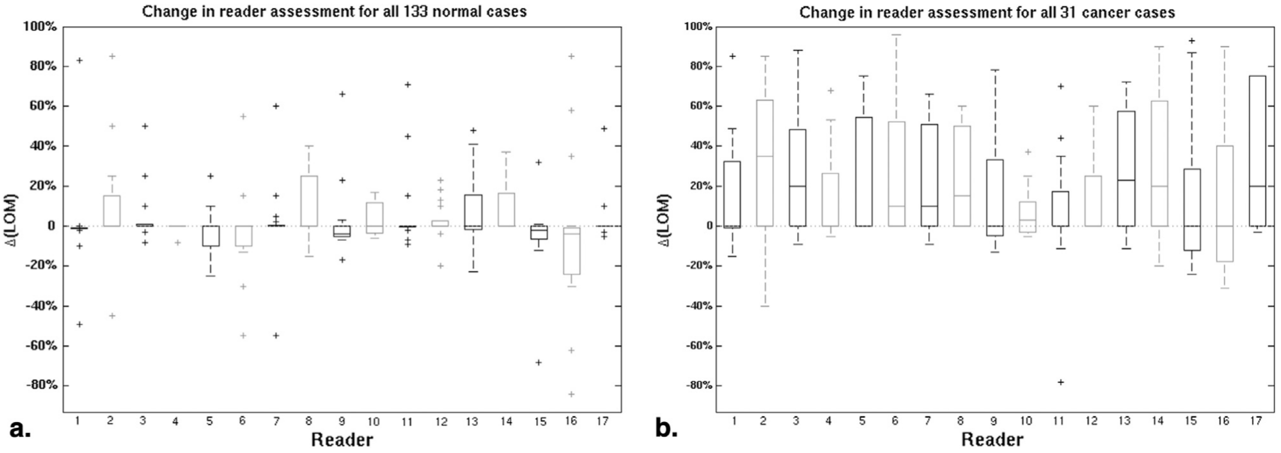

Figure 2.

The *change* in reader-assigned likelihood of malignancy, Δ(LOM), between the x-ray mammography (XRM)-alone and XRM + three-dimensional automated breast ultrasound (ABUS) conditions for the **(a)** actually normal cases and **(b)** actually cancerous cases. In all box plots in this report, the bottom and top of each box denote the 25th and 75th percentiles, respectively, while the horizontal line within denotes the median value. Whiskers extend to mark the range in values not considered outliers, while individual outliers are marked with a "+."

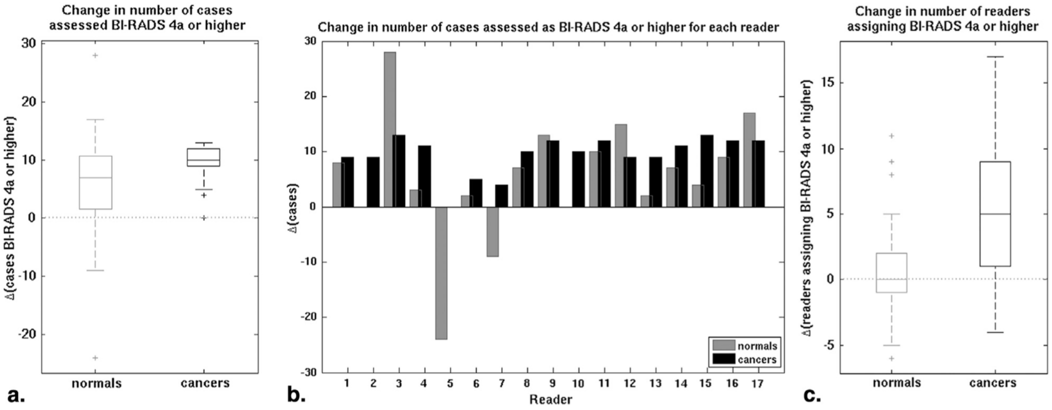

Figure 3.

The impact of three-dimensional automated breast ultrasound (ABUS) on the identification of breast cancer cases as **(a)** box plot of the *change* in the number of cases assigned a Breast Imaging Reporting and Data System (BI-RADS) category 4a or higher by a reader, **(b)** the *change* in number of cases assigned a BI-RADS 4a or higher per reader, and **(c)** box plot of the *change* in the number of readers assessing a case as BI-RADS 4a or higher.

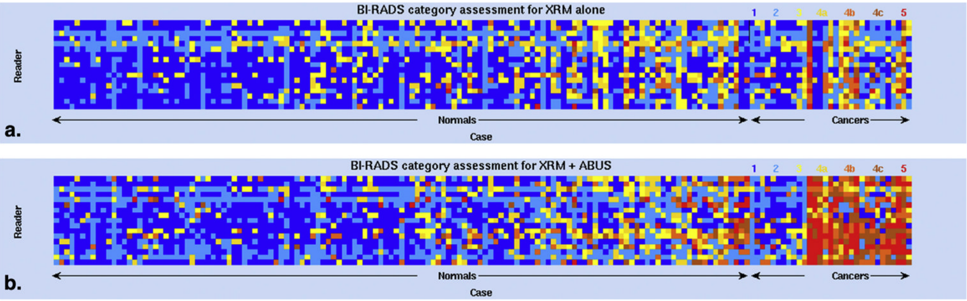

Figure 4.

The impact of three-dimensional automated breast ultrasound (ABUS) on the identification of breast cancer illustrated by color-coded Breast Imaging Reporting and Data System (BI-RADS) assessment categories for all cases and all readers. BI-RADS categories vary as indicated from blue (category 1), through shades of yellow and orange, to red (category 5). Cases are divided into actually normal cases and actually cancerous cases and then ordered by increasing change in reader-assigned likelihood of malignancy (LOM). Readers are again ordered by increasing $\mathrm{AUC}_{\mathrm{XRM}}$.

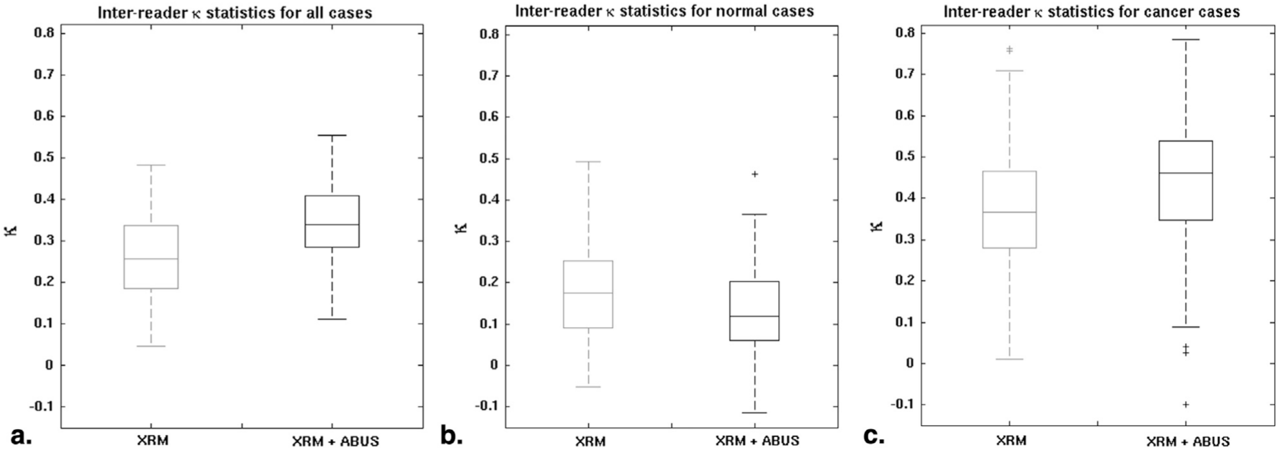

Figure 5.

Interreader agreement as indicated by the interreader κ. The box plots are of the $N_R(N_R - 1)/2$, with $N_R = 17$ readers, κ values for the x-ray mammography (XRM)-alone and XRM + three-dimensional automated breast ultrasound (ABUS) conditions for **(a)** all cases, **(b)** the normal cases, and **(c)** the cancer cases.

TABLE 1

Summary of Multicase Multireader Analysis (18) Results Obtained in (15) That Are Relevant to the Work Presented Here: AUC Values (with Standard Error in Parentheses), Sensitivity, and Specificity

	XRM-Alone	XRM + ABUS	P Value
AUC	0.65 (0.033)	0.77 (0.035)	<.001
Overall sensitivity [*]	27.1%	57.7% [†]	<.001
Overall specificity [*]	88.1%	84.0% [‡]	.86

ABUS, three-dimensional automated breast ultrasound; AUC, area under the ROC curve; XRM, x-ray mammography.

[*] Breast Imaging Reporting and Data System cutoff of 4a.

[†] Range for individual readers 32.3% to 71.0%.

[‡] Range for individual readers 67.8% to 97.0%.

TABLE 2

Overview of Sensitivities and Specificities in A Posteriori Simulated Double Reading Approaches (by Pairs of Radiologists) for XRM + ABUS (median [95% CI]) and the Changes with Respect to Single Reading Conditions

	XRM + ABUS Double Reading	Change wrt XRM + ABUS Single Reading	Change wrt XRM Single Reading
Aggressive			
Sensitivity	71.0% [61.3%; 77.4%]	13.8% [11.2%; 17.3%]	44.4% [35.2%; 56.3%]
Specificity	79.0% [64.7%; 90.2%]	−11.6% [−12.8%; − 10.5%]	−16.7% [−18.2%; −13.3%]
Conservative			
Sensitivity	48.4% [29.0%; 54.8%]	−13.9% [−17.2%; − 11.1%]	**16.7%** [9.1%; 26.1%]
Specificity	97.0% [95.4%; 99.3%]	11.6% [10.6%; 12.8%]	**7.6%** [6.0%; 9.2%]

ABUS, three-dimensional automated breast ultrasound; XRM, x-ray mammography.