Published in final edited form as:

AJR Am J Roentgenol. 2011 April ; 196(4): 971–981. doi:10.2214/AJR.10.5081.

Interpretation of Positron Emission Mammography and MRI by Experienced Breast Imaging Radiologists: Performance and Observer Reproducibility

Deepa Narayanan^{1,2}, Kathleen S. Madsen³, Judith E. Kalinyak¹, and Wendie A. Berg^{4,5}

¹Naviscan, San Diego, CA

³Certus International, Chesterfield, MO

⁴American College of Radiology Imaging Network, Lutherville, MD

Abstract

OBJECTIVE—In preparation for a multicenter trial of positron emission mammography (PEM) and MRI in women with newly diagnosed cancer, the two purposes of this study were to validate training of breast imagers in standardized interpretation of PEM and to validate performance of the same specialists interpreting MRI.

MATERIALS AND METHODS—A 2-hour didactic module was developed to train Mammography Quality Standards Act–qualified radiologist observers to interpret PEM images, consisting of a sample feature analysis lexicon analogous to BI-RADS and 12 sample cases. Observers were then asked to review separate interpretive skills tasks for PEM (49 breasts, 20 [41%] of which were malignant) and MRI (32 breasts, 11 [34%] of which were malignant), describe findings, and give assessments analogous to BI-RADS (category 1, 2, 3, 4A, 4B, 4C, or 5). Demographic experience variables were collected for 36 observers from 15 sites. Performance against histopathologic truth was determined, and interobserver agreement for classifying features and final assessments was evaluated using kappa statistics.

RESULTS—Across 36 observers, mean sensitivity, specificity, and area under the curve (AUC) for PEM were 96% (range, 75–100%), 84% (range, 66–97%), and 0.95 (range, 0.82–1.0), respectively. Mean sensitivity, specificity, and AUC for the MRI task were 82% (range, 45–100%), 67% (range, 38–91%), and 0.80 (range, 0.48–0.96), respectively. Interobserver agreement for PEM findings ranged from moderate to substantial, with kappa values of 0.57 for lesion type and 0.63 for final assessments.

© American Roentgen Ray Society

Address correspondence to D. Narayanan (narayanand@mail.nih.gov).

²Present address: National Cancer Institute, National Institutes of Health, 31 Center Dr, Rm 10A19, Bethesda, MD 20892-2580.

⁵Present address: Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA.

D. Narayanan was an employee of Naviscan and owns stock options in the company, J. E. Kalinyak is an employee of Naviscan, Inc. K. S. Madsen is a consultant statistician to Naviscan, Inc. with compensation based on fair market values and not tied to outcomes, and W. A. Berg is a paid consultant to Naviscan, Inc. with compensation based on the usual and customary rates and not tied to outcomes.

Presented at the 2007 RSNA Scientific Assembly.

CONCLUSION—With minimal training, experienced breast imagers showed high performance in interpreting PEM images. Performance in MRI interpretation by the same observers validated expected clinical practice.

Positron emission mammography (PEM) is an emerging molecular imaging technology that produces high-resolution tomographic 12-slice images of ^{18}F -FDG uptake in the breasts [1]. Indications for PEM include initial staging evaluation of patients with newly diagnosed cancer (i.e., determining extent of disease, but not including axillary node staging), distinguishing recurrent carcinoma from scar or restaging, and monitoring response to neoadjuvant chemotherapy. In early work, PEM showed a sensitivity of 90% and, importantly, high specificity of 86% in evaluating known breast cancer or suspicious lesions [1]. These results, together with further improvements in the PEM equipment and image display, prompted the development of a prospective multicenter trial to compare the performance and clinical utility of PEM to that of contrast-enhanced breast MRI in assessing disease extent in women with newly diagnosed cancer anticipating breast-conserving surgery [2].

Accurate interpretation of contrast-enhanced breast MRI is complex and requires experience [3]. PEM is a new technique, and most investigators in this multicenter trial had little or no prior experience in interpreting PEM. Even though even physicians with very little PEM experience showed high interpretive performance in previous work [1] to participate as investigators in the PEM-MRI multicenter comparison trial, breast imagers were required to successfully complete reader qualification tasks for both PEM and MRI.

The BI-RADS has proven to be successful in promoting consistent lesion description and management recommendations for mammography [4], ultrasound [5], and breast MRI [6]. For adoption of PEM or other molecular breast imaging techniques, it is similarly necessary that the terminology be easy to understand and consistently used by interpreting physicians. We developed a lexicon similar to BI-RADS for the standardized reporting for PEM findings such as lesion location, size, features, and qualitative and quantitative FDG uptake in background breast parenchyma and lesions [7]. A 2-hour self-paced tutorial in PEM image display, feature analysis, and interpretation based on this lexicon was developed, and each participating physician was required to complete this tutorial before performing the PEM interpretive skills task.

In preparation for a multicenter trial of PEM and MRI in women with newly diagnosed cancer, the two purposes of this study were to validate training of breast imagers in standardized interpretation of PEM and to validate performance of the same specialists interpreting MRI.

Materials and Methods

Observers

All participating radiologists had agreed to participate in an institutional review board–approved multicenter clinical trial to compare PEM and MRI for evaluating the extent of disease [2]. To qualify to interpret PEM and MRI examinations for the study, all radiologists

had to complete a self-paced PEM training module as well as both PEM and MRI interpretive skills tasks and agree to have their results analyzed. To be eligible to participate as observers, all radiologists were required to meet experience requirements of the Mammography Quality Standards Act (MQSA) for mammogram-interpreting physicians and to have interpreted at least 50 contrast-enhanced breast MRI examinations. Observers were recruited from 15 community hospitals and academic medical centers throughout the United States.

Demographic variables of observer experience were collected, including number of years in breast imaging; current percentage of time spent in clinical breast imaging; years spent interpreting breast MRI; and number of mammograms, breast ultrasounds, and breast MRI examinations interpreted per week. The total number of PEM scans interpreted by observers before the start of the study was recorded.

PEM Training Module

All eligible observers underwent a 2-hour didactic PowerPoint (Microsoft) presentation that included 12 proven cases, a sample feature analysis lexicon [7], and guidelines for interpreting PEM images. The module described the PEM imaging technique (with positioning analogous to mammography) and the standard 12-slice tomographic image display. Observers were shown how to draw regions of interest to obtain quantitative PEM FDG uptake values and were given guidance on utilizing the values in interpreting PEM images. The PEM lexicon [7] is based on the BI-RADS for MRI [6] and includes classification of the lesion type as a solitary focus (≤ 4 mm), multiple foci, mass, or nonmass uptake. Mass uptake is classified as round, oval, lobular, or irregular shape, whereas nonmass distribution is classified as scattered or diffuse uptake, focal area, regional uptake, linear or ductal uptake, or segmental uptake.

FDG uptake within a finding is judged qualitatively (homogeneous or heterogeneous) and quantitatively (with tumor-to-background standardized PEM uptake values). The lexicon also includes nomenclature for reporting location (side, quadrant, or clock-face), associated features or findings, lesion size, both quantitative and qualitative (none or minimal, mild, moderate, or intense) background parenchymal FDG uptake, and homogeneity of background FDG uptake (homogeneous, patchy/heterogeneous). The lexicon includes a 7-point assessment scale (equivalent to the expanded BI-RADS scales for mammography [4], ultrasound [5], and MRI [6]) for the imaging impression and associated management recommendation: 1, negative, routine follow-up; 2, benign, routine follow-up; 3, probably benign, short-interval follow-up; 4A, low suspicion, biopsy; 4B, intermediate suspicion, biopsy; 4C, moderate suspicion, biopsy; or 5, highly suggestive of malignancy, biopsy. An assessment of 0, incomplete, with recommendation for additional imaging, can be used clinically but was not included in the interpretive skills tasks for either PEM or MRI.

PEM Interpretive Skills Task

A PEM interpretive skills task was developed that included 49 breasts in 26 subjects with a mean age of 56 years (median, 56 years; range, 40–89 years). Cases were chosen to represent the spectrum of imaging findings found in clinical practice. Representative PEM

images from the 12-slice PEM displays were provided for all cases, and low-resolution mammograms were also supplied for 31 cases. Malignancies were present in 20 (41%) of 49 breasts, including nine invasive and intraductal carcinomas, three cases of pure ductal carcinoma in situ (DCIS), three infiltrating ductal carcinomas (IDCs), three infiltrating lobular carcinomas (ILCs), and one case each of mixed IDC-ILC and ILC-DCIS. Tumor sizes ranged from 5 to 38 mm with a median of 15 mm (mean, 17 mm). Twenty-one breasts were deemed negative on imaging and subsequent clinical follow-up. Eight breasts had biopsy-proven benign lesions: two fat necrosis, three fibrocystic changes, and one each of fibroadenoma, stromal sclerosis, and cyst. For each breast, observers were asked to classify qualitative background uptake (minimal, mild, moderate homogeneous, moderate patchy heterogeneous, or intense), as well as to describe findings using the PEM lexicon terminology (none, focus, multiple foci, mass, or nonmass). For masses, observers described the shape (oval, round, lobular, or irregular) and, for nonmass uptake, the distribution (diffuse, focal area, regional, linear or ductal, or segmental). For each breast, observers provided a final assessment using the same terminology as the expanded BI-RADS lexicon (category 1, 2, 3, 4A, 4B, 4C, or 5). An assessment of 4A or higher was considered test positive. The threshold for successful completion of the PEM interpretive skills task was sensitivity of 75% or greater and specificity of 45% or greater, according to expert consensus review. Quantitative lesion PEM FDG uptake was not part of the interpretive skills task.

MRI Interpretive Skills Task

No specific training in MRI was administered, though investigators were encouraged to review the BI-RADS for MRI [6]. A separate interpretive skills task was developed for breast MRI as a qualification task for the American College of Radiology Imaging Network 6666 MRI substudy [8] and was used for qualification of observers in this study. MRI scans from 32 breasts in 30 patients with a mean age of 51 years (median, 51 years; range, 32–72 years) were chosen to represent the spectrum of findings in usual practice. Mammograms were not provided. One to three selected subtraction images (obtained 2 minutes after contrast agent administration, with fat suppression), together with either inversion recovery or T2-weighted images or descriptions of T2 signal intensity (hypointense or hyperintense) were provided for each case, and maximum-intensity-projection images were provided for 19 cases. Kinetic curves were provided for three cases, and images from three time points were provided for two other cases to allow visual assessment of kinetics. Of the 11 (34%) malignancies, six were DCIS, two were IDC, and one each was IDC-DCIS, ILC, and tubular carcinoma. The median invasive tumor size was 12 mm (mean, 20 mm; range, 4–55 mm). Twenty-one (66%) breasts had known benign findings by biopsy or 2 years of follow-up. Of the 21 benign cases, five were fibrocystic change; three were benign lymph nodes; two each were benign breast tissue, fibroadenoma, and scar; and one each was fat necrosis, fibrosis, hematoma, and ruptured cyst. One MRI was negative and two had findings presumed benign (stable for 4 years). Observers were asked to describe artifacts if present, including motion artifacts, large breast abutting the coil, inhomogeneous fat suppression, or any clips or sutures. MRI description of findings followed standard BI-RADS MRI descriptors [6]: none, focus, mass, nonmass enhancement, or postsurgical scar. Masses were further classified by shape (round, oval, lobulated, or irregular), margins (smooth, irregular, or spiculated),

internal enhancement (homogeneous, heterogeneous, rim-enhancement, dark internal septations, enhancing internal septations, and central internal enhancement), and we also asked whether the mass contained fat. Nonmass enhancement was classified by distribution (focal area, linear or ductal, segmental, regional, multiple regional, or diffuse uptake) and enhancement (not applicable, homogeneous, heterogeneous, stippled or punctate, clumped, or reticular or dendritic). Observers were asked to provide BI-RADS MRI final assessments (1, 2, 3, 4A, 4B, 4C, or 5), and 4A or higher was considered test positive. The threshold for reader qualification for MRI assessment was sensitivity of 65% or greater and specificity of 45% or greater on the basis of early experience with observers for the ACRIN 6666 protocol.

Analysis

All observers were asked to complete the PEM and MRI assessment tasks independently (i.e., without consulting their colleagues). Data were entered into a Microsoft Excel worksheet and validated by dual data entry. An assessment of BI-RADS 4A or higher was considered a true-positive finding if malignancy was known to be present in the breast from which the image was obtained (i.e., observers were not required to mark the lesion of interest). An assessment of 1, 2, or 3 with negative or benign truth was considered as true-negative. Sensitivity, specificity, accuracy, positive predictive value, and negative predictive value were calculated for all observers. Receiver operating characteristic curves were calculated for all readings and for individual observers, and the area under the curve (AUC) was determined using the Web-based program JROCFIT [9]. Performance on the interpretive tasks for MRI and performance for PEM were compared for each of the diagnostic performance characteristics by two-sample Student *t* tests using Stata software (StataCorp) and 95% CIs were determined. A *p* value less than 0.05 was considered statistically significant.

Performance of PEM and MRI assessment tasks was analyzed as a function of each demographic variable by using Student *t* tests.

The degree of interobserver agreement for feature classification and final assessments among all observers was calculated by generalized kappa tests [10] separately for the PEM and MRI tasks. Agreement of observer performance with expert consensus for the PEM task, or expert opinion for the MRI task, was also calculated. For all calculations of agreement for final assessments, categories 1 and 2 were grouped, as were categories 4 and 5. Kappa statistics were calculated using Stata software to assess the proportion of interobserver agreement beyond that expected by chance alone. A kappa value of 1.0 corresponds to complete agreement, 0 to no agreement, and less than 0 to disagreement. A kappa value of equal to or less than 0.20 indicates slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–0.99, almost perfect agreement [11].

Results

PEM Interpretive Performance

Thirty-six observers from 15 sites throughout the United States completed both PEM and MRI interpretive skills tasks and were included in this analysis. The median sensitivity and specificity for PEM assessment tasks were 100% (mean, 96%; range, 75–100%; 95% CI, 0.94–0.98) and 83% (mean, 84%; range, 66–97%; 95% CI, 0.82–0.87) (Table 1), respectively. The median empirical AUC for PEM was 0.96 (mean, 0.95; range, 0.82–1.0; 95% CI, 0.94–0.97) for consolidated readings for all observers (Fig. 1). Of 36 observers, 22 (61%) had 100% sensitivity on the PEM task. All observers successfully qualified as PEM-interpreting physicians for the multicenter trial.

For the PEM task, 17 (35%) of 49 cases were read correctly by all 36 observers (Fig. 2). Eleven (55%) of 20 malignancies were scored as benign by at least one of the 36 observers (Figs. 3 and 4), and five observers misclassified more than two malignancies each. Of the 30 instances where a cancer was interpreted as negative for malignancy, 24 (80%) were given an assessment score of 3, probably benign, with recommendation for 6-month follow-up. Of the benign cases presented, 21 (72%) of 29 were interpreted as positive by at least one observer, with one benign fibroadenoma (Fig. 5) interpreted as suspicious (assessment 4A, $n = 2$ observers; 4B, $n = 7$; 4C, $n = 12$) or highly suggestive of malignancy (assessment 5, $n = 15$) by all participants.

MRI Interpretive Performance

The median sensitivity and specificity on the MRI assessment task were 82% (mean, 82%; range, 45–100%; 95% CI, 0.78–0.86) and 69% (mean, 67%; range, 38–91%; 95% CI, 0.62–0.71), respectively. The median AUC was 0.81 (mean, 0.80; range, 0.48–0.96; 95% CI, 0.77–0.83) (Fig. 1). For all performance parameters, observers performed better on the PEM task ($p < 0.001$) (Table 1), though these were different cases. Four observers had 100% sensitivity on the MRI assessment task. A total of seven (19%) of 36 observers did not meet required thresholds for the MRI task: four had sensitivity less than 65%; two had specificity less than 45%; and one observer was unsuccessful in both sensitivity and specificity thresholds.

There was only one MRI case (1/32 [3.1%]) that was read correctly by all 36 observers. Nine observers misclassified three or more malignancies as benign; eight malignancies were classified as benign by at least one observer; and all 11 malignancies in the MRI task were misclassified as probably benign by at least one observer. Of the 73 instances where a malignancy was read as benign, 49 (67%) were noted as benign. Of the benign cases, 20 (95%) of 21 were read as positive by at least one observer and all malignant MRI cases were read as negative by at least one observer. A fibroadenoma with atypia was the benign lesion most commonly read as suspicious with 33 (92%) of 36 participants rating it as BI-RADS 4A or higher, whereas a 21-mm grade II IDC was frequently read as negative for malignancy (17/36 [47%]; Fig. 6).

Experience Variables and Demographics

The majority (20/36 [56%]) of observers had more than 10 years of experience in breast imaging, with 25 (69%) of 36 spending at least 75% of their time interpreting clinical breast imaging studies (Table 2). Thirteen observers (36%) had more than 5 years of experience interpreting breast MRI, and 12 (33%) observers reviewed 10 or more breast MRI examinations per week, with all having career experiences interpreting at least 50 MRI examinations. Thirty-two (89%) observers had interpreted zero to 20 PEM scans before this task, with only one observer having read more than 100 PEM examinations.

PEM performance did not differ significantly with increasing years in breast imaging, with similar AUCs for all categories of breast imaging experience ($p = 0.38$, not significant). There was no difference in PEM performance between observers who had read none or fewer than 20 PEM images before the assessment skills task compared with those who had read more than 20 PEM images ($p = 0.62$). MRI performance was also not affected by years of breast imaging experience ($p = 0.62$), years of breast MRI experience ($p = 0.50$), or average number of breast MRI examinations interpreted per week ($p = 0.46$). There were no significant differences in demographics for those who passed the MRI interpretive skills thresholds as compared with those who did not. There was no difference between PEM or MRI performance based on number of mammograms or ultrasound examinations interpreted per week or percentage of time spent in breast imaging per week.

Interobserver Agreement

PEM—Fair to moderate agreement was seen describing background FDG uptake among all observers, with kappa values ranging from 0.27 for mild uptake to 0.57 for intense uptake, for an overall kappa value of 0.33 (Table 3). There was moderate agreement ($\kappa = 0.57$) regarding type of finding (focus, mass, nonmass, or no uptake) across all 36 observers, with masses having the greatest agreement ($\kappa = 0.70$). There was moderate agreement on mass shape ($\kappa = 0.49$). There was only fair agreement across observers describing type of nonmass uptake ($\kappa = 0.28$), with the greatest agreement seen for linear or ductal uptake ($\kappa = 0.36$); all other categories showed only slight agreement ($\kappa < 0.2$). Moderate interobserver agreement was seen for PEM final assessments (overall $\kappa = 0.63$; 95% CI, 0.61–0.65). For a 1 or 2 assessment, agreement was moderate ($\kappa = 0.69$; 95% CI, 0.60–0.77); for 3, probably benign, agreement was poor ($\kappa = 0.16$; 95% CI, 0.12–0.20); and for 4 or 5 assessment, agreement was moderate ($\kappa = 0.73$; 95% CI, 0.64–0.83).

There was reduced agreement with consensus truth for the type of background FDG uptake on PEM images ($\kappa = 0.17$) (Table 3). For all other descriptors and assessments, agreement with consensus was similar to interobserver agreement (Table 3).

MRI—Overall agreement among all 36 observers in identifying artifacts was fair ($\kappa = 0.38$) (Table 4), with highest agreement found for recognizing artifacts due to clips or sutures ($\kappa = 0.57$). There was substantial agreement ($\kappa = 0.64$) in describing lesion type, with masses and postsurgical scars showing the most agreement ($\kappa = 0.70$ and 0.72 , respectively). There was moderate agreement across observers with regard to mass shape ($\kappa = 0.47$) and margins ($\kappa = 0.55$). Nonmass linear enhancement and multiple regional enhancement distributions

showed the least agreement ($\kappa = 0.11$ and 0.10 , respectively). When “linear” and “ductal” were grouped, the kappa value improved to 0.45 for the combined term (with $\kappa = 0.48$ overall for type of nonmass distribution, not different from 0.47 originally). Interobserver agreement was fair for MRI assessments (overall $\kappa = 0.32$; 95% CI, 0.29 – 0.34). For a 1 or 2 assessment, agreement was fair ($\kappa = 0.36$; 95% CI, 0.28 – 0.43); for 3, probably benign, agreement was slight ($\kappa = 0.15$; 95% CI, 0.08 – 0.22); and for 4 or 5 assessment, agreement was moderate ($\kappa = 0.41$; 95% CI, 0.29 – 0.53).

Comparing observer descriptions and assessments to expert-defined truth, agreement was similar to interobserver agreement (Table 4).

Discussion

The high diagnostic performance of all 36 observers from 15 different institutions after a 2-hour tutorial shows that accurately interpreting PEM images is quickly learned and reproducible for breast imaging radiologists, independent of years of clinical experience. Although 32 of 36 observers had reviewed between zero and 20 PEM examinations before this training, a 2-hour tutorial including the PEM lexicon [7] allowed them to successfully identify malignancy and complete the PEM skills assessment task.

The development and use of a PEM lexicon [7] provided a standardized reporting format for PEM images similar to BI-RADS lexicons for mammography [4], ultrasound [5], and MRI [6], which contributed to observer success. We found moderate agreement among observers in identifying and classifying FDG uptake as a focus, multiple foci, mass, or nonmass lesion. There was moderate agreement in classifying the shape of mass uptake and the type of nonmass distribution. Observers had difficulty consistently describing the distribution of nonmass uptake. This variability can be attributed in part to the fact that there were 11 nonmass cases in this dataset and that there were five different categories of nonmass distribution from which to select. Unlike the BI-RADS for MRI [6], we did not distinguish “linear” from “ductal” in describing FDG uptake; we found that 62% of known malignancies so described were due to DCIS and that 43% of lesions prospectively described as linear or ductal were malignant in the multicenter trial [7].

Although there was moderate agreement among observers for final assessments, especially for classifying lesions as either benign or suspicious, agreement among observers in probably benign assessments was only slight. Indeed, in clinical use, probably benign assessments may be problematic on PEM because lesions seen on PEM have a higher likelihood of malignancy than those seen on MRI [2].

Interobserver agreement for MRI was moderate for most features, with kappa values between 0.44 and 0.64 . The greatest agreement was seen for describing lesion type as a mass. Ikeda et al. [12] showed similar interobserver agreement for MRI features, with moderate agreement for most overall categories, such as lesion type. Within each mass category, Ikeda et al. [12] showed moderate agreement for overall categories of shape, margin, and internal enhancement characteristics, similar to our observations. Nonmass enhancement on MRI has a lower positive predictive value for malignancy than masses,

with most malignant nonmass enhancement due to DCIS [13]. We found only fair agreement in describing the distribution and type of nonmass enhancement. As with PEM, this was probably the result of having only two cases in each nonmass category, with no cases having either diffuse distribution or reticular or dendritic nonmass like enhancement.

Final assessments on MRI showed only fair agreement in our series, similar to results reported by Stoutjesdijk et al. [14]. As with PEM and as has been seen with mammography [15], only slight agreement was seen for probably benign assessments on MRI. Although the use of a BI-RADS 3 assessment on MRI has been validated to have less than 2% risk of malignancy [16–18], clarification and increased physician education as to which lesions are most appropriately so characterized appear to be needed in clinical practice.

Although prior studies have shown improved performance in breast imaging with greater specialization [19, 20], we did not see variations in PEM or MRI interpretive performance as a function of observer experience. The lack of a difference in our study may be due to the small number of observers in each experience subgroup. Of seven readers who did not succeed in the MRI task, six had more than 10 years of clinical breast imaging experience, which suggests that breast MRI is a challenging technique to interpret, though the relatively small number of cases in the interpretive skills task makes this an arbitrary result.

In the two similar tasks developed, PEM interpretive performance was significantly better than that of MRI, though comparison of performance on the two tasks should be interpreted with caution. Different cases were used for the two tasks. There were no mammograms provided for the MRI task, though they were supplied for 31 (63%) of 49 PEM cases. Furthermore, the contralateral breast was not supplied for comparison with the MRI task, but representative contralateral images were available for 46 (94%) of 49 PEM cases. Only selected images were provided for either task, but this may be more problematic for MRI, where behavior of the lesion on multiple different pulse sequences is necessary for appropriate interpretation. Kuhl et al. [21] showed that T2-weighted sequences can be used to help distinguish benign and malignant breast tumors. Although we did provide inversion recovery images or T2 signal intensity, it may not have been as effective as if additional MRI scans had been available. Finally, all observers underwent a 2-hour training module in PEM but no focused training in MRI before the task.

It would have been preferable to use identical cases for both PEM and MRI tasks, but such cases were not available before the prospective PEM and MRI trial [2] for which the PEM tutorial and skills assessment task were developed and used to qualify investigators. To compensate for this deficiency, care was taken to ensure that the same range and size of malignant and benign findings and normal breasts were included in each skills assessment task.

There are several limitations to this work. In addition to challenges comparing results between PEM and MRI tasks, full clinical historical information was not provided to readers for either task. Neither MRI nor PEM assessment tasks were completed on the respective workstations. Kinetic information has proven to be useful in distinguishing malignant lesions from benign lesions [22] but was only available for four (12.5%) of 32 MRI cases

and there was no computer-assisted detection. The inability to quantify FDG uptake for PEM provided a similar interpretation challenge. For both imaging assessment tasks, observers lacked the ability to scroll through multiple images, because only static slice presentation was available. We did not assess consistency of measurement of lesions on PEM or MRI, though we have recently reported on accuracy of such: for the subset of invasive tumors with less than 10% DCIS component, Pearson's correlation coefficients with final tumor size at histopathology were 0.55 for PEM and 0.81 for MRI ($p < 0.001$ that MRI was more accurate) [2]. Finally, this was a skills assessment task, not real clinical interpretation for a clinical case. The lack of true impact on the clinical management of a patient may reduce the care of observers in interpreting the images and may have contributed to reduced accuracy, though the artificial test scenario would have affected each of the PEM and MRI tasks equally. Several reports indicate that performance of mammographers in a true clinical situation is significantly better than that in an artificial test scenario [23, 24].

Although there are no guidelines in the literature for optimal dose of FDG for PEM scanning, the imaging protocol for the multicenter study [2] required a dose of 10 mCi for each subject. This results in an approximately 700 mrem whole-body radiation dose to the patient, which is less than the radiation dose for whole-body PET or PET/CT [25]. Preliminary studies suggest that the dose can be decreased by varying the scan time without adversely affecting image quality or lesion detection [26], though definitive studies are required. PEM can also be performed the same day as PET/CT, using the same injected dose of FDG, and further study of such an approach is warranted.

In summary, accurate PEM interpretation was easy to learn by breast imagers, regardless of experience. At least moderate agreement was seen for most major categories of the PEM lexicon; most important, moderate agreement was seen for grouped PEM final assessment categories. Use of probably benign, category 3, may be problematic on PEM, with very little agreement on such an assessment. Our results are otherwise promising and validate the use of the PEM lexicon across multiple observers; we have separately reported on prospective use of the PEM lexicon in clinical practice [7]. It is important to recognize that all observers in this study were experienced radiologists who met all requirements of MQSA for mammogram-interpreting physicians; review of the training module and completion of the interpretive skills task is encouraged for any physician planning to interpret PEM examinations.

Acknowledgments

We thank Sherry Bullwinkel, Certus International, for overseeing distribution, collection, and initial scoring of observer interpretive skills tasks, and Joel Miller, Certus International, for statistical assistance. We also thank each of the breast imaging specialists who participated as observers.

This work was funded by Naviscan, Inc. and the National Institutes of Health (grant 5 R44 CA103102-05); development of the MRI qualification task was supported by grants from The Avon Foundation and the National Cancer Institute (grants U01 CA079778 and U01 CA89008) through the American College of Radiology Imaging Network.

References

1. Berg WA, Weinberg IN, Narayanan D, et al. High-resolution fluorodeoxyglucose positron emission tomography with compression ("positron emission mammography") is highly accurate in depicting primary breast cancer. *Breast J.* 2006; 12:309–323. [PubMed: 16848840]
2. Berg WA, Madsen KS, Schilling K, et al. Breast cancer: comparative effectiveness of positron emission mammography and MR imaging in pre-surgical planning for the ipsilateral breast. *Radiology.* 2011; 258:59–72. [PubMed: 21076089]
3. Mussurakis S, Buckley DL, Coady AM, Turnbull LW, Horsman A. Observer variability in the interpretation of contrast enhanced MRI of the breast. *Br J Radiol.* 1996; 69:1009–1016. [PubMed: 8958017]
4. D'Orsi, CJ.; Bassett, LW.; Berg, WA., et al. BI-RADS: mammography. In: D'Orsi, CJ.; Mendelson, EB.; Ikeda, DM., et al., editors. *Breast Imaging Reporting and Data System: ACR BI-RADS—breast imaging atlas.* 4th ed.. Reston, VA: American College of Radiology; 2003.
5. Mendelson, EB.; Baum, JK.; Berg, WA., et al. BI-RADS: ultrasound. In: D'Orsi, CJ.; Mendelson, EB.; Ikeda, DM., et al., editors. *Breast Imaging Reporting and Data System, ACR BI-RADS—breast imaging atlas.* 1st ed.. Reston, VA: American College of Radiology; 2003.
6. Ikeda, DM.; Hylton, NM.; Kuhl, CK., et al. BI-RADS: magnetic resonance imaging. In: D'Orsi, CJ.; Mendelson, EB.; Ikeda, DM., et al., editors. *Breast Imaging Reporting and Data System, ACR BI-RADS—breast imaging atlas.* 1st ed.. Reston, VA: American College of Radiology; 2003.
7. Narayanan D, Madsen KS, Kalinyak JE, Berg WA. Interpretation of positron emission mammography: feature analysis and rates of malignancy. *AJR.* 2011; 196:956–970. [PubMed: 21427350]
8. Berg WA, Mendelson EB, Merritt CRB, Blume J, Schleinitz M. ACRIN 6666: screening breast ultrasound in high-risk women. acrinc.org/Portals/0/Protocols/6666/Protocol-ACRIN%206666%20Admin%20Update%2011.30.07.pdf. Published November 9, 2007.
9. Eng J. ROC analysis: web-based calculator for ROC curves. www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html.
10. Fleiss J. Measuring the nominal scale agreement among many raters. *Psychol Bull.* 1971; 76:378–382.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33:159–174. [PubMed: 843571]
12. Ikeda DM, Hylton NM, Kinkel K, et al. Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies. *J Magn Reson Imaging.* 2001; 13:889–895. [PubMed: 11382949]
13. Liberman L, Morris EA, Lee MJ, et al. Breast lesions detected on MR imaging: features and positive predictive value. *AJR.* 2002; 179:171–178. [PubMed: 12076929]
14. Stoutjesdijk MJ, Futterer JJ, Boetes C, van Die LE, Jager G, Barentsz JO. Variability in the description of morphologic and contrast enhancement characteristics of breast lesions on magnetic resonance imaging. *Invest Radiol.* 2005; 40:355–362. [PubMed: 15905722]
15. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter-and intraobserver variability in feature analysis and final assessment. *AJR.* 2000; 174:1769–1777. [PubMed: 10845521]
16. Eby PR, DeMartini WB, Gutierrez RL, Saini MH, Peacock S, Lehman CD. Characteristics of probably benign breast MRI lesions. *AJR.* 2009; 193:861–867. [PubMed: 19696303]
17. Eby PR, Demartini WB, Peacock S, Rosen EL, Lauro B, Lehman CD. Cancer yield of probably benign breast MR examinations. *J Magn Reson Imaging.* 2007; 26:950–955. [PubMed: 17896380]
18. Weinstein SP, Hanna LG, Gatsonis C, Schnall MD, Rosen MA, Lehman CD. Frequency of malignancy seen in probably benign lesions at contrast-enhanced breast MR imaging: findings from ACRIN 6667. *Radiology.* 2010; 255:731–737. [PubMed: 20501712]
19. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology.* 2002; 224:861–869. [PubMed: 12202726]

20. Smith-Bindman R, Chu P, Miglioretti DL, et al. Physician predictors of mammographic accuracy. *J Natl Cancer Inst.* 2005; 97:358–367. [PubMed: 15741572]
21. Kuhl CK, Klaschik S, Mielcarek P, Gieseke J, Wardelmann E, Schild HH. Do T2-weighted pulse sequences help with the differential diagnosis of enhancing lesions in dynamic breast MRI? *J Magn Reson Imaging.* 1999; 9:187–196. [PubMed: 10077012]
22. Kuhl CK, Mielcarek P, Klaschik S, et al. Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology.* 1999; 211:101–110. [PubMed: 10189459]
23. 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]
24. Rutter CM, Taplin S. Assessing mammographers’ accuracy: a comparison of clinical and test performance. *J Clin Epidemiol.* 2000; 53:443–450. [PubMed: 10812315]
25. Wu TH, Huang YH, Lee JJ, et al. Radiation exposure during transmission measurements: comparison between CT- and germanium-based techniques with a current PET scanner. *Eur J Nucl Med Mol Imaging.* 2004; 31:38–43. [PubMed: 14534833]
26. Lu, X.; Lu, W.; Kalinyak, JE. Radiation dose reduction for personalized breast PET imaging. Salt Lake City, UT: Society of Nuclear Medicine; 2010. p. 358

Fig. 1. Receiver operating characteristic curves for positron emission mammography (PEM) (dashed line) and MRI (solid line). Mean area under the curve for PEM was 0.95 (95% CI, 0.94–0.97) and that of MRI was 0.80 (95% CI, 0.77–0.83; $p < 0.0001$). ROI = region of interest.

Fig. 2.

46-year-old woman with palpable mass due to 12-mm grade II infiltrating ductal carcinoma.

A, Craniocaudal mammogram was unremarkable except for vague asymmetry at site of palpable abnormality (marked by radiopaque marker).

B, Craniocaudal positron emission mammography image (obtained 92 minutes after IV injection of 14.5 mCi FDG; slice thickness, 5.8 mm) showed round mass (*arrow*) with intense FDG uptake. All 36 observers correctly identified finding to be a mass and

suspicious for malignancy. Thirty-five of 36 observers described shape of mass as oval or round.

Fig. 3. 63-year-old woman with multiple foci seen on lateromedial positron emission mammography image (obtained 116 minutes after IV injection of 15.3 mCi FDG; slice thickness, 2.8 mm), proven at pathology to be grade III infiltrating ductal carcinoma and ductal carcinoma in situ. Thirty of 36 observers identified finding (*arrows*) as multiple foci, with three each describing it as mass or nonmass. Thirty-two observers correctly assessed finding as malignant, with three observers calling it probably benign and one observer classifying finding as benign.

Fig. 4.

70-year-old woman with 5-mm infiltrating ductal carcinoma with ductal carcinoma in situ read as negative on positron emission mammography (PEM) by five of 36 observers. Four observers assessed this as probably benign and recommended short-term follow-up, although they correctly identified finding, and remaining 27 observers considered this suspicious.

A, No abnormality is seen on mediolateral oblique mammogram.

B, Two serial 4.6-mm-thick mediolateral oblique PEM images (obtained 74 minutes after IV injection of 11.6 mCi FDG) show focal area of mild FDG uptake (*arrows*) corresponding to malignancy.

Fig. 5.

55-year-old woman with fibroadenoma in right breast read as suspicious for malignancy by all 36 observers (assessment 4A, $n = 2$; 4B, $n = 7$; 4C, $n = 12$; 5, $n = 15$) in positron emission mammography (PEM) assessment task.

A, Craniocaudal mammogram shows lobulated mass (*arrow*).

B, Craniocaudal PEM image (obtained 86 minutes after injection of 13 mCi FDG; slice thickness, 5.4 mm) shows intense FDG uptake in oval mass (*arrow*) in outer right breast.

Fig. 6.

39-year-old woman with 2.1-cm grade II infiltrating ductal carcinoma variably described as benign ($n = 6$ observers), probably benign ($n = 11$ observers), suspicious ($n = 18$ observers), or highly suggestive of malignancy ($n = 1$ observer).

A–D, MRI scans, including maximum-intensity-projection image (**A**), T1-weighted image (**B**), STIR (**C**), and fat-suppressed spoiled gradient-echo T1-weighted image (**D**) obtained 90 seconds after injection of 0.1 mmol/kg gadolinium-based contrast agent show lobulated mass (*arrow*) with irregular margins and rim enhancement. All 36 observers identified

finding as mass, with 35 observers identifying it as lobular mass. Internal enhancement was classified as dark internal septations by 27 observers, rim enhancement by seven observers, and heterogeneous enhancement or enhancing internal septations each by one observer.

Comparison of Positron Emission Mammography (PEM) and MRI Performance on Qualification Tasks for 36 Experienced Breast Imaging Observers

TABLE 1

Technique Score	PEM		MRI		p
	Mean (Range)	95% CI	Mean (Range)	95% CI	
Sensitivity (%)	96 (75–100)	0.94–0.98	82 (45–100)	0.78–0.86	<0.001
Specificity (%)	84 (66–97)	0.81–0.87	67 (38–91)	0.62–0.71	<0.0001
Accuracy (%)	89 (71–98)	0.87–0.91	72 (44–91)	0.69–0.75	<0.0001
Positive predictive value (%)	82 (62–95)	0.79–0.85	58 (29–79)	0.54–0.61	<0.0001
Negative predictive value (%)	97 (83–100)	0.95–0.99	88 (60–100)	0.85–0.91	<0.0001
Area under the curve	0.95 (0.82–1.00)	0.94–0.97	0.80 (0.48–0.96)	0.77–0.83	<0.0001

TABLE 2

Mean Area Under the Curve (AUC) for Positron Emission Mammography (PEM) and MRI as a Function of Demographic Characteristics for 36 Observers

Characteristic	No. (%) of Observers	Mean AUC		p	
		PEM	MRI	PEM	MRI
Experience in breast imaging (y)					
< 2	3 (8.3)	0.93	0.81	0.38	0.62
2-5	2 (5.6)	0.96	0.77		
6-10	11 (31)	0.95	0.83		
> 10	20 (56)	0.96	0.78		
Time in clinical breast imaging (%)					
< 25	1 (2.8)	0.98	0.74	0.73	0.77
25-49	5 (14)	0.96	0.83		
50-74	5 (14)	0.96	0.82		
75-100	25 (69)	0.95	0.79		
Mammograms read per week (no.)					
50-99	6 (17)	0.97	0.82	0.44	0.91
100-149	7 (19)	0.97	0.82		
150-199	11 (31)	0.95	0.78		
200-299	8 (22)	0.96	0.81		
300	4 (11)	0.94	0.78		
Breast ultrasounds read per week (no.)					
1-20	6 (17)	0.97	0.80	0.19	0.57
21-39	20 (56)	0.94	0.80		
40-60	6 (17)	0.97	0.82		
70-99	2 (5.6)	0.97	0.86		
100	2 (5.6)	0.98	0.71		
Experience in breast MRI (y)					
< 1	2 (5.6)	0.96	0.75	0.56	0.50
1-2	6 (17)	0.94	0.84		
2-5	15 (42)	0.95	0.78		

Characteristic	No. (%) of Observers	Mean AUC		<i>p</i>	
		PEM	MRI	PEM	MRI
> 5	13 (36)	0.96	0.81	0.61	0.46
Breast MRI scans read per week (no.)					
1–4	9 (25)	0.95	0.78		
5–9	15 (42)	0.96	0.78		
10–14	5 (14)	0.94	0.87		
15–24	4 (11)	0.96	0.81		
25	3 (8.3)	0.98	0.81		
Total no. of PEM cases ever interpreted				0.62	0.52
0–20	32 (89)	0.96	0.80		
21–39	2 (5.6)	0.96	0.83		
40–99	1 (2.8)	0.97	0.74		
100	1 (2.8)	0.91	0.92		

TABLE 3

Interobserver Agreement for Positron Emission Mammography (PEM) Features Among 36 Experienced Breast Imaging Observers and Agreement With Consensus Truth for 49 Breasts

PEM Feature	No. of Breasts ^a	Interobserver Agreement Among 36 Experienced Observers, κ (95% CI)	Agreement Between 36 Experienced Observers and Consensus Truth, Mean κ (Median) [Range]
Background uptake		0.33 (0.32–0.35)	0.17 (0.14) [–0.10 to 0.64]
None	14	0.34 (0.30–0.39)	
Mild	20	0.27 (0.20–0.34)	
Moderate-to-homogeneous	10	0.16 (0.12–0.21)	
Moderate, heterogeneous, or patchy	3	0.40 (0.33–0.47)	
Intense	2	0.57 (0.53–0.60)	
Lesion type		0.57 (0.55–0.58)	0.55 (0.54) [0.41–0.74]
None	21	0.69 (0.61–0.76)	
Focus	4	0.39 (0.36–0.43)	
Multiple foci	3	0.33 (0.29–0.37)	
Mass	13	0.70 (0.64–0.77)	
Nonmass	8	0.40 (0.35–0.45)	
Mass shape		0.49 (0.44–0.54)	0.50 (0.53) [0.27–0.64]
Not applicable	36	0.67 (0.53–0.81)	
Oval or round	4	0.36 (0.32–0.40)	
Lobulated	2	0.36 (0.32–0.39)	
Irregular	7	0.34 (0.29–0.38)	
Type of nonmass uptake		0.28 (0.24–0.33)	0.28 (0.29) [0.07–0.43]
Not applicable, no nonmass uptake	38	0.40 (0.28–0.52)	
Scattered or diffuse	1	0.11 (0.09–0.15)	
Focal areas	4	0.15 (0.11–0.19)	
Regional	3	0.15 (0.12–0.18)	
Linear or ductal	2	0.36 (0.32–0.39)	
Segmental	1	0.21 (0.18–0.24)	
Final assessment ^b		0.63 (0.61–0.65)	0.70 (0.69) [0.44–0.93]
Benign (assessment 1 or 2)	25	0.69 (0.60–0.77)	
Probably benign (assessment 3)	3	0.16 (0.12–0.20)	
Suspicious (assessment 4 or 5)	21	0.73 (0.64–0.83)	

^aNumber of breasts with given feature by expert consensus for the PEM task description.

^bGrouped as 1 or 2, 3, and 4 or 5.

TABLE 4

Interobserver Agreement for MRI Features Among 36 Experienced Breast Imaging Observers and Agreement With Consensus Truth for 32 Breasts

Feature	No. of Breasts ^a	Interobserver Agreement Among 36 Experienced Observers, κ (95% CI)	Agreement Among 36 Experienced Observers and Consensus Truth, Mean κ (Median) [Range]
Artifacts		0.38 (0.30–0.46)	0.40 (0.43) [–0.01 to 0.72]
None	26	0.43 (0.23–0.62)	
Motion	2	0.17 (0.12–0.21)	
Large breast abuts coil	1	0.51 (0.46–0.55)	
Inhomogeneous fat suppression	0	0.07 (0.04–0.11)	
Clips or sutures	3	0.57 (0.53–0.61)	
Lesion type		0.64 (0.61–0.67)	0.73 (0.74) [0.40–0.91]
No finding	1	0.58 (0.55–0.62)	
Focus or foci	4	0.50 (0.46–0.55)	
Mass	15	0.70 (0.58–0.82)	
Nonmass	10	0.63 (0.55–0.71)	
Postsurgical scar	2	0.72 (0.68–0.76)	
Mass shape		0.47 (0.44–0.50)	0.52 (0.51) [0.34–0.68]
Not applicable, no mass	17	0.70 (0.58–0.82)	
Round	1	0.32 (0.28–0.37)	
Oval	6	0.24 (0.19–0.29)	
Lobulated	2	0.47 (0.41–0.53)	
Irregular	6	0.28 (0.23–0.33)	
Mass margin		0.55 (0.52–0.58)	0.51 (0.51) [0.25–0.71]
Not applicable, no mass	17	0.70 (0.58–0.82)	
Smooth	6	0.57 (0.49–0.64)	
Irregular	7	0.43 (0.36–0.48)	
Spiculated	2	0.18 (0.14–0.22)	
Mass internal enhancement		0.50 (0.47–0.54)	0.44 (0.46) [0.11–0.62]
Not applicable, no mass	17	0.70 (0.58–0.82)	
Homogeneous	7	0.29 (0.23–0.34)	
Heterogeneous	5	0.35 (0.30–0.41)	
Rim enhancement	2	0.56 (0.52–0.61)	
Dark internal septations	1	0.48 (0.43–0.52)	
Enhancing internal septations	0	0 (–0.04 to 0.04)	
Central internal enhancement	0	0 (–0.06 to 0.06)	
Mass contains fat		0.44 (0.40–0.48)	0.34 (0.27) [–0.03 to 0.86]
Not applicable or unable to assess	22	0.47 (0.33–0.62)	
No	7	0.37 (0.29–0.45)	
Yes	3	0.49 (0.45–0.54)	
Nonmass type ^{b,c}		0.47 (0.40–0.53)	0.50 (0.49) [0.18–0.69]

Feature	No. of Breasts ^a	Interobserver Agreement Among 36 Experienced Observers, κ (95% CI)	Agreement Among 36 Experienced Observers and Consensus Truth, Mean κ (Median) [Range]
Not applicable	22	0.64 (0.47–0.82)	0.52 (0.54) [0.08–0.82]
Focal area	2	0.31 (0.27–0.35)	
Linear	2	0.11 (0.08–0.15)	
Ductal	0	0.32 (0.29–0.35)	
Segmental	2	0.30 (0.27–0.34)	
Regional	2	0.42 (0.37–0.46)	
Multiple regions	2	0.10 (0.07–0.13)	
Diffuse	0	0.48 (0.44–0.51)	
Nonmass enhancement		0.45 (0.39–0.51)	0.43 (0.41) [0.02–0.68]
Not applicable	21	0.65 (0.47–0.82)	
Homogeneous	1	0.18 (0.14–0.22)	
Heterogeneous	3	0.36 (0.31–0.41)	
Stippled or punctuate	1	0.24 (0.21–0.27)	
Clumped	6	0.38 (0.33–0.42)	
Reticular or dendritic	0	0.10 (0.07–0.13)	
Overall assessment of malignancy		0.32 (0.29–0.34)	
Benign (assessment 1 or 2)	9	0.36 (0.28–0.43)	
Probably benign (assessment 3)	6	0.15 (0.08–0.22)	
Suspicious (assessment 4 or 5)	17	0.41 (0.29–0.53)	

^aNumber of breasts with given feature by expert opinion.

^bIf linear and ductal nonmass enhancement are combined for interobserver agreement, the kappa value is 0.45 (95% CI, 0.42–0.49), and the overall kappa value for nonmass enhancement is 0.48 (95% CI, 0.42–0.55).

^cIf linear and ductal nonmass enhancement are combined for agreement between observers and consensus, mean kappa value for agreement with expert opinion is 0.54 (range, 0.24–0.76) with a median value of 0.54.