JAOA

Original contribution

Objectivity and accuracy of mammogram interpretation using the BI-RADS final assessment categories in 40- to 49-year-old women

CLAIRE McKAY, DO
CURTIS L. HART, EdD
GEORGE ERBACHER, DO

To determine if use of the five final assessment categories of the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS) improved objectivity or accuracy of mammographic evaluation in 40- to 49-year-old women, fifty mammograms of 40- to 49-year-old women that were obtained at a tertiary referral teaching hospital were classified according to those five final assessment categories. The mammograms were blinded to six American Osteopathic Board of Radiology–certified radiologists who were asked to classify each mammogram within the five final BI-RADS categories based on the mediolateral oblique and craniocaudal views presented. No history was allowed. Use of the BI-RADS five final assessment categories provided moderate interobserver objectivity, moderately high agreement among the radiologists' interpretation (reliability), and moderate accuracy of interpretation (validity) when compared to criterion. Moderate interobserver reliability and accuracy has been previously identified; however, no scientific review of the BI-RADS five final assessment categories in 40- to 49-year-old females was discovered in the current literature. No overall improvement of objectivity or accuracy was demonstrated using the five final assessment categories of the BI-RADS lexicon in 40- to 49-year-old women.

(Key words: mammography, Breast Imaging Reporting and Data System, breast cancer)

Breast cancer remains the most frequently diagnosed female malignancy in the United States.[1] It is the single leading cause of death for women aged 40 to 49 years in the United States, with more than 40% of the lost years of life from breast cancer diagnosed before the age of 50.[2,3] The number of deaths caused by breast cancer can be reduced by early detection and intervention.[4,5] The interpretation and management recommendation of mammographic studies can influence the stage and progression of detected breast cancer and thus affect mortality.[6]

To standardize mammographic reporting, reduce confusion in breast imaging interpretation, and facilitate outcome monitoring, the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS). The BI-RADS lexicon includes evaluation of appropriateness and accuracy of examination interpretation to provide peer review and data for quality assurance in an effort to improve overall patient breast healthcare.[7,8] The mammographic details guide the classification and recommendation by the interpreting radiologist. The BI-RADS is believed to achieve reduction in description ambiguity of mammography.[9] The five final assessment categories of the BI-RADS lexicon are outlined in *Table 1*. An "assessment incomplete" category is included in the BI-RADS lexicon, but is not a focus of this study.

The purpose of this study was to describe the objectivity and accuracy of radiologists' interpretation of mammograms in 40- to 49-year-old women using the five final assessment categories described within the BI-RADS lexicon in a tertiary referral teaching hospital.

Methods

Fifty screening and diagnostic mammograms obtained at a tertiary referral teaching hospital from 1993 to 1997 were selected and classified according to the five BI-RADS final assessment categories. The hospital's mammography department is accredited by the Food and Drug Administration according to the Mammography Quality Standards Act requirements. The mammograms included in this study were either stable for at least 2 years or had histopathologic diagnosis. A biopsy was not required to verify lack of cancer.

The patients were 40- to 49-year-old females. Women with previously benign biopsies were included in this study. Patients with a history of breast cancer, large breasts requiring multiple or special film, previous cosmetic surgery, or mastectomy or augmentation, as well as films of inadequate technical quality were excluded. Standard mediolateral oblique and craniocaudal views of original films were available for each breast. Mammography was performed with a Bennett MF-150 unit (1992) using Kodak MIN-R screen and Kodak MIN-RE single emulsion film. All were processed with a dedicated mammographic processor using a 3-minute development process. The mammograms were randomly coded by number for confidentiality, blinded to the radi-

Dr McKay is an assistant professor at the University of Texas Health Science Center at San Antonio; Mr Hart is an assistant professor of Kinesiology and Health at the University of Texas at San Antonio, San Antonio, Tex; Dr Erbacher is section chief of Interventional Radiology at the Tulsa Regional Medical Center, Tulsa, Okla.

Correspondence to Claire McKay, DO, 7703 Floyd Curl Dr, San Antonio, TX 78284.

Table 1
ACR Breast Imaging Reporting and Data System (BI-RADS): Final Assessment Categories

Category	Assessment	Description and recommendation
1	Negative (N)	Nothing on which to comment, annual mammogram
2	Benign (B)	Definitely benign finding described, annual mammogram
3	Probably benign (P)	High probability of being benign. Short-term follow-up recommended to establish stability.
4	Suspicious abnormality (S)	Not characteristic, but has a reasonable probability of being malignant. Biopsy urged.
5	Highly suggestive of malignancy	High probability of cancer. Appropriate action should be taken.

Table 2
Subject Distribution of Diagnostic Criterion Using Five Final Assessment BI-RADS* Categories

BI-RADS category	n	%
1	14	28
2	20	40
3	10	20
4	2	4
5	4	8

*BI-RADS = Breast Imaging Reporting and Data System.

ologists, and placed on an appropriate alternator with masking for viewing.

Six radiologists certified by the American Osteopathic Board of Radiology from the same hospital agreed to participate. The radiologists were given no patient history and were unaware of the number of cases in which cancer was diagnosed. Each radiologist was instructed to independently classify each mammogram into a BI-RADS final assessment category based on the mammographic findings presented. The five BI-RADS categories were reviewed with the radiologist and available at the reading site for review.

The distribution of the final diagnostic criterion for the mammograms selected for this study is presented in *Table 2*. Criterion is defined as placement of the mammographic examination retrospectively into a BI-RADS final assessment category, 1 to 5, based on the known stability or tissue diagnosis. The five categories were also condensed and recoded into three categories representing similar patient management recommendations: benign (BI-RADS 1 and 2), probably benign (BI-RADS 3), and suspicious (BI-RADS 4 and 5). The distribution in the management categories is shown in *Table 3*.

Reliability is the comparison of two radiologists' interpretations. The five final assessment BI-RADS categories and the three condensed management recommendations were assessed for reliability using Pearson correlation coefficients (r) for comparison of all possible paired combinations of radiologists. A concordance coefficient (W) was calculated to determine the general agreement among the radiologists and was the measure of interobserver objectivity for the five category and three management recommendations. Thus, interobserver objectivity is defined as the overall agreement of the group of six radiologists.

Validity is the correlation of the individual radiologists' categorical placement of a mammogram versus the final diagnostic criterion, determined with Pearson correlation coefficient. For each radiologist, the Pearson correlation coefficient was also used to determine validity for the three patient management categories. Group validity was determined by calculating a concordance coefficient for the group of radiologists versus the final diag-

Table 3
Subject Distribution of Diagnostic Criterion Using Three Recommended Management Categories

Category	Recommendation	n	%
Benign	Annual mammography	34	68
Probably benign	6-month follow-up to establish stability	10	20
Suspected malignancy	Recommend tissue sampling	6	12

Table 4
Means and Standard Deviation of BI-RADS* Five Category Score for Diagnostic Criterion and Participating Radiologists

Variable	Mean	SD
Criterion	**2.24**	**1.15**
Radiologist 1	1.74†	1.10
Radiologist 2	2.38	0.86
Radiologist 3	2.42	1.05
Radiologist 4	2.16	1.02
Radiologist 5	2.00	1.14
Radiologist 6	1.80†	1.14

*BI-RADS = Breast Imaging Reporting and Data System.
†Significantly different from criterion ($P<.01$).

Radiologist	Criterion	1	2	3	4	5	6
1	0.74	1.00					
2	0.69	0.69	1.00				
3	0.61	0.61	0.68	1.00			
4	0.70	0.75	0.70	0.41	1.00		
5	0.53	0.49	0.59	0.48	0.53	1.00	
6	0.80	0.73	0.66	0.55	0.71	0.47	1.00

◀ **Figure 1.** *Correlation matrix for radiologists' comparisons using the five BI-RADS categories.*

Radiologist	Criterion	1	2	3	4	5	6
1	0.74	1.00					
2	0.57	0.71	1.00				
3	0.55	0.63	0.66	1.00			
4	0.72	0.83	0.70	0.53	1.00		
5	0.49	0.33	0.47	0.33	0.49	1.00	
6	0.72	0.76	0.74	0.58	0.80	0.41	1.00

◀ **Figure 2.** *Correlation matrix for radiologists' comparisons using the three patient management recommendation categories.*

nostic criterion of the five BI-RADS categories and again for the three patient management categories. Descriptors of correlation and concordance statistics include no correlation, <0.20; low, 0.20 to 0.39; moderate, 0.40 to 0.59; moderately high, 0.60 to 0.79; and high, >0.79.

One-way analysis of variance (ANOVA) with repeated measures on the subjects with the Dunnett test post hoc compared each radiologist with the criterion to assess significant differences in interpretation for the five BI-RAD categories and three management recommendations. The repeated-measures ANOVA was used to reduce variability attributed to subject differences. Sensitivity and specificity were not included due to the error created by the limited number of subjects in this study.

Results

The mean score of the five-category BI-RADS criterion was 2.24±1.15 (mean ± SD). The radiologists' mean score ranged from 1.74±1.10 to 2.42±1.05 (*Table 4*). Radiologists 1 and 6 mean scores were significantly lower than the criterion ($P<.01$). This was due primarily to disagreement within categories 1 and 2. When condensed into three categories based on similar patient manage-

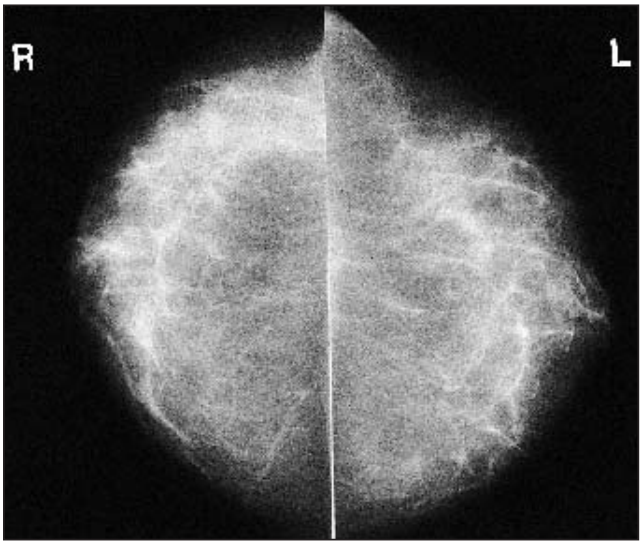


Figure 3. *BI-RADS category 1 mammogram, for which radiologists were in complete agreement.*

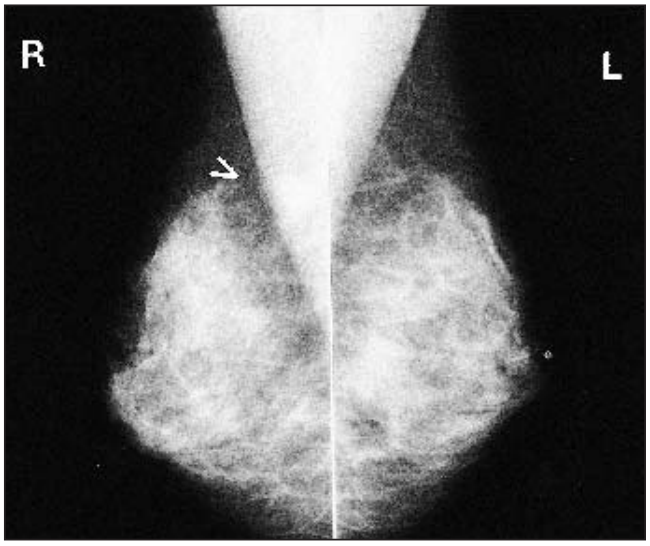


Figure 4. *Mammogram indicating invasive ductal adenocarcinoma, for which half of radiologists reported as category 4 and half as category 5.*

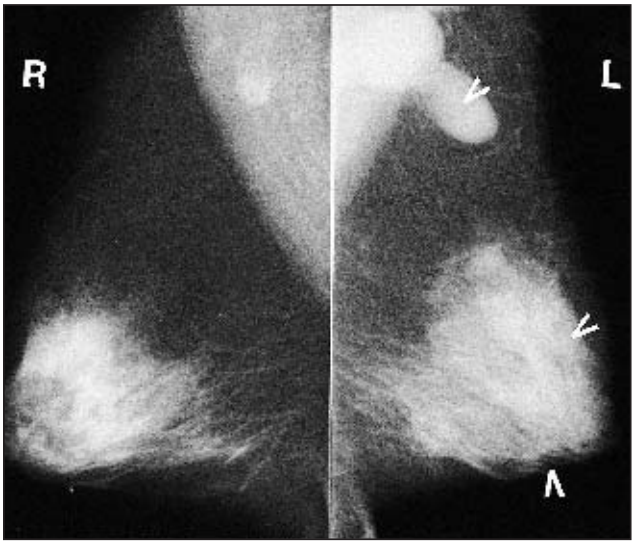


Figure 5. *Mammogram indicating invasive ductal adenocarcinoma, for which one third of radiologists reported as category 4 and two thirds as category 5.*

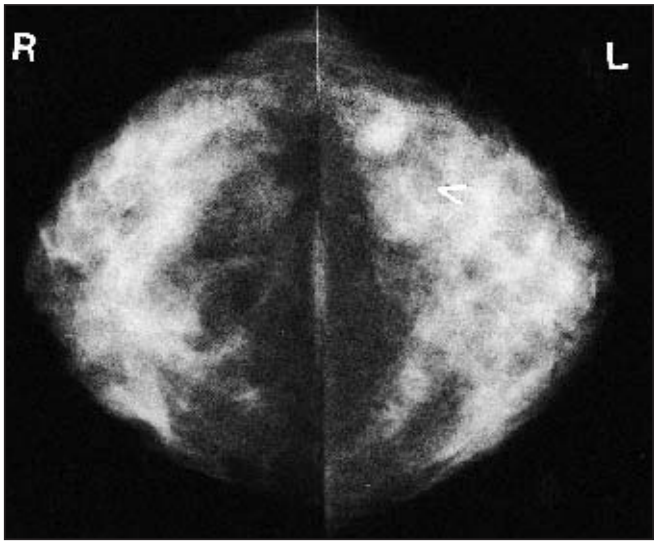


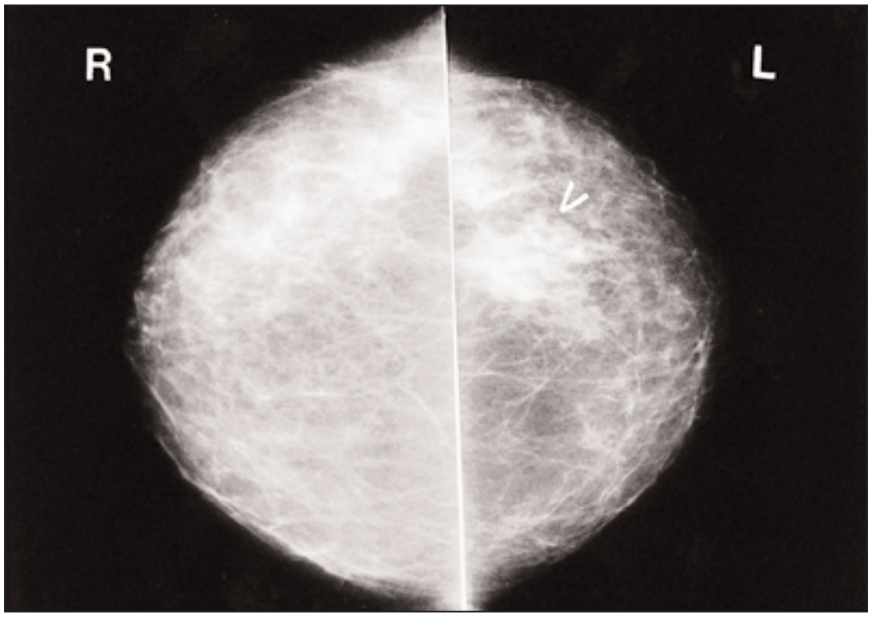
Figure 6. *Mammogram indicating radial scar, for which 17% of radiologists reported as category 2, 50% as category 3, and 33% as category 4.*

ment, no significant difference between radiologist and criterion was observed ($P>.05$).

Mean reliability for paired radiologist agreement for the five BI-RADS categories was r=0.60 (range, r=0.41 to 0.75; $P<.01$). The concordance coefficient demonstrated moderate interobserver objectivity among the radiologists (W=0.52) and moderate validity (W=0.58) compared with the five-category criterion (*Figure 1*). Individual radiologist validity ranged from r=0.53 to r=0.80 (moderate to high).

When condensed into three categories based on patient management recommendations, mean reliability remained r=0.60 (range, r=0.41 to 0.83), interobserver objectivity among radiologists remained moderate (W=0.58), and validity was W=0.60. There was no significant improvement in interobserver objectivity or validity when comparing the five BI-RADS categories to the three patient management categories ($P>.05$; *Figure 2*). An example of a mammogram with complete agreement between radiologists is demonstrated in *Figure 3*. The breasts are heterogeneous but without focal densities or calcifications. All radiologists interpreted this exam as BI-RADS Category 1.

Figure 4 and *Figure 5* illustrate mammograms for which all of the radiologists were suspicious, placing them into categories 4 and 5. Although there was



◀ **Figure 7.** *Mammogram indicating asymmetric benign parenchyma, for which 50% of radiologists reported as category 2, 17% as category 3, and 33% as category 4.*

disagreement on the specific category, all agreed on recommended management. Each demonstrates invasive ductal adenocarcinoma.

Considerable variability occurred with the mammograms in *Figure 6* and *Figure 7*. *Figure 6*, a biopsy-proven benign radial scar, received category ratings of 2, 3, and 4. *Figure 7* demonstrates focal parenchymal density that has remained stable for 3 years; however, without prior mammograms or patient history, this example received category ratings of 2, 3, and 4.

Comments

The management recommendation from the radiologist to the referring physician is the most significant component of the mammogram report. For that reason we included performance based not only on the five final BI-RADS assessment categories, but on the three management recommendation categories. Two radiologists demonstrated significant underscoring within categories 1 and 2. Radiologists interpret findings differently based on variable thresholds of concern; however, the agreement and accuracy remained moderate to moderately high, comparable to previously published literature.[10-14] It is important for the literature to continue to demonstrate only moderate to moderately high accuracy, but no study design will ever simulate radiologists' interpretation conditions. Accuracy is improved when mammography is combined with physical examination. Improved interpretation skills have been demonstrated with high-quality comparison exams and complete diagnostic workup of mammographic abnormalities,[15,16] but these were not components of this study design.

This study was not designed to simulate clinical conditions of mammographic interpretation. The fifty mammograms were selected to provide an adequate spectrum of variable findings. Many of the women had not had annual mammography performed in the past and, in preparing for this study, were variable in the timing of their mammograms. It was necessary to provide at least 2 years' stability, several patients having 4 years between examinations. The goal was to determine interobserver objectivity and accuracy based solely on the presenting mammographic characteristics.

One scientific review of the ACR BI-RADS final assessment categories was discovered in the literature[14]; however, no review specifically addressed 40- to 49-year-old women. Baker and colleagues[9] reported moderate success of the BI-RADS lexicon providing standardization of mammographic interpretation and reporting, but did not address the final assessment categories. While our interobserver objectivity and accuracy remained moderate, and reliability moderately high, the five BI-RADS final assessment categories did not improve interpretation or management recommendation.

Use of the BI-RADS five final assessment categories gave moderate interobserver objectivity, moderately high agreement among the radiologists' interpretation (reliability), and moderate accuracy of interpretation (validity) as compared to criterion. When the five final assessment categories were reduced to three categories based on similarity of recommended management, moderate interobserver objectivity was again demonstrated. Based on comparison to previous literature, no improvement of objectivity or accuracy was demonstrated using the five final assessment categories of the BI-RADS lexicon in 40- to 49-year-old women; however, due to the limited number of subjects of this study, further investigation is recommended.

References

1. Smigel K. Breast cancer death rates decline for white women. *J Natl Cancer Inst* 1995;87:173.

2. NIH Consensus Statement. Breast cancer screening for women ages 40-49. *NIH Consens Statement* 1997;15:1-35.

3. Shapiro S, Venet W, Strax P, Venet L. *Periodic Screening of Breast Cancer: The Health Insurance Plan Project and Its Sequelae, 1963-1986*. Baltimore, Md: The Johns Hopkins University Press, 1988.

4. Bassett LW, Cardenosa G, D'Orsi CJ, Dempsey PJ, Dershaw DD, Destouet JM, et al. Risk of risk-based mammography screening, ages 40-49. American College of Radiology Task Force on Breast Cancer. *J Clin Oncol* 1999;17:735-738.

5. Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of International Workshop on

Screening for Breast Cancer. *J Natl Cancer Inst* 1993;85:1644-1656.

6. Boyd NF, Wolfson C, Moskowitz M, Carlile T, Petitclerc M, Ferri HA, et al. Observer variation in the interpretation of xeromammograms. *J Natl Cancer Inst* 1982;68:357-363.

7. American College of Radiology. *Breast Imaging Reporting and Data System (BI-RADS)*. Reston, Va: American College of Radiology, 1993.

8. Feig SA, D'Orsi CJ, Hendrick RE, Jackson VP, Kopans DB, Monsees B, et al. American College of Radiology guidelines for breast cancer screening. *Am J Roentgenol* 1998;171:29-33.

9. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *Am J Roentgenol* 1996;166:773-778.

10. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994;331:1493-1499.

11. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Arch Intern Med* 1996;156:209-213.

12. Ciccone G, Vineis P, Figerio A, Segman N. Interobserver and intraobserver variability of mammogram interpretation. *Eur J Cancer* 1992;28A:1054-1058.

13. Howard DH, Elmore JG, Lee CH, Wells CK, Feinstein AR. Observer variability in mammography. *Trans Assoc Am Physicians* 1993;106:96-100.

14. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. *J Natl Cancer Inst* 1998;90:1801-1809.

15. Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology* 1992;184:39-43.

16. Roux S, Markle L, Diamond A. False positive rate of screening mammography. *N Engl J Med* 1998;339:561.

JAOA

Coming in...

The D.O.

The November issue of *The D.O.* will outline the AOA Campaign for Osteopathic Unity's new advertising initiative, as well as offer early critiques of the AOA's new ads from members of the profession.

Future issues of *JAOA*

- Annual education issue
- "Adjunctive osteopathic manipulation for the treatment of pneumonia in the hospitalized elderly"
- "Thoracic lymphatic pumping and the efficacy of influenza vaccination in healthy young and elderly populations"
- "Minor depression in primary care: what the generalist should know"
- "Collaboration between osteopathic medicine and pharmacy to teach via the Internet"
- "Predicting factors of successful recovery from lumbar spine surgery in workers' compensation patients"
- "Testing osteopathic medical school graduates for licensure: Is the COMLEX–USA the most appropriate examination?"
- "Correlation of scores for the COMLEX–USA with osteopathic medical school grades"
- "Black widow bites in children"
- "Prediction of student performances on COMLEX–USA Level 1 examination based on admission data and course performance"
- "Clinical experience using intracorporeal lithotripsy with the Swiss lithoclast"
- "Weaning from mechanical ventilation: an update"
- "Occupational and environmental medicine in a family medicine residency"
- "The cranial rhythmic impulse related to the Traube-Hering-Mayer oscillation: comparing laser-Doppler flowmetry and palpation"
- "A decline in structural examination compliance in the hospital medical record with advancing level of training"
- "Adjunctive osteopathic manipulative treatment in women with depression: a pilot study"
- "The muscle hypothesis: a model of chronic heart failure appropriate for osteopathic medicine"
- "The primary care physician's role in caring for internationally adopted children"