Original Article

Interobserver Variability and Diagnostic Performance with the Fifth Edition of the ACR BI-RADS Lexicon for Ultrasound; Validation from Multi-institutions

Sung Hun Kim[1], Dong Wook Kim[2], Ji Hyun Youk[3], Jung Hyun Yoon[4], Sun Hye Jeong[5],
You Me Kim[6], Eun Hye Lee[5], Min Jung Kim[4]

[1]*Department of Radiology, Seoul St. Mary's Hospital, The Catholic University of Korea,*
[2]*Department of Policy Research Affairs, National Health Insurance Corporation Ilsan Hospital,*
[3]*Department of Radiology, Gangnam Severance Hospital,*
Research Institute of Radiological Science, Yonsei University, College of Medicine,
[4]*Department of Radiology, Severance Hospital, Research Institute of Radiological Science,*
Yonsei University, College of Medicine,
[5]*Department of Radiology, Soonchunhyang University Bucheon Hospital,*
[6]*Department of Radiology, Dankook University Hospital, Dankook University, College of Medicine*

Objective: To evaluate the interobserver agreement of radiologists from multiple institutions when breast masses obtained from multi-institutions were characterized using the revised Breast Imaging Reporting and Data System (BI-RADS) lexicon and to investigate whether agreement levels and diagnostic performances differed according to years of experience.

Materials and Methods: US images of 80 breast masses in 74 women were obtained from five institutions (51 benign and 29 malignant masses). Five readers (two from the same institution and 3 from different academic institutions; less than 10 years of experience in two readers and more than 10 years in the remaining 3) independently reviewed the two orthogonal images of each case. Each reader described and recorded the findings of the breast masses, and assigned a final category according to the revised BI-RADS lexicon. Interobserver agreement was determined using Cohen's kappa and diagnostic performances were compared according to the years of experience in breast imaging.

Results: Overall agreement for the five radiologists varied from fair to substantial with the following kappa values: orientation ($\kappa = 0.63$), shape ($\kappa = 0.54$), final assessment ($\kappa = 0.40$), posterior features ($\kappa = 0.30$), margin ($\kappa = 0.29$) and echo pattern ($\kappa = 0.28$). There were no significant differences for US

This work was supported by a research fund from the Korean Society of Breast Imaging & Korean Society For Breast Screening (KSBI&KSFBS-2013- No.001).

Correspondence to: Min Jung Kim, M.D.
Department of Radiology, Yonsei University College of Medicine,
50-1 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea
Tel: 82-2-2228-7413, Fax: 82-2-393-3035, e-mail:mines@yuhs.ac

descriptors except for posterior features according to institution or experience level. The overall sensitivity, specificity, positive predictive value, negative predictive value and accuracy were 96.4% (106/110), 52.8% (153/290), 43.6% (106/243), 97.5% (153/157), 64.7% (259/400) without significant difference according to years of experience.

Conclusion: The overall interobserver agreement using multi-institutional cases was good, which validates the fifth edition of the BI-RADS lexicon, regardless of differences in years of experience. The overall diagnostic performances of reviewers from multi-institutions were excellent without significant difference according to years of experience.

Index words: Ultrasonography; Breast neoplasms; Breast diseases; Observer variation

Introduction

Ultrasonography is a useful adjunct to mammography and an important diagnostic tool in the evaluation and management of breast disease (1). As for the interpretation and communication of ultrasound (US) findings, the US section was first included in the fourth edition of the Breast Imaging Reporting and Data System (BI-RADS) to standardize sonographic reporting and to communicate results clearly and efficiently to physicians and patients (2). It has been used worldwide and validated in several studies (3-6). Recently, a fifth edition of BI-RADS was released (1). There has been a few key changes in the lexicon for US such as addition of the terms, “complex cystic and solid” and “heterogeneous echo texture”, as well as changes to the subcategories of category 4 (1).

As the limitations of mammography in the diagnosis of dense breasts has become known, supplemental screening using ultrasonography has been proposed to overcome lowered mammographic sensitivity in women with dense breast tissue (7-9). Accordingly, ultrasonography can be widely used nationwide in the future as a promising adjunctive screening modality to mammography, but only after interobserver agreement for lexicon and inter-institutional diagnostic performances are validated and the degree of proficiency needed by radiologists is understood. Several studies have indicated that the interobserver agreement of the BI-RADS lexicon is good, although the agreement level varies according to the categories using the fourth edition of BI-RADS (10-13). In a review of previous literature on interobserver variability with the fourth edition of BI-RADS, we found that the study populations in these studies consisted of cases from a single institution (10-13), and that the observers also consisted of radiologists who had continuously worked at the same institution and who even had trained at the very same institution (10, 11, 13). To our knowledge, the interobserver variability associated with multi-institutional cases using the fifth edition of BI-RADS has not been studied yet.

We postulated that variability was more common in a multi-institution group and in a group with shorter experience in comparison with a single institution group and a group with longer experience. Prior to breast US screening being performed nationwide, a validation process is inevitable. Our study had two purposes: To evaluate the interobserver agreement of radiologists from multiple institutions when breast masses obtained from multi-institutions were characterized using the revised BI-RADS lexicon and to investigate whether agreement levels and diagnostic performances differed according to years of experience.

Materials and Methods

This was a retrospective study approved by the institutional review boards (IRB) of the five institutions included in this study (Severance Hospital, Yonsei University; St. Mary's Hospital, The Catholic University; Gangnam Severance Hospital, Yonsei University; Bucheon Soonchunhyang Hospital, Soonchunhyang University; Dankook University Hospital). Informed consent for this study was waived. For image data used in this study, informed consent was obtained from all patients of all 5 institutions for the construction of an educational US-image database.

Patients

The database of breast lesions from five institutions was built for the education and evaluation of interpretation of breast lesions using sonography, and was supported by the Korean Society of Breast Imaging. Each institution contributed US images of breast masses that had undergone percutaneous US-guided biopsy, been stable for more than 2 years or decreased in extent compared with a prior exam after benign pathology results had been obtained, or had been described as having characteristically benign features in the BI-RADS lexicon such as simple cyst, intramammary lymph nodes, and a stable postsurgical scar (1). For each case, patient age, final pathology, method used to obtain pathology samples, US image, and follow-up period were recorded. For the final pathology, benignity was defined according to benign biopsy results at percutaneous US-guided biopsy with imaging-histologic concordance, and the lesion being stable for more than 2 years or with decreased size compared with a prior exam performed after obtaining benign pathology results. Malignancy was defined by either malignant biopsy results or malignancy on surgical pathology. In cases with atypical ductal hyperplasia or atypical cell, the final pathology was based on the subsequent surgical pathology. US images of 80 breast masses in 74 women were included in this study from September to October 2013. The mean size of the breast masses for which US images were obtained was 14.0 ± 7.2 mm (range, 4 to 50 mm). The mean age of the 74 women was 47.5 ± 10.2 years (range, 21 to 80 years).

US examination and image acquisition

For image acquisition, various US machines equipped with high-frequency linear array transducers were used (iU22, Philips Medical Systems, Bothell, WA, USA; GE LOGIQ E9, GE Medical Systems, Milwaukee, WI, USA; Supersonic Imagine, Aix en Provence, France, EUB-8500; Hitachi Medical, Tokyo, Japan). Seven radiologists dedicated to breast imaging with more than 3 years of experience (range, 3 to 13 years) were involved in patient collection and image acquisition. When a breast mass was detected on US examination, routine scanning protocols were used to measure size, including transverse and longitudinal images of the mass with and without calibers. Two representative transverse and longitudinal images were stored as a dicom file, which was included in the database with clinicopathologic information. One investigator (J.S.H.) randomly arranged 160 images of 80 cases in a Microsoft Power Point 2010 (ppt) file. The two orthogonal ultrasonographic images of each case were embedded in a slide of the ppt file. No clinicopathologic information was displayed on the ppt file.

Image analysis

Five breast radiologists (K.M.J., Y.J.H., K.S.H., Y.J.H., K.Y.M.), with 6 to 13 years of experience for breast imaging from 4 institutions, independently reviewed the images in the ppt file. Each radiologist contributed to data collection for the database. To reduce bias due to the radiologist's memory of the previously performed ultrasonography, images were reviewed 3 months after data collection. Clinical information, mammographic findings and pathologic results were not available during image review. Among the 5 readers, two readers (K.M.J., Y.J.H.) underwent residency and fellowship training and were practicing faculty members in an academic breast imaging department and three (K.S.H., K.Y.M., Y.J.H.) had been trained and were working in different academic breast imaging departments.

Among the 5 readers, two readers (Y.J.H., Y.J.H) had less than 10 years of experience and the remaining three (K.M.J., K.S.H., K.Y.M.) had more than 10 years of experience.

The readers evaluated the breast lesions according to the US descriptors of the fifth edition of BI-RADS (1) (Table 1). Calcifications, associated findings and special cases were not included in the analysis. Each reader chose the most appropriate descriptor and categorized the final assessment (category 2: benign, category 3: probably benign, category 4A: low suspicion for malignancy, category 4B: moderate suspicion for malignancy, category 4C: high suspicion for malignancy, category 5: highly suggestive of malignancy).

Statistical analysis

The standard reference for pathology was based on the final pathology registered in the database. For image interpretation, category 2 and 3 lesions were considered negative, while category 4 and 5 lesions were considered positive. Diagnostic performances in terms of sensitivity, specificity, positive predictive value, negative predictive value and accuracy were calculated. The area under the curve for the diagnostic performance was also evaluated for all the 5 radiologists and for each radiologist. Data analysis was performed by a statistician (K.D.W.). Five readers were grouped according to years of experience as the group with under 10 years of experience and the group with more than 10 years of experience. Interobserver agreement was assessed using generalized kappa statistics and compared according to years of experience. We used the following definition to interpret kappa coefficients: kappa value (κ) equal to or less than 0.20 indicated slight agreement; values from 0.21–0.40, fair agreement; from 0.41–0.60, moderate agreement; from 0.61–0.80, substantial agreement; and from 0.81–1.00, almost perfect agreement (14). Diagnostic performance was also compared according to years of experience. Generalized Estimating Equation models were performed to compare diagnostic performances between the two groups. The Generalized Estimating Equation is a statistical

Table 1. US Descriptors of the Fifth Edition of the BI-RADS

Shape	Oval
	Round
	Irregular
Orientation	Parallel
	Not parallel
Margin	Circumscribed
	Non-circumscribed
	Indistinct
	Angular
	Microlobulated
	Spiculated
Echo pattern	Anechoic
	Hyperechoic
	Complex cystic and solid
	Hypoechoic
	Isoechoic
	Heterogeneous
Posterior features	No posterior features
	Enhancement
	Shadowing
	Combined pattern
Category	Category 2: Benign
	Category 3: Probably Benign
	Category 4A: Low suspicion for malignancy
	Category 4B: Moderate suspicion for malignancy
	Category 4C: High suspicion for malignancy
	Category 5: Highly suggestive of malignancy

method that fits parameters to a generalized linear model when unknown correlation is present (15).

Statistical analysis was performed with SAS (version 9.2, SAS Institute Inc., Cary, NC, USA). A P value less than 0.05 was considered statistically significant.

Results

Of the 80 breast masses, 51 (63.8%) were benign

Table 2. Interobserver Variability for the US Descriptors According to Years of Experience

US descriptors	Percentage of Lesions* (n=80)	κ-value (SE)			p value
		5 Readers	Years of Experience		
			<10	≥10	
Shape	Oval (57.7)	0.58 (0.14)	0.69 (0.25)	0.62 (0.33)	0.23
	Round (12.0)	0.37 (0.84)	0.45 (0.16)	0.34 (0.30)	
	Irregular (30.3)	0.58 (0.09)	0.65 (0.16)	0.46 (0.30)	
	Overall (100.0)	0.54 (0.04)	0.63 (0.07)	0.49 (0.08)	
Orientation	Parallel (67.0)	0.63 (0.17)	0.71 (0.27)	0.61 (0.36)	0.46
	Not parallel (33.0)	0.63 (0.10)	0.71 (0.16)	0.61 (0.31)	
	Overall (100.0)	0.63 (0.04)	0.71 (0.08)	0.61 (0.11)	
Margin	Circumscribed (28.7)	0.56 (0.10)	0.51 (0.17)	0.58 (0.29)	0.60
	Non-circumscribed (71.3)				
	Indistinct (24.3)	0.06 (0.09)	0.01 (0.15)	−0.03 (0.29)	
	Angular (9.0)	0.14 (0.08)	0.07 (0.16)	0.12 (0.32)	
	Microlobulated (25.0)	0.20 (0.09)	0.22 (0.16)	0.14 (0.29)	
	Spiculated (13.0)	0.39 (0.08)	0.39 (0.15)	0.37 (0.34)	
	Overall (100.0)	0.29 (0.02)	0.27 (0.03)	0.23 (0.06)	
Echo pattern	Anechoic (0.3)	−0.007 (0.15)	−0.01 (0.31)		0.73
	Hyperechoic (1.2)	0.11 (0.14)	−0.004 (0.51)	0.66 (0.61)	
	Complex cystic and solid (10.0)	0.31 (0.08)	0.39 (0.16)	0.11 (0.30)	
	Hypoechoic (60.5)	0.31 (0.15)	0.25 (0.25)	0.52 (0.33)	
	Isoechoic (17.8)	0.29 (0.08)	0.15 (0.16)	0.26 (0.29)	
	Heterogeneous (10.2)	0.19 (0.08)	0.37 (0.16)	−0.08 (0.36)	
	Overall (100.0)	0.28 (0.04)	0.27 (0.07)	0.31 (0.08)	
Posterior features	No features (53.5)	0.30 (0.15)	0.29 (0.26)	0.09 (0.34)	0.002
	Enhancement (22.8)	0.37 (0.08)	0.38 (0.15)	0.11 (0.30)	
	Shadowing (13.2)	0.42 (0.07)	0.54 (0.15)	0.03 (0.31)	
	Combined pattern (10.5)	−0.02 (0.08)	−0.01 (0.27)	−0.22 (0.30)	
	Overall (100.0)	0.30 (0.04)	0.36 (0.07)	0.01 (0.08)	
Category	Category 2 (6.2)	0.58 (0.08)	0.46 (0.18)	0.74 (0.35)	0.33
	Category 3 (33.0)	0.57 (0.10)	0.58 (0.17)	0.58 (0.29)	
	Category 4A (31.3)	0.35 (0.10)	0.32 (0.17)	0.37 (0.30)	
	Category 4B (8.5)	0.10 (0.08)	−0.08 (0.17)	0.27 (0.336)	
	Category 4C (10.5)	0.21 (0.08)	0.19 (0.16)	0.22 (0.32)	
	Category 5 (10.5)	0.43 (0.08)	0.44 (0.16)	0.51 (0.37)	
	Overall (100.0)	0.40 (0.02)	0.37 (0.04)	0.44 (0.06)	

Note- SE-standard error

* Data are combined for all five readers

(15 fibroadenoma, 6 duct ectasia, 5 intraductal papilloma, 4 fibroadenomatoid hyperplasia, 4 fibrocystic change, 3 hamartoma, 3 postoperative fibrosis and 11 other) and 29 (36.2%) were malignant (4 ductal carcinoma in situ, 20 invasive ductal carcinoma, 2 mucinous carcinoma and 3 other).

Interobserver variability

The interobserver variability, percentage of US descriptors and final assessment category are

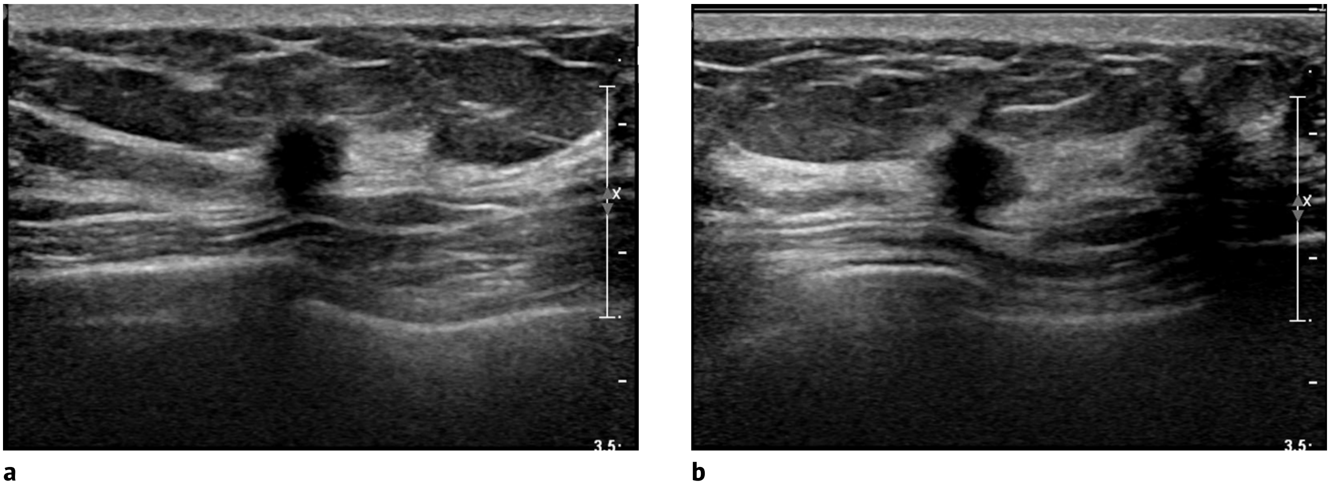

Fig. 1. Transverse (**a**) and longitudinal (**b**) images from a 45-year-old woman with invasive lobular carcinoma.
In the single institution group, reviewers agreed on mass orientation (not parallel), margin (spiculated), echo pattern (hypoechoic) and final assessment (category 5). Reviewers did not agree on mass shape (irregular in one and round in one) or posterior features (no features in one and shadowing in one). In the multi-institution group, reviewers agreed on mass shape (irregular), mass orientation (not parallel), margin (spiculated), echo pattern (hypoechoic) and posterior features (no features). Reviewers did not agree on mass shape (irregular in two and round in one) or final assessment (category 5 in two and category 4C in one).

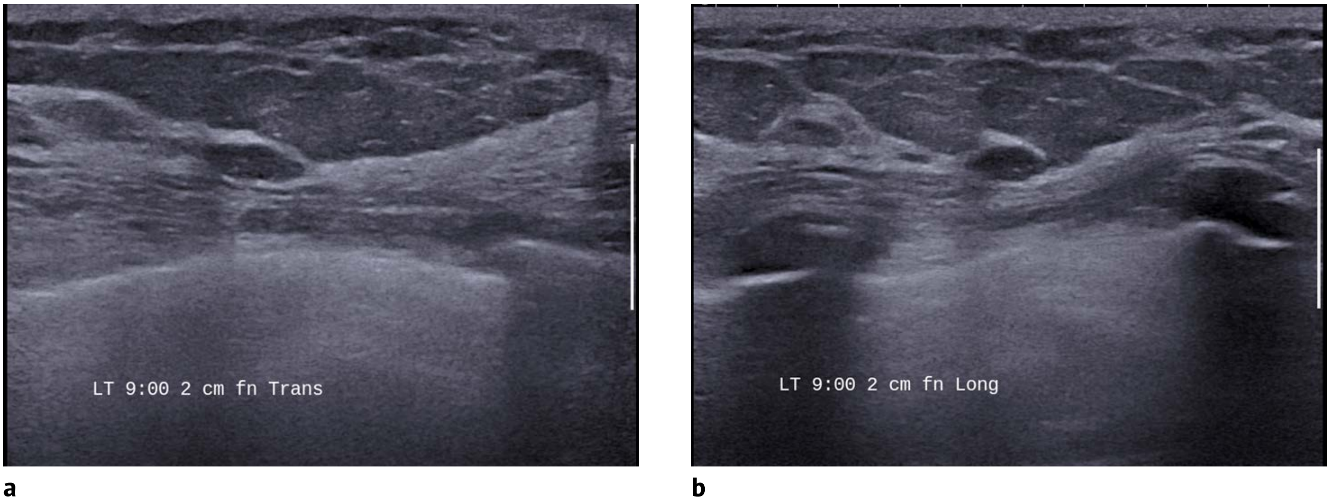

Fig. 2. Transverse (**a**) and longitudinal (**b**) images from a 58-year-old woman with a benign mass which was stable for more than 2 years.
In the single institution group, reviewers agreed on mass shape (oval), mass orientation (parallel), margin (circumscribed), echo pattern (hypoechoic), posterior features (no features) and final assessment (category 3). In the multi-institution group, reviewers agreed on mass shape (oval), mass orientation (not parallel), margin (circumscribed), posterior features (no features) and final assessment (category 3). Reviewers did not agree on echo pattern (isoechoic in two and hypoechoic in one).

summarized in Table 2. Representative cases are shown in Figures 1 and 2.

For the five readers, the overall agreement for mass orientation was substantial (κ= 0.63). The overall agreement for mass shape was moderate (κ= 0.54). Fair agreement was seen when the descriptors of mass margin, echo pattern, posterior features and final assessment category were used (κ= 0.29, 0.28, 0.30 and 0.40, respectively).

Greater agreement for the final assessment category was seen with assessments of categories 2, 3 and 5 (κ= 0.58,0.57 and 0.43, respectively) and lesser agreement was seen with assessments of category 4A, 4B and 4C (κ= 0.35, 0.10 and 0.21, respectively).

The overall agreement levels between the group with under 10 years of experience and the group with more than 10 years of experience were similar; substantial for mass orientation(κ= 0.71 vs. 0.61), and fair for both margin (κ= 0.27 vs. 0.23) and echo pattern (κ= 0.27 vs. 0.31). While, the agreement for the assessment category was higher in the group with more than 10 years of experience than in the group with under 10 years of experience (κ= 0.44 vs. κ= 0.37), the agreements for mass shape and posterior features were lower in the group with more than 10 years of experience (κ= 0.49 and 0.01, respectively) than in the group with under 10 years of experience (κ= 0.63 and 0.36, respectively). There were no significant differences for other descriptors between the two groups except for posterior features.

Diagnostic performance

Table 3 shows the diagnostic performance of the readers and of the groups according to years of experience. The overall sensitivity, specificity, positive predictive value, negative predictive value and accuracy were as follows; 96.4% (106/110), 52.8% (153/290), 43.6%(106/243), 97.5%(153/157), and 64.7%(259/400), respectively. The area under the curve of the 5 radiologists was 0.885 in overall and ranged from 0.876 to 0.898. There were no statistically significant differences in diagnostic performance according to years of experience (range of p values, 0.31 to 0.46).

Discussion

Our results show fair to substantial agreement for the sonography-based description of breast lesions, which is similar to results from previous literature (10-13, 16) (Table 4). Thus, this study validates the fifth edition of the US lexicon. In the fifth edition of BI-RADS, what was previously known as the complex descriptor was updated to the complex cystic and solid descriptors. Although the terminology indicated identical features, it was updated to clarify what the lexicon defined. Heterogeneous echogenicity, a mixture of echogenic patterns within a solid mass, was newly added to the lexicon and was used frequently in the descriptions

Table 3. Diagnostic Performance of Readers According to Years of Experience

		Reader					Years of Experience		p
	Overall	1	2	3	4	5	<10	≥10	value
Sensitivity	96.4	95.45	90.91	95.45	95.45	95.45	93.94	95.45	0.31
Specificity	52.8	51.72	53.45	60.34	50.00	51.72	55.17	51.72	0.34
PPV	43.6	42.86	42.55	47.73	42.00	42.86	44.29	42.86	0.46
NPV	97.5	96.77	93.94	97.22	96.67	96.77	96.00	96.77	0.43
Accuracy	64.7	63.75	63.75	70.00	62.50	63.75	65.83	63.75	0.43

Note - PPV, positive predictive value. NPV, negative predictive value.
Numbers are percentages.

Table 4. Interobserver Variability According to the BI-RADS 4th Edition and 5th Edition

		κ-value					
		BI-RADS 5th edition	BI-RADS 4th edition				
		Present study	Lazarus' (13)	Park's (12)	Lee's (11)	Abdullah's (10)	Berg's (16)
Case number		80	62	314	150	267	70
Reader number		5	5	4	4	5	35
Shape	Oval	0.58 (0.14)	0.71		0.47	0.68	0.59
	Round	0.37 (0.84)	0.29		0.32	0.70	0.30
	Irregular	0.58 (0.09)	0.70		0.58	0.63	0.67
	Overall	0.54 (0.04)	0.66	0.42	0.49	0.64	0.58
Orientation	Parallel	0.63 (0.17)	0.61		0.56	0.70	0.45
	Not parallel	0.63 (0.10)	0.61		0.56	0.70	0.48
	Overall	0.63 (0.04)	0.61	0.61	0.56	0.70	0.46
Margin	Circumscribed	0.56 (0.10)	0.71		0.42	0.68	0.54
	Non-circumscribed						0.50
	Indistinct	0.06 (0.09)	0.22		0.20	0.39	
	Angular	0.14 (0.08)	0.22		0.21	0.45	
	Microlobulated	0.20 (0.09)	0.25		0.25	0.33	
	Spiculated	0.39 (0.08)	0.26		0.66	0.52	
	Overall	0.29 (0.02)	0.40	0.32	0.33	0.36	0.51
Echo pattern	Anechoic	−0.007 (0.15)	−0.01		0.00	0.95	NA
	Hyperechoic	0.11 (0.14)	0.16		0.00	0.90	0.77
	Complex cystic and solid*	0.31 (0.08)	0.40		0.13	0.70	0.44
	Hypoechoic	0.31 (0.15)	0.29		0.41	0.49	0.44
	Isoechoic	0.29 (0.08)	0.05		0.38	0.59	0.27
	Heterogeneous **	0.19 (0.08)					
	Overall	0.28 (0.04)	0.29	0.36	0.37	0.58	0.41
Posterior features	No features	0.30 (0.15)	0.31		0.47	0.40	0.63
	Enhancement	0.37 (0.08)	0.39		0.50	0.54	0.60
	Shadowing	0.42 (0.07)	0.66		0.59	0.74	0.84
	Combined	−0.02 (0.08)	0.09		0.14	0.40	
	Overall	0.30 (0.04)	0.40	0.53	0.49	0.47	0.63
Category	Category 2	0.58 (0.08)	0.27	0.66		0.91	0.37
	Category 3	0.57 (0.10)	0.32	0.26	0.58	0.68	0.32
	Category 4A	0.35 (0.10)	0.14	0.30	0.57	0.33	0.19
	Category 4B	0.10 (0.08)	0.16		0.09	0.32	0.17
	Category 4C	0.21 (0.08)	0.26		0.38	0.17	0.12
	Category 5	0.43 (0.08)	0.56	0.54	0.71	0.60	0.52
	Overall	0.40 (0.02)	0.28	0.49	0.53	0.30	0.46

Note - * 'complex echoic descriptor' has been changed to 'complex cystic and solid' in the BI-RADS 5th edition.
** heterogeneous is now included as an echo pattern in the BI-RADS 5th edition.
NA (not applicable)

in this study (10.2%). However, heterogeneous echogenicity showed only slight agreement, which was a low value among the lexicons of echogenicity and had the lowest value among the lexicons used more than 10%. There were a few potential reasons for the lower agreement of heterogeneous echogenicity; first, radiologists might still be learning how to apply this new lexicon and might not be accustomed to using this term yet. Second, variability may be inevitable when using heterogeneous echogenicity. With the fourth edition of BI-RADS, radiologists would select a single dominant echogenicity in cases with lesions showing a mixture of echogenic patterns. However, in the updated edition, some radiologists are likely to try to choose a single dominant echogenicity while others choose to assign heterogeneous echogenicity. There was no cutoff value for the percentage of minor echogenicity for heterogeneity. Moreover, the echogenic halo was deleted in this edition. As a result, radiologists could choose heterogeneous echogenicity for hypo- or isoechoic lesions with echogenic halos. In this study, the positive predictive value was not included and needs to be calculated through consensus for usage in this lexicon. Another update was in the subcategories of category 4. However, this subcategorization has been widely used in previous studies on interobserver agreement with broad variability (0.14 to 0.57 of κ value) (10-13, 16). This study showed fair agreement, within the range of results from previous literature. Besides echogenicity and the subcategorization of category 4, sections on tissue composition, associated features, and elastography were added to the fifth edition of BI-RADS, but were not covered in this study (1).

Among the retained lexicons, the overall agreement for mass margin was fair in this study, which was relatively low, but similar to that of previous studies (10-13). The lower rate of agreement for mass margin was likely due to the multitude of possible terms for describing margins (11-13). The agreement of "circumscribed" margin was fair to moderate, which is consistent with several previous studies (11-13, 16). However, whenever the margin was categorized by one of the 'non-circumscribed' descriptors, the mass would categorized as category 4 or more, which indicates that the clinical impact from the interobsever variability for non-circumscribed margins may be obscured. Non-circumscribed margins may affect how category 4 lesions are subcategorized for final assessment, an issue with high variability in this study and previous literatures (10, 11, 13, 16).

We saw a relatively lower agreement for posterior features of this study, especially the combined pattern. The combined pattern had the lowest agreement among posterior features in this study (κ = -0.02) as well as in previous literature (0.09 to 0.40) (10-13). Agreement tended to be higher in the study that did not include the combined pattern in posterior features (κ = 0.63) (16) compared to the studies that did (0.40 to 0.49) (10, 11, 13). In addition, the agreement for posterior features was significantly different according to years of experience, a result which may be due to a preference for the combined pattern after longer experience with breast imaging. However, there were no significant differences in diagnostic performance between the two groups in spite of variability.

The agreements for mass shape, orientation, margin, echo pattern and category did not differ according to years of experience in this study. These results suggest that the BI-RADS lexicon can be widely used with similar interobserver agreement.

The cases included in our study were obtained from several institutions which used 4 different types of equipment from 4 different manufacturers. Although the images reviewed might not have been optimized to each radiologist' preference and using different equipment might play a role in lower agreement for posterior features and margin, the diagnostic performance of US-BI-RADS was excellent, showing 0.885 for the area under the curve. Diagnostic performance is a tradeoff between sensitivity and specificity, which, in our study, was within the range of previous literatures (11, 12, 16). Therefore, our results support the use of US-BI-RADS in the US evaluation of breast lesions and its use in medical audit for screening projects.

Our study had some limitations. First, assessment was limited to representative static images, which may not reflect actual practice. Second, calcifications, associated findings and special cases were not included in this analysis. Third, the sample size was small and there might be a selection bias. Fourth, all readers were experts in breast imaging and worked in academic institutions. Therefore, the results found here cannot be generalized to other breast radiologists.

In conclusion, the interobserver agreement of radiologists between multi-institutions with the fifth edition of the BI-RADS lexicon is similar to that with the fourth edition of the BI-RADS lexicon. Agreement according to years of experience was not significantly different. Diagnostic performances with cases from multi-institutions were excellent and not affected by variability.

References

1. American College of Radiology. Breast Imaging Reporting and Data System Atlas . 5th edition. Reston VA:American College of Radiology 2013.
2. American College or Radiology. Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas) 4th edition. Reston VA: American College of Radiology 2003. 2003
3. Hong AS, Rosen EL, Soo MS, Baker JA. BI-RADS for sonography: positive and negative predictive values of sonographic features. AJR Am J Roentgenol 2005;184:1260-1265
4. Kim EK, Ko KH, Oh KK, Kwak JY, You JK, Kim MJ, et al. Clinical application of the BI-RADS final assessment to breast sonography in conjunction with mammography. AJR Am J Roentgenol 2008;190:1209-1215
5. Raza S, Chikarmane SA, Neilsen SS, Zorn LM, Birdwell RL. BI-RADS 3, 4, and 5 lesions: value of US in management--follow-up and outcome. Radiology 2008;248:773-781
6. Satake H, Nishio A, Ikeda M, Ishigaki S, Shimamoto K, Hirano M, et al. Predictive value for malignancy of suspicious breast masses of BI-RADS categories 4 and 5 using ultrasound elastography and MR diffusion-weighted imaging. AJR Am J Roentgenol 2011;196:202-209
7. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Velez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA 2008;299:2151-2163
8. Hooley RJ, Greenberg KL, Stackhouse RM, Geisel JL, Butler RS, Philpotts LE. Screening US in patients with mammographically dense breasts: initial experience with Connecticut Public Act 09-41. Radiology 2012;265:59-69
9. Moon HJ, Jung I, Park SJ, Kim MJ, Youk JH, Kim EK. Comparison of Cancer Yields and Diagnostic Performance of Screening Mammography vs. Supplemental Screening Ultrasound in 4394 Women with Average Risk for Breast Cancer. Ultraschall Med 2015;36:255-263
10. Abdullah N, Mesurolle B, El-Khoury M, Kao E. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. Radiology 2009;252:665-672
11. Lee H-J, Kim E-K, Kim MJ, Youk JH, Lee JY, Kang DR, et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. Eur J Radiology 2008;65:293-298
12. Park CS, Lee JH, Yim HW, Kang BJ, Kim HS, Jung JI, et al. Observer agreement using the ACR Breast Imaging Reporting and Data System (BI-RADS)-ultrasound, First Edition (2003). Korean J Radiol 2007;8:397-402
13. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 2006;239:385-391
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174
15. Gwet K. Computing Inter-Rater Reliability With the

SAS System. Statistical Methods For Inter-Rater Reliability Assessment 2002;3:1-16

16. Berg WA, Blume JD, Cormack JB, Mendelson EB. Training the ACRIN 6666 Investigators and effects of feedback on breast ultrasound interpretive performance and agreement in BI-RADS ultrasound feature analysis. AJR Am J Roentgenol 2012;199:224-235

대한유방검진의학회지 2016;13:93-103

ACR BI-RADS 초음파 용어집 5판을 적용한 다기관 판독자 일치도와 진단능 분석

김성헌[1] · 김동욱[2] · 육지현[3] · 윤정현[4] · 정혜선[5] · 김유미[6] · 이은혜[5] · 김민정[4]

[1]가톨릭대학교 서울성모병원 영상의학과, [2]국민건강보험공단 일산병원 연구소 정책연구부,
[3]연세대학교 강남세브란스병원 영상의학과, [4]연세대학교 신촌세브란스병원 영상의학과,
[5]순천향대학교 부천순천향병원 영상의학과, [6]단국대학교병원 영상의학과

목적: BI-RADS 개정판을 적용하여 유방 종괴 판독시 다기관 판독자 간의 일치도를 평가하였고 유방영상 판독경력에 따른 일치도와 진단능을 분석하였다.

대상 및 방법: 5개 기관에서 74명 환자, 80개 유방 종괴의 초음파 영상을 얻었다 (양성 51개, 악성 29개). 5명의 판독자가 각 증례의 직교 영상을 독립적으로 분석하였다; 동일 기관 판독자 2명, 다른 기관 판독자 3명과 유방영상 판독경력 10년 미만 판독자 2명, 10년 이상 판독자 3명. 각 판독자는 개정된 BI-RADS 에 따라 유방 종괴의 소견을 기술하고 최종범주를 평가하였다. 판독자간 일치도는 Cohen's kappa 를 사용하여 분석하였고 진단능은 유방영상 판독경력에 따라 비교 분석하였다.

결과: 5명의 판독자간 일치도는 약간에서 상당한 정도까지 일치했다: 방향 (κ=0.63), 모양 (κ =0.54), 최종범주 (κ=0.40), 후방음영 (κ=0.30), 변영 (κ=0.29), 에코양상 (κ=0.28). 후방음영외에는 기관과 경력에 따라 판독자가 일치도에는 차이가 없었다. 5명 판독자의 민감도, 특이도, 양성예측도, 음성예측도 및 정확도는 다음과 같았고 판독경력에 따른 차이는 없었다; 96.4% (106/110), 52.8% (153/290), 43.6% (106/243), 97.5% (153/157), 64.7% (259/400).

결론: 판독자간 일치도는 판독 경력에 따른 차이 없이 우수하였고 이를 통해 BI-RADS 용어집 5판의 유용성이 입증 되었다. 다기관 판독자의 진단능은 판독 경력에 따른 차이 없이 우수하였다.

Index words: Ultrasonography; Breast neoplasms; Breast diseases; Observer variation

Corresponding author: Min Jung Kim, M.D.