BI-RADS (Breast Imaging Reporting and Data System) for breast ultrasound : Inter- and intraobserver variability

Hye-Jeong Lee

Department of Medicine

The Graduate School, Yonsei University

BI-RADS (Breast Imaging Reporting and Data System) for breast ultrasound : Inter- and intraobserver variability

Directed by Professor Eun-Kyung Kim

The Master's Thesis submitted to the Department of Medicine the Graduate School of Yonsei University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Hye-Jeong Lee

June 2006

This certifies that the Master's Thesis
of Hye-Jeong Lee is approved.

Thesis Supervisor : Eun-Kyung Kim

Jong Tae Lee : Thesis Committee Member

Byeong-Woo Park : Thesis Committee Member

The Graduate School
Yonsei University

June 2006

Acknowledgement

I would like to express my gratitude to my supervisor Prof. Eun-Kyung Kim for her warm personality and tireless support during my Master's Degree course.

I would also like to extent my gratitude to all my committee members, Prof. Jong Tae Lee and Prof. Byeong-Woo Park. They have helped me with sincere directions, encouragements, advices and revisions from the starting of my research to the finishing of writing this thesis.

I want to thank Prof. Min Jung Kim, M.D., Ji Hyun Youk, M.D. Ji Young Lee for their helps in reviewing the cases for this thesis. I want to thank Ph.D. Dae Ryong Kang and Ph.D. Chung Mo Nam for their helps in statistical analysis for this thesis. If it were not for their help, it would have been impossible for this thesis to come to light.

Finally, I deeply appreciate my parents, my sister and my husband for being beside me with love.

Lee Hye-Jeong

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Abstract

To retrospectively evaluate inter- and intra-observer variability in sonographic feature analysis and management, using the fourth edition of the Breast Imaging Reporting and Data System (BI-RADS). We included 136 patients with 150 breast lesions who underwent breast ultrasound (US) and core needle biopsy. A pathologic diagnosis was available for all 150 lesions: 77 (51%) malignant and 73 (49%) benign. The size of the lesions ranged from 3mm to 45 mm (mean, 14.1 mm). Four radiologists retrospectively reviewed sonographic images of lesions twice within an eight-week interval. The observers described each lesion, using BI-RADS descriptors and final assessment. Inter- and intra-observer variability was assessed with Cohen's kappa statistic. Sensitivity, specificity, positive predictive value, and negative predictive value were also calculated. Inter-observer agreements for sonographic descriptors were as follows: substantial agreement was obtained for lesion calcification and final assessment (κ = 0.61 for both), moderate agreement was obtained for lesion shape, orientation, boundary and posterior acoustic features (κ = 0.49, 0.56, 0.59 and 0.49, respectively), and fair agreement was achieved for lesion margin and echo pattern (κ = 0.33 and 0.37, respectively). For intra-observer agreement, substantial to perfect agreement was found for almost all lesion

descriptors and final assessments. Diagnostic indices were as follows: sensitivity, 98%; specificity, 33%; and negative predictive value, 95%. Positive predictive values of malignancy for the sonographic BI-RADS final assessment categorized as 4 or 5 were as follows: category 4a, 26%; category 4b, 89%; category 4c, 90%; and category 5, 97%. The high level of inter- and intra-observer agreement for BI-RADS descriptors and final assessment with sonography validates the use of BI-RADS lexicon for breast US.

Key words : Breast ultrasound, BI-RADS lexicon,

Inter-observer variability, Intra-observer variability

BI-RADS (Breast Imaging Reporting and Data System) for Breast Ultrasound : Inter- and Intraobserver Variability

Hey-Jeong Lee

Department of Medicine

The Graduate School, Yonsei University

Directed by Professor Eun-Kyung Kim

I. INTRODUCTION

In an effort to standardize mammographic reporting, the American College of Radiology and experts in mammography developed the Breast Imaging Reporting and Data System (BI-RADS) lexicon[1,2]. Until recently, BI-RADS had only been applied to mammography and not to other breast imaging techniques. Although

mammography is recognized as the best method of screening for breast cancer, breast ultrasound (US) has become well established as a valuable imaging technique[3]. In light of the widespread use of US, the ACR recently developed a BI-RADS lexicon for breast US to standardize the sonographic characterization of lesions[4,5] This lexicon includes descriptors of features such as mass shape, orientation, margin, posterior acoustic features, vascularity, and other sonographic features.

Although there has been some controversy regarding the utility of US for determining the likelihood of malignancy of solid breast masses[6,7], several studies have suggested that sonographic appearance can be useful in differentiating benign from malignant solid breast masses[8-10]. Because sonographic interpretation is subjective, observer variability is inherent in breast US analysis. The variability in US interpretation is attributable to differences in lesion detection and to variation in lesion characterization and subsequent management. However, to our knowledge, variability in the use of BI-RADS terminology for breast US has not been widely studied.

Therefore, the purpose of this study was to retrospectively assess inter- and intra-observer variability in sonographic feature analysis and management of breast lesions.

II. MATERIALS AND METHODS

1. Patient Population

Institutional review board approval was obtained for this retrospective study, and informed patient consent was not required.

Between January 2003 and May 2003 and 295 consecutive women with 324 breast lesions underwent US with core needle biopsy. Of these 324 lesions, we included only those that were pathologically proven to be malignant or those that were benign with no change in size by US for at least two years. Thus, a total of 136 patients with 150 lesions met the selection criteria and were included in this study. In 14 patients, more than one lesion had been detected.

The 136 patients ranged in age from 21 to 76 years (mean, 46.7 years). Of the 150 lesions, 72 had undergone core needle biopsy only, 7 had undergone both core needle biopsy and directional vacuum-assisted removal, and 71 lesions had undergone core needle biopsy and surgical excision. The size range of the lesions was 3mm to 45 mm (mean, 14.1 mm), 77 of the lesions were malignant, and 73 were benign. The breast lesions in our study are shown in Table 1. The most frequently observed malignant breast lesion in the patients was infiltrating ductal cell carcinoma and the most common benign lesion was fibroadenoma.

Table 1. Distribution of breast lesions in this study

Malignant lesions	77	Benign lesions	73
Infiltrating ductal cell carcinoma	62	Fibroadenoma	19
Ductal cell carcinoma in situ	12	Fibrocystic change	18
Medullary carcinoma	3	Fibroadenomatous hyperplasia	15
		Stromal fibrosis	12
		Fat necrosis	3
		Adenosis	3
		Ductal epithelial hyperplasia	2
		Papillary epithelial hyperplasia	1

2. Retrospective Reviews of Breast Ultrasound

One radiologist (L.H.J, radiology resident) selected two representative transverse and longitudinal sonographic images for each lesion and converted the sonographic images into TIFF (Tag Image File Format) files with 300 dpi (dots per inch). Then she arranged them in random order in Microsoft Power Point XP. Four experienced breast radiologists (Y.J.H, K.M.J, K.E.K, and L.J.Y) with 1, 4, 10, and 1 years of breast imaging experience, respectively, reviewed the sonographic images. To evaluate inter-observer agreement, each radiologist independently evaluated the images. To assess the intra-observer agreement, each radiologist reevaluated the

sonographic images in random order. To reduce learning effects, the interval between the two evaluations was at least eight weeks. All observers were blinded to the clinical information and pathologic results of each case as well as the ratio of malignant to benign lesions included in the study. The radiologists evaluated the images according to the BI-RADS lexicon for US. The contents of the BI-RADS descriptors are shown in Table 2.

Table 2. The contents of BI-RADS descriptors for breast sonography

BI-RADS lexicon	
Shape	oval round irregular
Orientation	parallel non parallel
Margin	circumscribed indistinct angular
	microlobulated spiculated
Lesion Boundary	abrupt interface echogenic halo
Echo pattern	anechoic hyperechoic complex
	hypoechoic isoechoic
Posterior acoustic features	absent enhancement shadowing combined
Calcification	macrocalcification
	microcalcification in a mass
	microcalcification out of a mass
	microcalcification in and out of a mass
Final assessment	Category 3 Category 4 (a, b, c) Category 5

Each observer was instructed to choose the most appropriate term to describe each

lesion. BI-RADS final assessments were combined with corresponding recommendations (Table 3). Although there are seven categories (from 0 to 6) in BI-RADS lexicon, we only used BI-RADS category 3, 4a, 4b, 4c, and 5 in this study. The BI-RADS categories of 0, 1, 2, and 6 were excluded in this study.

Table 3. BI-RADS categories with recommendations

BI-RADS category	Assessment	Recommendations
Category 3	Probably benign	Initial short interval follow-up
Category 4	Suspicious malignancy	Biopsy should be considered
4a	Low likelihood of malignancy	
4b	Intermediate likelihood of malignancy	
4c	Moderate likelihood of malignancy	
Category 5	Highly suggestive of malignancy	Appropriate action should be taken

3. Statistical Analysis

Kappa statistics were calculated, using the SAS system (MAGREE SAS Macro program) to assess the proportion of inter- and intra-observer agreement beyond that expected by chance[11]. The method for estimating an overall kappa value in the case of multiple observers and multiple categories is based on the work of Landis and Koch [12] as follows: for each category *j*, a kappa statistic (for multiple raters) is

calculated comparing category *j* with the other categories pooled. A weighted average is used to combine these kappa values, where the weight for a given kappa value is the product of *pj*, the proportion of ratings in category *j*, and (1 – *pj*), the proportion of ratings not in category *j*. A value of $\kappa = 1.0$ corresponds to complete agreement, 0 to no agreement, and less than 0 to disagreement. Landis and Koch [12] have suggested that a kappa value (κ) 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 0.81–1.00, almost perfect agreement.

III. RESULTS

1. Inter-observer Variability

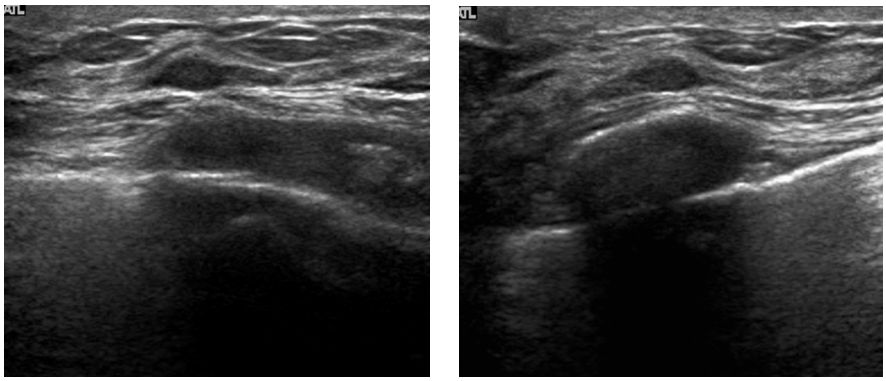
The summary of inter-observer variability for each descriptor and final assessment category is shown in Table 4. We obtained a relatively high degree of inter-observer agreement for BI-RADS lexicon (Figures 1 and 2).

Table 4. Inter-observer variability in BI-RADS descriptors for breast US.

BI-RADS lexicon for US	Descriptors	κ	SE	Prob>Z
Shape	oval	0.47	0.03	<0.0001
	round	0.32	0.03	<0.0001
	irregular	0.58	0.03	<0.0001
	overall	0.49	0.03	<0.0001
Orientation	parallel	0.56	0.03	<0.0001
	not parallel	0.56	0.03	<0.0001
	overall	0.56	0.03	<0.0001
Margin	circumscribed	0.42	0.03	<0.0001
	indistinct	0.2	0.03	<0.0001
	angular	0.21	0.03	<0.0001
	microlobulated	0.25	0.03	<0.0001
	spiculated	0.66	0.03	<0.0001
	overall	0.33	0.02	<0.0001
Boundary	abrupt interface	0.59	0.03	<0.0001
	echogenic halo	0.59	0.03	<0.0001
	overall	0.59	0.03	<0.0001

Echo Pattern	anechoic	-0.00	0.03	0.52
	hyperechoic	0.00	0.03	0.52
	complex	0.13	0.03	<0.0001
	hypoechoic	0.41	0.03	<0.0001
	isoechoic	0.38	0.03	<0.0001
	overall	0.37	0.03	<0.0001
Posterior Acoustic Features	absent	0.47	0.03	<0.0001
	enhancement	0.5	0.03	<0.0001
	shadowing	0.59	0.03	<0.0001
	combined	0.14	0.03	<0.0001
	overall	0.49	0.02	<0.0001
Calcification	absent	0.64	0.03	<0.0001
	macrocalcification	-0.00	0.03	0.54
	Microcalcification in a mass	0.6	0.03	<0.0001
	microcalcification out of a mass	-0.00	0.03	0.54
	micro in and out of a mass	0.53	0.03	<0.0001
	overall	0.61	0.03	<0.0001
Final Assessment	category 3	0.58	0.03	<0.0001
	category 4a	0.57	0.03	<0.0001
	category 4b	0.09	0.03	0.0024
	category 4c	0.38	0.03	<0.0001
	category 5	0.71	0.03	<0.0001
	overall	0.53	0.02	<0.0001

Shape – Moderate agreement among our radiologists was seen in describing breast shape, with an overall kappa value of 0.49. Moderate agreement was seen in the use of the terms “oval” (κ=0.47) and “irregular” (κ=0.58). Fair agreement was seen in use of the term “round” (κ=0.32).

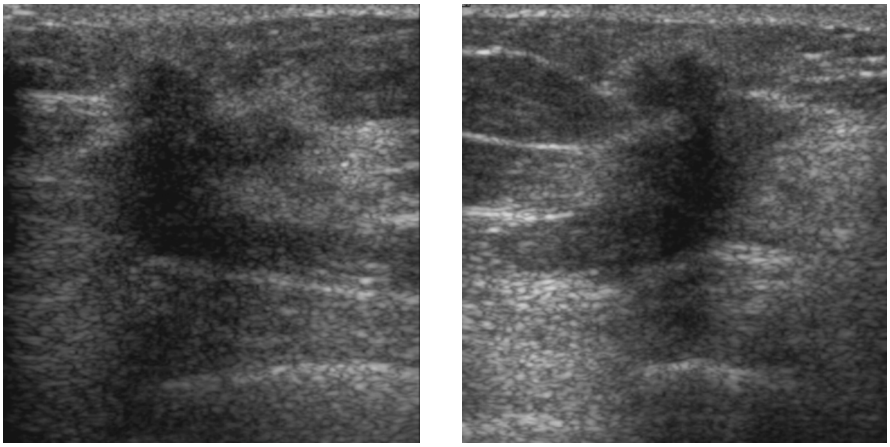


A. B.

Figure 1. This is a representative case that showed perfect inter-observer agreement for BI-RADS lexicon : transverse image (A) and longitudinal image (B). All observers described the mass as having oval shape, parallel orientation, circumscribed margin, abrupt interface for lesion boundary, hypoechoic echogenicity, no posterior acoustic feature, and no calcification. The mass was finally categorized as “category 3” by all observers and was pathologically confirmed as a benign mass.

Orientation – In describing the lesion orientation, overall agreement was moderate (***κ***=0.56). Moderate agreement was seen when the mass orientation was characterized as “parallel” (***κ***=0.56) or “not parallel” (***κ***=0.56).

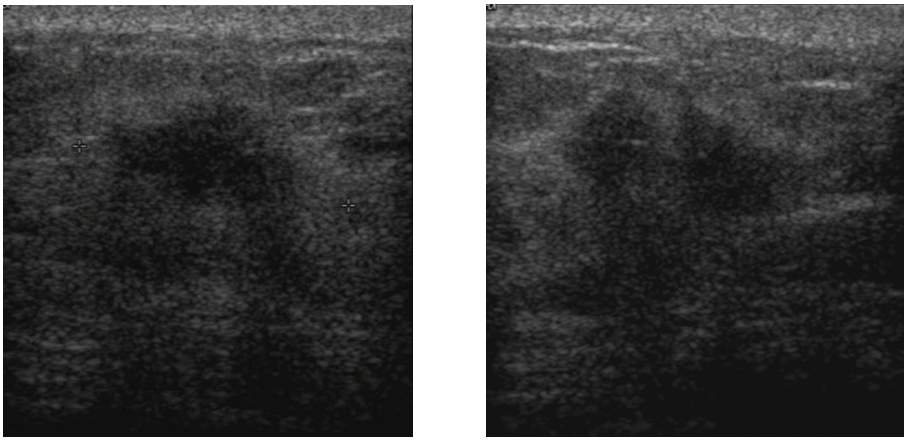
Margin – Overall agreement for the margins of the mass was fair (***κ***=0.33) (Figure 3). Agreement was substantial when mass margin was described as “spiculated” (***κ***=0.66). Moderate agreement was seen with lesions that were considered



A. B.

Figure 2. This is a representative case that showed perfect inter-observer agreement for BI-RADS lexicon: transverse image (A) and longitudinal image (B). All observers described the mass as having irregular shape, non-parallel orientation, speculated margin, echogenic halo for lesion boundary, hypoechoic echogenicity, posterior acoustic shadowing, and no calcification. The mass was finally categorized as “category 5” by all observers and was pathologically confirmed as a malignant mass.

“circumscribed” (κ=0.42). Fair agreement was seen in the use of the term “angular” (κ=0.21) and “microlobulated” (κ=0.25). Agreement was slight when mass margin was characterized as “indistinct” (κ=0.20). We also evaluated inter-observer agreement for margin, classified as circumscribed or not circumscribed. Overall



A. B.

Figure 3. This is a representative case that showed fair inter-observer agreement for the mass margin : transverse image (A) and longitudinal image (B). Each of the four observers used a different descriptor for the margin. The descriptive terms that were used were spiculated, angular, indistinct, and microlobulated. However, no one used the descriptor "circumscribed". The mass was pathologically confirmed as malignant.

Table 5. Inter-observer variability for mass margin with only "circumscribed" and " not circumscribed"

BI-RADS lexicon	**Descriptors**	***κ***	**SE**	**Prob>Z**
Margin	Circumscribed	0.48	0.03	<0.0001
	Not circumscribed	0.48	0.03	<0.0001
	Overall	0.48	0.03	<0.0001

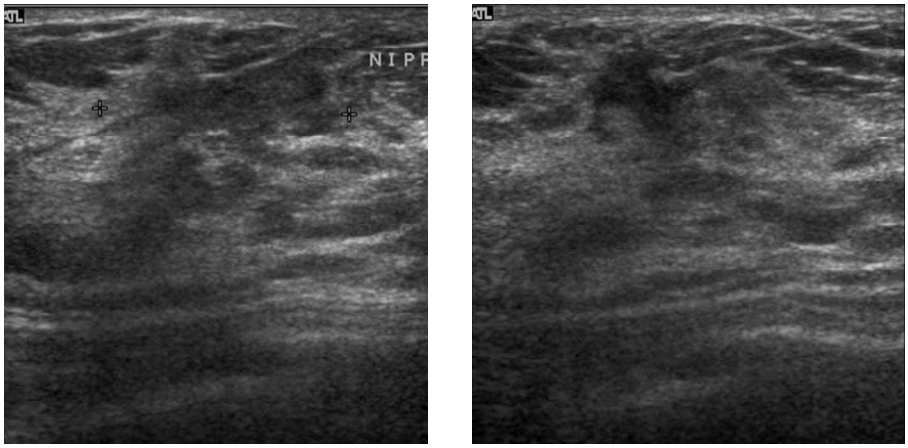
agreement (κ=0.48) and agreement for "circumscribed" (κ=0.48) and "not circumscribed" (κ=0.48) were moderate (Table 5).

Boundary – Moderate agreement was achieved for the evaluation of the lesion boundary (κ=0.59), "abrupt interface" (κ=0.59), and "echogenic halo" (κ=0.59).

Echo pattern – In assessing the echo pattern, overall agreement was fair (κ=0.37) (Figure 4). Agreement was poor for use of the descriptors "anechoic" (κ=-0.00), "hyperechoic" (κ=0.00), and "complex" (κ=0.13). The agreement was moderate, however, for use of the descriptor "hypoechoic" (κ=0.41) and fair when using "isoechoic" (κ=0.38).

Posterior acoustic features – Overall moderate agreement was achieved in describing posterior acoustic features (κ=0.49). Agreement was also moderate for descriptors "absent" (κ=0.47), "enhancement" (κ=0.50), and "shadowing" (κ=0.59), although agreement was poor for the descriptor "combined" (κ=0.14).

Calcification – Overall substantial agreement was found for evaluation of the calcification (κ=0.61). Agreement between observers for the "absence" of calcification was substantial (κ=0.64). Moderate agreement was seen in the "microcalcification in a mass" (κ=0.60) and "microcalcification both in and out of a mass" (κ=0.53). Poor agreement was seen for "macrocalcification" (κ=-0.00) and "microcalcification out of a mass" (κ =-0.00).



A. B.

Figure 4. This is a representative case that showed fair inter-observer agreement for the echo pattern of the mass : transverse image (A) and longitudinal image (B). There was a discrepancy between observers in describing the echo pattern for this mass. The descriptive terms that were used were complex, hypoechoic and isoechoic. The mass was pathologically confirmed as malignant.

Final assessment Category – In assessing the final category, overall agreement was moderate (0.53). Substantial agreement was achieved when the mass was considered a "Category 5" (κ=0.71), and moderate agreement was seen for "Category 3" (κ=0.58) and "Category 4a" (κ=0.57). Agreement was fair (κ=0.38) for "Category 4c" and was slight (κ=0.19) for "Category 4b". When we classified the BI-RADS final assessment category as 3, 4, or 5, (Table 6), the overall

agreement was substantial (κ=0.62). Agreement for “Category 3” and “Category 4” was moderate (κ=0.56 and 0.51, respectively). Substantial agreement was seen when the mass was considered a “Category 5” (κ=0.71), as well.

Table 6. Inter-observer variability for final assessment without consideration for subcategories

Final assessment	κ	**SE**	**Prob>Z**
Category 3	0.56	0.03	<0.0001
Category 4	0.51	0.03	<0.0001
Category 5	0.71	0.03	<0.0001
Overall	0.62	0.03	<0.0001

2. Intra-observer Variability

Intra-observer variability for each descriptor and final assessment category are summarized in Table 7. When lesions were reevaluated by the same radiologist, relatively substantial agreement was achieved.

Table 7. Intra-observer variability in BI-RADS descriptors and final assessment categories.

Observer	BI-RADS lexicon	κ	SE	Prob>Z
Observer 1	shape	0.71	0.06	<0.0001
	orientation	0.83	0.08	<0.0001
	margin	0.59	0.08	<0.0001
	boundary	0.85	0.08	<0.0001
	echo pattern	0.67	0.07	<0.0001
	posterior acoustic features	0.82	0.06	<0.0001
	calcification	0.8	0.08	<0.0001
	final assessment	0.77	0.06	<0.0001
Observer 2	shape	0.72	0.06	<0.0001
	orientation	0.8	0.08	<0.0001
	margin	0.53	0.04	<0.0001
	boundary	0.56	0.08	<0.0001
	echo pattern	0.81	0.08	<0.0001
	posterior acoustic features	0.68	0.06	<0.0001
	calcification	0.9	0.07	<0.0001
	final assessment	0.72	0.06	<0.0001
Observer 3	shape	0.66	0.06	<0.0001
	orientation	0.81	0.08	<0.0001
	margin	0.53	0.04	<0.0001
	boundary	0.8	0.08	<0.0001
	echo pattern	0.74	0.08	<0.0001
	posterior acoustic features	0.69	0.06	<0.0001
	calcification	0.84	0.08	<0.0001
	final assessment	0.79	0.06	<0.0001
Observer 4	shape	0.56	0.07	<0.0001
	orientation	0.75	0.08	<0.0001
	margin	0.61	0.05	<0.0001

	boundary	0.75	0.08	<0.0001
	echo pattern	0.72	0.07	<0.0001
	posterior acoustic features	0.67	0.06	<0.0001
	calcification	0.73	0.07	<0.0001
	final assessment	0.73	0.06	<0.0001

Observer 1 – Agreement was perfect for mass orientation (***κ***=0.83), boundary (***κ***=0.85), acoustic features (***κ***=0.82) and calcification (***κ***=0.90). Moderate agreement was seen for the mass shape (***κ***=0.71) and echo pattern (***κ***=0.67). For the margin, moderate agreement was seen (***κ***=0.59). Substantial agreement was achieved for the final assessment category (***κ***=0.77).

Observer 2 – Perfect agreement was found for the mass echo pattern (***κ***=0.81) and calcification (***κ***=0.90). For mass shape, orientation, posterior acoustic features, and final assessment category, agreement was substantial (***κ***=0.72, 0.80, 0.68, and 0.72, respectively). Moderate agreement was achieved for margin (***κ***=0.53) and boundary (***κ***=0.56)

Observer 3 – For mass orientation (***κ***=0.81) and calcification (***κ***=0.84), perfect agreement was seen. Substantial agreement was seen in evaluating mass shape (***κ***=0.66), echo pattern (***κ***=0.74), boundary (***κ***=0.80), acoustic features (***κ***=0.69), and final assessment category (***κ***=0.79). Moderate agreement was seen in the mass margin (***κ***=0.53).

Observer 4 –Assessment of most descriptors, except for mass shape, orientation,

margin, boundary, echo pattern, posterior acoustic features, calcification, and final assessment category (κ=0.75, 0.61, 0.75, 0.72, 0.67, 0.73 and 0.73, respectively) revealed substantial agreement. For mass shape, moderate agreement was seen (κ=0.53).

3. Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value

Diagnostic indices are summarized in Table 8 and Table 9. With the results of all observers combined, four had a sensitivity of 98% and a specificity of 33%. Positive predictive values of malignancy with the US BI-RADS final assessment categorized as 4 or 5 were as follows: category 4a, 26%; category 4b, 89%; category 4c, 90%; and category 5, 97%. Negative predictive value was 95% in our study.

Table 8. Diagnostic indices obtained in this study.

Observer	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
1	97	40	62	94
2	99	29	59	95
3	99	37	62	96
4	99	26	58	95
Overall	98	33	60	95

Table 9. Positive predictive values for category 4 and 5.

Category	4	4a	4b	4c	5
PPV (%)	51	26	89	90	97

IV. DISCUSSION

Breast US is now considered a fundamental method for evaluating a breast mass detected by physical examination and mammography[10]. However, there is inherent variability in breast US interpretation due to variation in lesion characterization and subsequent management. Baker et al[13] examined inter-observer variability in the description and assessment of solid breast masses by US and speculated that descriptor standardization for breast US is needed... In light of the increasing use of US in clinical practice, a standardized BI-RADS lexicon for US was developed in 2003 by the ACR[4]. Use of BI-RADS descriptors for breast US may help with lesion characterization and the determination of malignancy likelihood. However, until recently, there have been few reports that show the inter- and intra-observer variability for breast US with BI-RADS lexicon.

Our study results show a relatively high degree of inter- and intra-observer agreement in the use of BI-RADS descriptors and final assessment categories (Table 4, 7). Representative cases are shown in Figure 1 and Figure 2.

Substantial agreement was seen when the mass was evaluated for calcification (κ=0.61) and final assessment category (κ=0.62). Moderate agreement was achieved for assessment of lesion shape (κ=0.49), orientation (κ=0.56), boundary (κ=0.59) and posterior acoustic features (κ=0.49) (Table 4). However, fair

agreement was found for the mass margin (κ=0.33) (Figure 3). Lazarus et al[14] also reported that fair agreement was achieved in the evaluation of the lesion margin using sonographic descriptors due to the multitude of terms available to describe the mass margin. In this study, we also evaluated inter-observer agreement for classifying the mass margin as “circumscribed” or “not circumscribed” and the overall agreement (κ=0.48), agreement for “circumscribed” (κ=0.48), and agreement for “not circumscribed” (κ=0.48) were moderate (Table 5). Hong et al [3] reported that a circumscribed margin is predictive of a benign lesion with a negative predictive value of 90%. We suggest simplifying the available descriptive terms for the mass margin to “circumscribed” and “not circumscribed” in BI-RADS lexicon.

For echo pattern, fair agreement among radiologists was also achieved. Lazarus et al [14] speculated that the low rate of agreement for lesion echo pattern on sonogram suggests that observers had difficulty in choosing a single descriptor. Actually, many breast solid lesions, especially large lesions, can be described by more than one echo pattern descriptor. In such cases, it is difficult to describe the echo pattern with the preexisting BI-RADS lexicon (Figure 4). We suggest the addition of the term “combined” for heterogeneous solid breast mass to the BI-RADS lexicon for mass echo pattern. This may improve inter-observer agreement for echo pattern, although echo pattern is not very useful in differentiating between benign and

malignant masses [10,13] .

When lesions were reevaluated eight weeks later by the same radiologists, relatively substantial intra-observer agreement was achieved (Table 7). This suggests that each observer has a clear concept of the BI-RADS descriptors. The initial training on the proper use of BI-RADS descriptors for US may be important for breast radiologists. The high degree of intra-observer agreement in our study validates the utility of the sonographic BI-RADS lexicon for interpreting breast US images.

Orel et al [15] reported that mammographic BI-RADS categories have a positive predictive value of 30% for category 4 lesions and 97% for category 5 lesions. In this study, we had a positive predictive value of 51% for category 4 lesions and 97% for category 5 lesions. For the subcategories, there was a positive predictive value of 26% for category 4a lesions, 89% for category 4b lesions, and 90% for category 4c lesions. Lazarus et al[14] speculated that the use of subcategories is helpful in communicating the level of suspicion to referring physicians and patients. However, in this study, the four radiologists infrequently categorized breast lesion as category 4b. Because category 4a and 4c were unlikely, the category 4b was indeterminate. Dividing category 4 into only two sub-categories (i.e. 4a and 4b) may be simpler and more useful in clinical practice. This would also facilitate the decision with concordant or discordant imaging findings for category 4 lesions.

There are some limitations in our study. First, this study only contained benign masses that underwent core needle biopsy and showed no size change based on a follow-up US 2 years after biopsy. As a result, typical benign lesions, cysts or masses that are clearly benign based on US findings were excluded from this study. This exclusion results in a low specificity and negative predictive value. In clinical practice, the specificity and negative predictive value of benign descriptors may be much higher with the addition of typical benign findings.

Second, the four observers assigned a final BI-RADS category for each lesion using only sonographic features. In actuality, the decision for categorization is made using the highest level of both mammographic and US findings. For instance, for a lesion that is categorized as benign by US but is suspicious by mammography, final assessment is made using the most suspicious feature. However, we evaluated only US images. Therefore, sensitivity and PPV would be slightly higher if we evaluated by both US and mammography. But we suggest this effect may be very minimal because our sensitivity and PPV is already suitably high.

The third limitation is that some of the cases had undergone US with core needle biopsy by two of the radiologists prior to this retrospective study. Therefore, there may be a learning effect associated with impressive cases. However, because there were only a few impressive cases in our study, this probably had little effect on our results.

Fourth, the agreement among observers for the presence of associated findings and special cases could not be evaluated due to the rarity of cases in which associated findings or special cases were present. These associated findings can provide a high predictive value for malignancy. Thus, further study of associated findings and special cases of breast US is needed.

The fifth limitation is that the four radiologists reviewed the sonographic images using a software program, not a workstation. However, this was unlikely to be a problem because the monitor resolution was comparable.

V. CONCLUSION

In conclusion, because inter- and intra-observer agreement with the BI-RADS lexicon for US is good, the use of BI-RADS lexicon can provide accurate and consistent description and assessment for breast US.

Reference

1. American College of Radiology. Breast imaging Reporting and Data System, 2nd ed. Reston, VA: American College of Radiology; 1995.
2. American College of Radiology. Breast imaging Reporting and Data System (BI-RADS), 3rd ed. Reston, VA: American College of Radiology; 1998.
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. American College of Radiology. BI-RADS: ultrasound, 1st ed. In: Breast Imaging Reporting and Data System: BI-RADS atlas, 4th ed. Reston, VA: American College of Radiology; 2003.
5. Mendelson EB, Berg WA, Merritt CR. Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound. Semin Roentgenol. 2001;36:217-225.
6. Hall FM. Sonography of the breast: controversies and opinions. AJR Am J Roentgenol. 1997;169:1635-1636.
7. Jackson VP. Management of solid breast nodules: what is the role of sonography? Radiology. 1995;196:14-15.
8. Harper AP, Kelly-Fry E, Noe JS, Bies JR, Jackson VP. Ultrasound in the evaluation of solid breast masses. Radiology. 1983;146:731-736.
9. Rahbar G, Sie AC, Hansen GC, Prince JS, Melany ML, Reynolds HE, Jackson VP, Sayre JW, Bassett LW. Benign versus malignant solid breast masses: US differentiation. Radiology. 1999;213:889-894.
10. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology. 1995;196:123-134.
11. Fleiss JL. Statistical Methods for Rates and Proportions, Second Edition. New York: John Wiley & Sons Inc; 1981.

12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.
13. Baker JA, Kornguth PJ, Soo MS, Walsh R, Mengoni P. Sonography of solid breast lesions: observer variability of lesion description and assessment. AJR Am J Roentgenol. 1999;172:1621-1625.
14. 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
15. Orel SG, Kay N, Reynolds C, Sullivan DC. BI-RADS categorization as a predictor of malignancy. Radiology. 1999;211:845-850.

국문요약

유방 초음파 병변 분석에 BI-RADS (Breast Imaging Reporting and Data System)의 이용 : 관찰자간 및 관찰자내 일치도에 대한 평가

유방 촬영술이 유방 병변의 진단에 중요한 위치를 차지하고 있지만, 최근 유방초음파의 역할도 중요시 되고 있다. 유방촬영술에서 유방 병변의 평가 및 분석에 유용하게 사용되었던 BI-RADS가 2003년부터 유방초음파에서도 이용되기 시작하였다. 유방초음파의 유방 병변 분석에 대해 논란이 많지만, BI-RADS를 이용한 분석이 유방의 양성 및 악성 병변을 분류하는데 유용한 것으로 보고 되고 있다. 이 연구의 목적은, BI-RADS를 이용하여 유방초음파에서 보이는 병변을 분석함에 있어, 관찰자내 및 관찰자간 일치도가 어떠한지를 알아보고 유방초음파에서의 BI-RADS 이용에 정당성이 있는가를 알아보기 위한 것이다. 2003년 1월부터 2003년 5월까지 총 136명의 환자가 150개의 유방 병변에 대해 유방초음파와 함께 조직검사를 받았다. 낭성 병변을 제외하였고, 고형성 병변 중에서도 악성이 나왔거나, 양성이지만 2년의 추적검사에서 변화가 없는 경우만을 포함시켜, 연구하였다. 한 명의 방사선과 의사가 환자들의 초음파 사진을 정리한 다음, 다른 4명의 방사선과 전문의가 분석을 하였다. 첫번째 분석 후, 8주 뒤에 같은 사진을 가지고 두번째

분석을 하였다. 분석은 BI-RADS에 대한 표를 만들고, 유방초음파에서 보이는 병변에 대해, 가장 합당한 항목을 표시하는 식으로 이루어졌다. 분석이 끝난 후, BI-RADS descriptors와 final assessment category에 대해 관찰자간 및 관찰자내 일치도를 조사하였다. 본 연구에서, 유방초음파에 대한 BI-RADS descriptors와 final assessment category의 관찰자간 및 관찰자내 일치도는 중등도 이상으로, 비교적 높은 것으로 나타났다. 민감도는 98%, 특이도는 33%, 음성예측도는 95%, 양성예측도는 category 4a는 26%, category 4b는 89%, category 4c는 90%로 나타났다. Category 5는 97%의 양성예측도를 보였다. 결론적으로 유방초음파에 대한 BI-RADS lexicon에 대한 관찰자간 및 관찰자내 일치도는 비교적 높으며, 이는 유방초음파에 있어서 BI-RADS lexicon의 이용을 정당화할 수 있다.

핵심되는 말 : 유방초음파, BI-RADS lexicon, 관찰자간 일치도,
관찰자내 일치도