# Observer Agreement Using the ACR Breast Imaging Reporting and Data System (BI-RADS)-Ultrasound, First Edition (2003)

Chang Suk Park, MD[1]
Jae Hee Lee, MD[2]
Hyeon Woo Yim, MD[3]
Bong Joo Kang, MD[4]
Hyeon Sook Kim, MD[5]
Jung Im Jung, MD[6]
Na Young Jung, MD[7]
Sung Hun Kim, MD[2]

**Index terms:**
Breast, US
Breast neoplasms, US
Breast, abnormalities

***Korean J Radiol 2007;8:397-402***
Received August 16, 2006; accepted after revision March 30, 2007.

[1]Department of Radiology, Our Lady of Mercy Hospital, College of Medicine, The Catholic University of Korea; [2]Department of Radiology, Kangnam St. Mary's Hospital, College of Medicine, The Catholic University of Korea; [3]Department of Preventive Medicine, College of Medicine, The Catholic University of Korea; [4]Department of Radiology, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea; [5]Department of Radiology, St. Paul's Hospital, College of Medicine, The Catholic University of Korea; [6]Department of Radiology, St. Mary's Hospital, College of Medicine, The Catholic University of Korea; [7]Department of Radiology, Holy Family Hospital, College of Medicine, The Catholic University of Korea

**Address reprint requests to:**
Jae Hee Lee, MD, Department of Radiology, Kangnam St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Banpo-dong, 505, Seocho-gu, Seoul 137-701, Korea.
Tel. (822) 590-2944
Fax. (822) 599-6771
e-mail: heerad@catholic.ac.kr

**Objective:** This study aims to evaluate the degree of inter- and intraobserver agreement when characterizing breast abnormalities using the Breast Imaging Reporting and Data System (BI-RADS)-ultrasound (US) lexicon, as defined by the American College of Radiology (ACR).

**Materials and Methods:** Two hundred ninety three female patients with 314 lesions underwent US-guided biopsies at one facility during a two-year period. Static sonographic images of each breast lesion were acquired and reviewed by four radiologists with expertise in breast imaging. Each radiologist independently evaluated all cases and described the mass according to BI-RADS-US. To assess intraobserver variability, one of the four radiologists reassessed all of the cases one month after the initial evaluation. Inter- and intraobserver variabilities were determined using Cohen's kappa (*k*) statistics.

**Results:** The greatest degree of reliability for a descriptor was found for mass orientation ($k = 0.61$) and the least concordance of fair was found for the mass margin ($k = 0.32$) and echo pattern ($k = 0.36$). Others descriptive terms: shape, lesion boundary and posterior features ($k = 0.42$, $k = 0.55$ and $k = 0.53$, respectively) and the final assessment ($k = 0.51$) demonstrated only moderate levels of agreement. A substantial degree of intraobserver agreement was found when classifying all morphologic features: shape, orientation, margin, lesion boundary, echo pattern and posterior feature ($k = 0.73$, $k = 0.68$, $k = 0.64$, 0.68, $k = 0.65$ and $k = 0.64$, respectively) and rendering final assessments ($k = 0.65$).

**Conclusion:** Although BI-RADS-US was created to achieve a consensus among radiologists when describing breast abnormalities, our study shows substantial intraobserver agreement but only moderate interobserver agreement in the mass description and final assessment of breast abnormalities according to its use. A better agreement will ultimately require specialized education, as well as self-auditing practice tests.

Although mammography remains the 'gold' standard for breast screening, the use of ultrasound (US) can improve the specificity of mammography, reduce the number of false negatives for breast cancer diagnosis in dense breasts, and reduce the number of false positive recommendations for a biopsy (1–3). However, there are several inherent disadvantages of US examinations. The most important drawbacks are that US examinations are highly operator-dependent and their lack of reproducibility. Another continuing problem is a lack of a standardized method for lesion characterization and recommendations, which creates confusion among physicians, radiologists and patients. To address this shortcoming, a US lexicon was created and published in the fourth edition of the Breast Imaging Reporting and Data System (BI-RADS), based on the success of BI-RADS with mammography (1).

Although many published reports have evaluated and demonstrated the benefits of standardizing the interpretation of mammography using the BI-RADS mammography lexicon (4–7), there are few reports that have evaluated the efficacy of the BI-RADS-US lexicon. Skaane et al. (8) demonstrated that US compared to mammography has a lower agreement rate on diagnosis when the two modalities are interpreted independently, and Baker et al. (9) have shown a substantial lack of consistency when applying terms as defined by Stavros and colleagues and a moderate level of agreement for final assessments (10, 11). Recently, Lazarus and colleagues have reported that the addition of the BI-RADS lexicon for US is helpful and can be used with good agreement among radiologists (12). As there is an overall inconsistency in the reported results, more studies are needed to reach a more definitive conclusion about the efficacy of BI-RADS-US. Therefore, we evaluated both inter- and intraobserver variabilities when using the newly developed ACR BI-RADS-US lexicon for characterizing breast abnormalities.

## MATERIALS AND METHODS

### *Case Selection*

Cases evaluated in this study were selected from patients who underwent biopsies during a 12-month period. Two hundred ninety three female patients with 314 lesions underwent US-guided biopsies at a single facility. The mean patient age was 39.9 years with a range of 17−81 years. The vast majority of the masses had diameters in the range of 0.4−2.6 cm. The diameters measured were 0.4−1 cm in 112 (36%) cases, 1.1−1.5 cm in 101 (32%) cases, and 1.6−2.6 cm in 101 (32%) cases.

The pre-biopsy final assessments were performed by a single radiologist with five years of experience in breast imaging. Seven cases were category 2, 104 cases were category 3, 158 cases were category 4, and 45 cases were category 5.

Of the 314 lesions, 88 (28%) were confirmed to be malignant by histology. Of the malignant lesions, the most common diagnosis was invasive ductal carcinoma, which was found in 77 of the 88 cancer cases. Other diagnoses included tubular carcinoma (3/88), mucinous carcinoma (3/88) and ductal carcinoma in situ (5/88). The remaining 226 (72%) of the 314 total lesions were benign, including eight lesions associated with atypical ductal hyperplasia. Follow-up ultrasonography was performed for 117 of the 226 benign cases and the mean duration time was 15 months (range: 3 to 24 months). Four cases in category 2 had a mean duration time of 11 months (range: 6–17 months), 58 cases in category 3 had a mean duration time of 15 months (range: 4–24 months), and 55 cases in category 4 had a mean duration time of 14 months (range: 3–24 months). The lesions were stable in 65 cases (56%) and decreased in size in 52 (54%) cases.

### *Evaluation of Ultrasonographic Images*

Two radiologists at one facility performed a US-guided percutaneous biopsy or localization, and pre-biopsy static sonographic images of each breast lesion were acquired. All images were obtained with high-resolution US equipment (HDI 5000 or HDI 3000, Advanced Technology Laboratories, Bothell, WA) using a 12 MHz linear transducer. Four radiologists with experience in breast imaging (3–11 years) and who work at outside hospitals reviewed the images. Evaluation and designations were made according to the BI-RADS-US. For each case, at least two static images including radial and antiradial images or transverse and longitudinal images with and without caliper measurements were provided. Other US images including Doppler, color Doppler, and power Doppler images were not provided. Mammographic imaging and medical histories were also not provided to eliminate the possibility of introducing bias into the description and assessment of the US images.

The BI-RADS-US included the assessment of masses, calcification, special cases and the final assessment. However, we focused on US features of masses such as shape, orientation, margin, lesion boundary, echo pattern and posterior acoustic features in this study. Each of the four radiologists independently evaluated all of the cases and selected the most suitable single term from each group of the lexicon, and then decided the final category of the lesion.

To assess intraobserver variability, one of the four radiologists with experience in breast imaging for four years re-evaluated all cases one month after the initial evaluation.

### *Statistical Analysis*

The sensitivity and specificity was calculated for each radiologist by means of a binary outcome; categories 2−3 were grouped as negative, and categories 4−5 were grouped as positive.

Inter- and intraobserver variabilities in choosing sonographic descriptors and final assessment according to the BI-RADS-US lexicon were determined using Cohen's kappa statistics (13). Cohen's kappa statistics measures the proportion of decisions where observers agree while accounting for the possibility of agreements based on chance alone. Perfect agreement results in a kappa value of 1.0, and a kappa value of 0 indicates the level of

agreement expected based on chance alone. Agreement less than that expected by chance alone results in a negative kappa value. Although no definitive scale exists, prior reports have suggested a scale for kappa values and their level of agreement between observers: $\leq 0.2$ 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 indicates almost perfect agreement (14).

## RESULTS

The sensitivity and specificity of the US interpretations by the four radiologists are summarized in Table 1. The sensitivities of observers were high (96–100%), but the specificities were low and variable (8–43%).

### *Interobserver Variability*

Statistical analysis of agreement among observers when choosing lesion descriptions showed a range from fair to substantial concordance. The greatest reproducibility was found among observers determining the mass orientation ($k$ = 0.61). However, only moderate levels of interobserver agreement were found for three of the six descriptive groups: mass shape, boundary and posterior feature ($k$ = 0.42, $k$ = 0.55 and $k$ = 0.53, respectively). The lowest levels of concordance occurred when observers determined the mass margin ($k$ = 0.32) and echo pattern ($k$ = 0.36) (Table 2). Figure 1 shows an image for which observers used variable terms to describe the margin but had good agreement for the final assessment.

The reproducibility of the final assessment when assigning lesions as category 2, 3, 4, or 5 was moderate ($k$ = 0.49) (Table 2). When assigning lesions as category 2, the greatest reproducibility was found ($k$ = 0.66) and moderate agreement degree was found for category 5 ($k$ = 0.54). Only fair reproducibility was found for determining category 3 ($k$ = 0.26) and for category 4 ($k$ = 0.30). When choosing between a follow-up evaluation (category 2 and 3) or recommending a biopsy (category 4 and 5), the level of agreement was lower ($k$ = 0.33) than when assigning lesions as category 2, 3, 4, or 5. Figure 2 shows a lesion for which the observers disagreed on the final assessment and recommendation.

### *Intraobserver Variability*

Substantial intraobserver agreement was found in selecting all of the morphologic features (Table 3). Substantial agreement was achieved for the final assessment category ($k$ = 0.74) with all final categories (2, 3, 4 and 5). Perfect agreement was found with lesions categorized as category 2 ($k$ = 1.0). Substantial agreement was obtained for categories 3 and 5 ($k$ = 0.61 and $k$ = 0.68, respectively). There was moderate agreement among observers for category 4 ($k$ = 0.59). When choosing between follow-up care or recommending a biopsy, for all final categories the

**Table 1. Sensitivity and Specificity of US Interpretations by the Four Observers**

| Observer | Sensitivity % | Specificity % | PPV % | NPV % |
|---|---|---|---|---|
| 1 | 100 (88/88) | 8 (18/226) | 30 (88/296) | 100 (18/18) |
| 2 | 96 (84/88) | 34 (76/226) | 36 (84/234) | 95 (76/80) |
| 3 | 100 (88/88) | 15 (34/226) | 31 (88/280) | 100 (34/34) |
| 4 | 98 (86/88) | 43 (98/226) | 40 (86/214) | 98 (98/100) |

Note.— PPV = positive predictive value, NPV = negative predictive value

**Table 2. Interobserver Variability in Description According to BI-RADS-US Lexicon**

| Descriptors & Final assessments | This study *k*-value | Lazarus' study (12) *k*-value | Baker's study (9) *k*-value |
|---|---|---|---|
| Shape | 0.42 | 0.66 | 0.8 |
| Orientation | 0.61 | 0.61 | |
| Margin | 0.32 | 0.4 | 0.43 |
| Boundary | 0.55 | 0.69 | |
| Echo pattern | 0.36 | 0.29 | 0.4 |
| Posterior feature | 0.53 | 0.4 | 0.55 |
| Final category | 0.49 | 0.28 | 0.51 |

Note.— Level of agreement between observers: *k*-value $\leq$ 0.2 indicates slight (SL) agreement, 0.21–0.40 fair (F), 0.41–0.60 moderate (M), 0.61–0.80 substantial (S), and 0.81–1.00 indicates almost perfect (P)


**Fig. 1.** US in a 57-year-old woman with an invasive ductal carcinoma. The observers described its margins using variable terms: indistinct (2 observers), angular (1), and spiculated (1). All of the observers agreed that the lesion belonged to category 4.

level of agreement was lower ($k$ = 0.62) than the assessment category ($k$ = 0.74). The degrees of agreements for all descriptors except echo pattern were similar to the study of Baker et al. (9); in the current study, the degree of agreement for the final assessment was higher.

## DISCUSSION

Many studies have reported significant inter- and intraobserver variabilities in lesion description and assessment on mammography (4−8). For US, Baker et al. (9) reported a lack of uniformity among observers use of descriptive terms for breast masses using the lexicon described and further defined by Stavros et al. (10). A standardized lexicon similar to that of the BI-RADS was proposed (9). Recently, a study by Lazarus et al. examined observer variability using the new BI-RADS lexicon (12). In spite of using a standardized lexicon, these investigators showed a similar level of consistency using terminology and lower level of consistency for final assessment among observers compared to the Baker et al. study (9).

**Table 3. Intraobserver Variability in Description According to BI-RADS-US lexicon**

| Descriptors & Final Assessments | This study *k*-value | Baker's study (9) *k*-value |
|---|---|---|
| Shape | 0.73 | 0.79 |
| Orientation | 0.68 | |
| Margin | 0.64 | 0.62 |
| Boundary | 0.68 | |
| Echo pattern | 0.65 | 0.24 |
| Posterior feature | 0. 64 | 0.63 |
| Final category | 0.74 | 0.66 |

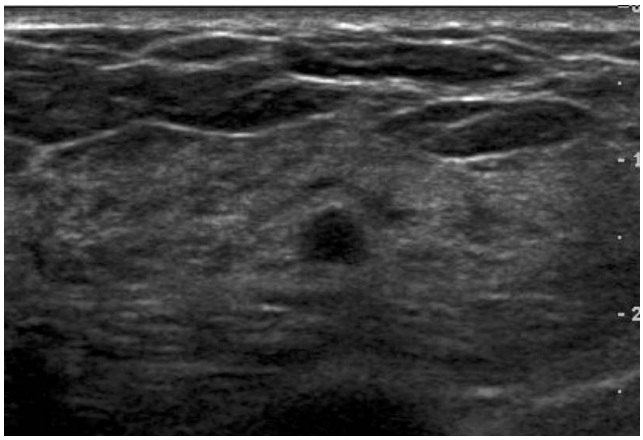

**Fig. 2.** US in a 40-year-old-woman with fibrocystic disease. Observers arrived at different final assessments and recommendations: two assigned this lesion to category 3 and recommended close follow-up, while the others considered it as category 4 and recommended a biopsy.

Compared to the study by Lazarus et al. (12), degrees of agreement for shape, margin and boundary were lower and higher for posterior features, echo pattern and final categories, and were similar for orientation (Table 2). In the Baker et al. study (9), a greater reproducibility was obtained in determining the shape, margin, echo pattern, posterior feature and final category of a mass ($k$ = 0.8, $k$ = 0.43, $k$ = 0.4, $k$ = 0.55 and $k$ = 0.51, respectively) than in the Lazarus et al. (12) study and our study (Table 2).

In our study, the greatest consistency was found in determining the orientation of a mass ($k$ = 0.61). In the Lazarus et al. study (12), good consistency was also seen ($k$ = 0.61). The determination of parallel or not parallel orientation is generally easily measured, explaining the high degree of observer agreement. Moderate agreement was obtained for shape, boundary, and posterior feature. For shape, some difficulties arose when trying to classify abnormalities containing five or six more gentle lobulations as oval or irregular, which may have contributed to the inter-observer variability (Fig. 3). In addition, questions arose among the observers when the lesion was elliptical-shaped with not-parallel orientation as to whether it could be deemed as having an "oval shape." This situation is not trivial as the designation of an oval shape can influence a radiologist to conclude that the lesion is benign. When determining the posterior acoustic feature, we also had some difficulties because four observers evaluated only static images. Furthermore, the use of good US equipment can compensate for posterior acoustic features and posterior shadowing becomes less conspicuous.

Only fair degrees of agreement were obtained for the echo pattern and margin. This finding was similar to that

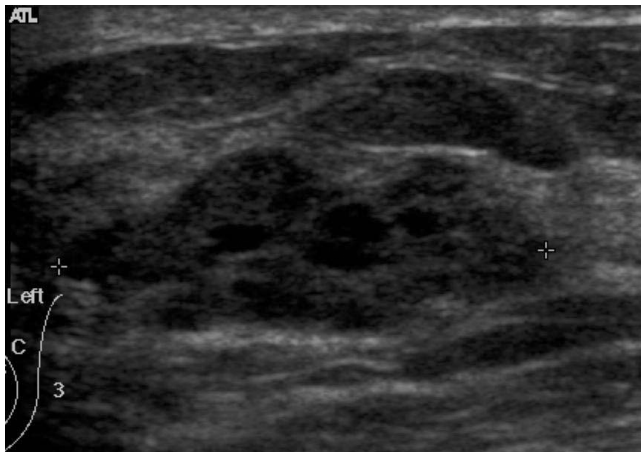


**Fig. 3.** US in a 47-year-old woman with fibrocystic disease. The mass shows six or seven macrolobules with a circumscribed margin and complex echogenecity. Three observers agreed on the irregular shape and circumscribed margin. One observer denoted this mass as having an oval shape and microlobulated margin.

found in the studies by Lazarus et al. (12) and Baker et al. (9). We could not find out why the agreement for echo pattern was so low, but it seems to have little effect on determining a final category. The greatest variation amongst observers was found when labeling a mass margin ($k$ = 0.32). As mass margin is a critical feature for determining whether a lesion should be biopsied or not, this determination alone can have a substantial effect on the final assessment. The margins of a lesion may be heterogeneous, which may make it difficult to accurately label it using only one term. In this study when more than one type of margin existed in a lesion, we resolved this issue by choosing the term having the greatest positive predictive value, according to Stavros and colleagues criteria (10, 11).

The variability in the observers description of breast masses resulted in an inconsistent final assessment using BI-RADS-US. Only a moderate level of agreement ($k$ = 0.49) was found for the final assessment. When assigning lesions category 2, negative for malignancy and category 5, highly suggestive of malignancy, greater agreements were achieved ($k$ = 0.66 and $k$ = 0.54, respectively). This means that observers have similar conceptions for benign (category 2) and malignant lesions (category 5) and variable for probably benign (category 3, $k$ = 0.26) and suspicious abnormalities (category 4, $k$ = 0.3). Thus, the numbers of category 3 or 4 lesions will influence observer variability for each study. The level of agreement in determining a lesion as belonging to category 2 to 5 ($k$ = 0.40) was higher than when trying to determine whether close follow-up or a biopsy would be recommended ($k$ = 0.33). This inconsistency is expected to have a negative effect on patient management.

As expected, the agreement of evaluation within a single observer is better than that among multiple observers. We found substantial intraobserver agreement in characterizing each descriptive term and the final categorization. This result is also similar to that in the Baker et al. study (9). We believe that radiologists have their own internally established standardization of lesion descriptions and criteria of the final categorization, but that the difference in choosing from a fixed set of lesion descriptors between observers may be the result of each individual having different cut-off points for determining whether one description or another is applicable.

Our study has several limitations. First, the number of category 2 or 3 lesions was small and the radiologists may have had a greater tendency to interpret a lesion as being worrisome as the cases of this study were selected from those performed for biopsy. Second, observers did not perform and evaluate real time US, but interpreted only static images. Thus the observers in this study did not have the opportunity to take advantage of certain real time US benefits such as manual compensation of the posterior feature. Third, the radiologists studied BI-RADS-US by themselves and made their own criteria. If they had had a more standardized education for BI-RADS-US, we assume that interobserver reproducibility would be higher than reported.

Upon issuing BI-RADS-US, the ACR states that agreement on terminology and assessment categorization was reached by consensus of an expert working group and agreement among both experienced and novice breast imagers for most terms (1). Yet our study shows only moderate agreement for most descriptive terms and the final assessment, although the accuracies of observers were similar with one another. Of special note, the level of agreement in determining the margin and echo patterns was commonly low in our study, as well as in studies by others.

In conclusion, although the use of ACR BI-RADS-US as a unified descriptor system was made, observer agreements were only fair to substantial in the mass description and final assessment of breast abnormalities. The achievement of better agreement will ultimately require specialized education, as well as periodic performance assessments and self-auditing practice tests.

## References


1. American College of Radiology. Breast imaging reporting and data system, Breast imaging atlas, forth ed. Reston, VA, American College of Radiology 2003
2. 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
3. Mendelson EB, Berg WA, Merritt CR. Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound. *Semin Roentgenol* 2001;36:217-225
4. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. *Eur J Cancer* 1992;28A:1054-1058
5. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretation of mammograms. *N Engl J Med* 1994;331:1493-1499
6. Vineis P, Sinistrero G, Temporelli A, Azzoni L, Bigo A, Burke P, et al. Inter-observer variability in the interpretation of mammograms. *Tumori* 1988;74:275-279
7. Baker JA, Kornguth PJ, Floyd CE Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *AJR Am J Roentgenol* 1996;166:773-778
8. Skaane P, Engedal K, Skjennald A. Interobserver variation in the interpretation of breast imaging. *Acta Radiol* 1997;38:497-502
9. 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
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. Stavros AT. *Breast Ultrasound*, 1st ed. Philadelphia: Lippincott Williams & Wilkins, 2004:455-527
12. 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
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174
14. Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. *APMIS* 1989;97:689-698