ULTRASONOGRAPHY

ORIGINAL ARTICLE

https://doi.org/10.14366/usg.18012
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2019;38:172-180

Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study

Eun Jung Choi[1], Eun Hye Lee[2], You Me Kim[3], Yun-Woo Chang[4], Jin Hwa Lee[5], Young Mi Park[6], Keum Won Kim[7], Young Joong Kim[7], Jae Kwan Jun[8], Seri Hong[8] on the behalf of the Alliance for Breast Cancer Screening in Korea (ABCS-K)

*Author affiliations appear at the end of this article.

Purpose: The purpose of this study was to record and evaluate interobserver agreement as quality control for the modified categorization of screening breast ultrasound developed by the Alliance for Breast Cancer Screening in Korea (ABCS-K) for the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial.

Methods: Eight breast radiologists with 4–16 years of experience participated in 2 rounds of quality control testing for the MUST-BE trial. Two investigators randomly selected 125 and 100 cases of breast lesions with different ratios of malignant and benign lesions. Two versions of the modified categorization were tested. The initially modified classification was developed after the first quality control workshop, and the re-modified classification was developed after the second workshop. The re-modified categorization established by ABCS-K added size criteria and the anterior-posterior ratio compared with the initially modified classification. After a brief lecture on the modified categorization system prior to each quality control test, the eight radiologists independently categorized the lesions using the modified categorization. Interobserver agreement was measured using kappa statistics.

Results: The overall kappa values for the modified categorizations indicated moderate to substantial degrees of agreement (initially modified categorization and re-modified categorization: κ=0.52 and κ=0.63, respectively). The kappa values for the subcategories of category 4 were 0.37 (95% confidence interval [CI], 0.24 to 0.52) and 0.39 (95% CI, 0.31 to 0.49), respectively. The overall kappa values for both the initially modified categorization and the re-modified categorization indicated a substantial degree of agreement when dichotomizing the interpretation as benign or suspicious.

Conclusion: The preliminary results demonstrated acceptable interobserver agreement for the modified categorization.

Keywords: Screening; Neoplasms; Breast neoplasms; Ultrasonography; Interobserver variability

Received: March 15, 2018
Revised: September 20, 2018
Accepted: September 22, 2018

Correspondence to:
Eun Hye Lee, MD, PhD, Department of Radiology, Soonchunhyang University Bucheon Hospital, Soonchunhyang University College of Medicine, 170 Jomaru-ro, Wonmi-gu, Bucheon 14584, Korea

Tel. +82-32-621-5851
Fax. +82-32-621-5018
E-mail: grace@schmc.ac.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2019 Korean Society of Ultrasound in Medicine (KSUM)

How to cite this article:
Choi EJ, Lee EH, Kim YM, Chang YW, Lee JH, Park YM, et al. Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study. Ultrasonography. 2019 Apr;38(2):172-180.

Introduction

Ultrasonography (US) has been regarded as an effective complementary imaging adjunct to mammography in breast cancer screening [1,2]. Although breast US is widely used in standard practice, it has well-known drawbacks, such as operator dependency and a lack of standardization and reproducibility [3,4]. To minimize variability in the characterization and final assessment of breast masses identified on breast US, the American College of Radiology (ACR) developed the Breast Imaging Reporting and Data System (BI-RADS) US lexicon in 2003 [5]. After a decade of clinical practice, the ACR presented an updated second version of the BI-RADS US lexicon with the addition of new sections and changes in terminology in 2013 [6].

Although the BI-RADS lexicon standardizes the characterization of breast lesions, the final BI-RADS assessment for breast lesions is determined by the subjective decision of radiologists, taking multiple factors into account. As a result, potential interobserver variability compared with other imaging modalities is unavoidable. In particular, with regard to the screening criteria, there are some limitations of the final BI-RADS assessment for screening breast US reports that result in increases in the short-term follow-up of breast lesions and the false-positive biopsy rate. To the best of our knowledge, no studies have defined the most appropriate criteria for breast US categorization in breast cancer screening.

The Alliance for Breast Cancer Screening in Korea (ABCS-K) initiated the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial in 2016 to compare the cost-effectiveness of combined mammography and US screenings versus conventional digital mammography screening alone for women 40–59 years of age. In order for this trial to be successful and to achieve reliable results, quality control of the modified categorization system of screening breast US is needed to evaluate interobserver agreement among participating radiologists. Therefore, the purpose of this study was to record and evaluate interobserver agreement as quality control for the modified categorization developed for screening breast US by the ABCS-K for the MUST-BE trial.

Materials and Methods

General Design and Included Cases

This study was approved by the Institutional Review Board of each of the 14 participating institutions with a waiver of informed consent (Chonbuk National University Hospital, Soonchunhyang University Bucheon Hospital, Dankook University Cheonan Hospital, Soonchunhyang University Seoul Hospital, Dong-A University Busan Hospital, Inje University Busan Paik Hospital, Konyang University Hospital, Wonkwang University Hospital, Ulsan University Hospital, Kyungpook National University Chilgok Hospital, Chungnam National University Hospital, Jeju National University Hospital, Gyeongsang National University Changwon Hopital, and Gyeongsang National University Jinju Hospital). Eligible patients were women with visible lesions on breast US examinations performed at two different institutions from March 2014 to February 2016 who had undergone US-guided biopsy or vacuum-assisted excision or surgery, had been stable for at least 2 years of US follow-up after the lesions were classified as BI-RADS category 3, or showed typically benign findings, such as simple cysts, clustered microcysts, complicated cysts, and intramammary lymph nodes [6].

Ultrasound Examination

Various US machines equipped with high-frequency linear array transducers were used for image acquisition (iU22, Philips Medical Systems, Bothell, WA, USA; GE LOGIQ E9, GE Medical Systems, Milwaukee, WI, USA; SuperSonic Imagine, Aix-en-Provence, France; ACUSON S2000, Siemens Medical Solution, Mountain View, CA, USA). Two breast radiologists with 4 and 16 years of experience, respectively, were involved in image acquisition. Spatial compounding was used in all units during scanning. Two representative orthogonal B-mode and color Doppler images were selected, and the mass size (2 orthogonal dimensions of the mass on breast US) was recorded. On the color Doppler US examinations, the color box was adjusted to include the target lesion with a minimal amount of normal surrounding tissue. Minimal pressure was applied with the transducer to avoid obliterating small vessels in the lesion. The color gain was set to a level that could identify low-velocity flow in the lesion, while minimizing background noise.

Image Preparation and Review

Two consecutive observers performed quality control tests in March 2016 and August 2016, and two investigators selected 125 and 100 breast lesions for each test, respectively. There were 44 malignant lesions (35.2%) and 81 benign lesions (65.8%) based on the initially modified categorization and 46 malignant lesions (46.0%) and 54 benign lesions (54.0%) based on the re-modified categorization. All images were saved as TIFF files with a resolution of 300 dpi and arranged in random order in Microsoft PowerPoint (ver. 14, Microsoft, Redmond, WA, USA), showing two representative orthogonal B-mode and color Doppler images per lesion for review. Fourteen breast radiologists participated in a quality control workshop for the MUST-BE trial in January 2016. After a brief lecture on the modified categorization during the quality control workshop, the radiologists individually reviewed the images on a PACS monitor (*m*-view, Marotech, Seoul, Korea). No clinical information related to the mammographic images or pathologic results was given. The radiologists assessed each lesion using the modified categorization provided by ABCS-K during every quality control workshop. To compare the observers' categorizations pre- and post-quality control, the radiologists were asked to assess the same breast lesions according to the fifth edition of BI-RADS and the initially modified categorization. Among the 14 radiologists, eight with varying experience in breast US (ranging from 4–16 years with a mean experience of 10.1 years) completed all tests. The participating radiologists were divided into two groups of observers: three of the eight radiologists who completed the tests had less than average experience (<10 years) with breast US, and the remaining five radiologists had more than average experience interpreting breast imaging (≥10 years) in an academic setting (Table 1).

Modified Categorization System Established by the ABCS-K

All radiologists involved in this study were fully aware of the ACR BI-RADS lexicon for breast US, and the ACR BI-RADS lexicon and final assessments were used in the radiology reports [6]. We independently categorized the breast lesions from category 2 to category 5 based on the fifth edition of BI-RADS as 2 (benign), 3 (probably benign), 4 (suspicious; 4a, low suspicion; 4b, intermediate suspicion; 4c, moderate suspicion), or 5 (highly suggestive of malignancy).

After the first quality control workshop, we established an initially modified categorization for screening breast US based on a brief lecture and discussion among the participating observers. For the initially modified categorization, we categorized the benign breast lesions as follows: category 2 (benign: absence of any suspicious findings, including simple cysts, intramammary lymph nodes, skin lesions, silicone granuloma, fat-containing lesions, non-simple cysts in the setting of multiple or bilateral cysts, and hyperechoic masses) or category 3 (probably benign: neither category 2 nor category 4 or 5, i.e., isolated complicated cysts, solid masses [defined as space-occupying lesions without an anechoic component in two different planes] with a round or oval shape that is predominantly circumscribed with parallel orientation, clustered microcysts, fat necrosis). Furthermore, we defined suspicious US features based on BI-RADS and previous studies [6–8]. We segregated the suspicious findings into major and minor findings to distinguish between category 4 and 5 lesions. Major suspicious findings were irregular shapes with spiculated margins and microcalcifications. Minor suspicious findings included a round shape; microlobulated, indistinct, or angular margins; complex cystic and solid composition; posterior shadowing; a non-parallel orientation; and duct extension. Category 4a was defined as lesions showing 1 or more minor suspicious findings, category 4b as lesions showing more than three minor suspicious findings, category 4c as lesions showing one major suspicious finding with or without minor suspicious findings, and category 5 as lesions showing more than two major suspicious findings.

After the second quality control workshop, we re-modified the appropriate criteria for breast US categorization to establish more clearly defined criteria for category 2, 3, and 4 lesions (Table 2, Figs.

Table 1. Characteristics of the radiologists who participated in the study

Characteristic	No. of radiologists (%)
Total	8 (100)
Years of experience interpreting breast ultrasound	
<10	3 (37.5)
≥10	5 (62.5)
Fellowship training in breast imaging	
Yes	4 (50.0)
No	4 (50.0)
Mean annual breast ultrasonography volume (no. of ultrasonography)	
<3,000	3 (37.5)
≥3,000	5 (62.5)

Table 2. The MUST-BE categorization for screening breast ultrasound

Category	Finding	Size
2	Simple cyst/intramammary lymph node/calcified fibroadenoma/fat-containing lesion	–
	Solitary, oval, circumscribed complicated cysts	≤5 mm
	Non-simple cysts in the setting of multiple or bilateral cysts (i.e., at least three cysts with at least one in each breast)	–
	Round, circumscribed, solid mass	≤5 mm
	Oval, circumscribed, parallel, solid mass	≤10 mm or ratio <0.7
3	Isolated complicated cyst	≥6 mm
	Round, circumscribed, solid mass	6–10 mm
	Oval, circumscribed, parallel, solid mass	≥11 mm or ratio ≥0.7
	Clustered microcysts/fat necrosis	–
4	Suspicious abnormality: one or more suspicious findings, not category 5	–
5	Solid mass with an irregular shape and spiculated margin	–

MUST-BE, Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness.

1, 2). Furthermore, we added criteria related to the lesion size and anterior-posterior ratio and revised the criteria for major and minor suspicious findings. Although it was excluded from the fifth edition of BI-RADS, we added the finding of an echogenic halo, defined as a blurred irregular hyperechoic rim around the lesion. An echogenic halo is seen in malignancies and abscesses, so an echogenic halo on screening breast US is likely to indicate a malignancy. In addition, we added internal vascularity to the minor suspicious findings and moved microcalcifications from the major suspicious findings to the minor suspicious findings (Table 3). The radiologists performed the final assessment by assessing the minor and major suspicious findings using the following definitions. Category 4a was defined as lesions showing one or more minor suspicious findings, category 4b as lesions showing more than three minor suspicious findings or 1 major suspicious finding with one or two minor suspicious findings, category 4c as lesions showing one major suspicious finding with or without minor suspicious findings, and category 5 as breast masses of irregular shape with a spiculated margin.

In addition, if there were unexplained findings based on the initially modified categorization and the re-modified categorization,

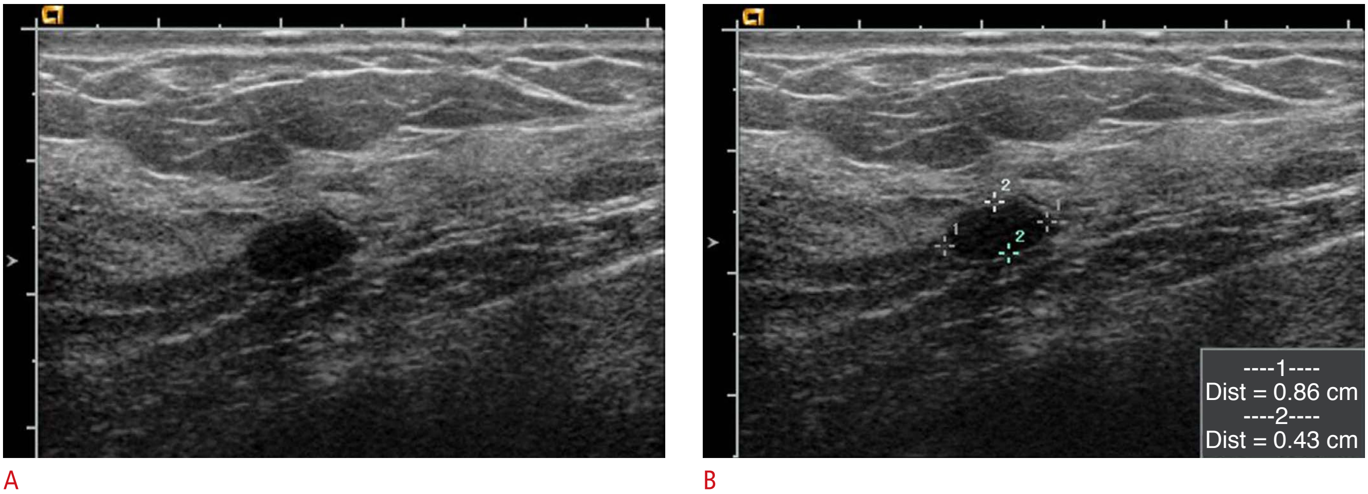

Fig. 1. Clinically proven benign lesion in a 43-year-old woman.
A, B. Ultrasonography showed an 8.6×4.3-mm, oval-shaped, circumscribed, hypoechoic mass that was classified as category 3 based on BI-RADS. However, this lesion was classified as category 2 using the modified categorization by the ABCS-K developed for the MUST-BE trial because it was smaller than 10 mm and had an anterior-posterior ratio less than 0.7. This lesion was not pathologically confirmed, but was clinically proven to be benign, as it remained stable for 2 years. BI-RADS, Breast Imaging Reporting and Data System; ABCS-K, Alliance for Breast Cancer Screening in Korea; MUST-BE, Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness.

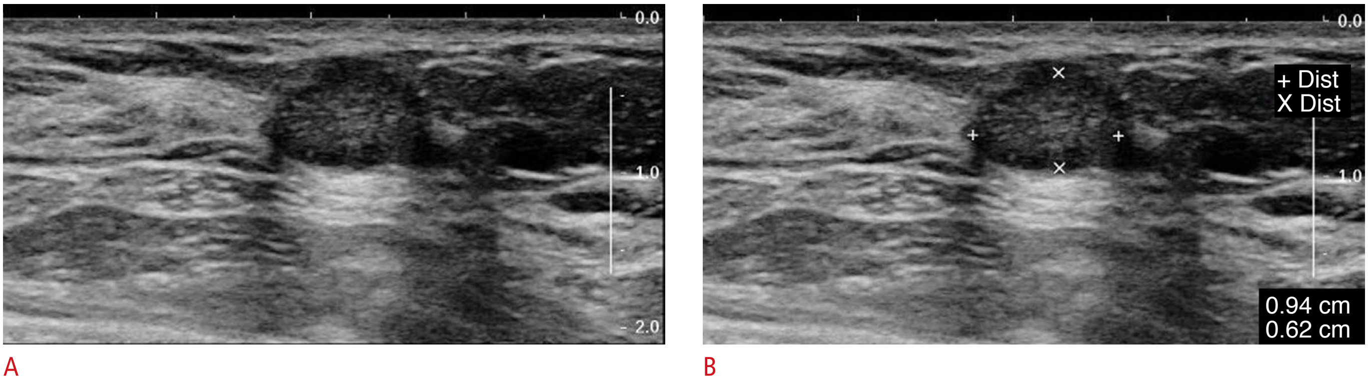

Fig. 2. Mucinous carcinoma in a 54-year-old woman.
A, B. Breast ultrasonography showed a 9.4×6.2-mm, oval-shaped, hypoechoic mass with a microlobulated margin and parallel orientation, which was classified as category 4a using the modified categorization by ABCS-K for the MUST-BE trial because the lesion had one minor suspicious finding. This lesion was pathologically confirmed as mucinous carcinoma. ABCS-K, Alliance for Breast Cancer Screening in Korea; MUST-BE, Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness.

observers were permitted to discuss them with the principal investigator of this study after assigning the most appropriate category based on his or her subjective assessment.

Statistical Analysis

The reference standard was a combination of pathologic results and 2 years of clinical follow-up data. Agreement on the US categorization of breast lesions according to the modified categorization system, as well as the dichotomized categorization (positive [categories 4a, 4b, 4c, and 5] and negative assessments [categories 2 and 3]) among radiologists was evaluated. The interobserver agreement on both the initially modified and the re-modified categorization was assessed using kappa and weighted kappa statistics with Stata software (StataCorp., College Station, TX, USA). In addition, agreement on both the initially modified and the re-modified categorization according to the observer's amount of experience in interpreting breast US was analyzed using kappa and weighted kappa statistics.

Table 3. Classification of suspicious ultrasonography findings in the re-modified categorization

Classification	Finding
Minor	Round shape ≥11 mm[a]
	Non-circumscribed margin: microlobulated, indistinct, or angular
	Echogenic halo[a]
	Complex cystic and solid
	Posterior shadowing
	Nonparallel orientation
	Microcalcification: within or outside mass/intraductal[a]
	Duct extension
	Internal vascularity[a]
Major	Irregular shape[a]
	Spiculated margin[a]

Category 4: Suspicious abnormality-one or more suspicious findings, not category 5. Category 4a was defined as lesions showing one or more minor suspicious findings, category 4b as lesions showing more than three minor suspicious findings or one major suspicious finding with one or two minor suspicious findings, and category 4c as lesions showing one major suspicious finding with or without minor suspicious findings.

[a]Revised criteria for major and minor suspicious findings from the initially modified categorization to the re-modified categorization. Size was added to the minor suspicious findings for breast masses, along with round shape, an echogenic halo, and internal vascularity. Microcalcifications were moved from the major suspicious findings to the minor suspicious findings. The major suspicious findings consisted of an irregular shape and spiculated margins.

The overall kappa value for multiple observers and multiple categories was determined based on the work of Landis and Koch [9]. We used the following definitions to interpret the kappa coefficients (κ): less than 0.20 indicated poor agreement, 0.21–0.40 indicated fair agreement, 0.41–0.60 indicated moderate agreement, 0.61–0.80 indicated substantial agreement, and 0.81–1.00 indicated nearly perfect agreement.

Results

We found a moderate to substantial degree of interobserver agreement for both the initially modified and re-modified categorizations. Tables 4 and 5 summarize the interobserver agreement for the modified categorization and the subcategory classification.

The overall kappa value for the BI-RADS categorization (categories 2, 3, 4, and 5) was 0.50 (95% confidence interval [CI], 0.32 to 0.65). The overall kappa value was 0.51 (95% CI, 0.45 to 0.59) when dichotomizing the interpretation as benign (categories 2 and 3) or suspicious (categories 4 and 5) (Table 4).

Table 4. Interobserver agreement in three quality control tests performed in the first year of the MUST-BE trial

Final assessment	κ-value[a]		
	BI-RADS categorization	Initially modified categorization	Re-modified categorization
Category 2	0.416 (0.325–0.522)	0.485 (0.359–0.612)	0.781 (0.694–0.859)
Category 3	0.393 (0.257–0.555)	0.435 (0.338–0.537)	0.511 (0.381–0.632)
Category 4	0.441 (0.250–0.604)	0.556 (0.484–0.643)	0.592 (0.513–0.668)
Category 5	0.455 (0.261–0.609)	0.674 (0.477–0.820)	0.511 (0.327–0.649)
Overall	0.495 (0.318–0.651)	0.521 (0.451–0.600)	0.626 (0.560–0.688)
Interpretation after dichotomization as benign[b] and suspicious[c] lesions	0.512 (0.450–0.589)	0.676 (0.599–0.751)	0.725 (0.648–0.797)

The numbers in parentheses are 95% confidence intervals.

MUST-BE, Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness; BI-RADS, Breast Imaging Reporting and Data System.

[a]Interobserver agreement was evaluated according to the BI-RADS categorization, the initially modified categorization, and the re-modified categorization including the size and anterior-posterior ratio criteria. [b]Benign lesions were category 2 or 3. [c]Suspicious lesions were category 4 or 5.

In the initially modified categorization, the overall kappa value (categories 2, 3, 4, and 5) was 0.52 (95% CI, 0.45 to 0.60). The overall kappa value was 0.68 (95% CI, 0.60 to 0.75) when dichotomizing the interpretation as benign (categories 2 and 3) or suspicious (categories 4 and 5) (Table 4). However, the kappa value for the subcategory classification of category 4 (4a, 4b, and 4c) was 0.37 (95% CI, 0.24 to 0.52) (Table 5).

The overall interobserver agreement was higher for the re-modified categorization (categories 2, 3, 4, and 5; re-modified categorization and initially modified categorization: κ=0.67; 95% CI, 0.56 to 0.69 and κ=0.52; 95% CI, 0.45 to 0.60, respectively) as well as when dichotomizing the interpretation as benign (categories 2 and 3) or suspicious (categories 4 and 5) than in the initially modified categorization (re-modified categorization and initially modified categorization: κ=0.73; 95% CI, 0.65 to 0.80 and κ=0.68; 95% CI, 0.60 to 0.75, respectively) (Table 4).

Among the subcategory classifications of category 4, category 4b showed the lowest interobserver agreement in both the initially modified categorization (κ=0.18; 95% CI, 0.09 to 0.32) and the re-modified categorization (κ=0.21; 95% CI, 0.14 to 0.30). Furthermore, there was no difference in interobserver agreement according to the radiologist's experience (Table 6).

Discussion

The aim of this preliminary study was to record interobserver agreement as part of quality control testing for the modified categorization of screening breast US developed by the ABCS-K for the MUST-BE trial. The interobserver agreement for the modified categorization in this study was higher than that for the fifth

Table 5. Interobserver agreement for the subcategorization of category 4 in the first year of the MUST-BE trial

Subcategorization of category 4	κ-value[a]		
	BI-RADS categorization	Initially modified categorization	Re-modified categorization
Category 4a	0.320 (0.197–0.472)	0.495 (0.325–0.710)	0.485 (0.366–0.621)
Category 4b	0.120 (0.014–0.244)	0.183 (0.094–0.316)	0.211 (0.138–0.301)
Category 4c	0.345 (0.124–0.546)	0.412 (0.205–0.612)	0.534 (0.354–0.696)
Overall subcategorization	0.254 (0.141–0.381)	0.366 (0.242–0.524)	0.394 (0.308–0.492)

The numbers in parentheses are 95% confidence intervals.

MUST-BE, Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness; BI-RADS, Breast Imaging Reporting and Data System.

[a]Interobserver agreement was evaluated according to the BI-RADS categorization, the initially modified categorization, and the re-modified categorization.

Table 6. Interobserver agreement according to the radiologist's experience in the tests performed in the first year of the MUST-BE trial

Final assessment	κ-value			
	Initially modified categorization		Re-modified categorization	
	Junior (<10 yr)	Senior (≥10 yr)	Junior (<10 yr)	Senior (≥10 yr)
Category 2	0.451 (0.288–0.616)	0.457 (0.311–0.599)	0.781 (0.675–0.878)	0.807 (0.717–0.890)
Category 3	0.401 (0.263–0.540)	0.410 (0.298–0.519)	0.557 (0.403–0.722)	0.503 (0.327–0.644)
Category 4	0.584 (0.464–0.716)	0.528 (0.439–0.627)	0.593 (0.476–0.717)	0.620 (0.524–0.708)
Category 5	0.790 (0.564–0.928)	0.612 (0.416–0.779)	0.586 (0.229–0.839)	0.542 (0.334–0.716)
Overall	0.528 (0.431–0.641)	0.490 (0.413–0.577)	0.643 (0.553–0.737)	0.648 (0.568–0.726)
Interpretation after dichotomization as benign[a] and suspicious[b] lesions	0.680 (0.571–0.794)	0.659 (0.571–0.742)	0.680 (0.579–0.800)	0.756 (0.669–0.841)
Subcategorization of category 4				
Category 4a	0.601 (0.342–0.880)	0.437 (0.244–0.681)	0.542 (0.349–0.730)	0.478 (0.335–0.625)
Category 4b	0.340 (0.088–0.648)	0.044 (0.040–0.167)	0.184 (0.030–0.374)	0.251 (0.129–0.388)
Category 4c	0.485 (0.164–0.762)	0.313 (0.118–0.531)	0.397 (0.105–0.650)	0.593 (0.392–0.770)
Overall subcategorization	0.477 (0.276–0.714)	0.270 (0.136–0.444)	0.375 (0.227–0.553)	0.423 (0.307–0.537)

Numbers in parentheses are the 95% confidence intervals.

MUST-BE, Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness.

[a]Benign lesions were category 2 or 3. [b]Suspicious lesions were category 4 or 5.

edition of BI-RADS. The interobserver agreement for the modified categorization ranged from moderate to substantial in the final assessment, resulting in higher interobserver agreement than seen in previous studies based on the fifth edition of BI-RADS, which were fair to moderate [2,10–13]. Interestingly, the interobserver agreement for the final assessment category in the re-modified categorization, including quantitative criteria such as size and the anterior-posterior ratio, was substantial. The interobserver agreement for the re-modified categorization was higher than for either the initially modified categorization or the BI-RADS categorization.

The interobserver agreement for category 2 was higher than in previous studies, and that for category 3 was similar to the values reported in previous studies [2,10–13]. Interobserver agreement for category 5 was found to be sufficient in the re-modified categorization (κ=0.51), consistent with the findings reported by Lazarus et al. [2] (κ=0.56), but lower than those reported by Abdullah et al. [14] and Elverici et al. [15] (κ=0.60 and κ=0.65, respectively). This difference might have been due to the presence of several small lesions leading to lower concordance in the margin and shape assessments, which are important factors for determining the level of suspicion because they are included as criteria for non-palpable masses found in screening breast US reports [16]. Furthermore, Lazarus et al. [2] suggested that the use of subcategories is helpful in communicating the level of suspicion to referring physicians and patients. However, previous studies have reported a fair degree of interobserver agreement for the subcategory classification of category 4, even among experienced observers [14,17]. Therefore, several studies have suggested the most appropriate criteria for the subcategories of category 4 [7,18]. In this study, the interobserver agreement was fair for both the initially modified and the re-modified categorization using the objective criteria, similar to previous studies [17,19,20]. There was no difference between the initially modified and re-modified categorization for the subcategories in this study. Although the interobserver agreement for the subcategory classification of category 4 in BI-RADS using subjective criteria was lower than was observed using both the initially modified and re-modified categorization, it was still fair. Therefore, in this study, the subcategory classification of category 4 was assessed by observers' subjective criteria due to the lack of differences between subjective criteria and objective criteria. However, after the MUST-BE trial, the most appropriate subcategory classification of category 4 in screening breast US must be determined.

There was a substantial degree of interobserver agreement in breast categorization regarding the decision to biopsy in all tests in this study when dichotomizing the interpretation as benign (categories 2 and 3) or suspicious (categories 4 and 5) (initially modified categorization and re-modified categorization: κ=0.68 and κ=0.73, respectively). The overall interobserver agreement for both the initially modified categorization and re-modified categorization in this study was higher than the agreement based on the fifth edition of BI-RADS categorization in this study and a previous study [20]. As a result, we decided to modify the categorization system for categories 2 and 3 for the MUST-BE trial by adding size and the anterior-posterior ratio as criteria due to their relatively high interobserver agreement compared with previous studies [2,14,19,20].

In this study, we evaluated interobserver agreement based on the modified categorization developed by the ABCS-K according to two groups of observers: senior (≥10 years' experience) and junior (<10 years' experience). There was no significant difference according to the radiologist's experience in overall interobserver agreement or when the assessment was dichotomized into benign or suspicious for the initially modified and re-modified categorizations (Table 6). A previous study evaluated 54 breast lesions assessed by the same two groups of observers [13]. The authors suggested that interobserver agreement was more dependent on case difficulty than on observer experience [13]. Other studies have reported that interobserver variation depended on lesion size rather than observer experience [16,21]. Therefore, we conclude that the modified breast categorization by ABCS-K is likely to be widely applicable in screening breast US because it is not dependent on the radiologist's experience.

Our study had several limitations. First, this study was based on a data review of static images of breast masses, which, while allowing for selection of representative lesion images, precluded the visualization of full lesions. Second, this study was retrospective, with a relatively small number of cases. Third, the enrolled cases were detected on both screening breast US and diagnostic breast US. Therefore, the results may not be generalizable to the true screening population. Fourth, biopsy-proven lesions or lesions with a follow-up interval of at least 2 years were included, which limited the number of BI-RADS category 2 and category 3 lesions included in the study. This selection bias may have contributed to an overcategorization of some masses despite the fact that the observers were blind to the biopsy results. Furthermore, diagnostic performance was not evaluated in this study because false-negative lesions could not be identified.

This preliminary study suggests the need to standardize breast US categorization for breast cancer screening in Korea. Even if the findings of this study do not reach the level of a fully-standardized categorization for breast cancer screening in Korea, this study provides an acceptable guide for quality control for radiologists participating in the MUST-BE trial for the first time. For this trial to be successful and to achieve reliable results, the interobserver agreement and accuracy of the participating radiologists should be periodically monitored using the modified categorization. After the MUST-BE trial, a retrospective comparative analysis of enrolled cases will be undertaken to standardize breast US categorization for breast cancer screening in Korea.

In conclusion, these preliminary results demonstrate acceptable interobserver agreement as quality control for the modified categorization of screening breast US developed by the ABCS-K for the MUST-BE trial.

ORCID: Eun Jung Choi: https://orcid.org/0000-0002-2339-4327; Eun Hye Lee: https://orcid.org/0000-0002-8773-700X; You Me Kim: https://orcid.org/0000-0001-5807-9012; Yun-Woo Chang: https://orcid.org/0000-0001-9704-8112; Jin Hwa Lee: https://orcid.org/0000-0003-0843-9862; Young Mi Park: https://orcid.org/0000-0001-7332-3853; Keum Won Kim: https://orcid.org/0000-0002-7312-5483; Young Joong Kim: https://orcid.org/0000-0002-7084-0289; Jae Kwan Jun: https://orcid.org/0000-0003-1647-0675; Seri Hong: https://orcid.org/0000-0002-2536-0606

*Author affiliations

[1]Department of Radiology and Research Institute of Clinical Medicine of Chonbuk National University-Biomedical Research Institute of Chonbuk National University Hospital, Chonbuk National University Medical School, Jeonju; [2]Department of Radiology, Soonchunhyang University Bucheon Hospital, Soonchunhyang University College of Medicine, Bucheon; [3]Department of Radiology, Dankook University Hospital, Dankook University College of Medicine, Cheonan; [4]Department of Radiology, Soonchunhyang University Seoul Hospital, Soonchunhyang University College of Medicine, Seoul; [5]Department of Radiology, Dong-A University Hospital, Dong-A University College of Medicine, Busan; [6]Department of Radiology, Inje University Busan Paik Hospital, Busan; [7]Department of Radiology, Konyang University Hospital, Konyang University College of Medicine, Daejeon; [8]National Cancer Control Institute, National Cancer Center, Goyang, Korea

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (1520200).

We would like to thank Hye-Mi Jo, MPH for data collection and management.

References

1. Chala L, Endo E, Kim S, de Castro F, Moraes P, Cerri G, et al. Gray-scale sonography of solid breast masses: diagnosis of probably benign masses and reduction of the number of biopsies. J Clin Ultrasound 2007;35:9-19.
2. 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.
3. Hashimoto BE, Morgan GN, Kramer DJ, Lee M. Systematic approach to difficult problems in breast sonography. Ultrasound Q 2008;24:31-38.
4. Youk JH, Kim EK. Supplementary screening sonography in mammographically dense breast: pros and cons. Korean J Radiol 2010;11:589-593.
5. American College of Radiology. Breast Imaging Reporting and Data System: BI-RADS atlas. 4th ed. BI-RADS: ultrasound. Reston, VA. American College of Radiology, 2003.
6. Mendelson EB, Bohn-Velez M, Berg WA, Whitman GJ, Feldman MI, Madjar H, et al. ACR BI-RADS ultrasound. In: D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, eds. ACR BI-RADS atlas: Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology, 2013;1-173.
7. 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.
8. 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.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
10. Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: past, present, and future. AJR Am J Roentgenol 2015;204:234-240.
11. Kim SJ, Chang JM, Cho N, Chung SY, Han W, Moon WK. Outcome of breast lesions detected at screening ultrasonography. Eur J Radiol 2012;81:3229-3233.
12. D'Orsi CJ, Sickles EA. To seek perfection or not? That is the question. Radiology 2012;265:9-11.
13. Shimamoto K, Sawaki A, Ikede M, Satake H, Naganawa S, Tadokoro M, et al. Interobserver agreement in sonographic diagnosis of breast tumors. Eur J Ultrasound 1998;8:25-31.
14. 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.
15. Elverici E, Zengin B, Nurdan Barca A, Didem Yilmaz P, Alimli A, Araz L. Interobserver and intraobserver agreement of sonographic BIRADS lexicon in the assessment of breast masses. Iran J Radiol 2013;10:122-127.
16. Del Frate C, Bestagno A, Cerniato R, Soldano F, Isola M, Puglisi F, et al. Sonographic criteria for differentiation of benign and malignant

solid breast lesions: size is of value. Radiol Med 2006;111:783-796.

17. 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.
18. Jales RM, Sarian LO, Torresan R, Marussi EF, Alvares BR, Derchain S. Simple rules for ultrasonographic subcategorization of BI-RADS(R)-US 4 breast masses. Eur J Radiol 2013;82:1231-1235.
19. Lee HJ, Kim EK, 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 Radiol 2008;65:293-298.
20. Berg WA, Blume JD, Cormack JB, Mendelson EB. Operator dependence of physician-performed whole-breast US: lesion detection and characterization. Radiology 2006;241:355-365.
21. Calas MJ, Almeida RM, Gutifilen B, Pereira WC. Interobserver concordance in the BI-RADS classification of breast ultrasound exams. Clinics (Sao Paulo) 2012;67:185-189.