More than interobserver agreement is required for comparisons of categorization systems

ULTRA SONO GRAPHY

Gloria Palazuelos, Sergio Alfonso Valencia, Javier Andres Romero

Department of Radiology, Fundación Santa Fe de Bogotá, Bogota, Colombia

LETTER

https://doi.org/10.14366/usg.19021
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2019;38:374-376

We read with interest the article by Choi et al. [1], titled "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" in the last issue of *Ultrasonography.* Their article evaluated the interobserver agreement of the modified categorization system established by the Alliance for Breast Cancer Screening in Korea (ABCS-K) and compared the results with the Breast Imaging Reporting and Data System (BI-RADS) categorization. Because the present data consist of preliminary results, it is crucial for us to clarify some points.

The authors used the kappa statistic to evaluate interobserver concordance, but they did not present a frequency table by categories for each categorization system. The kappa statistic has limitations depending on the prevalence of a condition. This is known as the kappa paradox, and if there are doubts about its presence, some other statistics can be used to determine levels of concordance [2].

It is interesting to see the good interobserver concordance of the re-modified ABCS-K categorization, but the interobserver concordance of the BI-RADS categorization differs from previous reports (κ-value of 0.495 vs. 0.51–0.53) [3,4], especially in BI-RADS category 5 (κ-value of 0.45 vs. 0.71) [1,4]. The authors should determine why these differences in the BI-RADS concordance occurred and should take into account the possibility that the discrepancies could have been due to the expertise of the radiologist. It would be also interesting to see a table that compares the κ-value of the BI-RADS categorization by the radiologist's years of experience.

Another point worth discussing is the methodology used in the ABCS-K categorization, because it is categorized according to major and minor findings, in contrast to BI-RADS, which uses the positive predictive value (PPV) of each finding; this difference can be meaningful, especially for subcategories 4a, 4b, and 4c. Some minor findings in ABCS-K have previously been proven to have a high PPV, such as the presence of calcification in the mass (PPV, 84.6%–100%), echogenic halo (PPV, 66.7%), and angular margin (PPV, 60%) [5]. For this reason, it is essential to compare the diagnostic performance of the ABCS-K categorization to that of BI-RADS. Although concordance is important when selecting a categorization system, the diagnostic performance of a categorization system is an essential factor affecting its suitability for clinical use. For example, the BI-RADS categorization system has shown good diagnostic performance, with an area under the receiver operating characteristic curve of 0.708 in the fourth edition and 0.690 in the fifth edition [5].

ORCID: Gloria Palazuelos: https://orcid.org/0000-0001-5245-6809; Sergio Alfonso Valencia: https://orcid.org/0000-0002-0605-411X; Javier Andres Romero: https://orcid.org/0000-0003-1193-9980

Received: April 19, 2019
Accepted: May 15, 2019

Correspondence to:
Sergio Alfonso Valencia, MD,
Department of Radiology, Fundación Santa Fe de Bogotá, 116 Street # 9-02, Bogotá, Colombia

Tel. +57-1-6030303
Fax.+57-1-6575714
E-mail: sevava92@gmail.com

How to cite this article:
Palazuelos G, Valencia SA, Romero JA. More than interobserver agreement is required for comparisons of categorization systems. Ultrasonography. 2019 Oct;38(4):374-376.

Conflict of Interest

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

References

1. 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;38:172-180.
2. Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. BMC Med Res Methodol 2014;14:100.
3. 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.
4. 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.
5. Yoon JH, Kim MJ, Lee HS, Kim SH, Youk JH, Jeong SH, et al. Validation of the fifth edition BI-RADS ultrasound lexicon with comparison of fourth and fifth edition diagnostic performance using video clips. Ultrasonography 2016;35:318-326.

Response

Eun Jung Choi[1], Eun Hye Lee[2]

[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, Korea

We thank you for your interest and comments on our article titled, "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."

First, using the initially modified categorization, there were 63 benign and 62 suspicious lesions on ultrasonography (US), and 81 benign and 44 breast cancers in the final results. In contrast, using the re-modified categorization, there were 43 benign and 57 suspicious lesions on US, and 54 benign lesions and 46 breast cancers in the final results.

As you mentioned, the kappa statistic is subject to limitations based on the prevalence of a condition [1]. We stated in the Materials and Methods that the proportion of breast cancers among the test series in this article was not low; in fact, the proportion in the test series of this article was 35.2% (44 of 125) using the initially modified categorization and 46.0% (46 of 100) using the re-modified categorization. Therefore, applying the kappa statistic to evaluate interobserver agreement for ultrasound screening in this article is acceptable. In contrast, the prevalence of breast cancers among the test series for screening mammography in the MUST-BE trial was low (1.2%) [2]. Therefore, to avoid the kappa paradox, we applied percent agreement as well as the kappa statistic when evaluating interobserver agreement for mammography, which was done as a part of a quality control program in the trial.

Although most radiologists participating in the MUST-BE trial were experienced in breast imaging (mean, 10.1 years) in an academic setting, the kappa values reported in this article were lower than those of other studies [3,4]. Our results might have been influenced by a larger number of cases and observers than other studies [3,4] because the kappa statistic is dependent on the number of categories and observers, and its value is generally higher if there are fewer categories and observers [1]. In spite of the lower interobserver agreement using the Breast Imaging Reporting and Data System (BI-RADS) categorization in this article, we believe that it is acceptable for real-world clinical practice because the interobserver agreement for dichotomous categories (whether to biopsy or not) was moderate and similar to those of other studies (Table 6 in the manuscript).

Regarding suspicious findings, some minor findings, including calcification in the mass and angular margin, are known to have high positive predictive values. We segregated the suspicious findings into major and minor findings to distinguish category 4 and 5 lesions with the goal of achieving both high reproducibility and convenience based on previous studies [5]. However, we did not achieve an acceptable value for interobserver agreement regarding category 4 subcategorization using the modified categorization

system. Therefore, we decided not to apply these criteria for the subcategorization of category 4 in the MUST-BE trial. Instead, we will perform a further analysis to classify the major and minor findings for the subcategorization of category 4 after completion of a research database including information about patients' breast cancer diagnoses.

References

1. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010;73:1167-1179.
2. Kim SH, Lee EH, Jun JK, Kim YM, Chang YW, Lee JH, et al. Interpretive performance and inter-observer agreement on digital mammography test sets. Korean J Radiol 2019;20:218-224.
3. 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.
4. 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.
5. 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.