Radiología. 2020;62(6):481–486

RADIOLOGÍA

www.elsevier.es/rx

ORIGINAL ARTICLES

Interobserver and intraobserver variability in determining breast density according to the fifth edition of the BI-RADS® Atlas☆

K. Pesce[a], M. Tajerian[b], M.J. Chico[a], M.P. Swiecicki[a], B. Boietti[b], M.J. Frangella[b,*], S. Benitez[b]

[a] *Servicio de Diagnóstico por imágenes, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina*
[b] *Departamento de Informática en Salud, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina*

Received 22 October 2019; accepted 15 April 2020

KEYWORDS
Breast density;
Intraobserver variability;
Interobserver variability;
BI-RADS® fifth edition;
Breast density on mammograms

Abstract
Objective: To determine the level of agreement within and between observers in the categorization of breast density on mammograms in a group of professionals using the fifth edition of the American College of Radiology's BI-RADS® Atlas and to analyze the concordance between experts' categorization and automatic categorization by commercial software on digital mammograms.
Methods: Six radiologists categorized breast density on 451 mammograms on two occasions one month apart. We calculated the linear weighted kappa coefficients for inter- and intra-observer agreement for the group of radiologists and between the commercial software and the majority report. We analyzed the results for the four categories of breast density and for dichotomous classification as dense versus not dense.
Results: The interobserver agreement among radiologists and the majority report was between moderate and nearly perfect for the analysis by category ($\kappa = 0.64$ to 0.84) and for the dichotomous classification ($\kappa = 0.63$ to 0.84). The intraobserver agreement was between substantial and nearly perfect ($\kappa = 0.68$ to 0.85 for 4 categories and $k = 0.70$ to 0.87 for the dichotomous classification). The agreement between the majority report and the commercial software was moderate both for the four categories ($\kappa = 0.43$) and for the dichotomous classification ($\kappa = 0.51$).
Conclusion: Agreement on breast density within and between radiologists using the criteria established in the fifth edition of the BI-RADS® Atlas was between moderate and nearly perfect. The level of agreement between the specialists and the commercial software was moderate.

☆ Please cite this article as: Pesce K, Tajerian M, Chico MJ, Swiecicki MP, Boietti B, Frangella MJ, et al. Estudio de la variabilidad inter- e intraobservador en la determinación de la densidad mamaria según la 5.ª edición del Atlas BI-RADS®. Radiología. 2020;62:481–486.

* Corresponding author.
E-mail address: maria.frangella@hospitalitaliano.org.ar (M.J. Frangella).

PALABRAS CLAVE
Densidad mamaria;
Variabilidad intraobservador;
Variabilidad interobservador;
BI-RADS® 5 edición;
Densidad mamográfica

Estudio de la variabilidad inter- e intraobservador en la determinación de la densidad mamaria según la 5.ª edición del Atlas BI-RADS®

Resumen
Objetivo: Determinar el acuerdo intra- e interobservador en la categorización de la densidad mamográfica entre un grupo de profesionales según la 5.ª edición del Atlas BI-RADS® - ACR y analizar la concordancia entre la categorización de los expertos y un software comercial de un mamógrafo digital para categorización automática.
Métodos: 6 médicos categorizaron la densidad mamográfica de 451 mamografías en dos oportunidades con un intervalo de 1 mes. Calculamos los coeficientes kappa ponderados lineales de acuerdo inter- e intraobservador para el grupo médico y la concordancia entre el *software* comercial y el reporte de la mayoría. Analizamos los resultados para las cuatro categorías de densidad mamaria y para el resultado dicotómico de mama densa/no densa.
Resultados: El acuerdo interobservador entre especialistas y el reporte de la mayoría fue moderado y casi perfecto para el análisis por categoría ($\kappa = 0{,}64$ a 0,84) y de manera dicotómica ($\kappa = 0{,}63$ a 0,84). El acuerdo intraobservador fue sustancial y casi perfecto ($\kappa = 0{,}68$ a 0,85 para 4 categorías y $k = 0{,}70$ a 0,87 para el análisis dicotómico). El acuerdo entre el reporte de la mayoría y el *software* comercial fue moderado tanto por categoría ($\kappa = 0{,}43$) como en el análisis dicotómico ($\kappa = 0{,}51$).
Conclusión: Hemos observado un acuerdo entre moderado y casi perfecto inter- e intraobservador entre los radiólogos, según los criterios establecidos en la 5.ª edición del Atlas BI-RADS®. El nivel de acuerdo entre el reporte de los especialistas y un *software* disponible comercialmente fue moderado.

Introduction

Breast density, evaluated by means of mammography, is defined as relative quantities of radiodense stromal and epithelial tissues in comparison to radiotransparent adipose tissue.[1] The 5th edition of the BI-RADS® Atlas of the American College of Radiology (ACR) defines four patterns of breast density, specifically ACR a (almost entirely fatty), ACR b (scattered areas of fibroglandular tissue), ACR c (heterogeneously dense) and ACR d (extremely dense).[2]

High breast density (categories c and d) is an independent risk factor for developing breast cancer[3–5] and a masking factor that decreases mammography's sensitivity for detecting this disease and other lesions.[6–10]

Breast density may be visually evaluated by a radiologist. Some authors have found that this method entails considerable intraobserver and interobserver variability.[11–13] A systematic literature review conducted by the Task Force working group in 2016 reported that, in community environments, 19–22% of mammography examinations initially classified as dense were subsequently classified as not dense, and 10–16% of examinations initially deemed not dense were reclassified as dense. Regarding sequential reproducibility, the group reported that in 20% of studies the category was changed in the subsequent round if the reading was performed by the same radiologist, and that this proportion rose to 33% if the evaluation was performed by a different radiologist.[13,14]

Moreover, automated breast density measuring systems are becoming increasingly common.[15–19] Consequently, in clinical practice, breast density is reported based on assessment performed by specialists aided by these systems. Yet few studies have analysed professionals' use of, concordance with and perception of the usefulness of these tools.

The objective of our study was to determine the degree of intraobserver and interobserver agreement in categorising breast density according to the 5th edition of the BI-RADS® Atlas of the ACR among a group of professionals specialised in breast imaging at a highly complex health institution. We also analysed concordance between categorisation by experts and categorisation by an automated categorisation method (a commercial software program for a digital mammography machine).

Methods

Setting

Our study was conducted in the Breast Diagnosis and Intervention Section of the Diagnostic Imaging Department at a tertiary hospital. The department has had digital imaging and an integrated radiology information system/picture archiving and communication system (RIS/PACS) since 2010.[20] The section is composed of ten specialists and two fellows, and reports an annual average of 30,000 mammograms. Mammograms are randomly assigned on a daily basis to radiologists for reporting; each receives 200–400 cases per month. Once mammogram reports are redacted, 10% of studies reported by specialists (approximately 300 studies per month) and all studies reported by fellows are submit-

ted for peer review. In addition, report quality audits are performed by the physicians who order the studies.

Study design

This cross-sectional study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Independent Ethics Committee at our institution. Patient consent was also obtained.

Categorisation of breast density according to the 5th edition of the BI-RADS® of the ACR

The team of professionals was made up of six physicians from the Breast Diagnosis and Intervention Section. The group had an average of nine years and a range of two to eighteen years of experience in breast imaging. A total of 451 mammograms from randomly selected asymptomatic patients 40–90 years of age, performed at the institution in February 2019, were included. One of four acquisitions was drawn from each study; it could be of the craniocaudal or mediolateral oblique view. Focalised and magnified incidents as well as mammograms for patients with a personal history of breast surgery (including breast implants) or gigantomastia were excluded. The latter was defined by the need to use more than one plate per incident.

A bioengineer and a radiologist, who did not take part in the subsequent categorisation, extracted the images to be evaluated from the hospital database and removed patient-identifying data. Consecutive sampling was used.

A week before the 451 mammography images were evaluated, the participating professionals reviewed the criteria for categorisation of breast density with images in the 5th edition of the BI-RADS®[2] Atlas. Next, the specialists categorised breast density in each mammogram. The images were evaluated at 5-megapixel workstations.

The evaluators were not aware of the patients' demographic data or the category assigned in the original report for each mammogram. They were also unaware of the density assigned by the other participants in the study or the assessment of the commercial software program. Two readings of the same mammograms were performed a month apart. The order of the studies in each of the two readings was random. All this information was recorded in an electronic database.

We used a commercially available breast density measurement software program for the AMULET Innovality (3000AWS7.0 option) Fujifilm® mammography machine for automated categorisation of breast density. This software program calculates breast density based on the ratio of fibroglandular tissue to total breast area.[21] It uses this proportion to estimate breast density on a continuous scale and automatically assigns a category.

Endpoints of interest and statistical methods

We evaluated intraobserver and interobserver variability for each participating professional by calculating linear weighted kappa coefficients, reported with a 95% confidence interval (CI). The coefficient was calculated for the four categories of breast density and for the dichotomous result of dense versus non-dense breast.

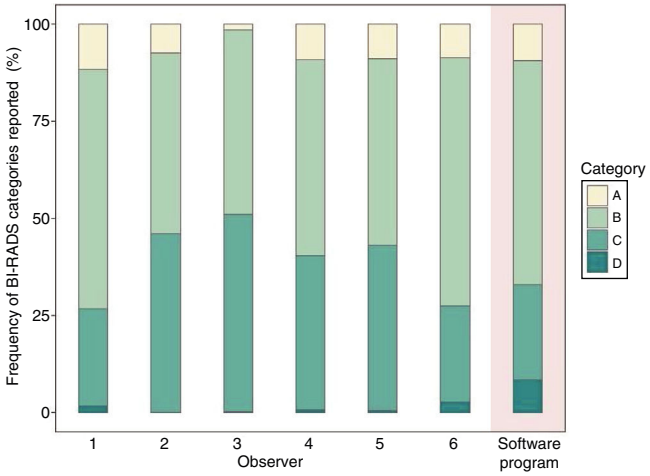

Figure 1 Distribution of frequencies of breast density categories by specialist.

We also selected the majority report measure, defined as the statistical mode of the observers' reports, consistent with pre-existing literature.[22] We calculated the level of agreement between each evaluator and the majority report. In the event of a tie (non-unimodal categorisation distribution), a seventh imaging specialist categorised the mammogram to reach an agreement.

For the calculation of the linear weighted kappa coefficient (κ) we used the method described by Cohen and Fleiss.[23,24] We reported kappa coefficients of interobserver and intraobserver agreement, which for their part were classified according to their level of agreement in accordance with the kappa coefficient breakdown proposed by Landis and Koch (0: ''poor''; 0–0.2: ''slight''; 0.21–0.4: ''fair''; 0.41–0.6: ''moderate''; 0.61–0.8: ''substantial''; 0.81–1: ''almost perfect'').[25]

Finally, we reported the concordance between the commercial software program and the majority report.

We used the statistics software programs STATA version 14 and R version 3.6.0. A p-value <0.05 was considered significant.

The participating investigators declare that they have no conflicts of interest.

Results

The frequency of the categories assigned by each specialist is shown in Fig. 1. The linear weighted kappa values for overall interobserver agreement (between each specialist and the majority report) at the time of the first evaluation are shown in Table 1. These values attained a range of 0.64 (CI: 0.58–0.70) to 0.84 (CI: 0.80–0.89) for the category analysis. In other words, a moderate to almost perfect agreement was reached between each observer and the majority report. Similar values were obtained when the results were analysed dichotomously from a clinical perspective and the images were categorised as ''dense breast'' or ''non-dense breast''.

Table 1 Linear weighted kappa coefficient between each specialist and the majority report.

	Linear weighted kappa coefficient (95% CI)	
	Interobserver On 4-category scale[a]	Interobserver Dichotomised (dense/non-dense)[b]
Observer 1	0.66 (0.60—0.72)	0.63 (0.55—0.70)
Observer 2	0.77 (0.72—0.83)	0.76 (0.70—0.82)
Observer 3	0.64 (0.58—0.70)	0.72 (0.66—0.78)
Observer 4	0.84 (0.80—0.89)	0.84 (0.79—0.89)
Observer 5	0.83 (0.78—0.87)	0.80 (0.75—0.86)
Observer 6	0.67 (0.61—0.73)	0.66 (0.59—0.73)
Commercial software	0.46 (0.39—0.52)	0.51 (0.43—0.59)

[a] Interobserver agreement for the first step of observation in the 4 density categories of the BI-RADS®.
[b] Interobserver agreement for the first step of observation for the dichotomised result (dense/non-dense breast).

Table 2 Linear weighted kappa coefficient for each specialist in the readings a month apart.

	Linear weighted kappa coefficient (95% CI)	
	Intraobserver On 4-category scale[a]	Intraobserver Dichotomised (dense/non-dense)[b]
Observer 1	0.76 (0.71—0.81)	0.75 (0.68—0.82)
Observer 2	0.70 (0.64—0.76)	0.70 (0.64—0.77)
Observer 3	0.85 (0.80—0.89)	0.87 (0.83—0.92)
Observer 4	0.72 (0.66—0.77)	0.72 (0.66—0.79)
Observer 5	0.68 (0.63—0.74)	0.73 (0.67—0.77)
Observer 6	0.73 (0.68—0.79)	0.77 (0.70—0.83)
Majority report	0.80 (0.76—0.85)	0.85 (0.80—0.90)

[a] Intraobserver agreement for the six radiologists according to the 4 categories.
[b] Intraobserver agreement for the dichotomised result (dense/non-dense breast).

The linear weighted kappa values for intraobserver agreement between the first and the second observation are shown in Table 2. The results attained a range of 0.68 (CI: 0.63—0.74) to 0.85 (CI: 0.80—0.89). For the dichotomous analysis, the results were 0.70 (CI: 0.64—0.77) to 0.87 (CI: 0.83—0.92). For the above-mentioned results, the values correspond to a level of agreement between substantial and almost perfect.

The agreement between the majority report and the commercial software program was moderate for the four categories and for the dichotomous analysis, with kappa values of 0.46 (CI: 0.39—0.52) and 0.51 (CI: 0.43—0.59), respectively.

Discussion

Breast density is an independent risk factor for developing breast cancer. Categorisation of breast density is essential for performing a personalised risk assessment and efficiently supplementing population screening efforts with higher-sensitivity studies in patients with dense patterns, such as magnetic resonance imaging and breast ultrasound.[26] This factor may be used to select the most appropriate method for diagnosis of each patient. This prevents clinicians from performing too many or not enough complementary tests, thus preventing delays in due access to said tests.

This study evaluated interobserver and intraobserver variability in a diagnostic imaging department in categorising breast density, based on the 5th edition of the BI-RADS® Atlas.

Interobserver agreement for breast density was substantial in the concordance between the six radiologists and the majority report, both for the four-category scale and for the dichotomous (dense/non-dense) categorisation. When we analysed similar studies in the scientific literature, we found highly variable results. Some groups have reported more mixed observations. Some studies have shown interobserver concordance between pairs of radiologists ranging from slight to substantial, with kappa values ranging from 0.02 to 0.72 (mean = 0.46; 95% CI: 0.36—0.55).[27] Other studies, by contrast, have reported values similar to those presented in this study.[22,27–30] The fact that the group of physicians participating in the study engage in regular academic activities, grand rounds, updates, report audits and so on could account for these results.

In its evaluation of intraobserver variability, our study recorded kappa values with substantial and almost perfect agreement,[31] both on a four-category scale and on a dichotomous scale, with no statistically significant difference between them. However, we should mention the potential risk of bias due to the brief period that elapsed between the two readings.

In relation to the above, while we found studies that reported a statistically significant difference for observers with more than 10 years of experience in mammogram diagnosis,[28] our results diverged from theirs. This may be explained by the fact that the medical team belongs to a specialised section with uniform diagnostic criteria, regular update meetings and an exclusive focus on breast diagnosis. As mentioned, the evaluators participated in an update session on criteria for breast density categorisation prior to the start of the study. Therefore, our results could be generalised to centres with the same characteristics.

The main advantages of an automated diagnostic tool are its consistency over time and lack of variability.[32] Hence, an observer-independent automated system enables reproducible measurements and should be more appropriate for a reliable, standardised evaluation. In this study, agreement between the majority of the physicians and the commercially available software program was moderate. This could be attributed, in the first place, to the fact that the criteria associated with the two characterisation methods are different. Visual categorisation is primarily based on professional knowledge and experience, whereas the automated method uses a quantitative strategy to determine the ACR category. Second, development and validation processes for a software tool are determinant of its performance in different scenarios. It would be useful to explore the reasons for this difference in future studies.

Although this study was conducted at a single institution, it is a leading institution and receives referrals from all over Argentina on a daily basis. A multi-centre study would be a good idea for evaluating new technologies. The sample used had a limited number of cases with a breast density pattern classified as extremely dense (ACR d), consistent with the prevalence thereof according to department reports over the past five years, which hovers around 1–2%. Even so, the total prevalence of high density (categories c and d) in the sample was 41%, also consistent with the prevalence in our hospital population. The number of mammograms used and the randomised order of the studies avoided memory biases due to potential effects of familiarisation with the sample of mammograms between the two observation periods. Finally, our design ensured that the physicians were blinded to the reports issued by the automated classification software and the diagnoses made by all the other evaluators.

Conclusion

Although there is general variability between observers and even within a single operator, qualitative classification of breast density is an acceptable method with moderate to almost perfect interobserver and intraobserver agreement according to the criteria established in the 5th edition of the BI-RADS® Atlas. We found a moderate level of agreement between the reports of the specialists and a commercially available software program. Future studies will be able to examine and characterise agreement between specialists and automated classification methods in greater depth.

Authors

Study integrity: SB, KP.
Study concept: MT, KP, JF.
Study design: BB, JF.
Data acquisition: MJC, MPS.
Data analysis and interpretation: BB, JF, MT, KP.
Statistical processing: BB, MT.
Literature search: JF, MT, KP, MJC.
Drafting of the paper: JF, MT, KP, BB, MPS.
Critical review of the manuscript with intellectually significant contributions: SB.
Approval of the final version: JF, MT, KP, BB, MPS, SB.

Conflicts of interest

The authors declare that they have no conflicts of interest.

References

1. Winkler NS, Raza S, Mackesy M, Birdwell RL. Breast density: clinical implications and assessment methods. Radiographics. 2015;35:316–24, http://dx.doi.org/10.1148/rg.352140134.
2. Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS® Mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology; 2013.
3. Ciatto S, Visioli C, Paci E, Zappa M. Breast density as a determinant of interval cancer at mammographic screening. Br J Cancer. 2004;90:393–6, http://dx.doi.org/10.1038/sj.bjc.6601548.
4. Wanders JOP, Holland K, Karssemeijer N, Peeters PHM, Veldhuis WB, Mann RM, et al. The effect of volumetric breast density on the risk of screen-detected and interval breast cancers: a cohort study. Breast Cancer Res. 2017;19:67, http://dx.doi.org/10.1186/s13058-017-0859-9.
5. Strand F, Azavedo E, Hellgren R, Humphreys K, Eriksson M, Shepherd J, et al. Localized mammographic density is associated with interval cancer and large breast cancer: a nested case-control study. Breast Cancer Res. 2019;21:8, http://dx.doi.org/10.1186/s13058-019-1099-y.
6. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2006;15:1159–69, http://dx.doi.org/10.1158/1055-9965.EPI-06-0034.
7. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007;356:227–36, http://dx.doi.org/10.1056/NEJMoa062790.
8. Swann CA, Kopans DB, McCarthy KA, White G, Hall DA. Mammographic density and physical assessment of the breast. AJR Am J Roentgenol. 1987;148:525–6, http://dx.doi.org/10.2214/ajr.148.3.525.
9. Mousa DSAL, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? Clin Radiol. 2014;69:333–41, http://dx.doi.org/10.1016/j.crad.2013.11.014.
10. Carreira Gómez MC, Estrada Blan MC. What we need to know about dense breasts: implications for breast cancer screening. Radiologia. 2016;58:421–6, http://dx.doi.org/10.1016/j.rx.2016.08.002.

11. Sprague BL, Conant EF, Onega T, Garcia MP, Beaber EF, Herschorn SD, et al. Variation in mammographic breast density assessments among radiologists in clinical practice: a multicenter observational study. Ann Intern Med. 2016;165:457–64, http://dx.doi.org/10.7326/M15-2934.
12. Eom H-J, Cha JH, Kang J-W, Choi WJ, Kim HJ, Go E. Comparison of variability in breast density assessment by BI-RADS category according to the level of experience. Acta Radiol. 2018;59:527–32, http://dx.doi.org/10.1177/0284185117725369.
13. Alikhassi A, Esmaili Gourabi H, Baikpour M. Comparison of inter- and intra-observer variability of breast density assessments using the fourth and fifth editions of Breast Imaging Reporting and Data System. Eur J Radiol Open. 2018;5:67–72, http://dx.doi.org/10.1016/j.ejro.2018.04.002.
14. Melnikow J, Fenton JJ, Whitlock EP, Miglioretti DL, Weyrich MS, Thompson JH, et al. Supplemental screening for breast cancer in women with dense breasts: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med. 2016;164:268–78, http://dx.doi.org/10.7326/M15-1789.
15. Jeffers AM, Sieh W, Lipson JA, Rothstein JH, McGuire V, Whittemore AS, et al. Breast cancer risk and mammographic density assessed with semiautomated and fully automated methods and BI-RADS. Radiology. 2017;282:348–55, http://dx.doi.org/10.1148/radiol.2016152062.
16. Ciatto S, Bernardi D, Calabrese M, Durando M, Gentilini MA, Mariscotti G, et al. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. Breast. 2012;21:503–6, http://dx.doi.org/10.1016/j.breast.2012.01.005.
17. Alonzo-Proulx O, Jong RA, Yaffe MJ. Volumetric breast density characteristics as determined from digital mammograms. Phys Med Biol. 2012;57:7443–57, http://dx.doi.org/10.1088/0031-9155/57/22/7443.
18. Martínez Gómez I, Casals El Busto M, Antón Guirao J, Ruiz Perales F, Llobet Azpitarte R. Semiautomatic estimation of breast density with DM-Scan software. Radiologia. 2014;56:429–34, http://dx.doi.org/10.1016/j.rx.2012.11.007.
19. Gao J, Warren R, Warren-Forward H, Forbes JF. Reproducibility of visual assessment on mammographic density. Breast Cancer Res Treat. 2008;108:121–7, http://dx.doi.org/10.1007/s10549-007-9581-0.
20. Luna D, Plazzotta F, Otero C, González Bernaldo de Quirós F, Baum A, Benítez S, Available from: http://hdl.handle.net/11362/3959, 2012.
21. https://www.dma.no/files/298/fujifilm_amulet_innovality.pdf.
22. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS Atlas. AJR Am J Roentgenol. 2016;206:1119–23, http://dx.doi.org/10.2214/AJR.15.15049.
23. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46. Available from: http://journals.sagepub.com/doi/10.1177/001316446002000104
24. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33:613–9. Available from: http://journals.sagepub.com/doi/10.1177/001316447303300309
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74 https://www.ncbi.nlm.nih.gov/pubmed/843571
26. Berg WA, Zhang Z, Lehrer D, Jong RA, Pisano ED, Barr RG, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA. 2012;307:1394–404, http://dx.doi.org/10.1001/jama.2012.388.
27. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol. 2000;174:1769–77, http://dx.doi.org/10.2214/ajr.174.6.1741769.
28. Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH, Rutter CM. Misclassification of breast imaging reporting and data system (BI-RADS) mammographic density and implications for breast density reporting legislation. Breast J. 2015;21:481–9, http://dx.doi.org/10.1111/tbj.12443.
29. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, et al. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. Breast. 2005;14:269–75, http://dx.doi.org/10.1016/j.breast.2004.12.004.
30. Redondo A, Comas M, Macià F, Ferrer F, Murta-Nascimento C, Maristany MT, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br J Radiol. 2012;85:1465–70, http://dx.doi.org/10.1259/bjr/21256379.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
32. Tourassi GD, Floyd CE. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. Med Decis Making. 1997;17:186–92, http://dx.doi.org/10.1177/0272989X9701700209.