Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms

[1,2]A REDONDO, MD, MPH, [1,3,4]M COMAS, MSc, PhD, [1,3,4]F MACIÀ, MD, MPH, [5]F FERRER, MD, PhD, [1,3,4]C MURTA-NASCIMENTO, MD, PhD, [5]M T MARISTANY, MD, [1]E MOLINS, MSc, [1,3,4]M SALA, MD, PhD and [1,3,4]X CASTELLS, MD, PhD

[1]*Servei d'Epidemiologia i Avaluació, Hospital del Mar, Barcelona, Spain,* [2]*Unitat Docent de Medicina Preventiva i Salut Pública Hospital del Mar-UPF-ASPB, Barcelona, Spain,* [3]*CIBER Epidemiología y Salud Pública (CIBERESP), Spain,* [4]*Grup de Recerca en Epidemiologia i Avaluació, IMIM (Institut de Recerca Hospital del Mar), Barcelona, Spain, and* [5]*Servei de Radiodiagnòstic, IDIMAS-CRC-Mar, Barcelona, Spain*

Objective: The aim of this study was to evaluate reader variability in screening mammograms according to the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) assessment and breast density categories.
Methods: A stratified random sample of 100 mammograms was selected from a population-based breast cancer screening programme in Barcelona, Spain: 13 histopathologically confirmed breast cancers and 51 with true-negative and 36 with false-positive results. 21 expert radiologists from radiological units of breast cancer screening programmes in Catalonia, Spain, reviewed the mammography images twice within a 6-month interval. The readers described each mammography using BI-RADS assessment and breast density categories. Inter- and intraradiologist agreement was assessed using percentage of concordance and the kappa (κ) statistic.
Results: Fair interobserver agreement was observed for the BI-RADS assessment [$\kappa = 0.37$, 95% confidence interval (CI) 0.36–0.38]. When the categories were collapsed in terms of whether additional evaluation was required (Categories III, 0, IV, V) or not (I and II), moderate agreement was found ($\kappa = 0.53$, 95% CI 0.52–0.54). Intra-observer agreement for BI-RADS assessment was moderate using all categories ($\kappa = 0.53$, 95% CI 0.50–0.55) and substantial on recall ($\kappa = 0.66$, 95% CI 0.63–0.70). Regarding breast density, inter- and intraradiologist agreement was substantial ($\kappa = 0.73$, 95% CI 0.72–0.74 and $\kappa = 0.69$, 95% CI 0.68–0.70, respectively).
Conclusion: We observed a substantial intra-observer agreement in the BI-RADS assessment but only moderate interobserver agreement. Both inter- and intra-observer agreement in mammographic interpretation of breast density was substantial.
Advances in knowledge: Educational efforts should be made to decrease radiologists' variability in BI-RADS assessment interpretation in population-based breast screening programmes.

Received 2 February 2011
Revised 27 December 2011
Accepted 26 March 2012

DOI: 10.1259/bjr/21256379

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death in females worldwide, accounting for 23% of the total new cancer cases and 14% of the total cancer-related deaths in 2008 [1]. In Spain, more than 20 000 cases are diagnosed and approximately 6000 females die each year because of this tumour [2].

Breast cancer screening by mammography is the only evidence-based screening procedure currently available to reduce breast cancer mortality. Accuracy of screening mammography depends on various factors, such as the protocols for mammogram reading, the characteristics of the female and of the breast, and the experience of radiologists [3–5]. Great efforts have been made to improve its accuracy. One of these is the implementation of double reading as it increases the cancer detection rate and reduces the further assessment rate [6]. Furthermore, the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) in order to reduce discordance in the interpretation of mammographic findings, to standardise mammographic reporting and to facilitate follow-up [7, 8].

A limited number of studies have analysed observer variability in mammography interpretation using BI-RADS assessment as well as breast density categories [9–13]. Thus, the aim of this study was to assess the inter- and intra-observer agreement regarding the assessment and breast density in a breast cancer screening

Address correspondence to: Dr Xavier Castells, Servei d'Epidemiologia i Avaluació, Hospital del Mar–Parc de Salut Mar, Passeig Marítim, 25-29, 08003 Barcelona, Spain. E-mail: xcastells@hospitaldelmar.cat

This work was partially funded by the Fondo de Investigación Sanitaria, Instituto de Salud Carlos III, Spain (grant no. PI/020708); and by the Agència d'Avaluació de Tecnologia i Recerca Mèdiques, Servei Català de la Salut (grant no. 087/18/2004).

programme in the city of Barcelona, Spain. Furthermore, we wanted to investigate the association between female characteristics and BI-RADS discordance.

Methods and materials

A stratified random sample of 100 mammograms was selected from a consecutive sample of 33 435 examinations performed from 1996 to 1999 within the Barcelona breast cancer screening programme at the Hospital del Mar in Barcelona, Spain. Screening procedures have been described previously [5]. At that time, females aged 50–64 years were invited biennially to undergo a cost-free mammogram. For each female, two-view mammograms were undertaken (mediolateral–oblique and craniocaudal). In the screening programme, all mammograms are interpreted independently by two radiologists according to the BI-RADS classification and, in case of disagreement, a third reader is asked to arbitrate.

The mammogram sample selection took into account four possible screening results: true positive, true negative, false negative and false positive. These results were validated by comparing the original interpretation obtained in the screening programme with the result of the mammogram performed in the next screening round, 2 years later. The composition of the sample was based on an earlier publication by Kerlikowske et al [14]. Thus, of the 100 mammograms selected, 13 corresponded to females with histopathologically confirmed breast cancer (12 true positive and 1 false negative) and the remaining 87 corresponded to females with a definitive result of absence of cancer (51 with true-negative and 36 with false-positive results). The mean age of the subjects was 57.6 years (standard deviation 4.8 years).

All mammograms achieved the following minimum quality criteria. For the mediolateral–oblique view, the pectoral muscle was visible at least to the level of the nipple; all glandular tissue was included; the nipple was seen in profile; and the inframammary angle was visualised. For the craniocaudal view, the breast was situated centrally with the nipple in profile and the maximum amount of breast tissue was visualised. Females with previous cosmetic breast surgery, those with breast implants and those with radio-opaque skin markers on the breast were not eligible to be included in this study. Exclusions represented a small number of mammograms. All mammograms were original, for which both views were obtained using standard screen–film units [Toshiba^{TM} MGU-10A (Toshiba Medical System Corporation, Otawara-shi, Tochigi-ken, Japan) and Bennett^{TM} Profile Mammography System, model M-PRO (Bennet X-Ray Corp., Copiagne, NY)] using Agfa^{TM} Mamoray HT film (Afga-Gevaert N.V., Mortsel, Belgium).

A sample of 21 radiologists with experience in routinely reading mammograms was selected from the radiology unit services of distinct healthcare centres in Catalonia, Spain. To evaluate radiologists' experience in mammographic interpretation, a questionnaire was sent to them as previously reported [5]. In brief, radiologists had a mean of 12 years of clinical experience reading mammograms (range 4–22 years), had interpreted a mean of 5773 mammograms in the year prior to the study (range 1890–13 230) and had devoted an average of 56% of their working hours to reading mammograms (range 15–100%).

Participating radiologists did not have access to the subjects' characteristics or to prior mammograms. To standardise the criteria to report mammogram data, a training session was held with all readers before the start of the study. Screening mammograms were read independently by the radiologists, who were blinded to the original mammographic interpretation and cancer status, although they were notified that cancer cases were oversampled.

All mammograms were read twice by the participating radiologists. There was an interval of 6 months between the first reading and the second, and the reading order was changed. For each breast, readers provided information on the following variables: BI-RADS assessment and breast density (fatty, scattered fibroglandular densities, heterogeneously dense and extremely dense). The standardised assessment categories of the third edition of the BI-RADS assessment (1998) include: Category I, negative; Category II, benign finding; Category III, probably benign finding; Category 0, need additional imaging evaluation; Category IV, suspicious abnormality; and Category V, highly suggestive of malignancy. Additionally, we collapsed the BI-RADS assessments into two categories based on whether further investigation was required (Categories III, 0, IV and V) or not (Categories I or II).

For this study, the following variables relating to the females were considered: age at screening, family history of breast cancer, personal history of benign breast diseases and menopausal status.

The proportion of disagreement was calculated as follows: the numerator was the number of disagreements among pairs of radiologists and the denominator was the overall number of possible pairwise comparisons for the 21 radiologists: 210 pairs for 1 mammogram (21 × 20 divided by 2), and 21 000 pairs (210 × 100) for all 100 mammograms.

Statistical analyses

Cohen's kappa coefficient (κ) and its 95% confidence interval (95% CI) were calculated to measure inter- and intra-observer variability for both assessment and breast density. Since both variables are ordinal scales, we used weighted κ with quadratic weights [15]. Macro !KAPPA for SPSS [JM Domenech, A Bonillo and R Granero; Bellaterra (Barcelona), Spain; v.04.23.2002; available at www.metodo.uab.cat/macros] was used to calculate the weighted κ coefficients. BI-RADS assessment categories were ordered according to increasing positive predictive value [16]. Interpretation of the κ values was based on the Landis and Koch guidelines [17]: poor agreement, $<$0.01; slight agreement, 0.01–0.20; fair agreement, 0.21–0.40; moderate agreement, 0.41–0.60; substantial agreement, 0.61–0.80; and almost perfect agreement, 0.81–1.00.

The association between the females' characteristics and the interobserver disagreement was analysed for the BI-RADS result. Disagreement was considered when one radiologist assigned Category I or II and the other assigned Category III, 0, IV or V. Agreement was considered when both radiologists agreed to recall for further assessment (Categories III, 0, IV or V) or not (Categories I or II).

The present study was approved by the local research ethics committee.

Results

Inter- and intra-observer concordance

Table 1 shows the distribution of pairs of results for all the combinations of radiologists and mammograms. The most frequent interobserver disagreement was between Categories I and II (7.8%), followed by pairs of Categories 0 and I (6.7%), Categories III and I (6.0%), Categories III and IV (4.8%), and Categories III and II (3.9%). For intra-observer combinations, the most frequent disagreement was between Categories I and II (5.8%), followed by pairs of Categories 0 and I (5.3%), Categories III and 0 (4.3%), Categories III and I (3.7%) and Categories III and II (3.6%).

For breast density (Table 2), the highest discordant combinations were between adjacent categories in the following order: fatty and fibroglandular, fibroglandular and heterogeneously dense, and heterogeneously dense and extremely dense with 16.3%, 14.0% and 6.9%, respectively, of interobserver discordance and 10.8%, 9.2% and 4.0%, respectively, of intra-observer discordance.

Table 3 shows that intra-observer agreement was higher than interobserver agreement in all the items studied. The lowest agreement was found for BI-RADS assessment when the six categories were compared separately. The rest of the κ values indicated moderate interobserver agreement and substantial intra-observer agreement. Weighted κ increased the agreement indices, indicating that disagreement was more frequent between adjacent categories.

Table 4 shows the percentages of interobserver disagreement in recall according to the subjects' characteristics. We observed a higher percentage of disagreement between Categories II and III, especially in the cancer group. The highest discordance was found between Categories I or II and Category III for all age groups, for females after the menopause, for those with previous benign breast diseases and for those without a previous family history of breast cancer. The percentage of discordance was lower between Categories I or II and Categories IV or V in the screening result and for all female characteristics.

Discussion

In this study, we found a moderate degree of interobserver agreement in the use of the BI-RADS assessment and breast density categories. Intra-observer agreement was substantial. Although the most frequent

Table 1. Inter and intra-observer concordance and discordance for American College of Radiology Breast Imaging Reporting and Data System final assessment

Category	I	II	III	0	IV	V
Interobserver						
I	**7655 (36.62%)**	1623 (7.76%)	1251 (5.98%)	1403 (6.71%)	569 (2.72%)	26 (0.12%)
II		**616 (2.95%)**	812 (3.88%)	376 (1.80%)	311 (1.49%)	26 (0.12%)
III			**1922 (9.19%)**	690 (3.30%)	993 (4.75%)	50 (0.24%)
0				**357 (1.71%)**	629 (3.01%)	72 (0.34%)
IV					**959 (4.59%)**	420 (2.01%)
V						**143 (0.68%)**
Intra-observer						
I	**804 (40.30%)**	116 (5.81%)	74 (3.71%)	106 (5.31%)	24 (1.20%)	1 (0.05%)
II		**108 (5.41%)**	72 (3.61%)	32 (1.60%)	14 (0.70%)	1 (0.05%)
III			**195 (9.77%)**	86 (4.31%)	47 (2.36%)	4 (0.20%)
0				**78 (3.91%)**	56 (2.81%)	2 (0.10%)
IV					**125 (6.27%)**	29 (1.45%)
V						**21 (1.05%)**

Concordant data are in bold.

Table 2. Inter- and intra-observer concordance and discordance for breast density

Category	Fatty	Fibroglandular	Heterogeneously dense	Extremely dense
Interobserver				
Fatty	**3675 (17.76%)**	3376 (16.31%)	117 (0.57%)	1 (0.00%)
Fibroglandular		**5342 (25.81%)**	2894 (13.98%)	168 (0.81%)
Heterogeneously dense			**2850 (13.77%)**	1431 (6.92%)
Extremely dense				**840 (4.06%)**
Intra-observer				
Fatty	**440 (22.29%)**	213 (10.79%)	15 (0.76%)	1 (0.05%)
Fibroglandular		**601 (30.45%)**	181 (9.17%)	8 (0.41%)
Heterogeneously dense			**328 (16.62%)**	78 (3.95%)
Extremely dense				**109 (5.52%)**

Concordant data are in bold.

Table 3. Inter- and intra-observer variability in final assessment and breast density

	Interobserver variability			Intra-observer variability		
Category	% agreement	κ[a]	95% CI	% agreement	κ[a]	95% CI
BI-RADS assessment categories						
6 categories	55.74	0.37	(0.36–0.38)	66.72	0.53	(0.50–0.55)
Weighted	92.16	0.58	(0.56–0.59)	94.95	0.72	(0.69–0.75)
Recall *vs* no recall	77.16	0.53	(0.52–0.54)	83.76	0.66	(0.63–0.70)
Density[b]	61.40	0.44	(0.43–0.45)	74.87	0.64	(0.61–0.67)
Weighted	95.25	0.73	(0.72–0.74)	96.77	0.82	(0.80–0.84)

BI-RADS, American College of Radiology Breast Imaging Reporting and Data System; CI, confidence interval.

[a] κ-values: poor agreement, <0.01; slight agreement, 0.01–0.20; fair agreement, 0.21–0.40; moderate agreement, 0.41–0.60; substantial agreement, 0.61–0.80; and almost perfect agreement, 0.81–1.00.

[b] Four categories: fatty, fibroglandular, heterogeneous and extremely dense.

Table 4. Interobserver discordance of American College of Radiology Breast Imaging Reporting and Data System according to screening results and subjects' characteristics

		Discordance						
		Categories I or II *vs* Category III ($n = 2063$)		Categories I or II *vs* Category 0 ($n = 1779$)		Categories I or II *vs* Categories IV or V ($n = 932$)		
Characteristic	Total number of combinations	*n*[a]	%[b]	*n*[a]	%[b]	*n*[a]	%[b]	Concordance on recall % (95% CI)
Final result								
Cancer (*n*=13)	2730	405	14.8	134	4.9	88	3.2	77.1 (75.4–78.6)
True negative (*n*=51)	10 710	1045	9.7	975	9.1	584	5.4	75.8 (74.8–76.5)
False positive (*n*=36)	7560	613	8.1	670	8.8	260	3.4	79.7 (78.7–80.0)
Age group								
50–54 years (*n*=32)	6720	425	6.3	554	8.2	187	2.8	82.7 (81.7–83.6)
55–59 years (*n*=28)	5880	751	12.7	581	9.9	300	5.1	72.3 (71.1–73.4)
60–64 years (*n*=40)	8400	887	10.5	644	7.7	445	5.3	76.5 (75.6–77.4)
Menopause status								
Yes (*n*=80)	16 800	1595	9.5	1492	8.9	795	4.7	76.9 (76.2–77.5)
No (*n*=19)	3990	358	9.0	287	7.2	137	3.4	80.4 (79.2–81.6)
Personal history of benign breast disease								
Yes (*n*=15)	3150	461	14.6	305	9.7	180	5.7	70.0 (68.3–71.6)
No (*n*=85)	17 850	1602	9.0	1474	8.3	752	4.2	78.5 (77.9–79.1)
Family history of breast cancer								
Yes (*n*=9)	1890	126	6.7	105	5.6	12	0.6	87.1 (85.6–88.7)
No (*n*=91)	19 110	1937	10.1	1674	8.7	920	4.8	76.4 (75.7–76.9)

CI, confidence interval.

[a] Number of discordant combinations.

[b] Number of discordant combinations/total number of combinations.

disagreement in BI-RADS assessment was found between Categories I or II and Category III, a non-negligible discordance between Categories I or II and Categories IV or V was found, especially in cases of cancer.

Previous studies also reported fair to moderate variability in interobserver mammographic interpretation [9, 14, 18, 19]. However, it should be borne in mind that these studies were performed in different settings, they sometimes included diagnostic and screening mammograms, the breast disease status varied greatly, only some studies used BI-RADS classification, and they presented variations in the sample sizes and the experience of participating radiologists. The current analysis was conducted on a subsample of an earlier study [5]; however, we attempted to balance a high number of experienced radiologists with a sufficient number of mammograms and a period between readings long enough to avoid memory bias.

Weighted κ for both BI-RADS assessment and density indicates that disagreement between categories has a gradient, that is, there is more agreement between adjacent categories than between distant categories. This may be positively interpreted for breast density, but disagreement between adjacent categories of BI-RADS assessment may be very relevant. BI-RADS assessment categories were ordered according to their positive predictive value but, for example, disagreement between Categories II and III means that one of the radiologists has detected a benign lesion and finds no reason to recall and the other has found a probably benign lesion and recommends further assessment. In fact, the information that females get, which is what modifies the next step, is recall for further assessment.

When analysing concordance between no recall (categories I and II) and recall we found that, although they are the lowest, the frequencies of disagreement between no recall and Categories IV and V (suspicious of malignancy) are non-negligible. This is especially important in the group of 13 females with cancer: of 2730 possible pairs of radiologists, 88 pairs would have disagreed on recall. This clearly supports the need for double reading in population-based screening programmes. Other studies have shown that the cancer detection rate increased when double reading of screening mammograms was used [20, 21].

Regarding the subjects' characteristics, which were not available to the reading radiologists, slight differences were found: the most remarkable was a higher discordance between Categories I or II and Category III among females with a personal history of benign breast disease. These results indicate a higher percentage of agreement on recall for younger females (50–54 years) and those not yet experiencing menopause. However, it is important to mention that this analysis is exploratory and further research including larger samples sizes and taking into account confounding variables should be considered.

Regarding mammographic breast density, in our study we obtained a moderate agreement using unweighted κ (0.44) and substantial agreement using weighted κ (0.73). Our findings are similar to those in previous publications. Three previous studies observed moderate interobserver agreement using the BI-RADS breast density categories [9, 11, 14]. Berg et al [9] found moderate agreement (κ=0.43) in a study including 103 screening mammograms reviewed by 5 experienced radiologists who were not specifically trained in BI-RADS assessment. In addition, Ciatto et al [11] observed a κ value of 0.54 using the BI-RADS breast density, among 12 breast radiologists reading 100 mammograms. However, a more recent study [13], using 57 mammograms read by 4 experienced breast radiologists, reported substantial agreement (κ=0.77, 95% CI 0.69–0.85). Finally, Kerlikowske et al [14] reported moderate agreement and they also found that there was more variability in film interpretation among females with more dense breasts. This is an important point if we take into account that mammographic breast density has been consistently associated with breast cancer risk. A recent meta-analysis has shown that females with >75% breast density have more than four times greater risk of breast cancer than females with <5% [22].

To our knowledge, this is the first study performed in our setting investigating observer concordance of the BI-RADS assessment assigned to screening mammograms. The number of radiologists participating in the study and the number of mammograms were high. However, the present study has limitations. First, the observed agreement might be different from the overall agreement within a population-based breast cancer screening programme as breast cancer cases are overrepresented. Second, as we did not measure the agreement before and after the BI-RADS instruction session, we do not know whether it had any effect on agreement. Furthermore, we do not know whether agreement could be affected by the lack of information on the subjects' characteristics or access to prior mammograms, which readers may have had in the context of a population-based screening programme. Finally, it would have been interesting to know the agreement among radiologists taking into account the different types of lesions and breast density. Unfortunately, because of the design of our study, it was not possible to perform this analysis.

In conclusion, our study shows a substantial intraobserver level of agreement for the BI-RADS assessment but only moderate interobserver agreement. Both inter- and intra-observer agreement in mammographic interpretation of breast density was substantial. Double reading should be recommended in the context of population-based screening programmes while radiologists' variability in the BI-RADS assessment interpretation is not reduced.

Acknowledgments

We would like to thank all radiologists who gave their time to make this study possible: J Albanell, E Alert, JM Bantulá, M Castilla, JS Civil, A Cordon, R Dinerstein, JA Escobar, B Fábregas, M Fonoll, F García, A Mariscal, J Martinez, JL Munuera, C Nadal, A Palacio, J Planas, L Prieto, L Ramos, M Reñe, J Royo, ML Sánchez, A Solano, C Solduga, JM Teixidó, F Tous, V Villalba and I Vollmer. We also wish to acknowledge Montse Rué and Ruth Carlos, for their valuable assistance in this project.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61:69–90.
2. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. GLOBOCAN 2008, cancer incidence and mortality worldwide. IARC CancerBase No. 10. Lyon, France: IARC; 2010. Available from: http://globocan.iarc.fr
3. Smith-Bindman R, Ballard-Barbash R, Miglioretti DL, Patnick J, Kerlikowske K. Comparing the performance of mammography screening in the USA and the UK. J Med Screen 2005;12:50–4.
4. Yankaskas BC, Klabunde CN, Ancelle-Park R, Renner G, Wang H, Fracheboud J, Pou G, et al. International Breast Cancer Screening Network. International comparison of performance measures for screening mammography: can it be done? J Med Screen 2004;11:187–93.
5. Molins E, Macià F, Ferrer F, Maristany MT, Castells X. Association between radiologists' experience and accuracy in interpreting screening mammograms. BMC Health Serv Res 2008;8:91.
6. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. Breast 2001;10:455–63.
7. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS®). 3rd edn. Reston, VA: American College of Radiology; 1998.
8. American College of Radiology. Illustrated Breast Imaging Reporting and Data System (BI-RADS®). 3rd edn. Reston, VA: American College of Radiology; 1998.
9. 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.
10. Lehman C, Holt S, Peacock S, White E, Urban N. Use of the American College of Radiology BI-RADS guidelines by community radiologists: concordance of assessments and recommendations assigned to screening mammograms. AJR Am J Roentgenol 2002;179:15–20.
11. 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.

12. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, et al. Reader variability in reporting breast imaging according to BI-RADS assessment categories (the Florence experience). Breast 2006;15:44–51.
13. Ooms EA, Zonderland HM, Eijkemans MJ, Kriege M, Mahdavian Delavary B, Burger CW, et al. Mammography: interobserver variability in breast density assessment. Breast 2007;16:568–76.
14. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998;90:1801–9.
15. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability. Educ Psychol Meas 1973;33:613–19.
16. Orel SG, Kay N, Reynolds C, Sullivan DC. BI-RADS categorization as a predictor of malignancy. Radiology 1999;211:845–50.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
18. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–9.
19. 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;239:385–91.
20. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol 2003;180:1461–7.
21. Perry NM, Broeders M, de Wolf C. European guidelines for quality. Assurance in the mammography screening. Luxembourg: European Commission—Europe Against Cancer; 2002.
22. 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.