February 2016

To asses inter- and intra-observer variability for breast density and BIRADS assessment categories in mammographic reporting

Imrana Masroor
Aga Khan University, imrana.masroor@aku.edu

Mahreen Rasoor
Aga Khan University

Shaista Afzal Saeed
Aga Khan University

Saba Sohail
Aga Khan University, saba.sohail@aku.edu

ORIGINAL ARTICLE

To asses inter- and intra-observer variability for breast density and BIRADS assessment categories in mammographic reporting

Imrana Masroor, Mahreen Rasool, Shaista Afzal Saeed, Saba Sohail

Abstract

Objective: To evaluate the inter- and intra-observer variability among radiologists in the characterisation of mammograms according to Breast Imaging Reporting and Data System assessment and breast density categories.

Methods: The descriptive cross-sectional study was conducted at Aga Khan University Hospital, Karachi, from January 2014 to June 2014. Using non-probability purposive sampling, all mammograms in the study were interpreted by three radiologists on the basis of Breast Imaging Reporting and Data System categories and by assessing the breast density composition. The inter-observer variability was recorded by comparing the difference in the interpretation and categorisation of each case. Intra-observer variability was noted by comparing the differences in the two sets of results from reading the same mammogram three months apart.

Results: A total of 254 mammograms were reviewed and the mean age of patients was 55.2±11.6 years. In the first round of diagnostic imaging, there was moderate agreement among all three possible pairs of observers regarding breast density (k= 0.50-0.41), but for Breast Imaging Reporting and Data System categories the agreement was less (k=0.27-0.13). After 3 months, variability of observer 1 showed substantial agreement (k=0.65).Variability between observer 2 and observer 3 showed moderate agreement (k=0.13).In terms of categories, intra-observer differences were variable: observer 1 (?=0.61; observer 2(?=0.17); observer 3 (k=0.45).

Conclusion: Despite standardised guidelines for reporting density and assessment categories, observer variability continues to exist.

Keywords: BIRADS, Breast, Density, Mammography, Inter-observer variability, Intra-observer variability. (JPMA 66: 194; 2016)

Introduction

Breast cancer is the most common malignant tumour and a cause of death due to cancer in women in Pakistan[1] apart from being the leading cause of worldwide cancer deaths in women. Breast cancer accounted for 28% of total new cancer cases and 15% of total cancer-related deaths in 2010.[2] It is the second most common cancer in the United States, second only to lung cancer as the leading cause of death.[3]

The reported incidence of breast carcinoma in Karachi is 32%,[4] therefore early diagnosis and treatment can reduce the high mortality rate associated with carcinoma.

To reduce breast cancer mortality, the only evidence-based procedure available for screening is mammography.[5] For standardised reporting of the mammographic findings Breast Imaging Reporting and Data System (BIRADS) is routinely used.[6] However, a varying degree of agreement is seen among the radiologists for the assessment of breast density.[7,8] Breast composition can be measured by area-based method in which the image of compressed and projected breast are taken and breast density is assessed by segmenting areas of the mammographic image.[9] Because of the subjective nature of analysis, there is substantial inter-and intra-observer variability and it requires additional time even by an experienced radiologist.[10]

Accuracy of mammography depends upon various factors such as protocols for mammographic reading, characteristics of the patient, type of breast and the expertise of the radiologist.[11,12] Efforts have been made to improve its accuracy; one of them is double reading as it increases the cancer detection rate and reduces further assessment rate.[13]

BIRADS has been developed by the American College of Radiology (ACR) in order to classify the breast density,[12] to reduce discordance in the interpretation of mammographic findings, to standardise the reporting and to facilitate the follow-up.[14,15]

Mammographic density is considered to be an important independent risk factor for breast cancer.[16-19] It is important for breast cancer risk prediction as well. Therefore, the reproducibility of breast density is extremely important.[20]

Department of Radiology, Aga Khan University Hospital, Karachi.
Correspondence: Imrana Masroor. Email: imrana.masroor@aku.edu

Limited studies have analysed observer variability in mammographic interpretation using BIRADS assessment as well as breast density categories.[21-23] These studies done on Western population have shown insufficient inter- and intra-observer variability. In our country more and more centres are adopting BIRADS for mammographic reporting, but no such study has been conducted in Pakistan to the best of our knowledge.

The current study was planned to evaluate the inter- and intra-observer variability among radiologists in assessing breast density and BIRADS characterisation.

Subjects and Methods

The descriptive cross-sectional study was conducted at Aga Khan University Hospital, Karachi, from January 2014 to June 2014, after obtaining exemption from the institutional ethical review committee. Using Open EPI calculator, the sample size was calculated taking breast cancer prevalence of 18%[24] with 95% confidence interval (CI).

Non-probability purposive sampling was done while including mammographic images of female patients referred to the Radiology Department.

Women falling in the age range of 35-65 years were included in the study. All patients had no prior history of breast cancer, or risk factors for breast cancer. Patients with previous history of mastectomy, chemotherapy or radiation therapy were excluded.

Two view mammograms were taken; medio-lateral oblique view, and cranio-caudal view (Siemen MAMMOMAT NOVA 3000). All the mammograms were of good quality and were checked by the mammography quality control supervisor. They were interpreted by three radiologists having 5 to 10 years' experience each in mammographic reporting and breast imaging and who were specifically involved in mammogram reading on a weekly basis. All the readers interpreted the mammograms in sets of 10 for each session, to avoid fatigue and overwork bias. Each session was of half-an-hour duration. The readers described each mammogram using BIRADS categories and by assessing the breast composition, in both, the first round of reading and in a similar session repeated three months later. All three radiologists were blinded to the identity of the patients, and the sequence of mammograms was changed for each reading session. The inter-observer variability was recorded by comparing the difference in the interpretation and categorisation of cases among the three specialists. The same method was repeated three months later with the same set of mammograms to obtain another set of results.

Intra-observer variability was noted by comparing the differences in two sets of results from the same readers three months apart, which was done for all the three specialists.

For the sake of statistical analysis all three specialists later were referred to as the following variables: A and D: Observer 1; B and E: Observer 2; C and F: Observer 3. A,B and C referred to the first round, and D,E and F referred to the second round of reading for radiologists 1,2 and 3 respectively. For all mammograms, breast density and final BIRADS category were recorded by all the three readers.

Data analysis was done using SPSS 20, and inter- and intra-radiologist agreement were assessed using percentage of concordance and Kappa statistic, Kappa coefficient and its 95% CI. Interpretation of Kappa values was done by using standard method of interpretation[25] (Table-1).

Results

As against the sample size requirement of 227, the study comprised all the 254 mammograms done during the period. The mean age of patients was 52.2±11.6 years.

In the first round of reading, the agreement between observer 1 and 2 was moderate (k=0.50; 68.5%), while the agreement between observers 1 and 3 (k=0.43; 63.8%) and observers 2 and 3 (k=0.44; 63.4%) was almost similar.

In terms of BIRADS category, the agreement was fair between observers 1 and2 (k=0.27l; 44.9%) and observers 1and3 (k=0.24; 48.0%). Poor agreement was recorded between observers 2 and3 (k=0.13; 28.0%) (Table-2).

After three months, the overall kappa value increased

Table-1: Standard Kappa Values.

Value of K	Strength of agreement
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

Table-2: Inter-observer variability (First Round).

DENSITY		BIRADS	
Between A and B	κ=0.50(n=63.77%)	Between A and B	κ=0.27 (n=58.66%)
Between A and C	κ=0.43(n=97.24%)	Between A and C	κ=0.24 (n=70.47%)
Between B and C	κ=0.41(n=84.64%)	Between B and C	κ=0.13(n= 40.15%)

BIRADS: Breast Imaging Reporting and Data System.

Table-3: Inter-observer variability (Second Round).

Density		BIRADS Category	
Between D and E	κ=0.55 (n=50.39%)	Between D and E	κ=0.34(n=68.50%)
Between D and F	κ=0.50(n=31.49%)	Between D and F	κ=0.18 (n=56.29%)
Between E and F	κ=0.53 (n=89.76%)	Between E and F	κ=0.42(n=81.49%)

BIRADS: Breast Imaging Reporting and Data System.

Table-4: Intra-observer variability.

Density		BIRADS Category	
Between A and D	κ=0.65(n=88.97%)	Between A and D	κ=0.61 (n=94.48%)
Between B and E	κ=0.59 (n=96.85%)	Between B and E	κ=0.17(n=43.30%)
Between C and F	κ=0.49 (n=95.66%)	Between C and F	κ=0.45(n=86.22%)

BIRADS: Breast Imaging Reporting and Data System.

among all the three pairs (k=0.50) compared to the value in the first round (k=0.45) (Table-3).

There was a detectable difference in the results of BIRADS categorisation from the previous attempt. There was fair agreement between observers 1 and 2 (k=0.34; l53.1%), slight agreement between observers 1 and3 (k=0.18l; 43.9%), and moderate agreement between observers 2 and3 (k=0.42; 69.2%).

After the specified time period, the variability of observer 1 showed substantial agreement with the highest kappa value achieved throughout any part of the study (k=0.65; 78.0%). Variability of observer 2 and 3 showed moderate agreement (k=0.59; l74.8%and k=0.49; 67.3%) (Table-4).

As for BIRADS category, the intra-observer differences were highly variable, with observer 1 attaining the substantial agreement (k=0.61; 73.2%), observer 2 having a drastic drop and attaining poor agreement (k=0.17; 31.9%), while observer 3 attained moderate agreement (k=0.45; 72.3%).

Discussion

In this study we used BIRADS categories to determine inter-observer and intra-observer variability for breast density and assessment for recommending management. Minimising the variability in mammographic film interpretation is a priority in order to maximise the accuracy of screening mammography which in turn affects breast cancer detection.

The inter-observer agreement for breast density was moderate among all three readers with higher kappa values for moderate agreement in the second round of film viewing. A remarkably better agreement was seen for intra-observer variability for breast density for observer 1 and moderate agreement was noted in intra-observer variability for breast density for observers 2 and 3.

The breast density interpretation results are almost similar to a study that took a stratified random sample of 100 mammograms, and obtained substantial intra-observer agreement for breast density interpretation.[5] In that study, however, only screening mammograms were taken and the second set of images was reviewed after a gap of six months.

In terms of BIRADS assessment categories, our study showed less agreement amongst readers with almost fair agreement in the first round with low Kappa values. This showed a detectable increase to fair inter-observer agreement, according to the Kappa values in the second round. Our results of BIRADS assessment categories are comparable with a study that reported kappa value of 0.37.[5]

These results are also comparable to a study in which lesion management or final BIRADS assessment categories showed highly variable results with final kappa values of 0.37.[8] However, that study also interpreted screening mammograms only. This particular study had evaluated mammographers who were not trained in BIRADS categorisation. The mammographers interpreted the mammograms instead of trained radiologists, and the interval between the first and second set of readings was two months.

The results of intra-observer variability are in substantial and moderate agreement for observers 1 and 3, comparable to the results of the study conducted by Redondo et al[5] (kappa=0.55)and Berg et al[8] (kappa=0.60). The intra-observer variability for observer 2 showed moderate agreement for breast density assessment, but a drastic drop in BIRADS category assessment with kappa value of 0.17, thus showing slight agreement with previous results. On further reviewing of data, it was noted that observer 2 had given a more definitive category to all mammograms. This was attributed to the fact that observer 2 had attended a hands-on workshop on BIRADS reporting during the 3-month period, and thus became more confident in assigning final BIRADS assessment categories. This may also suggest that training in BIRADS lexicon has an effect on the confidence level of the radiologist.

Overall, the experience and training in BIRADS reporting do affect reporting skills and in providing feature analysis and final assessment categories using BIRADS.[26] In addition feedback on reporting from breast physician and patients' follow-up findings may inadvertently and unintentionally change the observer's interpretation of BIRADS category of a particular mammographic appearance. These factors may have affected the

reporting format and skills of radiologists in the second round and thus may have influenced the results of our study.

There are a few limitation of our study which may have affected the results. We took both screening and diagnostic mammograms. Secondly the time period between the first and second round was three months. This short period may have had an element of recall bias from the radiologists themselves which may in turn have affected the inter- and intra-observer variability values.

Conclusion

Despite standardised guidelines for reporting density and assessment categories, personal observer variability continues to exist, and continued training in BIRADS lexicon and assessment categories may result in increase in the confidence level of radiologists and result in decreased variability in image interpretation.

References

1. Rasheed R.Breast Cancer.J Coll Physicians Surg Pak. 2013; 23:766-7.
2. Jemal A, Siegel R, Xu J, Ward E. Cancer Statistics, 2010. CA Cancer J Clin. 2010; 60:277-300.
3. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global Cancer Statistics. CA Cancer J Clin. 2011; 61:69-90.
4. Bhurgari Y, Bhurgari A, Hasan S H, Usman A, Faridi N, Malik J et al. Cancer pattern in Karachi Division. J Pak Med Assoc. 2002; 52:244-6.
5. Redondo A, Comas M, Macià F, Ferrer F, Murta-Nascimento C, MaristanyMT et al. Inter- and intra-radiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br J Radiol. 2012; 85:1465-70.
6. Timmers JMH, Van Doorne-Nagtegaal HJ, Zonderland HM, Van Tinteren H, Visser O, Verbeek ALM, et al. Breast Imaging Reporting and Data System (BIRADS) in Dutch breast cancer screening programme: its role as an assessment and stratification tool. European Radiology. 2012; 22:1717-23.
7. Martin KE, Helvie MA, Zhou C, Roubidoux MA, Baily JE, Paramagul C. et al. Mammographic density measured with quantitative computer aided method: comparison with radiologists' estimates and BI-RADS categories. Radiology. 2006;240:656-65.
8. Berg WA, Compassi 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.
9. Sivaramakrishna R, Obuchowski NA, Chilcote WA, Powell KA. Automatic segmentation of mammographic density. Acad Radiol. 2001; 8:250-6.
10. Prevrhal S, Shepherd JA, Smith-Bindman R, Cummings SR, Kerlikowske K. Accuracy of mammographic breast density analysis: results of formal operator training. Cancer Epidemiol Biomarkers Prev. 2002;11:1389-93.
11. Yankaskas BC, Klabunde CN, Ancelle Park R, Renner G, Wang H, Fracheboud J, 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.
12. Molins E, MaciaF, FerrerF, Maristany MT, Castells X. Association between radiologists' experience and accuracy in interpreting screening mammogram. BMC Health Serv Res. 2008; 8:91.
13. 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 systemic review. Breast. 2001; 10:455-63.
14. Sujata VG, Jay AB,Connie EK, Karen S, Johnson, Ruth W, etal. Using BIRADS Lexicon in a Restrictive Form of Double Reading as a Strategy for Minimizing Screening Mammography Recall Rates.AJR Am J Roentgenol 2012; 198:962-70
15. Brisson J, Diorio C, Masse B. Wolfe's parenchymal pattern and percentage of the breast with mammographic densities: redundant or complimentary classification? Cancer Epidemiol Biomarkers Prev. 2003;12:728-32.
16. Vanchon CM, Kuni CC, Anderson K, Anderson VE, Sellers TA. Association of mammographically defined percent breast density with epidemiologic risk factors for breast cancer (United States). Cancer Causes Control. 2000; 11:653-62.
17. Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR, et al. Heritability of mammographic breast density, a risk factor for breast cancer. N Engl J Med 2002;347:886-94.
18. Chen Z, Wu AH, Gauderman WJ, Bernstein L, Ma H, Pike MC, et al.Does mammographic densities reflect ethnic differences in breast cancer incidence rates? Am J Epidemiol. 2004; 159:140-7.
19. Boyd NF, Guo H, Martin LJ, Sun L,StoneJ,Fishell E et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007;356:227-36.
20. Lehman C, Holt S, Peacock S, White E, Urban N. Use of 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.
21. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, et al. Categorizing breast mammographic density: intra- and interobserver variability of BI-RADS density categories. Breast. 2005;14:269-75.
22. 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.
23. Ooms EA, Zonderland HM, Eijkemans MJ, Kriege M, MahdavianD. B, Burger CW, et al. Mammography: interobserver variability in breast density assessment. Breast. 2007;16 :568-76.
24. Mcpheron K, Steel CM,Dixon JM. ABC of breast diseases. Breast Cancer epidemiology,risk factors and genetics. BMJ. 2000; 321:624-28.
25. Landis JR, Koch GG.The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159-74.
26. Timmers JM, van Doorne-Nagtegaal HJ, Verbeek AL, den Heeten GJ, Broeders MJ. A dedicated BIRADS training programme: Effect on the inter-observer variation among screening radiologist. Eur J Radiol. 2012; 81:2184-8.