IJSRM
INTERNATIONAL JOURNAL OF SCIENCE AND RESEARCH METHODOLOGY
An Official Publication of Human Journals

Human Journals
Research Article
April 2021 Vol.:18, Issue:2

Performance, Inter-Observer and Intra-Observer Variability of Radiologists Versus Residents Using Soft Copy Reading Mammography

IJSRM
INTERNATIONAL JOURNAL OF SCIENCE AND RESEARCH METHODOLOGY
An Official Publication of Human Journals

Khaled A. H Fahid*[1], Nik Munirah Nik Mahdi[2], Meher Angez Rahman[1], Mizanul Hasan[3]

[1]Department of Radiology and Imaging, Suri Seri Begawan Hospital, Brunei.

[2]Department of Radiology, Hospital University Sains Malaysia, Malaysia. [3]Popular Diagnostic Centre Limited, Dhaka, Bangladesh.

Submitted: 23 March 2021
Accepted: 31 March 2021
Published: 30 April 2021

Keywords: Mammography, Inter, and Intra-observer Variability

ABSTRACT

Background: Mammography is a well-validated screening tool to detect breast cancer globally, and it has been proven to reduce the mortality rate associated with breast cancer, typically by detecting breast cancer during its very early stage. **Aim:** To assess performance, inter-observer, and intra-observer variability of radiologists and 4th-year radiology residents in reading soft copy mammograms. **Methods and Materials:** This retrospective study was carried out in Hospital University Sains Malaysia, for two years from January 2010 to December 2011. A total of 104 samples were obtained, which included BIRADS 2 and above. Mammograms were interpreted by four observers, two radiologists, and two residents, for the presence of any breast mass and calcification. Three weeks interval was given between a review of a total of 104 mammogram images and another reading of 24 randomly selected mammograms from a total of 104 mammograms. A 5-scale BIRADS category (BIRADS 1-5) was used to categorize the findings. Agreements were analyzed using Kappa analysis. **Result:** The interobserver variability using Kappa agreement for detection of breast cancer was in the range of moderate among the radiologists and between the radiologists and residents; however the agreement was fair among the residents. The variability was greater for the characterization of breast masses or calcifications. The Intraobserver variability was not significant for the readers in the detection of breast cancer except for resident 1, who had a fair agreement. **Conclusion**: We concluded that there was greater sensitivity and specificity in breast cancer detection of the specialist radiologist with less degree for the general radiologist and radiology residents. Also, there was interobserver variability, while there is no intraobserver variability for the detection of breast cancer among the observers.

INTRODUCTION

Breast cancer is the commonest cancer in women in most parts of the world. The incidence in Malaysia is lower than in the developed countries. The difference may be attributable to the difficulty in getting accurate statistics and to the under-reporting of cases. Mammography is a well-validated screening tool to detect breast cancer globally and has been proven to reduce the mortality rate associated with breast cancer, typically by detecting breast cancer during its very early stage. Inevitably, not all breast cancer will be detected using mammography. Conventional screen-film mammography (SFM) is proven to have a high sensitivity and specificity for the detection of breast cancer. The specificity ranged from 90% to 98% and the sensitivities ranged from 83% to 95%[1]. However, it is less sensitive for women under age 50 years, women with radiographically dense breasts, and premenopausal or perimenopausal women[2]. For improving the sensitivity and specificity of screening mammography in picking up clinically silent breast cancer, various researches and technologies have been utilized. The most recent advancement would be digital mammography. The discrepancy in assessments of reading mammograms by two radiologists is often being described as "interobserver variability" in literature. Interobserver variability could decrease the effectiveness of breast cancer detection and therefore, should be kept to the minimum by using some techniques like computer assisted detection (CAD), or by reading soft copy digital mammograms using special high resolution monitors or dedicated mammography workstation.

MATERIALS AND METHODS:

This was a retrospective cross-sectional study, carried out in Hospital University Sains Malaysia (HUSM), for two years from January 2010 to December 2011. By using HUSM Radiology Department's picture archives and communication system (PACS), a total of 104 samples were obtained, which included BIRADS 2 and above. The diagnostic viewing workstation monitor was Kodak with an optimum resolution of 2048 x1536. Mammograms in the workstation were evaluated by readers independently, who were blinded to the patient's information and clinical history, as well as the diagnosis. The findings recorded by the readers using standardized forms. The readers were two radiologists; one breast radiologist with 10 years experience in breast imaging (Radiologist 1), and one general radiologist with two years experience in mammogram interpretation (Radiologist 2), and two 4th year radiology residents. A 5-scale BIRADS category

(BIRADS 1-5) was used to categorize the findings. Agreements were analyzed using Kappa analysis. Observers evaluated the mammograms for the presence of calcification and breast mass, according to the data collection form. We characterized the location, shape, and margin of the mass. Regarding the calcification, the size, morphology, and distribution of calcifications were seen. Associated findings like architectural distortion, nipple retraction, axillary nodes, and skin retraction were also noted. If there was more than one lesion seen in the mammogram, the area deemed the highest possibility of breast cancer was marked and described in the form. Based on these findings, a BIRADS category was assigned. Recommendations for further assessment of the lesion were taken in form of performing a biopsy or just follow up by ultrasound or mammography. To assess intraobserver variability, all observers rereviewed 24 randomly selected cases, from the total sample size of 104 after three weeks had elapsed since the first interpretation, to avoid recall bias. The mammograms were evaluated for the same previous assessment. The findings were recorded using standardized forms. The outcomes were assessed by sensitivity, specificity, interobserver variability, and intraobserver variability.

RESULTS AND DISCUSSION:

RESULTS:

Based on data collected over 2 years (January 2010 to December 2011), a total of 889 digital mammographies were done in HUSM, A total of 104 patients were included in this study. The age ranged from 30 to 70 years with a mean age of 49.85 years. The study population consisted of Malays (86.54%), Chinese (11.54%), and others (1.92%). The 104 samples had breast mass or calcification, and some samples were having both mass and calcifications. Based on official histopathological examination (HPE) reports, breast samples were reported as benign breast lesions 58, whereas 46 samples were reported as malignant. Nine samples relied on the official mammogram reports because of absences of histopathological reports, all those cases were reported as benign on the radiological reports. From a total of 104 samples, Radiologist 1 had detected 58 benign lesions on mammograms, while 46 were malignant. Radiologist 2 was able to detect 41 benign breast lesions, and 63 were malignant. For Resident 1, the benign breast lesions were detected in 56, while 48 were malignant. Resident 2 was able to detect 41 benign breast lesions, while 63 were malignant (Table 1).

Table No. 1: Frequency of breast lesions detection by the Radiologists and Residents

	Benign	Malignant
Radiologist 1	58	46
Radiologist 2	41	63
Resident 1	56	48
Resident 2	41	63
HPE	58	46

Sensitivity and specificity: Sensitivity for detection of breast cancer was 95.7% for Radiologist 1, while the specificity measured 96.6%. While the values of false positive and false negative rates were 3.44%, 4.35% respectively. For Radiologist 2, sensitivity and specificity were 80.4% and 55.2% respectively. While the values of false positive and false negative rates were 44.8%, 19.5% respectively. Resident 1 had a sensitivity of 73.9%, and a specificity of 75.9%. While the values of false positive and false negative rates were 24.1%, 26.1% respectively. For Resident 2 sensitivity measured 80.4% and specificity 55.2% for detection of breast cancer. The values of false positive and false negative rates were 44.8%, 19.5% respectively (Table 2).

Table No. 2: Sensitivity and specificity of breast cancer detection by the Radiologists and Residents

	Sensitivity	**Specificity**	**FPR**	**FNR**
Radiologist 1	95.7%	96.6%	3.44%	4.35%
Radiologist 2	80.4%	55.2%	44.8%	19.5%
Resident 1	73.9%	75.9%	24.1%	26.1%
Resident 2	80.4%	55.2%	44.8%	19.5%

FPR = False Positive Rate FNR = False Negative Rate

Intra-observer variability: For Intra-observer agreement to detect breast cancer by reading soft copy mammograms, after three weeks from the first reading of 104 mammograms, the second session of reading 24 mammograms, Radiologist 1 was able to detect breast cancer with 1.000 Kappa agreement between first and second readings of soft copy mammograms. Radiologist 2 had 0.647 Kappa agreements, while Resident 1 measured 0.179 of Kappa agreement, and Resident 2 had 0.684 Kappa agreements (Table 3).

Table No. 3: Summary for inter-observer and intra-observer Kappa agreement of Radiologists and Residents for detection of breast cancer

	Radiologist 1	Radiologist 2	Resident 1	Resident 2
Radiologist 1	**1.000**	0.418	0.534	0.418
Radiologist 2	0.418	**0.647**	0.413	0.919
Resident 1	0.534	0.413	**0.179**	0.338
Resident 2	0.418	0.919	0.338	**0.684**

Inter-observer variability: Radiologists had a moderate Kappa agreement of 0.418 compared with residents who had a fair agreement of 0.338 for detection of breast cancer (Table 4).

Table No. 4: Comparison of Kappa agreement for cancer detection between Radiologists and Residents

Parameter	Kappa for Radiologists	Kappa for Residents
	0.418	0.338

Kappa agreement for mammographic findings ranging between 0.166 to 0.699 for Radiologists, however for Residents ranged between 0.111 to 0.492, both were in fair to a moderate agreement. The mass characterization agreement of mass location was substantial measuring 0.699, the mass shape was fair 0.313, and was moderate 0.443 for mass margin for Radiologists. On the other hand, for the Residents the mass characterization agreement of mass location was moderate measuring 0.492, the mass shape was fair 0.318 and the mass margin was fair 0.297 respectively. Radiologists had better agreements for calcification size which was 0.333, while both calcification distribution and calcification morphology were in fair agreement measuring

0.338 and 0.304 respectively. Residents had slight agreements for calcification size and calcification distribution measuring 0.128 and 0.151, while it was a fair agreement for calcification morphology measuring 0.234. However, the agreement for the associated findings among the Residents was found to be higher than that among Radiologists measuring 0.225 and 0.166 respectively. On the other hand for BIRADS classification, the agreement measured 0.019 for Radiologists and 0.069 for Residents. The agreement regarding recommendation was fair 0.289 for Radiologists and slight 0.161 for Residents (Table 5).

Table No. 5: Comparison of mammographic findings Kappa agreement between Radiologists and Residents

Parameter	Kappa for Radiologists	Kappa for Residents
Calcification Size	0.333	0.128
Calcification Distribution	0.338	0.151
Calcification Morphology	0.304	0.234
Mass Location	0.699	0.492
Mass Shape	0.313	0.318
Mass Margin	0.443	0.279
Associated Findings	0.166	0.225
Calcification within mass	0.420	0.117
Recommendations	0.289	0.161
Conclusion BIRADS	0.019	0.064
Diagnosis	0.418	0.338

DISCUSSION:

The difference noted in the detection of breast cancer between radiologists could be attributed to the difference in years of interpreting mammograms. A study was done by Sickles [3] evaluating the performance of specialist radiologists and general radiologists in reading mammograms on a total of 61.084 screening and diagnostic mammogram images. They found that specialist radiologists detect more cancers and more early-stage cancers than general radiologists, and they recommend more biopsies than general radiologists. The specialist radiologists interpret 10 times more mammographic studies per year than the general radiologists. Study done by Nascimento [4]

showed that the mammography sensitivity ranged from 68% to 87% between the observers for identification of malignant and benign breast lesions. In a study done by Norsa'adah et al [5] on 328 patients, only 136 (41.5%) of patients had a mammogram, and 8.8% of those were false negative, which contributed to the delayed diagnosis of breast cancer in this study. This rate was higher than the previous study done by Goodson[6], which was 7%. In our study, radiologist 1 had a good false-negative rate of 4.35%, while both, radiologist 2 and resident 2 were at 19.5%, which slightly higher than the acceptable rate. The acceptable false-negative rate of the mammogram was 10-15% according to Huynh et al [7].Mammographically missed cancers could be attributed to interpretation error (52%), observer error (30-43%), which could include overlooked heavy caseload or eye fatigue. A technical error in (5%), and tumor biology in form of failure to incite desmoplastic reaction [8].

Inter-observer variability between Radiologists: We found that the Kappa agreement between the radiologists in detecting breast cancer was moderate in value of 0.418. Less agreement was seen in the earlier work of Boyd et al [9], in which nine radiologists reviewed 100 mammograms. Between pairs of radiologists, kappa values ranged from 0.17 to 0.55 for diagnostic assessment. However similar moderate agreement to our study was concluded by Skaane et al [10], mammograms of 100 benign breast masses and 100 malignant ones in 200 patients were retrospectively analyzed by 4 radiologists. The overall kappa value was moderate 0.58 for mammography and concluded that radiologists differ substantially in their interpretations of mammograms. Another study had the same agreement done by Kerlikowske et al [11] and reported 0.58 agreement in the final assessments of two observers in the evaluation of 2616 mammograms. However, Kerlikowske had fair to a moderate agreement for a description of feature analysis, similarly to our study. This was explained by Carney et al [12] in a study that showed interobserver variability between radiologists in the interpretation of screening mammograms depended on the protocols for mammogram reading, and to large extent on the experience of the radiologists. Several differences in the experience and expertise of the specialist and general radiologists may explain the differences in their interpretation. The specialist radiologists have considerably more training and education in mammography. The general radiologists received the basic training in mammography during their diagnostic radiology training and have received minimum amount of continuing education in mammography interpretation. The variability that was seen in our study in describing

characteristics of masses and calcifications which could be explained that the radiologists were different in their understanding of the definitions of BI-RADS terms or because the terms provided did not adequately describe the lesion, making the choice of descriptor difficult. Management or recommendation variability is attributable to variation in intervention threshold for biopsy workup [13].

Inter-observer variability between Radiologists and Residents: The interobserver variability seen in detection of breast cancer was higher for residents than the specialist radiologist and by less degree with the general radiologist, could be attributed to the experience and volume of mammograms interpretation of specialist radiologist, however the similar ability of general radiologist and the residents for detection of breast cancer related to the less training and low volume of mammograms interpretation by them. The greater interobserver variability in the description of breast lesions or calcification, which was noted in a study done by Singh et al [14], which included Images of 50 breast lesions were individually interpreted by seven dedicated breast imagers and 10 radiology residents, Lesions were described with the use of descriptors from the Breast Imaging Reporting and Data System, and interobserver variability was calculated with the Cohen k statistic. A significant difference was observed for lesion features description. However, the radiology residents had greater interobserver variability in their selection of the lesion features than did dedicated breast imagers. It must be noted that in our study both residents and breast imagers showed little agreement regarding recommendations, fair 0.289 for radiologists and slight 0.161 for residents, it is slightly better than the agreement shown by Singh et al [14] (0.09 for residents, 0.21 for breast imagers). In his study where images of 50 breast lesions were individually interpreted by seven dedicated breast imagers and 10 radiology residents. Our results showed that radiologists could differ in their mammographic interpretation and recommendations for management. These results should not be regarded as casting doubt on the efficacy of mammography, the value of which has been well documented [15]. Reduction in variability will require more consistent criteria for diagnostic interpretation and standers. for recommending subsequent evaluation.

Inter-observer variability between Residents: There are few studies done before to compare the interobserver variability between the residents for detection of breast cancer. The fair agreement of residents for detection of breast cancer and the minimal agreement for

characterization of breast calcification and masses were noted and we cannot explain these observed differences in performance among the residents based on either experience or expertise, because there were no substantial differences in the amounts of initial or continuing education in mammography. The difference in observers' agreement could be attributed to the difference in interest, the familiarity of the mammography workstation system, reading environment, and human factors such as fatigue or inattention. In a study done by Bassett et al [16] explained the reason given by radiology residents for not going into the field of breast imaging was that breast imaging was not an interesting field and limited in its application of advanced technology compared with other imaging subspecialties.

Intra-observer variability: Both radiologists were agreeing with themselves in the detection of breast cancer, Kappa agreement was almost perfect, with less degree in mammographic features description. Our result was better than the result of a study done by Kerlikowske et al [11] reported 0.73 intraobserver Kappa agreement in the evaluation of 2616 mammograms which were evaluated by two radiologists in the detection of breast cancer. However our result was similar in a study done by Ciatto et al[17], an intraobserver agreement was almost perfect ($\kappa = 0.81$) in reporting according to Breast Imaging Reporting and Data System (BI-RADS), breast cancer detection was tested in 12 dedicated breast radiologists reading a digitized set of 100 two-view mammograms. Resident 2 was better than Resident 1, who was slightly agreeable with himself in the detection of breast cancer. The low agreement of Resident 1 could be related to eye fatigue and stress prior to the final exams of graduation.

CONCLUSION:

We concluded that there was greater sensitivity and specificity in breast cancer detection of the specialist radiologist with less degree for the general radiologist and radiology residents. However there was interobserver variability, while there was no intraobserver variability for detection of breast cancer. The interobserver agreement using Kappa agreement for detection of breast cancer was in the range of moderate among the radiologists, between the radiologists and residents, and fair among the residents. The variability was greater for the characterization of breast masses or calcifications. The study found that a greater volume of experience at interpreting mammograms was associated with better performance, especially for the specialist radiologist. The intraobserver variability was not significant for the readers in the detection of

breast cancer except for Resident 1, who had a fair agreement. The greater intraobserver variability in the description of breast mass and calcification was observed, which could be improved by more training and interpreting mammograms.

To improve screening mammography performance for residents and radiologists we need either better educational tools to communicate BI-RADS terms to radiologists or the development of more effective criteria for reporting mammographic findings and selecting assessment categories. Also, continuous training is a factor for developing interpretation skills [18]. Dedicated mammography courses have been shown to improve mammographic interpretation [19]. Double reading mammographs technique is recommended. It has been shown that double reading can increase radiologist sensitivity by 5–15% [20].

REFERENCES:

1. Mushlin AI, Kouides RW, Shapiro DE. Estimating the accuracy of screening mammography: a meta-analysis. American journal of preventive medicine. 1998;14(2):143-153.
2. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Annals of Internal Medicine. 2003;138(3):168-175.
3. Sickles EA, Wolverton DE, Dee KE. Performance Parameters for Screening and Diagnostic Mammography: Specialist and General Radiologists1. Radiology. 2002;224(3):861-869.
4. Nascimento JHRd, Silva VDd, Maciel AC. Accuracy of mammographic findings in breast cancer: correlation between BI-RADS classification and histological findings. RadiologiaBrasileira. 2010;43(2):91-96.
5. Norsa'adah B, Rampal KG, Rahmah MA, et al. Diagnosis delay of breast cancer and its associated factors in Malaysian women. BMC cancer. 2011;11(1):141.
6. Goodson WH, Moore DH. Causes of physician delay in the diagnosis of breast cancer. Archives of internal medicine. 2002;162(12):1343-1348.
7. Huynh PT, Jarolimek AM, Daye S. The false-negative mammogram. Radiographics. 1998;18(5):1137-1154.
8. Dähnert W. Radiology review manual: Lippincott Williams & Wilkins 2011.
9. Boyd NF, Wolfson C, Moskowitz M, et al. Observer variation in the classification of mammographic parenchymal patterns. Journal of chronic diseases. 1986;39(6):465-472.
10. Skaane P, Engedal K, Skjennald A. Interobserver variation in the interpretation of breast imaging: comparison of mammography, ultrasonography, and both combined in the interpretation of palpable noncalcified breast masses. Acta Radiologica. 1997;38(4):497-502.
11. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. Journal of the National Cancer Institute. 1998;90(23):1801-1809.
12. Carney PA, Elmore JG, Abraham LA, et al. Radiologist uncertainty and the interpretation of screening. Medical decision making. 2004;24(3):255-264.
13. Kopans DB. Accuracy of mammographic interpretation. The New England journal of medicine. 1994;331(22):1521-1522.
14. Singh S, Maxwell J, Baker JA, et al. Computer-aided classification of breast masses: Performance and interobserver variability of expert radiologists versus residents. Radiology. 2011;258(1):73-80.

15. Of UTOED, Group BC. First results on mortality reduction in the UK trial of early detection of breast cancer. The Lancet. 1988;332(8608):411-416.
16. Bassett LW, Monsees BS, Smith RA, et al. Survey of Radiology Residents: Breast Imaging Training and Attitudes1. Radiology. 2003;227(3):862-869.
17. Ciatto S, Houssami N, Apruzzese A, et al. Categorizing breast mammographic density: intra-and interobserver reproducibility of BI-RADS density categories. The Breast. 2005;14(4):269-275.
18. Molins E, Macià F, Ferrer F, et al. Association between radiologists' experience and accuracy in interpreting screening mammograms. BMC health services research. 2008;8(1):91.
19. Linver M, Paster S, Rosenberg R, et al. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. Radiology. 1992;184(1):39-43.
20. Murphy Jr W, Destouet J, Monsees B. Professional quality assurance for mammography screening programs. Radiology. 1990;175(2):319-320.