Received:
13 April 2020

Revised:
20 July 2020

Accepted:
21 July 2020

<https://doi.org/10.1259/bjr.20200363>

Cite this article as:

Trieu PDung(Yun), Lewis SJ, Li T, Ho K, Tapia KA, Brennan PC. Reader characteristics and mammogram features associated with breast imaging reporting scores. *Br J Radiol* 2020; **93**: 20200363.

FULL PAPER

Reader characteristics and mammogram features associated with breast imaging reporting scores

PHUONG DUNG(YUN) TRIEU, PhD, SARAH J LEWIS, PhD, TONG LI, PhD, KAREN HO, MSc, KRISCIA A TAPIA, PhD and PATRICK C BRENNAN, PhD

Discipline of Medical Imaging Sciences, Faculty of Medicine and Health, The University of Sydney 75 East street, Lidcombe, New South Wales, Australia 2141

Address correspondence to: Dr Phuong Dung(Yun) Trieu
E-mail: phuong.trieu@sydney.edu.au

Objectives: This study aims to explore the reading performances of radiologists in detecting cancers on mammograms using Tabar Breast Imaging Reporting and Data System (BIRADS) classification and identify factors related to breast imaging reporting scores.

Methods: 117 readings of five different mammogram test sets with each set containing 20 cancer and 40 normal cases were performed by Australian radiologists. Each radiologist evaluated the mammograms using the BIRADS lexicon with category 1 - negative, category 2 - benign findings, category 3 - equivocal findings (Recall), category 4 - suspicious findings (Recall), and category 5 - highly suggestive of malignant findings (Recall). Performance metrics (true positive, false positive, true negative, and false negative) were calculated for each radiologist and the distribution of reporting categories was analyzed in reader-based and case-based groups. The association of reader characteristics and case features among categories was examined using Mann-Whitney U and Kruskal-Wallis tests.

Results: 38% of cancer-containing mammograms were reported with category 3 which decreased to 32.3% with category 4 and 16.2% with category 5 while 16.6 and 10.3% of cancer cases were marked with categories 1 and 2. Female readers had less false-negative rates when using categories 1 and 2 for cancer cases than male readers ($p < 0.01$). A similar pattern as gender

category was also found in Breast Screen readers and readers completed breast reading fellowships compared with non-Breast Screen and non-fellowship readers ($p < 0.05$). Radiologists with low number of cases read per week were more likely to record the cancer cases with category 4 while the ones with high number of cases were with category 3 ($p < 0.01$). Discrete mass and asymmetric density were the two types of abnormalities reported mostly as equivocal findings with category 3 (47–50%; $p = 0.005$) while spiculated mass or stellate lesions were mostly selected as highly suggestive of malignancy with category 5 (26%, $p = 0.001$).

Conclusions: Most radiologists used category 3 when reporting cancer mammograms. Gender, working for BreastScreen, fellowship completion, and number of cases read per week were factors associated with scoring selection. Radiologists reported higher Tabar BIRADS category for specific types of abnormalities on mammograms than others.

Advances in knowledge: The study identified factors associated with the decision of radiologists in assigning a BIRADS Tabar score for mammograms with abnormality. These findings will be useful for individual training programs to improve the confidence of radiologists in recognizing abnormal lesions on screening mammograms.

INTRODUCTION

Screening mammography plays a central role in the early detection of breast cancer since mammograms are able to show changes in the breast up to 2 years before a patient or physician can feel them.¹ Breast Imaging Reporting and Data System (BIRADS) is often used for mammography reporting in most countries which have a breast cancer screening program. This system is designed to standardize the interpretation of mammographic examinations by

providing well-defined assessment categories and standard recommendations for follow-up. In Australia, BIRADS encompasses a lexicon of descriptors based on American College of Radiologists recommendations using a 5-tier scoring scheme called Tabar: category 1, normal; category 2, benign; category 3, indeterminate/equivocal; category 4, suspicious and category 5, radiologically malignant² with BIRADS 3, 4 and 5 having a likelihood of <2%, 2% to 95%, and more than 95% of a malignancy, respectively.³

The difference between the ACR BIRADS (America) and Tabar BIRADS (Australia) is that ACR BIRADS 3 is suggested for the recommendation of 6-month follow-up while further examinations followed by a biopsy is often indicated for Tabar BIRADS 3.

Prior to the implementation of the BIRADS lexicon in late 1980s, there was minimal uniformity in how mammogram reports were recorded. The absence of uniformity often resulted in ambiguous reports, which led to difficulties for the referring healthcare providers to decide which management strategy was required for patients. Moreover, the degree of concern (*i.e.*, probably benign *vs* indeterminate *vs* highly suspicious) could not be determined on the basis of the report in the majority of cases.⁴ The key impetus for the implementation of the BIRADS lexicon is to decrease the confusion to relating mammography reports so that the findings and the recommendations will be made clearer. By recording a final assessment category on the basis of the mammographic findings, patients and healthcare providers will receive reports that consist of both a succinct interpretation of the mammographic findings and a comprehensible clinical management recommendation.

However, even with BIRADS, differences in mammographic interpretation among radiologists are still observed⁵ and may be attributed to a variety of lesion characteristics and reader experience. Studies have shown that higher levels of experience and training of the radiologist increased diagnostic efficacy^{5,6} while younger and more recently trained radiologists had higher false-positive rates.⁷ The number of mammograms read per year was also reported to affect performance directly,^{6,8} and image factors, such as the subtlety of lesions or breast density influences interpretation⁹ resulted in higher rates of false-positives and false-negatives.¹⁰ Nevertheless, these studies focused on overall diagnostic efficacy with little attention paid to predictive factors related to specifically allocated BIRADS categories. When allocating the categories, decision-making confidence is highly complex due to the interaction between visual perceptions and clinical judgment, both of which can be impacted by such factors as the experience of interpreter or image characteristics.

This study therefore aims to detail conditions associated with radiologists and cancer case features that could determine Tabar BIRADS category employed in mammogram interpretation among Australian radiologists. Understanding this utilization of categories will assist in creating individual tailored training programs to improve the confidence of radiologists in recognizing suspicious findings on mammograms.

METHODS AND MATERIALS

Ethical approval for this study was obtained from the University of Sydney ethics committee [2017/028] with informed consent obtained from each participant.

Case collection

BreastScreen REader Assessment STRategy (BREAST), the official training program for radiologists in Australia to monitor and assess the performance of BreastScreen readers,¹¹ released five mammogram test sets between 2014 and 2018 with each set

including 20 cancer and 40 normal cases. Each case consisted of two-view craniocaudal (CC) and mediolateral oblique (MLO) bilateral screening mammograms. All mammograms were collected from the image database of BreastScreen services and were de-identified with patient names removed. Malignant cases were biopsy proven while negative cases were assessed by the consensus reading of at least two BreastScreen radiologists following the negative screen reports detected 2 years after. Cases were selected in order to be challenging to participants and thus offer both self-assessment as well as training value. The case collection was conducted by an expert panel of two senior radiologists who had more than 20 years in reading mammograms and were responsible for assessing training, support, and quality assurance of BreastScreen services. Ground truth of case features were recorded based on the details in radiology and pathology reports and confirmed by the expert panel. The rates of cases with high dense breast (>50% mammographic density) and low dense breast (≤50% mammographic density) were relatively similar at 55 and 45%, respectively. The cancer lesions were equally distributed with 51% in the left breast and 52% in the right breast. The cancer types included calcifications (20%), discrete mass (13%), spiculated mass (39%), architectural distortion (7%), asymmetric density (17%), and mix of lesion types (11%). The percentage of large size lesion (≥10 mm) was slightly over the small size lesions (<10 mm) with 54 and 49%, respectively. Regarding the cancer true location, 47% of lesions were found in the upper outer quadrant (Table 1).

Participants

Participants were Australian radiologists who registered as BREAST readers. In total, there were 117 readings of five BREAST mammogram test sets performed by 85 radiologists between 2014 and 2018. Each radiologist could read as many test sets as they wished but only results from when radiologists completed each set for the first time were included in the data analysis. 62 radiologists completed one test set, 16 radiologists completed two test sets, five radiologists completed three test sets, and two radiologists completed four test sets. Test sets had equal level of difficulty. Radiologist experience in mammography interpretation was obtained from the online survey embedded in the test set. The average age of participants was 51 years, with an average of mammography reading experience of 15 years. Among 85 radiologists, 58 were female (54 BreastScreen and four non-BreastScreen readers) and 27 were male (21 BreastScreen and six non-BreastScreen readers). 80% of all readings were conducted by BreastScreen readers while 31% of them had completed by radiologists having a minimum of 3-month training in reporting mammograms. Details of the demographic background and professional experience are described in Table 2.

Reading environments

Participants completed test sets either at a workshop or at their clinical service workplaces. At the workshop, viewing conditions were simulated to the usual screen-reading environment, and all readings were performed using high specification workstations in a dark environment (ambient light ranging from 20 to 30 lux) adhering to the national recommendation for soft-copy interpretation of mammograms.¹² Participants were given

Table 1. Details of cancer cases in test sets

Category		Cancer cases
		N (%)
Breast density		
	<25%	6 (6%)
	25–50%	39 (39%)
	51–75%	51 (51%)
	>75%	4 (4%)
Type of cancer appearance on mammograms		
	Architectural distortion	7 (7%)
	Calcification	21 (20%)
	Discrete mass	13 (13%)
	Asymmetric density	17 (17%)
	Spiculated mass	41 (39%)
	Mix of types ^a	4 (11%)
Side-of cancer		
	Left	51 (50%)
	Right	52 (50%)
Site of cancer		
	Axillary tail	1 (1%)
	Retro Areolar	5 (5%)
	Central	16 (16%)
	Lower Inner (inferior medial)	18 (17%)
	Lower Outer (inferior lateral)	4 (4%)
	Upper Inner (superior medial)	10 (10%)
	Upper Outer (superior lateral)	49 (47%)
Size of cancer lesions (mm)		
	<10	49 (48%)
	≥10	54 (52%)

^a *: Architectural distortion/Stellate/Spiculated Mass/Asymmetric density

a 2-hour reading time for each session. Reading of BREAST test sets is encouraged by BreastScreen Australia and readers receive continuing professional development points for their involvement.

Mammograms were displayed in DICOM format on a dual 5-megapixel, 21-inch monochrome liquid crystal display monitor. Web-based software was displayed on an additional 19-inch monitor which also showed the test set images for the purpose of recording all decisions made by participants and, subsequently, to provide feedback on the readers' performance after the completion of each set. Readers were asked to complete a brief questionnaire including their experience in reading

Table 2. Details of participants

Categories		Mean [*] / N (%) [^]
Age		51 [*]
Number of years reading mammogram		15 [*]
Gender	Female	81 (69%)
	Male	36 (31%)
Number of cases reading per week	<20	19 (16%)
	20–59	15 (13%)
	60–100	6 (5%)
	101–150	17 (15%)
	151–200	30 (26%)
	>200	30 (25%)
Number of hours reading mammograms per week	<4	27 (23%)
	5–10	61 (52%)
	11–15	19 (16%)
	16–20	6 (5%)
	21–30	1 (1%)
	>30	3 (3%)
Type of mammograms reading	Soft copy	71 (61%)
	Hard copy	26 (22%)
	Both	20 (17%)
BreastScreen readers	Yes	94 (80%)
	No	23 (20%)
Completion a fellowship at least 3–6 months in reading mammograms	Yes	36 (31%)
	No	81 (69%)
Wearing eye-glasses	Yes	72 (62%)
	No	45 (38%)

^a *: Mean value; [^]N (%): Number and percentage of readers

mammograms and reading volume before test set reading commenced.

Readers were given an instruction before commencing the test set reading and informed that the test sets had been cancer-enriched relative compared with a typical screening population, yet the specific cancer prevalence was not revealed to avoid bias. For each case, participants were asked to give a Tabar BIRADS category based on the definition¹³:

- Category 1: Negative, no findings observed;
- Category 2: Benign findings, no further imaging is required;
- Category 3: Indeterminate/equivocal findings, require further investigation, usually with percutaneous needle biopsy (fine needle aspiration (FNA) cytology);
- Category 4: Suspicious findings of malignancy, requires further investigation with percutaneous needle biopsy sampling;
- Category 5: Highly suggestive of malignant findings, require further investigation even if percutaneous needle biopsy sampling is benign.

Table 3. Radiologists' characteristics and the average percentage of cases selected by these readers among categories. These values are given for 100 cancer cases across five test sets

Characteristics		Reported as Normal (Cat. 1)	Reported as Benign (Cat. 2)	Reported as Equivocal (Cat. 3)	Reported as Suspicious malignancy (Cat. 4)	Reported as Highly suggestive malignancy (Cat. 5)
Age	≤50	15.4	10.2	37.4	33.3	18.2
	>50	17.6	10.4	36.3	31.3	14.3
Number of years reading mammograms	≤15	15.8	10.2	36.9	32.9	18.3
	>15	17.4	10.4	36.8	31.7	14.1
Gender	Female	14.3*	8.1**	38.1	32.3	16.9
	Male	21.6*	16.2**	33.6	32.4	13.4
Number of cases reading per week	<20	16.6	20.8	24.4***	43.5**	18.8
	20–59	25.4	7.9	23.7***	44.3**	7.9
	60–100	14.2	10.0	39.3***	32.1**	15.8
	101–150	12.9	8.1	40.6***	30.3**	16.7
	151–200	14.3	8.2	39.8***	28.7**	18.9
	>200	17.1	10.0	44.2***	25.0**	15.0
Types of mammograms reading	Soft copy	16.2	12.0	38.8	30.7	16.1
	Hard copy	18.7	9.1	31.9	36.3	16.7
	Both	15.0	8.0	36.4	32.8	15.8
Working for BreastScreen	No	23.5*	15.0	24.0***	44.4*	9.4*
	Yes	14.9*	9.1	39.7***	29.5*	16.9*
Complete a reading fellowship	No	17.3	12.6*	35.5	34.2	14.4
	Yes	15.0	6.8*	39.5	28.8	18.9
Wearing corrective lenses/ eyeglasses	No	15.5	9.0	36.4	34.9	16.0
	Yes	17.2	11.0	37.1	30.8	16.3

Notes: Tabar BIRADS Cat. 1 (Non-marking obtained) or Cat. 2 (marking for benign lesions) was considered as False Negative while Cat. 3, 4 or 5 was recorded as True Positive.

^a *, **, ***: Significant differences in detection rates of radiologists at level of $p < 0.05$, $p < 0.01$, $p < 0.001$

To simulate the Australian BreastScreen system, when the cases were given categories 3–5, it meant that the radiologist had recalled the female for further assessment. In addition, for abnormal lesions (Grade 3–5), readers were required to indicate the lesion location by using a mouse-controlled cursor to mark the coordinates (x, y) on the screen. On all monitors, readers were able to zoom and digitally manipulate the images including panning or adjusting window width and level. At the completion of test sets, readers were provided with immediate feedback for each case, which allowed comparison between the readers' selections and the truth.

Data analysis

As this study focused on the Tabar BIRADS usage of radiologists in detecting abnormal mammograms, performance metrics were calculated for each reader and defined as follows:

- True positive: The number of correctly identified positive or cancer mammograms with category 3, 4 or 5;

- True negative: The number of correctly identified negative or normal mammograms with category 1 or 2;
- False negative: The number of positive or cancer cases identified as negative or normal cases with category 1 or 2;
- False positive: The number of negative or normal cases identified as positive or cancer cases with category 3, 4 or 5;

Only the data of participants who completed the full test set were included in the data analysis. In the first step, the average proportion of cancer cases in each Tabar BIRADS category was calculated and its association with reader characteristics was analyzed using Mann-Whitney U test for independent variables with two levels/groups and Kruskal-Wallis test for those with more than two levels/groups. The second step, the average percentage of radiologists in each BIRADS category among cancer cases and its relationship with case features were explored using the same statistical tests. All analysis was conducted using SPSS software (v.22; SPSS, Chicago, IL, USA) and P value <0.05 is considered as a significant result.

Table 4. Case features and the average percentage of radiologists among categories. These values are given for 100 cancer cases across five test sets

Categories		Reported as Normal (Cat. 1)	Reported as Benign (Cat. 2)	Reported as Equivocal (Cat. 3)	Reported as Suspicious malignancy (Cat. 4)	Reported as Highly suggestive malignancy (Cat. 5)
Breast density	1	17.8	5.8	33.6	41	12.7
	2	19.9	8.9	42.7	34.9	20.3
	3	22.8	9	34	36.5	19.3
	4	25.7	0	38.8	40.3	16.3
Cancer type	Architectural Distortion	26.4	10.9	35.7**	27.5	11.1**
	Calcification	25.1	8.3	39.2**	39.8	9.4**
	Discrete mass	18.4	13	50.0**	36.4	18.2**
	Asymmetric density	24.7	8.9	46.7**	27.2	5.6**
	Spiculated mass/ Stellate	19	7.6	29.1**	38.8	26.1**
	Mix of types	15.7	5.2	39.1**	39.6	16.7**
Cancer side	Left	19	7.9	35.6	40.5	17
	Right	24.2	9.7	38.7	32.5	22.1
	Both	3.9	2	48.3	39.7	9.1
Cancer site	Central	26.4	9.6	38.2	32.8	21.8
	Lower Outer	22.5	4.5	31.8	31.2	19.5
	Retro Areolar	19.4	8.8	37.1	37.3	8.9
	Upper Inner	17	6.9	37.3	41.6	14.7
	Lower Inner	21.1	16	42.5	32.6	22.5
	Upper Outer	22	7.7	35.3	37.2	20.2
Lesion size	Axilla tail	3.9	2	44.3	44.6	9.1
	≤10 mm	21.3	7.7	38.8	38.1	18.8
	>10 mm	21.9	10.2	35.4	33.8	19.1

Notes: Tabar BIRADS Cat. 1 (Non-marking obtained) or Cat. 2 (marking for benign lesions) was considered as False Negative while Cat. 3, 4 or 5 was recorded as True Positive. Mix: Stellate/Spiculated mass/ Architectural distortion/ Asymmetric density.

^a **: Significant differences in the percentage of radiologists ($p < 0.01$)

RESULTS

Reader performance levels in specific Tabar BIRADS category

On average, 36.8% of mammograms with abnormal lesions were detected as equivocal findings with category 3. This rate reduced to 32.3% with category 4 as suspicious malignancy and 16.2% with category 5 as highly suggestive malignancy. Regarding false negatives, 16.6 and 10.3% of cancer cases were marked with categories 1 (reported as normal cases) and 2 (reported as benign cases).

The association of reader and case characteristics with Tabar BIRADS categories

The data showed that female radiologists recorded lower false-negative rates than male readers in category 1 (14.3% vs 21.6%,

$p = 0.02$) and category 2 (8.1% vs 16.2%, $p = 0.006$). Radiologists working for BreastScreen services had significantly lower rate for reporting cancer cases as normal cases (category 1) compared with non-BreastScreen readers (14.9% vs 23.5%, $p = 0.03$) and higher correct cancer detection rates with category 3 ($p < 0.0001$) and category 5 ($p = 0.02$). Radiologists with lower number of mammograms reading per week were more likely to report cancer cases with category 4 (suspicious malignancy) ($p < 0.0001$) while the ones with higher number of mammograms had a tendency to report with category 3 as equivocal findings ($p = 0.004$). Completing a 3–6-month fellowship in reading mammograms was significantly related to the lower rate of identifying cancer as benign mammograms with category 2 compared with non-fellowship readers (6.8% vs 12.6%; $p = 0.01$) (Table 3). There were no significant associations between reporting scores and

Figure 1. Mammograms were obtained the highest percentages of readers' selection in categories 3 and 5. (A) The mammogram with discrete mass reported as category 3 by 81.8% of readers; (B) the mammogram with asymmetric density reported as category 3 by 74.5% of readers; (C) the mammogram with spiculated mass reported as category 5 by 90.1% of readers.

other radiologists' characteristics such as age, number of years reading mammograms, types of mammograms, and wearing eyeglasses or contact lenses.

Cancer types were found to be the only case feature significantly associated with reporting categories. Discrete mass and asymmetric density were the two lesions graded mostly as indeterminate findings (category 3) (47–50%; $p = 0.005$) while spiculated mass or stellate lesions were reported by a large proportion of radiologists as highly suggestive malignancy with category 5 (26%) compared with other cancer types ($p = 0.001$). Other features such as breast density, lesion side, site and size were not also significantly associated with reporting categories (Table 4).

DISCUSSION

This is the first study providing details of features associated with radiologists and mammograms that could determine the Tabar BIRADS score given to an image. The study included a retrospective analysis with case categories assigned based on the mammographic report and terminology is paralleled to the mammogram reporting assessment categories currently applied in BreastScreen services where data were collected.

The findings from our study show that gender was significantly related to the performances of radiologists. Male readers reported more false negatives in abnormal mammograms with categories 1 and 2 than female readers. This implies that male readers were potentially more likely to miss cancer cases than females. Prior research showed that female doctors had been more likely to adhere to clinical guidelines, provide preventive care more often and use more patient-centred communication than male doctors. In psychological studies, females's recollections in episodic memory such as face recognition were also more vivid than those of males.^{14,15} However, female radiologists participating in the study (63%) and the numbers of test sets read by female (69%) outnumbered the number of males which reflects the fact that females were more willing to participate in mammogram set readings than males. Further investigation with larger size of samples is required to confirm whether female radiologists are more likely to capture the abnormalities (or less likely to miss cancers) on mammograms than male radiologists.

Working for BreastScreen services also played an important role in the performances of radiologists with Tabar BIRADS score as BreastScreen readers recorded lower false-negative rates by using Grade 1 in cancer cases than non-BreastScreen readers (14.9% vs 23.5%). These results were aligned with findings from other authors who found the significant association of experience and diagnosis as the most experienced radiologists were related to better performances in reading mammograms than the least experienced radiologists.^{6,16} Data from our study report that Australian radiologists who worked for BreastScreen services are exposed to more mammograms (approximately 10,000 cases per year) than non-BreastScreen radiologists (approximately 2,000 cases per year) and therefore the experience is likely to be a determining factor in the risk of missing abnormal cases, with less-experienced radiologists more likely to have higher false-negative and lower true-positive rates.^{17,18}

In addition, the results show an interesting finding that radiologists with lower number of mammograms reading per week were more likely to report cancer cases with category 4 as suspicious malignancy (44%) than the readers with higher number of mammograms who were found to have more cancers reported with category 3 as equivocal findings (40%). The literature well describes that radiologists with more experience in mammograms often obtained higher diagnostic accuracy in cancer detection.⁶ However, there were lack of information about how BIRADS is used by radiologists and the finding from this study presented that less experienced radiologists might be more confident in detecting cancer on mammograms than experienced radiologists. This may result from the fact that less experienced individuals are possibly more junior radiologists who have been trained to use the Tabar BIRAD reporting system to categorize a breast lesion. More mature readers with greater years of experience and high volume of readings may be more used to selecting a breast lesion to recall assessment without categorization according to BIRADS, using the options either normal or recall. This may explain the high percentage of category 3 in more experienced readers.

Regarding case features, lesion type appeared to be the only factor in this study that had a significant impact on the category that radiologists selected when reporting mammograms. Approximately 50% of radiologists used category 3 for discrete masses and asymmetric density while the largest proportion of readers (26%) used category 5 (the highest confident category for a malignant case) for mammograms with spiculated mass or stellate lesions (Figure 1). These are in accordance with the study of Dorrius et al.¹⁹ who reported that 89% of the mammograms with category 3 were non-calcified solid masses or asymmetric densities. In the Australian radiologic scoring system, category 3 is considered as the lowest confident score to recall the cases as it is equal to an indeterminate decision.

A notable factor in our study is that the recall rate of cancer cases (cases with category >2) was much higher in our study than is generally found in BreastScreen Australia (37% vs 8%).²⁰ This could be because of two main factors. First, the National Accreditation Standards in Australia¹³ recommended that screen readers limited their clinical recall rates to 10 and 5% in the first and subsequent screens, respectively, (which are frequently monitored via clinical audit reports) and hence, radiologists in the clinics may have a greater reluctance to recall mammograms that are most likely normal in comparison with test set observations. This recall restriction obviously does not exist in test set reading situation and might therefore lead to a higher recall rate.²¹ Second, prior to the reading, participants were informed that

the test sets contained cancer-enriched cases, and this increased prevalence of abnormality might encourage a greater inclination to recall cases when unusually larger numbers of positives are presented.²² The authors acknowledged these issues as possible limitations of this study, yet it could be argued that these circumstances are inevitable in this particular type of research design.

CONCLUSION

The results from this study showed that a large proportion of radiologists used Tabar BIRADS category 3 to report the abnormal mammograms. Gender, working for BreastScreen service, undertaking a fellowship in reading mammograms and number of cases read per week were found to associate with Tabar BIRADS categories. The radiologists were more likely to report highly suggestive malignancy (category 5) for mammograms containing of spiculated mass or stellate lesion while most of readers voted as equivocal (category 3) for cases with discrete mass and asymmetric density. These findings will be useful for creating individual tailored training programs for radiologists to improve their diagnostic efficacy in identifying abnormal findings on the mammograms.

ACKNOWLEDGMENT

BreastScreen REader Assessment STRategy (BREAST) program is funded by Australian Department of Health and New South Wales Cancer Institute.

REFERENCES

1. Svahn TM, Chakraborty DP, Ikeda D, Zackrisson S, Do Y, Mattsson S, et al. Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy. *Br J Radiol* 2012; **85**: e1074–82. doi: <https://doi.org/10.1259/bjr/53282892>
2. Centre NBC Breast imaging: a guide for practice. *Camperdown, NSW* 2002;.
3. 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. doi: <https://doi.org/10.1148/radiol.2392042127>
4. Taplin SH, Ichikawa LE, Kerlikowske K, Ernster VL, Rosenberg RD, Yankaskas BC, et al. Concordance of breast imaging reporting and data system assessments and management recommendations in screening mammography. *Radiology* 2002; **222**: 529–35. doi: <https://doi.org/10.1148/radiol.2222010647>
5. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 2009; **253**: 641–51. doi: <https://doi.org/10.1148/radiol.2533082308>
6. Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings. *Radiology* 2013; **269**: 61–7. doi: <https://doi.org/10.1148/radiol.13122581>
7. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst* 2002; **94**: 1373–80. doi: <https://doi.org/10.1093/jnci/94.18.1373>
8. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst* 2003; **95**: 282–90. doi: <https://doi.org/10.1093/jnci/95.4.282>
9. Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. *Acad Radiol* 2002; **9**: 531–40. doi: [https://doi.org/10.1016/S1076-6332\(03\)80330-6](https://doi.org/10.1016/S1076-6332(03)80330-6)
10. Laya MB, Larson EB, Taplin SH, White E. Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. *J Natl Cancer Inst* 1996; **88**: 643–9. doi: <https://doi.org/10.1093/jnci/88.10.643>
11. Trieu PDY, Tapia K, Frazer H, Lee W, Brennan P. Improvement of cancer detection on mammograms via breast test sets. *Acad Radiol* 2019; **26**: e341–7. doi: <https://doi.org/10.1016/j.acra.2018.12.017>
12. AIHW. *The Active Australia Survey: a guide and manual for implementation, analysis and reporting*. Canberra: Australian Institute of Health and Welfare; 2003.
13. BreastScreenAustraliaNationalAccreditation Standards BreastScreen Australia quality improvement program. 2008;.
14. Demchig D, Mello-Thoms C, Lee WB, Khurelsukh K, Ramish A, Brennan PC. Mammographic detection of breast cancer in a non-screening country. *Br J Radiol* 2018; **91**: 20180071. doi: <https://doi.org/10.1259/bjr.20180071>
15. Pillemer DB, Wink P, DiDonato TE, Sanborn RL. Gender differences in autobiographical memory styles of older adults. *Memory* 2003;

- 11: 525–32. doi: <https://doi.org/10.1080/09658210244000117>
16. Suleiman WI, Lewis SJ, Georgian-Smith D, Evanoff MG, McEntee MF. Number of mammography cases read per year is a strong predictor of sensitivity. *J Med Imaging* 2014; **1**: 015503. doi: <https://doi.org/10.1117/1.JMI.1.1.015503>
17. Reed WM, Lee WB, Cawson JN, Brennan PC. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. *Acad Radiol* 2010; **17**: 1409–13. doi: <https://doi.org/10.1016/j.acra.2010.06.016>
18. Hawley JR, Taylor CR, Cubbison AM, Erdal BS, Yildiz VO, Carkaci S. Influences of radiology trainees on screening mammography interpretation. *J Am Coll Radiol* 2016; **13**: 554–61. doi: <https://doi.org/10.1016/j.jacr.2016.01.016>
19. Dorrius MD, Pijnappel RM, Sijens PE, van der Weide MCJ, Oudkerk M. The negative predictive value of breast magnetic resonance imaging in noncalcified BIRADS 3 lesions. *Eur J Radiol* 2012; **81**: 209–13. doi: <https://doi.org/10.1016/j.ejrad.2010.12.046>
20. Farshid G, Downey P, Gill P, Pieterse S. Assessment of 1183 screen-detected, category 3B, circumscribed masses by cytology and core biopsy with long-term follow up data. *Br J Cancer* 2008; **98**: 1182–90. doi: <https://doi.org/10.1038/sj.bjc.6604296>
21. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008; **249**: 47–53. doi: <https://doi.org/10.1148/radiol.2491072025>
22. Brem RF, Tabár L, Duffy SW, Inciardi MF, Guingrich JA, Hashimoto BE, et al. Assessing improvement in detection of breast cancer with three-dimensional automated breast us in women with dense breast tissue: the SonoInsight study. *Radiology* 2015; **274**: 663–73. doi: <https://doi.org/10.1148/radiol.14132832>