Received:
20 January 2017

Revised:
31 May 2017

Accepted:
8 June 2017

<https://doi.org/10.1259/bjr.20170064>

Cite this article as:

Damases CN, Hogg P, McEntee MF. Intercountry analysis of breast density classification using visual grading. *Br J Radiol* 2017; **90**: 20170064.

FULL PAPER

Intercountry analysis of breast density classification using visual grading

^{1,2}CHRISTINE N DAMASES, MTech, PhD, ³PETER HOGG, Bsc (Hons), MPhil and ¹MARK F MCENTEE, BSc (Hons), PhD

¹Faculty of Health Sciences, University of Sydney, Discipline of Medical Radiation Sciences and Brain and Mind Research Center, Sydney, NSW, Australia

²Faculty of Health Sciences, University of Namibia, Allied Health Department, Windhoek, Namibia

³Department of Radiography, University of Salford, Salford, UK

Address correspondence to: Dr Christine N Damases

E-mail: cdam2504@uni.sydney.edu.au

Objective: Disagreement in mammographic breast density (MBD) assessment can impact breast cancer risk stratification, choices of further breast cancer screening intervals and pathways. This study examines whether intercountry MBD expectations and assessment approaches are associated with differences in MBD assessment.

Methods: 20 American Board of Radiology (ABR) examiners and 24 UK practitioners using the 4th edition BI-RADS® lexicon assessed 40 mammogram cases of 20 females. 26 Royal Australian and New Zealand College of Radiologists (RANZCR) registered radiologists also assessed the same cases. Interobserver correlation and agreement were assessed using Spearman's correlation (ρ) and weighted kappa (κ_w), respectively.

Results: Strong positive correlation was observed between the study cohorts on a binary scale (1–2 vs 3–4) [ABR examiners and RANZCR radiologists ($\rho = 0.950$);

ABR examiners and UK practitioners ($\rho = 0.940$); and RANZCR radiologists and UK practitioners ($\rho = 0.958$)]. ABR and RANZCR radiologists demonstrated slight agreement [$\kappa_w = 0.10$; 95% confidence interval (CI) = -1.13 – 0.43], whereas ABR and UK practitioners showed a fair agreement [$\kappa_w = 0.25$; 95% CI = -0.42 – 0.61], and an almost perfect agreement was observed between RANZCR radiologists and UK practitioners [$\kappa_w = 0.95$; 95% CI = 0.91 – 0.97].

Conclusion: Findings demonstrate wide international and interobserver variability in MBD assessment. This level of variability underscores the need for automation and standardization of MBD assessment.

Advances in knowledge: Intercountry analysis of MBD assessment shows variations, with less variation on the binary scale than on the 4-point scale. With this level of variation, automation and standardization of MBD assessment becomes more appropriate.

INTRODUCTION

Mammographic breast density (MBD) is an indicator of risk of breast cancer, with females reported to have a four- to six-fold increase in breast cancer risk if they have extremely dense breasts compared with females with predominantly fatty breasts.^{1–3} MBD is defined as the proportion of radio-opaque fibroglandular tissue in the breast as apparent on a mammogram.⁴ High MBD is associated with a decrease in the sensitivity of mammography due to the potential masking of a breast cancer in fibroglandular tissue.² The potential of cancer being missed in a breast with high MBD leads to adjunctive imaging of females with dense breast using ultrasound, digital breast tomosynthesis, MRI or more regular recall for mammogram imaging.^{5–7} Therefore, it is important to assess the causal factors contributing to national and international variability in MBD assessment in order to underscore the importance of standardization of breast density assessment.

Radiologist professional bodies have proposed ways of assessing MBD visually. In the USA, the American College of Radiology (ACR) developed the breast imaging reporting and data system (BI-RADS®) scheme to provide a standardized categorization system for reporting MBD. The fourth edition ACR BI-RADS scheme classifies breast density into four categories based on the percentage of fibroglandular tissue in the breast.⁸ MBD classification descriptors for BI-RADS and the Royal Australian and New Zealand College of Radiology (RANZCR) synoptic scales are presented in Table 1.

The BI-RADS classification scheme has been modified in the 5th edition, with 1–4 changed to A–D, and breasts having a higher amount of dense tissue behind the nipple rated as C or D to account for the masking effect of MBD.⁹ To enable contribution of females to decision making, regarding screening for early breast cancer detection, 27

Table 1. Mammographic breast density classification according to the breast imaging reporting and data system (BI-RADS®) and Royal Australian and New Zealand College of Radiologists (RANZCR) scales

4th edition BI-RADS	RANZCR
(1) The breast is almost entirely fatty (<25% glandular)	(1) 25% glandular
(2) There are scattered areas of fibroglandular densities (approximately 25–50% glandular)	(2) 25–50% glandular
(3) The breast is heterogeneously dense, which could obscure detection of small masses (approximately 51–75% glandular)	(3) 51–75% glandular
(4) The breast is extremely dense. This may lower the sensitivity of mammography (>75% glandular)	(4) >75% glandular

states in the USA now have legislations authorizing reporting of MBD by radiologists.¹⁰ Although there is no such legislation in other countries as yet, many practices now assess and include MBD information in mammography reports.

The Australian and New Zealand College of Radiology also proposed the RANZCR synoptic guidelines to categorize MBD into four categories.¹¹ The RANZCR synoptic scale describes the percentage of glandular tissue for each of the four categories as shown in Table 1. Currently, only two studies have investigated the assessment of MBD using the RANZCR synoptic scales.^{12,13}

Although the BI-RADS scheme forms the basis for a majority of studies on MBD,^{14,15} it is limited by observer subjectivity and prone to intra- and interobserver variability in MBD rating.^{15–17} The reported inter- and intraobserver variability in MBD assessment using the BI-RADS system ranged from a kappa of 0.27–0.94.^{16–18} RANZCR breast density assessment is visual and subjective and thus has similar challenges as BI-RADS.^{12,13} There is a paucity of data on the level of interradiologists' agreement in MBD assessment using the RANZCR synoptic scale. No data exist for the level of variability in MBD assessment between RANZCR radiologists and breast readers from other parts of the world. Also, no work has assessed how the MBD assessed using BI-RADS reflects that of the RANZCR synoptic scale. Further work is required to assess how MBD assessment using the RANZCR scale compares with that assessed using the BI-RADS scale and to investigate whether prevalence expectation impacts observers' density assessments.

Even though there is no literature on the impact of density assessment based on the expected density, performance studies have shown that prevalence expectation has an impact on radiologists' behaviour.^{19–24} Prevalence expectation is a phenomenon that has been shown to influence the performance of observers in mammography interpretation.^{19–24} It is referred to as “the relationship between the prevalence of a particular image appearance and observer performance.”²⁴ Considering that prevalence expectation holds true for mammography interpretation, it might influence the categorization of MBD by observers from different countries. Many mammography image readers do not perform this task often, and this may lead to a reduced reliability and validity of visual determination of breast density. No study has assessed intercountry differences in

the ability of observers to assign mammography images for breast density into categories; this study attempts to fill that gap.

Therefore, this study aims to assess the level of agreement in MBD assessment between BI-RADS and RANZCR assessment scales. It does this by comparing MBD assessment of American Board of Radiology (ABR) examiners, UK practitioners and RANZCR registered radiologists. Such international interobserver comparison may improve understanding of potential factors associated with variability in MBD assessment.

METHODS AND MATERIALS

Institutional review board ethical approval was provided for the study (IRB 2013/448). The study cohorts consisted of US radiologists, Australian radiologists and UK practitioners (radiographers). All 20 ABR examiners were Mammography Quality Standards Act certified; the 26 Australian radiologists were RANZCR certified; and the 24 UK radiography practitioners were Health and Care Professions Council registered and working at advanced and consultant levels. All participants consented to the study. Flyers and e-mails were used to recruit the study cohorts. The Louisville data were collected from volunteer ABR examining radiologists. Flyers were placed around the hotel, and the proposed studies were announced at the information sessions for the ABR examiners. The Melbourne data were collected from volunteer RANZCR registered radiologists at the RANZCR annual scientific meeting. Flyers were placed around the convention centre, and the proposed studies were announced at the information sessions for radiologists. For the Salford data, flyers were posted out to the breast screening centres in the Salford area. Once the lead radiographer granted permission, the flyer was circulated to reporting radiographers in the department. The UK radiography practitioners voluntarily participated in the mammogram reading study. These held a diagnostic radiography entry qualification such as Bachelor of Science (BSc), an additional mammography imaging qualification such as a postgraduate certificate (PGC) and further specific qualifications in images reading for mammography. These qualifications enable them to perform the same clinical roles as radiologists in full-field digital mammography (FFDM) imaging and to the same standard.^{25–29} In the UK, advanced practice/consultant radiographers perform FFDM reporting in the same way and to the same standard as a radiologist within the National Health Service Breast Screening Program. Given that a radiologist's

ability to assess density visually has already been determined, our study builds on that work by offering insight into a specific group of highly skilled radiographers, as it is possible that their scope could develop to include density scoring. BI-RADS lexicon is sometimes used in the UK as a subjective method for MBD assessment. More commonly, a rating of “fatty”, “mixed” or “dense” is given. The study cohorts had differing mean years of experience and the average number of mammograms read per year (Table 2).

Image selection and Volpara® density™ grading

A FFDM data set comprised of 40 cases was obtained from 20 normal cases. These were negative for cancer and had no obvious benign findings. The females were aged 42–89 years. These images were acquired at a single site in New York under the same protocol on GE Senographe Essential (or DS; GE, Fairfield, CT) and on Hologic Lorad Selenia (Hologic, Bedford, MA) imaging systems 1 year apart. The images were selected to enable a comparison of the Volpara density grades (VDGs) for females whose images were produced 1 year apart and also to have a comparable number of cases for each of the four VDG categories. A stratified sample was selected in an attempt to ensure similar numbers in each density category, with 22.5% images in VDG 1, 32.5% in VDG 2, 20% in VDG 3 and 25% in VDG 4. For each case, the images were displayed in the following order: first left craniocaudal followed by a left mediolateral oblique and then the combination of left craniocaudal and left mediolateral oblique presented together. To ensure observers could evaluate the images in a 15- to 20-min time period, only the left breast images were used for this study. Considering that this data set contained images of the same females taken 1 year apart and the MBD assessment scores were obtained within a single sitting, the same observer saw both cases from the same females. Although the observers were presented with the images of the same females, these images were not identical. Positioning and equipment changes sufficiently change images such that they did not appear to be of the same female. Also, observers had no reason to suspect that images of the same females would appear twice. Furthermore, the ability to remember an image is related to remarkable aspects of the image.³⁰ Additionally, the trace decay theory of forgetting suggests that short-term memory can only hold information between 15 and 30 s, unless it is rehearsed.³¹ Therefore, the fact that the observers were presented with the images of the same females taken 1 year apart within one sitting can not be considered as a confound in this study.

Automated volumetric breast density assessment of these cases was first performed using Volpara density v. 1.4.3 (Mātakina,

Wellington, New Zealand) to obtain Volpara density grades (VDGs). The pre-set VDG categories are as follows: VDG 1: <4.5%; VDG 2: 4.5–7.5%; VDG 3: 7.6–15.5%; and VDG 4: >15.5%. These VDG thresholds are used to represent BI-RADS and RANZCR 1–4 categories, respectively.

Image display and MBD quantification using BI-RADS and RANZCR

Images were displayed on a single EIZO® GS510 5-MP display (Tokyo, Japan). This was calibrated to the digital imaging and communications in medicine greyscale standard display function and the user interface was ViewDEX software v. 2.0.³² The monitor has been shown to demonstrate the required characteristics detailed in the Association of Physicists in Medicine Task Group 18 report.³³ The observers were able to adjust the window width and level and also could pan and zoom the images. The reading environment was standardized, with the ambient lighting kept constant between 25 and 35 lux as confirmed by a calibrated photometer (model 07–621, Nuclear Associates).³⁴

The mammogram cases were randomized using random integer generator³⁵ prior to MBD assessment by ABR examiners, Australian radiologists and UK practitioners, respectively. 20 ABR examiners and 24 UK practitioners assessed the same images using the BI-RADS MBD assessment scheme.^{8,36} The images were also assessed by 26 RANZCR registered radiologists using the RANZCR synoptic scale.¹¹ Since MBD Categories 3 and 4 have the potential to conceal small lesions and reduce the sensitivity of mammography, respectively, the assessments of observers were then grouped into two categories [low (1 and 2) and high (3 and 4)] for both BI-RADS and RANZCR. This was to assess the level of interobserver agreement on a binary scale to provide the potential level of variability with regard to screening individualization.

Data analysis

Statistical analyses were performed using the Statistical Package for Social Sciences (SPSS® v. 21.0; IBM Corp., New York, NY; formerly SPSS Inc., Chicago, IL). A non-parametric Spearman's analysis was used to assess the correlation between MBD assessments made using BI-RADS and RANZCR for all observers (20 ABR examiners, 26 RANZCR registered radiologists and 24 UK practitioners). A Wilcoxon signed-rank test was also used to compare the median scores of observers. A weighted kappa (κ_w) statistic was used to test for the degree of agreement between MBD assessment schemes and pairs of observers. A weighted kappa was used because it accounts for the

Table 2. Demonstrates the mean years of specialization and number of mammograms read per year (standard deviation) for each of the study cohort

Study cohorts	Mean years of specialisation	Average mammograms read per year
ABR examiners	24.9 (8.3)	7107 (5308)
RANZCR radiologists	17.28 (12.03)	2072 (2441)
UK practitioners	5.96 (6.31)	2583(2865)

ABR, American Board of Radiology; RANZCR, Royal Australian and New Zealand College of Radiologists.

Figure 1. Percentage distribution of cases assigned into mammographic breast density categories by each American Board of Radiology examiner (a), Royal Australian and New Zealand College of Radiologists registered Australian radiologists (b) and UK practitioners (c). BI-RADS®, the breast imaging reporting and data system.

level of disagreement between observers. A two-way mixed model, which allows for selection of cases randomly and for nesting the computation within observers, was used to calculate average absolute agreement for all the study cohorts, respectively. The interobserver agreement for ABR examiners, RANZCR radiologists and UK practitioners was assessed separately. MBD assessments of these observers were compared in pairs to assess their interreader agreement. For each cohort, the agreement between every possible pair combination of all the observers was performed off-line using a commercial software package MATLAB® v. 2009 (The MathWorks® Inc., Natick, MA) and confusion matrices were formulated. Cohen kappa algorithm was implemented, and a mean Kappa for each reader was computed. Then, overall kappa for each cohort was calculated by averaging the means of all the observers in a specific cohort. The level of agreement was examined both on a four-point (1–4) and binary (1 and 2 vs 3 and 4) classification scales. The results were considered to be statistically significant at $p < 0.05$.

RESULTS

Classification of images

The VDG classifications were regarded as the “truth categories” for this study. The percentage distribution of cases classified into MBD categories by individual ABR examiners, Australian radiologists and UK practitioners are shown in Figure 1a–c. Each of the observers in the three study cohorts assessed MBD and a majority report was generated from these assessments. The term majority report denotes the consensus of at least 51% of the cohort of the observers. The number of cases assigned to different MBD categories according to the majority reports of observer cohorts is shown in Table 3.

Comparison of median MBD values between ABR examiners, RANZCR radiologists and UK practitioners BI-RADS scores

For all observers, the median MBD scores obtained using BI-RADS and RANZCR were 2 and 2, respectively, and the median difference between ABR BI-RADS and RANZCR MBD score was not significant ($Z = -0.199$; $p < 0.843$). The median MBD scores by ABR examiners and UK practitioners using BI-RADS were 2 and 2, respectively ($Z = -0.788$; $p < 0.431$). RANZCR radiologists demonstrated a median MBD score of 2, and this was not statistically significantly different from the median score of the UK practitioners ($Z = -1.414$; $p < 0.157$).

Correlation between ABR examiners, Australian radiologists and UK practitioners

Spearman's correlation analysis demonstrated a weakly non-significant negative relationship between the MBD assessment of ABR examiners and RANZCR radiologists on a four-point scale ($\rho = -0.029$; $p < 0.859$). A strong positive correlation was demonstrated between MBD assessments of ABR examiners and RANZCR radiologists on a binary scale ($\rho = 0.950$; $p < 0.001$).

A weak positive correlation was observed between MBD assessments made by ABR examiners and UK practitioners on a four-point scale ($\rho = 0.148$; $p < 0.362$). Both groups of observers demonstrated a strong positive correlation on the binary scale ($\rho = 0.940$; $p < 0.001$).

The MBD assessed by Australian radiologists showed a strong positive relationship with that of UK practitioners on a four-point scale ($\rho = 0.916$; $p < 0.001$). A strong positive correlation was also noted on the binary scale ($\rho = 0.958$; $p < 0.001$).

Agreement between ABR examiners and RANZCR registered radiologists on the same images

All the cohorts of the study were presented with 40 cases. These images had MBD ratings from Volpara which were used as the ground truth. Where the majority report of the ABR and RANZCR radiologist concurs, this is counted as agreement. Overall, the ABR examiners and RANZCR registered radiologists agreed on 12/40 (30%) images. The ABR examiners generally graded cases into a higher MBD category than the Australian radiologists. Of the 40 cases in the data set, 4 images rated as BI-RADS 3 by the ABR examiners were rated RANZCR 1 by RANZCR radiologists, 3 images rated as BI-RADS 3 were rated RANZCR 2 (Table 4). The overall agreement (κ_w) between BI-RADS and RANZCR was 0.010 [95% confidence interval (CI) = -1.13 – 0.43].

Agreement between American Board of Radiology examiners and UK practitioners

The ABR examiners and UK practitioners agreed on 15/40 (38%) cases. Where the majority report of the ABR radiologist and UK practitioners concurs, this is counted as agreement. Again, the ABR examiners provided a higher MBD score than the UK practitioners. Four cases rated as BI-RADS 3 by ABR examiners were rated BI-RADS 1 by UK practitioners (Table 5). The overall agreement (κ_w) between ABR examiners' and UK practitioners' BI-RADS assessment was 0.25 (95% CI = -0.42 – 0.60).

Table 3. The percentages of cases assigned into different mammographic breast density categories according to each cohort's majority report

Volpara® density™	ABR examiners	RANZCR radiologists	UK practitioners
22.5%	34%	34%	29%
32.5%	26%	28%	28%
20%	31%	27%	28%
25%	9%	11%	15%

ABR, American Board of Radiology; RANZCR, Royal Australian and New Zealand College of Radiologists.

Table 4. A matrix of the number of cases allocated into the breast imaging reporting and data system (BI-RADS®) and Royal Australian and New Zealand College of Radiologists (RANZCR) categories based on the majority report, showing the number of cases in which the American Board of Radiology (ABR) examiners and RANZCR radiologists agreed for each category in bold

MBD assessment method		ABR BI-RADS				Total
		1	2	3	4	
RANZCR	1	5	3	4	2	14
	2	4	2	3	0	9
	3	3	4	5	1	13
	4	2	1	1	0	4
Total		14	10	13	3	40

Agreement between RANZCR registered radiologists and UK practitioners

RANZCR registered radiologists and UK practitioners agreed on 32/40 (80%) cases. Where the majority report of the RANZCR and UK practitioners concurs, this is counted as agreement. UK practitioners classified six cases into a higher MBD category than the Australian radiologists. Four of the cases that were rated RANZCR 1 were rated BI-RADS 2 by the UK practitioners (Table 6). The overall agreement between RANZCR and UK practitioners' BI-RADS was 0.95 (95% CI = 0.91–0.97).

Interobserver agreement for ABR examiners, RANZCR registered radiologists and UK practitioners

Generally, the UK practitioners and RANZCR radiologists tended to call the cases denser than the ABR radiologists. Table 7 shows the interobserver agreement in MBD assessment for each observer cohort. The overall interobserver agreement among ABR examiners was average [$(\kappa_w) = 0.57$; 95% CI = 0.52–0.61] on a four-point BI-RADS scale and ranged from a kappa of 0.33–0.67. On a binary scale, the overall interobserver agreement (κ_w) was 0.86; 95% CI = 0.82–0.87 and ranged from a kappa of 0.66–0.90 (Figure 2a,b).

Interobserver agreement using RANZCR four-point scale was 0.36 (95% CI = 0.31–0.41) and ranged from 0.078 to 0.499. RANZCR

Table 5. A matrix of the number of cases allocated into each breast imaging reporting and data system (BI-RADS®) category by American Board of Radiology (ABR) examiners and UK practitioners, with the number of cases in which these cohorts agreed for each category shown in bold

MBD assessment method		ABR BI-RADS				Total
		1	2	3	4	
UK BI-RADS	1	6	3	4	1	14
	2	2	3	3	2	10
	3	2	5	5	1	13
	4	1	1	0	1	4
Total		11	12	12	5	40

interobserver agreement on the binary scale was substantial [0.71; 95% CI = 0.66–0.77] and ranged from 0.22 to 0.89 (Figure 2c,d).

The interobserver agreement in MBD assessment amongst UK practitioners was 0.47 (95% CI = 0.43–0.50) on a four-point scale, with kappa values ranging from 0.24 to 0.58. A substantial interobserver agreement was observed for UK practitioners on a binary scale (0.78; 95% CI 0.74–0.82) and ranged from 0.48 to 0.85 (Figure 2e,f).

DISCUSSION

Reproducibility of MBD classification is important given the relevance of MBD information in breast cancer risk assessment and the tailoring of screening methods and frequency. It is important that the same cohort of females imaged under similar conditions have the same opportunity for personalized screening from MBD assessment. The current work explored the agreement in MBD of the same females assessed using different approaches and by different cohorts of observers. Findings demonstrate a wide variability in MBD categorization between observers, with ABR examiners demonstrating slight agreement with RANZCR radiologists and fair agreement with UK practitioners, and RANZCR radiologists demonstrating almost perfect agreement with UK practitioners. This wide variability was also noted among the same observer cohort, with ABR examiners and UK practitioners each demonstrating moderate interobserver agreement and RANZCR radiologists demonstrating fair interobserver agreement.

The interobserver agreement among RANZCR radiologists was lower than that among the ABR examiners and UK practitioners. Factors such as years of experience, number of mammograms read per year, training and the legislation framework governing reporting of breast density, might affect the classification of MBD by observers from different domains. The effect of legislation on MBD reporting has been demonstrated in a recent study, which showed change in the reporting patterns of radiologists after the implementation of density reporting legislation in the USA.³⁷ The study showed that 50% of the observer cohort assigned more cases to BI-RADS 2 than to BI-RADS 3. The remaining observers (44%) had equal ratings for BI-RADS 2 and 3 categories.³⁷ It should be noted that MBD

Table 6. A matrix of the number of cases allocated into breast imaging reporting and data system (BI-RADS®) and Royal Australian and New Zealand College of Radiologists (RANZCR) categories by RANZCR radiologists and UK practitioners, showing the number of cases in which they agreed for each category in bold

MBD assessment method		UK BI-RADS				Total
		1	2	3	4	
RANZCR	1	10	4	0	0	14
	2	1	8	0	0	9
	3	0	0	11	2	13
	4	0	0	1	3	4
Total		11	12	12	5	40

Table 7. The overall interobserver agreement (95% confidence interval) for each cohort of observers

Scale	Interobserver agreement		
	ABR examiners	RANZCR radiologists	UK practitioners
Four-point	0.57 (0.52–0.61)	0.36 (0.31–0.41)	0.47 (0.43–0.50)
Binary	0.86 (0.82–0.87)	0.71 (0.66–0.77)	0.78 (0.74–0.82)

ABR, American Board of Radiology; RANZCR, Royal Australian and New Zealand College of Radiologists.

legislations aim to facilitate shared decision making between screened females and their physicians regarding adjunctive screening. Some radiologists grade MBD taking into consideration age and clinical history of the patient.³⁷ Hence it is logical that these factors may significantly impact upon interreader variability in MBD assessment. This finding suggests that MBD classification may be influenced by systems requirement, legislation and individual perception of the potential impact of breast density. Therefore, further work should investigate whether these factors are

associated with wide international variability in the MBD classification of the same patient cohort observed in the current work.

A recent study reported a 32.4% disagreement between a pair of radiologists³⁸ and suggested that this level of disagreement limits use of qualitative assessments for recommending additional screening and risk management of females with dense breasts.³⁸ The current study demonstrates a 70% disagreement between ABR examiners and Australian radiologists and a 62%

Figure 2. Interobserver agreement for mammographic breast density assessment using the breast imaging reporting and data system (BI-RADS®) and the Royal Australian and New Zealand College of Radiologists (RANZCR) scales. (a) American Board of Radiology examiners' agreement on BI-RADS four-point scale and (b) agreement on BI-RADS binary scale. (c) RANZCR radiologists' agreement on RANZCR four-point scale and (d) agreement on RANZCR binary scale. (e) UK practitioners' agreement on BI-RADS four-point scale and (f) agreement on BI-RADS binary scale.

disagreement between ABR examiners and UK practitioners. The lowest disagreement was reported between Australian radiologists and UK practitioners (20%). These levels of disagreement are likely to change or influence individualized screening of females with dense breasts. The disagreement particularly becomes crucial when it affects the categorization that differentiates low (1 and 2) from high (3 and 4) MBD categories. This is because it determines the category of females who are likely to be referred for additional imaging with ultrasound or MRI.^{12,39} Encouragingly, the level of disagreement observed on a binary scale in the current study was less than that observed on a four-point scale. Nevertheless, the level of interobserver disagreement on a binary scale was still appreciable and underscores the need for standardization of breast density assessment. This may require the introduction of automated MBD assessment techniques in all screening programmes to more appropriately tailor adjunctive imaging and screening intervals for females with dense breasts.

The current work is based on cases taken from females in the USA. The results of the study suggest that radiologists' perception of MBD may be based on the normal MBD distribution seen within their local population. This finding is consistent with previous studies, which reported low interobserver variability for American radiologists.^{40,41} However, the question arises whether or not it causes a difference for a group of observers assessing density in females whom they are not accustomed to. Further work is required to examine whether observer rating of MBD is influenced by breast density distribution of a population that they are accustomed to.

Considering that the fourth edition BI-RADS and RANZCR classify breast density according to the same percentages (Table 1), it is logical that they would demonstrate a good level of agreement in the same patient cohort. The wide level of interreader and intercountry variability observed in our study is a cause for concern and shows perhaps the lack of understanding of or adherence guidelines for MBD assessment. It is unclear whether the negative correlation observed between ABR and RANZCR radiologists is due to prevalence expectation, where observers are accustomed to a certain MBD grade. Additional training for further assessment of performance could be beneficial. Our findings show how the same-females cohort could be classified into different risk strata and into screening regimens and pathways in different countries and among observers, thus limiting consistency in clinical use of MBD.

The BI-RADS system originated in the USA, and no difference has been shown in the range of interobserver agreement for studies based in the USA vs outside the USA.⁴² However, it is possible that interregional or intercountry differences in visual MBD assessment approaches would cause variation in MBD rating of the same female as demonstrated in the current work.

There is evidence that visual assessment of MBD has wide interreader disagreement.^{16,40,43,44} This variability was observed in previous work with interreader agreement (κ) ranging from 0.328 to 0.669 and 0.078–0.499, respectively.^{13,39} Given the reported intra- and interobserver variation with other visual assessment methods such as BI-RADS,^{45–48} there is a need to determine the range of agreement that can be expected between countries using the same criteria for MBD assessment. Importantly, the current study has provided insight to the level of variability in MBD assessment between observers from different practices and countries.

The strengths of our study include the large number of observers from different backgrounds. Secondly, this is the first international assessment of interobserver agreement in MBD assessment using BI-RADS and RANZCR synoptic scales. Data provided show for the first time how MBD of the same cohort of females can be classified differently by observers from different domains. There were several limitations to the study. Only the left breast was used for BI-RADS and RANZCR assessment. It is possible that including the right breast may have affected the results presented in this study. The observers may not have been familiar with the presentation state of the images and may be used to a different look. This may have affected their conclusions on density, and even though Volpara was used as the “ground truth” for all the cohorts of the study, BI-RADS and RANZCR scales are not designed to be exactly the same as Volpara. Therefore, the disagreement shown between these scales and Volpara might be expected. Furthermore, observers are familiar with using BI-RADS and RANZCR density assessment scales, therefore intercountry differences are also expected. Previous studies found in-country variations between observers.^{12,39} Therefore, different observers are likely to see different patient populations even compared with their in-country colleagues, as all these assessment methods are using a four-point scale. The UK cohort for the current study comprised of radiographers, therefore further work will investigate international interobserver comparisons for UK radiologists.

CONCLUSION

Data produced demonstrate wide international and interobserver disagreement in MBD assessment. In particular, the findings show poor agreement between ABR examiners and RANZCR and UK mammography image readers. The findings also showed moderate interobserver agreement in MBD assessment among ABR radiologists, fair agreement amongst RANZCR radiologists and moderate agreement amongst UK practitioners. The findings emphasize the need to improve reproducibility of MBD classification internationally in order to improve risk stratification and more appropriately tailor screening in females with dense breast.

REFERENCES

1. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med* 2007; **356**: 227–36. doi: <https://doi.org/10.1056/nejmoa062790>
2. Mandelson MT, Oestreich N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic

- detection: comparison of interval- and screen-detected cancers. *J Natl Cancer Inst* 2000; **92**: 1081–7. doi: <https://doi.org/10.1093/jnci/92.13.1081>
3. 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. doi: <https://doi.org/10.1158/1055-9965.epi-06-0034>
 4. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Res* 2011; **13**: 223.
 5. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Velez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA* 2008; **299**: 2151–63.
 6. Berg WA, Zhang Z, Lehrer D, Jong RA, Pisano ED, Barr RG, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA* 2012; **307**: 1394–404. doi: <https://doi.org/10.1001/jama.2012.388>
 7. ACRIN: Screening breast ultrasound in high-risk women, 2007 [database on the Internet] [Cited 23 June]. Available from: <http://www.acrin.org/Portals/0/Protocols/6666/Protocol-ACRIN>
 8. Radiology ACo. *BI-RADS Mammography 2013-ACR BI-RADS Atlas*, 5th edn. Reston, VA: American College of Radiology, 2013; [cited 2014 17 March]; Available from: <http://www.acr.org/Quality-Safety/Resources/BIRADS>
 9. Ekpo EU, Hogg P, Highnam R, McEntee MF. Breast composition: measurement and clinical use. *Radiography* 2015; **21**: 324–33. doi: <https://doi.org/10.1016/j.radi.2015.06.006>
 10. Durning MV. Breast density notification laws state-interactive map. [Cited 1 December 2016]; Available from: <http://www.diagnosticimaging.com/breast-imaging/breast-density-notification-laws-state-interactive-map>
 11. NBCC 2007. Synoptic breast imaging report, 2014 [Cited 20 August 2014]; Available from: <http://canceraustralia.nbcc.org.au/view-document-details/rsig-1-synoptic-breast-imaging-report-update>
 12. Damases CN, Mello-Thoms C, McEntee MF. Inter-observer variability in mammographic density assessment using Royal Australian and New Zealand College of Radiologists (RANZCR) synoptic scales. *J Med Imaging Radiat Oncol* 2016; **60**: 329–36. doi: <https://doi.org/10.1111/1754-9485.12451>
 13. Damases CN, Mello-Thoms C, McEntee MF. Inter-observer variability within BI-RADS and RANZCR mammographic density assessment schemes. *Med Imaging* 2016; **97**: 9787.
 14. McCormack VA, Highnam R, Perry N, Silva ID. Comparison of a new and existing method of mammographic density measurement: intramethod reliability and associations with known risk factors. *Cancer Epidemiol Biomarkers Prev* 2007; **16**: 1148–54. doi: <https://doi.org/10.1158/1055-9965.epi-07-0085>
 15. Ekpo EU, McEntee MF. Measurement of breast density with digital breast tomosynthesis—a systematic review. *Br J Radiol* 2014; **87**: 20140460.
 16. 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. doi: <https://doi.org/10.1016/j.breast.2004.12.004>
 17. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS atlas. *AJR Am J Roentgenol* 2016; **206**: 1119–23.
 18. Bernardi D, Pellegrini M, Di Michele S, Tuttobene P, Fanto C, Valentini M, et al. Interobserver agreement in breast radiological density attribution according to BI-RADS quantitative classification. *Radiol Med* 2012; **117**: 519–28. doi: <https://doi.org/10.1007/s11547-011-0777-3>
 19. Littlefair S, Mello-Thoms C, Reed W, Pietryk M, Lewis S, McEntee M, et al. Increasing prevalence expectation in thoracic radiology leads to overcall. *Acad Radiol* 2016; **23**: 284–9.
 20. Reed WM, Chow SL, Chew LE, Brennan PC. Assessing the impact of prevalence expectations on radiologists' behavior. *Acad Radiol* 2014; **21**: 1220–1.
 21. Reed WM, Chow SL, Chew LE, Brennan PC. Can prevalence expectations drive radiologists' behavior? *Acad Radiol* 2014; **21**: 450–6.
 22. Gur D, Rockette HE, Warfel T, Lacomis JM, Fuhrman CR. From the laboratory to the clinic: The “prevalence effect”. *Acad Radiol* 2003; **10**: 1324–6.
 23. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: changing the confidence ratings. *Acad Radiol* 2007; **14**: 49–53.
 24. Reed WM, Ryan JT, McEntee MF, Evanoff MG, Brennan PC. The effect of abnormality-prevalence expectation on expert observer performance and visual search. *Radiology* 2011; **258**: 938–43. doi: <https://doi.org/10.1148/radiol.10101090>
 25. Wivell G, Denton ER, Eve CB, Inglis JC, Harvey I. Can radiographers read screening mammograms? *Clin Radiol* 2003; **58**: 63–7. doi: <https://doi.org/10.1053/crad.2002.1087>
 26. Pauli R, Hammond S, Cooke J, Ansell J. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. *J Med Screen* 1996; **3**: 18–22. doi: <https://doi.org/10.1177/096914139600300106>
 27. Moran S, Warren-Forward H. A retrospective pilot study of the performance of mammographers in interpreting screening mammograms. *Radiography* 2010; **57**: 12. doi: <https://doi.org/10.1002/j.2051-3909.2010.tb00115.x>
 28. Bennett RL, Sellars SJ, Blanks RG, Moss SM. An observational study to evaluate the performance of units using two radiographers to read screening mammograms. *Clin Radiol* 2012; **67**: 114–21. doi: <https://doi.org/10.1016/j.crad.2011.06.015>
 29. Pauli R, Hammond S, Cooke J, Ansell J. Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions. *Br J Radiol* 1996; **69**: 10–14. doi: <https://doi.org/10.1259/0007-1285-69-817-10>
 30. Ryan JT, Haygood TM, Yamal JM, Evanoff M, O'Sullivan P, McEntee M, et al. The “memory effect” for repeated radiologic observations. *AJR Am J Roentgenol* 2011; **197**: W985–91. doi: <https://doi.org/10.2214/ajr.10.5859>
 31. Brown J. Some tests of the decay theory of immediate memory. *Q J Exp Psychol* 1958; **10**: 12–21. doi: <https://doi.org/10.1080/17470215808416249>
 32. Börjesson S, Håkansson M, Båth M, Kheddache S, Svensson S, Tingberg A, et al. A software tool for increased efficiency in observer performance studies in radiology. *Radiat Prot Dosim* 2005; **114**: 45–52. doi: <https://doi.org/10.1093/rpd/nch550>
 33. Ekpo EU, McEntee MF. An evaluation of performance characteristics of primary display devices. *J Digit Imaging* 2015. doi: <https://doi.org/10.1007/s10278-015-9831-3>
 34. Brennan PC, McEntee M, Evanoff M, Phillips P, O'Connor WT, Manning DJ. Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist. *AJR Am J Roentgenol* 2007; **188**: W177–80. doi: <https://doi.org/10.2214/ajr.05.2048>
 35. Random Integer Generator. [Cited 23 December 2016]; Available from: <https://www.random.org/integers/>
 36. D'Orsi CJ, Bassett L, Berg W. BI-RADS mammography. In: D'Orsi CJ, Mendelson FB, Ikeda DM, eds. *Breast imaging and*

- reporting and data system: ACR BI-RADS-breast imaging atlas. 4th edn. Reston, VA: American College of Radiology; 2003.
37. Gur D, Klynn AH, King JL, Bandos AI, Sumkin JH. Impact of the new density reporting laws: radiologist perceptions and actual behavior. *Acad Radiol* 2015; **22**: 679–83.
 38. van der Waal D, den Heeten GJ, Pijnappel RM, Schuur KH, Timmers JM, Verbeek AL, et al. Comparing visually assessed BI-RADS breast density and automated volumetric breast density software: a cross-sectional study in a breast cancer screening setting. *PLoS One* 2015; **10**: e0136667.
 39. Damases CN, Brennan PC, Mello-Thoms C, McEntee MF. Mammographic breast density assessment using automated volumetric software and breast imaging reporting and data system (BIRADS) categorization by expert radiologists. *Acad Radiol* 2016; **23**: 70–7.
 40. Redondo A, Comas M, Macia F, Ferrer F, Murta-Nascimento C, Maristany MT, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *Br J Radiol* 2012; **85**: 1465–70. doi: <https://doi.org/10.1259/bjr/21256379>
 41. Masroor I, Rasool M, Saeed SA, Sohail S. To assess inter- and intra-observer variability for breast density and BIRADS assessment categories in mammographic reporting. *J Pak Med Assoc* 2016; **66**: 194–7.
 42. Antonio AL, Crespi CM. Predictors of interobserver agreement in breast imaging using the breast imaging reporting and data system. *Breast Cancer Res Treat* 2010; **120**: 539–46. doi: <https://doi.org/10.1007/s10549-010-0770-x>
 43. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994; **331**: 1493–9. doi: <https://doi.org/10.1056/nejm199412013312206>
 44. Abdolell M, Tsuruda K, Payne JJ, Schaller G, Iles SE, Lightfoot CB, et al. Agreeing to disagree: assessing inter-rater variation in breast density measurement. In: *Radiology ESO*. Vienna, Austria: European Congress of Radiology; 2014. pp. 1–9.
 45. Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH, Rutter CM. Misclassification of breast imaging reporting and data system (BI-RADS) mammographic density and implications for breast density reporting legislation. *Breast J* 2015; **21**: 481–9. doi: <https://doi.org/10.1111/tbj.12443>
 46. Gweon HM, Youk JH, Kim JA, Son EJ. Radiologist assessment of breast density by BI-RADS categories versus fully automated volumetric assessment. *AJR Am J Roentgenol* 2013; **201**: 692–7. doi: <https://doi.org/10.2214/ajr.12.10197>
 47. Sauber N, Chan A, Highnam R. BI-RADS breast density classification—an international standard. *Eur Radiol* 2013.
 48. Wang K, Chan A, Highnam R. Robustness of automated volumetric breast density estimation for assessing temporal changes in breast density. *Eur Radiol* 2015.