



University of
Salford
MANCHESTER

Inter country analysis of breast density
classification using visual grading.
Analysis of mammographic breast density
using visual grading

Damases, C, Hogg, P and McEntee, M

<http://dx.doi.org/10.1259/bjr.20170064>

Title	Inter country analysis of breast density classification using visual grading. Analysis of mammographic breast density using visual grading
Authors	Damases, C, Hogg, P and McEntee, M
Publication title	British Journal of Radiology
Publisher	British Institute of Radiology
Type	Article
USIR URL	This version is available at: http://usir.salford.ac.uk/id/eprint/42552/
Published Date	2017

USIR is a digital collection of the research output of the University of Salford. Where copyright permits, full text material held in the repository is made freely available online and can be read, downloaded and copied for non-commercial private study or research purposes. Please check the manuscript for any further copyright restrictions.

For more information, including our policy and submission procedure, please contact the Repository Team at: library-research@salford.ac.uk.

Inter country analysis of breast density classification using visual grading

Analysis of mammographic breast density using visual grading

ABSTRACT

Purpose: 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 inter-country MBD expectations and assessment approaches are associated with differences in MBD assessment.

Methods: Twenty American Board of Radiology (ABR) examiners and 24 United Kingdom (UK) practitioners using the 4th edition BI-RADS® lexicon assessed 40 mammogram cases of 20 women. Twenty-six Royal Australian and New Zealand College of Radiologists (RANZCR) registered radiologists also assessed the same cases. Inter-observer correlation and agreement were assessed using a 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 ($\rho = 0.940$); RANZCR radiologists and UK practitioners ($\rho = 0.958$)]. ABR and RANZCR radiologists demonstrated slight agreement [$\kappa_w = 0.10$; 95% CI = -1.13 - 0.43] while 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 inter-observer variability in MBD assessment. This level of variability underscores the need for automation and standardisation of MBD assessment.

Key words: Breast density, inter-observer agreement, visual assessment methods.

Advances in knowledge:

*Inter country analysis of mammographic breast density assessment shows variations, with less variation on binary scale than on 4-point scale.

*With this level of variation automation and standardisation of mammographic breast density assessment becomes more appropriate.

Introduction

Mammographic breast density (MBD) is an indicator of risk of breast cancer, with women reported to have a four- to six-fold increase in breast cancer risk if they have extremely dense breasts, compared to women 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 women with dense breast using ultrasound, digital breast tomosynthesis (DBT), magnetic resonance imaging (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 standardisation of breast density assessment.

Radiologist professional bodies have proposed ways of assessing MBD visually. In the United States of America (USA), the American College of Radiology (ACR) developed the breast imaging reporting and data system (BI-RADS®) scheme to provide a standardised categorisation system for reporting MBD. The 4th 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.

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 women's contribution to decision-making, regarding screening for early breast cancer detection, 27 states in the USA now have legislations authorising 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 categorise MBD into 4 categories.¹¹ The RANZCR synoptic scale describes the percentage of glandular tissue for each of the 4 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 inter-observer variability in MBD rating.¹⁵⁻¹⁷ The reported inter and intra-observer variability in MBD assessment using the BI-RADS® system ranged from a Kappa 0.27 to 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 inter-radiologists' agreement in MBD assessment using the RANZCR synoptic

scale. No data exists 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 categorisation of MBD by observers from different countries. Many mammography image readers don't perform this task often and this may lead to a reduced reliability and validity of visual determination of breast density. No study has assessed inter country 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 ABR examiners, UK practitioners, and RANZCR registered radiologists. Such international inter-observer 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 USA radiologists, Australian radiologists and UK practitioners (radiographers). All 20 American Board of Radiology (ABR) examiners were Mammography Quality Standards Act (MQSA) certified, the 26 Australian radiologists were RANZCR certified, and 24 UK radiography practitioners were HCPC 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 was 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 center, and the proposed studies were announced at the information sessions for radiologists. For the Salford data, flyers were posted out to the breast screening centers 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 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 (NHSBSP). Given that radiologist ability to assess density visually has already been determined for radiologists, 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. BIRADS 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, see table 2.

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 women were aged 42-89 years. These images were acquired at a single site in New York, USA, under the same protocol on GE Senographe Essential (or DS) (GE Fairfield, CT and Hologic Lorad Selenia (Hologic Bedford MA) imaging systems one year apart. The images were selected to enable a comparison of the Volpara density grades (VDG) for women whose images were produced one year apart and also to have a comparable number of cases for each of the 4 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 (LCC) followed by a left mediolateral oblique (LMLO) and then the combination of LCC and LMLO presented together. To ensure observers could evaluate the images in 15-20 minute time period, only the left breast images were used for this study. Considering that this dataset contained images of the same women taken one year apart and the MBD assessment scores were obtained within a single sitting, the same observer saw both cases from the same women. Although the observers were presented with the images of the same women, these images were not identical. Positioning changes and equipment changes sufficiently changes images such that they did not appear to be of the same woman. Observers also had no reason to suspect that images of the same women would appear twice. Furthermore the ability to remember an image is related to remarkable aspects of the image³⁰. Additionally, the trace decay theory of forgetting suggests that short-term memory can only hold information between 15-30 seconds unless it is rehearsed³¹. Therefore, the fact that the observers were presented with the images of the same women taken one 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™ version 1.4.3 (Matakina, Wellington, New Zealand) to obtain Volpara Density Grades (VDGs). The preset VDG categories are as follows: VDG 1: <4.5%; VDG 2: 4.5 - 7.5%; VDG 3: 7.6 - 15.5%; 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, five-megapixel display (Tokyo, Japan). This was calibrated to the digital imaging and communications in medicine (DICOM) grayscale standard display function (GSDF) and the user interface was ViewDEX software (Version 2.0).³² The monitor has been shown to demonstrate the required characteristics detailed in the Association of Physicists in Medicine (AAPM) Task Group18 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 standardised, 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. Twenty 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&2), and high (3&4)] for both BI-RADS® and RANZCR. This was to assess the level of inter-observer agreement on a binary scale to provide the potential level of variability with regards to screening individualization.

Data analysis

Statistical analyses were performed using the Statistical Package for Social Sciences (SPSS) Version 21.0 (IBM, Chicago, IL, USA). 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 level of disagreement between observers. A two-way mixed model, which allows for selection of cases randomly and nesting the computation within observers was used to calculate average absolute agreement for all the study cohorts, respectively. The inter-observer agreement for ABR examiners, RANZCR radiologists, and UK practitioners was assessed separately. MBD assessments of these observers were compared in pairs to assess their inter-reader 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 version 2009 (The MathWorks Inc., Natick, MA, United States) 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 4-point (1 - 4) and binary (1&2 vs. 3&4)

classifications scales. 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 figures 1A, 1B and 1C. 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.

Table 3

Figure 1

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 4-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 4-point scale ($\rho = 0.148$; $p < 0.362$). Both groups of observers demonstrated a strong positive correlation on 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 4-point scale ($\rho = 0.916$; $p < 0.001$). A strong positive correlation was also noted on binary scale ($\rho = 0.958$; $p < 0.001$).

Agreement between ABR examiners and RANZCR registered radiologists on 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 concur 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 compared to the Australian radiologists. Of the 40 cases in the dataset, four images rated as BI-RADS® 3 by ABR examiners were rated

RANZCR 1 by RANZCR radiologists, 3 images rated BI-RADS® were rated RANZCR 2 (table 4). The overall agreement (κ_w) between BI-RADS® and RANZCR was 0.010 (95% CI = -1.13 – 0.43).

Table 4

Agreement between ABR 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 concur this is counted as agreement. Again, the ABR examiners provided a higher MBD score compared to 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 5

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 concur this is counted as agreement. UK practitioners classified six cases into a higher MBD category than the Australian radiologists. Four of the cases rated RANZCR 1 were rated BI-RADS® 2 by UK practitioners (table 6). The overall agreement between RANZCR and UK practitioners' BI-RADS® was 0.95 (95% CI = 0.91 – 0.97).

Table 6

Inter-observer agreement for ABR examiners, RANZCR registered radiologists and UK practitioners

Generally, the UK practitioners and RANZCR radiologists tended to call the cases denser than ABR radiologists. Table 7 shows the inter-observer agreement in MBD assessment for each observer cohort. The overall inter-observer agreement among ABR examiners was average [κ_w] = 0.57; 95% CI = 0.52– 0.61] on a 4-point BI-RADS® scale, and ranged from a Kappa of 0.33 to 0.67. On a binary scale, the overall inter-observer agreement (κ_w) was 0.86; 95% CI = 0.82 – 0.87, and ranged from a Kappa of 0.66 to 0.90 (Fig. 2A & 2B).

Inter-observer agreement using RANZCR four-point scale was 0.36 (95% CI = 0.31–0.41), and ranged from 0.078 to 0.499. RANZCR inter-observer agreement on binary scale was substantial [0.71; 95% CI = 0.66 – 0.77], and ranged from 0.22 to 0.89 (Fig. 2C & 2D).

The inter-observer agreement in MBD assessment amongst UK practitioners was 0.47 (95% CI = 0.43 – 0.50) on a 4-point scale, with Kappa values ranging from 0.24 to 0.58. A substantial inter-observer 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 (Fig. 2E & 2F).

Table 7

Figure 2

1 **DISCUSSION**

2 Reproducibility of MBD classification is important given the relevance of MBD
3 information in breast cancer risk assessment and the tailoring of screening
4 methods and frequency. It is important that the same cohort of women imaged
5 under similar conditions have the same opportunity for screening
6 personalisation from MBD assessment. The current work explored the
7 agreement in MBD of the same women assessed using different approaches and
8 by different cohort of observers. Findings demonstrate a wide variability in MBD
9 categorisation between observers, with ABR examiners demonstrating slight
10 agreement with RANZCR radiologists and fair agreement with UK practitioners,
11 and RANZCR radiologists demonstrating almost perfect agreement with UK
12 practitioners. This wide variability was also noted among the same observer
13 cohort, with ABR examiners and UK practitioners each demonstrating moderate
14 inter-observer agreement, and RANZCR radiologists demonstrating fair inter-
15 observer agreement.

16 The inter-observer agreement among RANZCR radiologists was lower than the
17 ABR examiners and UK practitioners. Factors such as years of experience,
18 number of mammograms read per year, training, and the legislation framework
19 governing reporting of breast density, might affect the classification of MBD by
20 observers from different domains. The effect of legislation on MBD reporting has
21 been demonstrated in a recent study, which showed change in the reporting
22 patterns of radiologists after the implementation of density reporting legislation
23 in the USA.³⁷ The study showed that 50% of the observer cohort assigned more
24 cases in BI-RADS® 2 than BI-RADS® 3. The remaining observers (44%) had equal

25 ratings for BI-RADS® 2 and 3 categories.³⁷ It should be noted that MBD
26 legislations aims to facilitate shared decision-making between screened women
27 and their physicians regarding adjunctive screening. Some radiologists grade
28 MBD, taking into consideration age and clinical history of the patient.³⁷ Hence it
29 is logical that these factors may significantly impact upon inter-reader variability
30 in MBD assessment. This finding suggests that MBD classification may be
31 influenced by systems requirement, legislation, and individual perception of the
32 potential impact of breast density. Therefore, further work should investigate
33 whether these factors are associated with the wide international variability in
34 the MBD classification of the same patient cohort observed in the current work.

35 A recent study reported a 32.4% disagreement between a pair of radiologists³⁸,
36 and suggested that this level of disagreement limits use of qualitative
37 assessments for recommending additional screening and risk management of
38 women with dense breasts.³⁸ The current study demonstrates a 70%
39 disagreement between ABR examiners and Australian radiologists, and 62%
40 disagreement between ABR examiners and UK practitioners. The lowest
41 disagreement was reported between Australian radiologists and UK
42 practitioners (20%). These levels of disagreement are likely to change or
43 influence the individualised screening of women with dense breasts. The
44 disagreement particularly becomes crucial when it affects the categorisation that
45 differentiates low (1 & 2) from high (3 & 4) MBD categories. This is because it
46 determines the category of women who are likely to be referred for additional
47 imaging with ultrasound or MRI.^{12, 39} Encouragingly, the level of disagreement
48 observed on a binary scale in the current study was less compared to that

49 observed on a four-point scale. Nevertheless, the level of inter-observer
50 disagreement on a binary scale was still appreciable, and underscores the need
51 for standardisation of breast density assessment. This may require the
52 introduction of automated MBD assessment techniques in all screening
53 programmes to more appropriately tailor adjunctive imaging and screening
54 intervals for women with dense breasts.

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

64 Considering that the 4th edition BI-RADS® and RANZCR classify breast density
65 according to the same percentages (table 1), it is logical that they would
66 demonstrate a good level of agreement in the same patient cohort. The wide
67 level of inter-reader and inter-country variability observed in the our study is a
68 cause for concern, and shows perhaps the lack of understanding of, or adherence
69 guidelines for MBD assessment. It is unclear whether the negative correlation
70 observed between ABR and RANZCR radiologists is due to prevalence
71 expectation, where observers are accustomed to a certain MBD grade. Additional
72 training for further assessment of performance could be beneficial. Our findings

73 show how the same women cohort could be classified into different risk strata,
74 and screening regimen and pathways in different countries and among
75 observers, thus limiting consistency in clinical use of MBD.

76 The BI-RADS® system originated in the USA, and no difference has been shown
77 in the range of inter-observer agreement for studies based in the USA versus
78 outside the USA.⁴² However, it is possible that inter-regional or inter-country
79 differences in visual MBD assessment approaches would cause variation in MBD
80 rating of the same woman as demonstrated in the current work. There is
81 evidence that visual assessment of MBD has wide inter-reader disagreement.^{16,}
82 ^{40, 43, 44} This variability was observed in previous work with inter-reader
83 agreement (κ) ranging from 0.328 to 0.669 and 0.078 – 0.499 respectively.^{13,39}
84 Given the reported intra and inter-observer variation with other visual
85 assessment methods such as BI-RADS®⁴⁵⁻⁴⁸, there is a need to determine the
86 range of agreement that can be expected between countries using the same
87 criteria for MBD assessment. Importantly, the current study has provided insight
88 to the level of variability in MBD assessment between observers from different
89 practices and countries.

90 The strengths of our study include the large number of observers from different
91 backgrounds. Secondly, this is the first international assessment of inter-
92 observer agreement in MBD assessment using BI-RADS® and RANZCR synoptic
93 scales. Data provided show for the first time how MBD of the same cohort of
94 women can be classified differently by observers from different domains. There
95 were several limitations to the study. Only the left breast was used for BI-RADS®
96 and RANZCR assessment. It is possible that including the right breast may have

97 affected the results presented in this study. The observers may not have been
98 familiar with the presentation state of the images and may be used to a different
99 look. This may have affected their conclusions on density, Even-though
100 Volpara™ was used as the 'ground truth' for all the cohorts of the study, BI-
101 RADS® and RANZCR scales are not designed to be exactly the same as Volpara™.
102 Therefore, the disagreement shown between these scales and Volpara™ might
103 be expected. Furthermore, observers are familiar with using BI-RADS® and
104 RANZCR density assessment scales therefore inter-country differences are also
105 expected. Previous studies found in-country variations between observers.^{12, 39}
106 Therefore, different observers are likely to see different patient populations even
107 compared to their in-country colleagues as all these assessment methods are
108 using a four point scale. The UK cohort for the current study comprised of
109 radiographers, therefore further work will investigate international inter-
110 observer comparisons for UK radiologists.

111 **Conclusion**

112 Data produced demonstrate wide international and inter-observer disagreement
113 in MBD assessment. In particular, the findings show poor agreement between
114 ABR examiners and RANZCR and UK mammography image readers. The findings
115 also showed moderate inter-observer agreement in MBD assessment among ABR
116 radiologists, fair agreement amongst RANZCR radiologists, and moderate
117 agreement amongst UK practitioners. The findings emphasise the need to
118 improve reproducibility of MBD classification internationally in order to improve
119 risk stratification and more appropriately tailor screening in women with dense
120 breast.

121

122 **References**

123

124 1. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic
125 density and the risk and detection of breast cancer. *N Engl J Med.*
126 2007;356(3):227-36.

127 2. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et
128 al. Breast density as a predictor of mammographic detection: comparison of
129 interval- and screen-detected cancers. *Journal of the National Cancer Institute.*
130 2000;92(13):1081-7.

131 3. McCormack VA, dos Santos Silva I. Breast density and parenchymal
132 patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiology,*
133 *Biomarkers & Prevention.* 2006;15(6):1159-69.

134 4. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast
135 cancer risk: current understanding and future prospects. *Breast Cancer Res.*
136 2011;13(6):223.

137 5. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Velez M,
138 et al. Combined screening with ultrasound and mammography vs mammography
139 alone in women at elevated risk of breast cancer. *Jama-Journal of the American*
140 *Medical Association.* 2008;299(18):2151-63.

141 6. Berg WA, Zhang Z, Lehrer D, Jong RA, Pisano ED, Barr RG, et al. Detection
142 of Breast Cancer With Addition of Annual Screening Ultrasound or a Single
143 Screening MRI to Mammography in Women With Elevated Breast Cancer Risk.
144 *Jama-Journal of the American Medical Association.* 2012;307(13):1394-404.

- 145 7. ACRIN 6666: SCREENING BREAST ULTRASOUND IN HIGH-RISK WOMEN
146 [database on the Internet]2007 [cited 23 June]. Available from:
147 <http://www.acrin.org/Portals/0/Protocols/6666/Protocol-ACRIN>.
- 148 8. Radiology ACo. BI-RADS Mammography 2013-ACR BI-RADS Atlas, 5th
149 Edition. Reston, VA: American College of Radiology,2013.; [cited 2014 17 March
150]; Available from: <http://www.acr.org/Quality-Safety/Resources/BIRADS>.
- 151 9. Ekpo EU, Hogg P, Highnam R, McEntee MF. Breast composition:
152 Measurement and clinical use. Radiography. 2015;21(4):324-33.
- 153 10. Durning MV. Breast Density Notification Laws State-Interactive Map.
154 [cited 2016 01 December]; Available from:
155 [http://www.diagnosticimaging.com/breast-imaging/breast-density-](http://www.diagnosticimaging.com/breast-imaging/breast-density-notification-laws-state-interactive-map)
156 [notification-laws-state-interactive-map](http://www.diagnosticimaging.com/breast-imaging/breast-density-notification-laws-state-interactive-map).
- 157 11. 2007 NBCC. "Synoptic breast imaging report:. 2014 [cited 2014 20
158 August]; Available from: [http://canceraustralia.nbocc.org.au/view-document-](http://canceraustralia.nbocc.org.au/view-document-details/rsig-1-synoptic-breast-imaging-report-update)
159 [details/rsig-1-synoptic-breast-imaging-report-update](http://canceraustralia.nbocc.org.au/view-document-details/rsig-1-synoptic-breast-imaging-report-update).
- 160 12. Damases CN, Mello-Thoms C, McEntee MF. Inter-observer variability in
161 mammographic density assessment using Royal Australian and New Zealand
162 College of Radiologists (RANZCR) synoptic scales. Journal of medical imaging and
163 radiation oncology. 2016;60(3):329-36.
- 164 13. Damases CN, Mello-Thoms C, McEntee MF, editors. Inter-observer
165 variability within BI-RADS and RANZCR mammographic density assessment
166 schemes. Medical Imaging 2016: Image Perception, Observer Performance, and
167 Technology Assessment; 2016; San Diego.
- 168 14. McCormack VA, Highnam R, Perry N, Silva ID. Comparison of a new and
169 existing method of mammographic density measurement: Intramethod

170 reliability and associations with known risk factors. *Cancer Epidemiology*
171 *Biomarkers & Prevention*. 2007;16(6):1148-54.

172 15. Ekpo EU, McEntee MF. Measurement of breast density with Digital Breast
173 Tomosynthesis- a systematic review. *Br J Radiol*. 2014:20140460.

174 16. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, et al.
175 Categorizing breast mammographic density: intra- and interobserver
176 reproducibility of BI-RADS density categories. *Breast*. 2005;14(4):269-75.

177 17. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of
178 Interradiologist Agreement Regarding Mammographic Breast Density
179 Classification Using the Fifth Edition of the BI-RADS Atlas. *AJR Am J Roentgenol*.
180 2016:1-5.

181 18. Bernardi D, Pellegrini M, Di Michele S, Tuttobene P, Fanto C, Valentini M,
182 et al. Interobserver agreement in breast radiological density attribution
183 according to BI-RADS quantitative classification. *Radiol Med*. 2012;117(4):519-
184 28.

185 19. Littlefair S, Mello-Thoms C, Reed W, Pietryzk M, Lewis S, McEntee M, et al.
186 Increasing Prevalence Expectation in Thoracic Radiology Leads to Overcall.
187 *Academic Radiology*. 2016;23(3):284-89.

188 20. Reed WM, Chow SLC, Chew LE, Brennan PC. Assessing the Impact of
189 Prevalence Expectations on Radiologists' Behavior. *Academic Radiology*.
190 2014;21(9):1220-21.

191 21. Reed WM, Chow SLC, Chew LE, Brennan PC. Can Prevalence Expectations
192 Drive Radiologists' Behavior? *Academic Radiology*. 2014;21(4):450-56.

- 193 22. Gur D, Rockette HE, Warfel T, Lacomis JM, Fuhrman CR. From the
194 laboratory to the clinic: The "prevalence effect". *Academic Radiology*.
195 2003;10(11):1324-26.
- 196 23. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The
197 prevalence effect in a laboratory environment: Changing the confidence ratings.
198 *Academic Radiology*. 2007;14(1):49-53.
- 199 24. Reed WM, Ryan JT, McEntee MF, Evanoff MG, Brennan PC. The Effect of
200 Abnormality-Prevalence Expectation on Expert Observer Performance and
201 Visual Search. *Radiology*. 2011;258(3):938-43.
- 202 25. Wivell G, Denton ERE, Eve CB, Inglis JC, Harvey I. Can Radiographers Read
203 Screening Mammograms? . *Clin Radiol*. 2003;58(1):63-67.
- 204 26. Pauli R, Hammond S, Cooke J, Ansell J. Comparison of
205 radiographer/radiologist double film reading with single reading in breast
206 cancer screening. *Journal of Medical Screening*. 1996;3:18-22.
- 207 27. Moran S, Warren-Forward H. A retrospective pilot study of the
208 performance of mammographers in interpreting screening mammograms.
209 *Radiographer: The Official Journal of the Australian Institute of Radiography*.
210 2010;57(1):12.
- 211 28. Bennett RL, Sellars SJ, Blanks RG, Moss SM. An observational study to
212 evaluate the performance of units using two radiographers to read screening
213 mammograms. *Clin Radiol*. 2012;67(2):114-21.
- 214 29. Pauli R, Hammond S, Cooke J, Ansell J. Radiographers as film readers in
215 screening mammography: an assessment of competence under test and
216 screening conditions. *The British Journal of Radiology*. 1996;69(817):10-14.

- 217 30. Ryan JT, Haygood TM, Yamal JM, Evanoff M, O'Sullivan P, McEntee M, et al.
218 The "Memory Effect" for Repeated Radiologic Observations. *American Journal of*
219 *Roentgenology*. 2011;197(6):W985-W91.
- 220 31. Brown J. Some tests of the decay theory of immediate memory. *Quarterly*
221 *Journal of Experimental Psychology*. 1958;10(1):12-21.
- 222 32. Börjesson S, Håkansson M, Båth M, Kheddache S, Svensson S, Tingberg A,
223 et al. A software tool for increased efficiency in observer performance studies in
224 radiology. *Radiat Prot Dosimetry*. 2005;114(1-3):45-52.
- 225 33. Ekpo EU, McEntee MF. An Evaluation of Performance Characteristics of
226 Primary Display Devices. *J Digit Imaging*. 2015.
- 227 34. Brennan PC, McEntee M, Evanoff M, Phillips P, O'Connor WT, Manning DJ.
228 Ambient lighting: Effect of illumination on soft-copy viewing of radiographs of
229 the wrist. *American Journal of Roentgenology*. 2007;188(2):W177-W80.
- 230 35. Random Integer Generator. [cited 2016 23 December]; Available from:
231 <https://www.random.org/integers/>.
- 232 36. D'Orsi. C.J, Bassett L, Berg W. *BI-RADS Mammography in, 4th edition:*
233 *D,Orsi CJ, Mendelson FB, Ikeda DM et al: Breast Imaging and Reporting and Data*
234 *System: ACR BI-RADS-Breast Imaging Atlas*. 4th ed. Reston, VA: American College
235 of Radiology; 2003.
- 236 37. Gur D, Klynn AH, King JL, Bandos AI, Sumkin JH. Impact of the New
237 Density Reporting Laws: Radiologist Perceptions and Actual Behavior. *Academic*
238 *Radiology*. 2015;22(6):679-83.
- 239 38. van der Waal D, den Heeten GJ, Pijnappel RM, Schuur KH, Timmers JMH,
240 Verbeek ALM, et al. Comparing Visually Assessed BI-RADS Breast Density and

241 Automated Volumetric Breast Density Software: A Cross-Sectional Study in a
242 Breast Cancer Screening Setting. Plos One. 2015;10(9).

243 39. Damases CN, Brennan PC, Mello-Thoms C, McEntee MF. Mammographic
244 Breast Density Assessment Using Automated Volumetric Software and Breast
245 Imaging Reporting and Data System (BIRADS) Categorization by Expert
246 Radiologists. Academic Radiology. 2016;23(1):70-77.

247 40. Redondo A, Comas M, Macia F, Ferrer F, Murta-Nascimento C, Maristany
248 MT, et al. Inter- and intraradiologist variability in the BI-RADS assessment and
249 breast density categories for screening mammograms. Br J Radiol.
250 2012;85(1019):1465-70.

251 41. Masroor I, Rasool M, Saeed SA, Sohail S. To asses inter- and intra-observer
252 variability for breast density and BIRADS assessment categories in
253 mammographic reporting. Journal of the Pakistan Medical Association.
254 2016;66(2):194-97.

255 42. Antonio ALM, Crespi CM. Predictors of interobserver agreement in breast
256 imaging using the Breast Imaging Reporting and Data System. Breast Cancer
257 Research and Treatment. 2010;120(3):539-46.

258 43. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in
259 radiologists interpretations of mammograms. New England Journal of Medicine.
260 1994;331(22):1493-99.

261 44. Abdolell M, Tsuruda K, Payne JI, Schaller G, Iles SE, Lightfoot CB, et al.
262 Agreeing to disagree: assessing inter-rater variation in breast density
263 measurement. In: Radiology ESo, editor. European Congress of Radiology Vienna,
264 Austria2014. p. 1-9.

265 45. Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH, Rutter CM.
266 Misclassification of Breast Imaging Reporting and Data System (BI-RADS)
267 Mammographic Density and Implications for Breast Density Reporting
268 Legislation. The Breast Journal. 2015;21(5):481-89.

269 46. Gweon HM, Youk JH, Kim JA, Son EJ. Radiologist Assessment of Breast
270 Density by BI-RADS Categories Versus Fully Automated Volumetric Assessment.
271 American Journal of Roentgenology. 2013;201(3):692-97.

272 47. Sauber N, Chan A, Highnam R. BI-RADS breast density classification - an
273 international standard. ECR; 2013.

274 48. Wang K, Chan A, Highnam R. Robustness of automated volumetric breast
275 density estimation for assessing temporal changes in breast density. ECR; 2015.

276

277 **Figures:**

278 **Figure 1:** Percentage distribution of cases assigned into MBD categories by each
279 ABR examiner (A), RANZCR registered Australian radiologists (B), and UK
280 practitioners (C).

281 **Figure 2:** Inter-observer agreement for MBD assessment using BI-RADS® and
282 RANZCR scales. (A) Shows the ABR examiners' agreement on BI-RADS® four-
283 point scale and (B) shows the agreement on BI-RADS® binary scale. (C) Shows
284 the RANZCR radiologists' agreement on RANZCR four-point scale and (D) Shows
285 the agreement on RANZCR binary scale. (E) Shows the UK practitioners'
286 agreement on BI-RADS® four-point scale and (F) shows the agreement on BI-
287 RADS® binary scale.