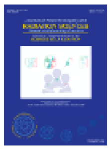
Journal of Medical Imaging and Radiation Sciences

Available online 5 January 2019

In Press, Corrected Proof



Variability of Breast Density Classification Between US and UK Radiologists

Wijdan Alomaim PhD [a], Desiree O'Leary PhD [b], John Ryan PhD [a], Louise Rainford PhD [a], Michael Evanoff PhD [c], Shane Foley PhD [a]

Show more

https://doi.org/10.1016/j.jmir.2018.11.002

INTRODUCTION:

Women with extremely dense breast tissue are at 4-6 times greater risk of developing breast cancer than women with fatty breast[1], meaning density, which can only be judged based on imaging, is an important factor in breast cancer prediction models[2]. Additionally, images with high mammographic density are difficult to evaluate due to the fact that density limits sensitivity and specificity when detecting lesions[3]. This has led to initiatives to include the breast density category as part of the mammographic report as an indicator of test sensitivity and/or to guide decisions regarding supplemental imaging[4,5]. Many studies have found that BI-RADS categorisation is prone to inconsistencies between radiologists[6,7]. This is due to the fact that the categorization system is based on readers' subjective evaluation of two-dimensional imaging[8–10]. This inconsistency increases the concern that a single patient may receive different breast density categorisation between screenings or for different patients with similar density to undergo differing diagnostic procedures. This has potential to have a number of adverse impacts, firstly on patients' experience, by increasing anxiety if incorrectly informed that the mammogram sensitivity is reduced or by putting them through additional unnecessary further imaging[9–11], while providing a lower category of density could create a false sense of security for this group[9,10]. Moreover, the use of supplemental screening requires additional resources, for which there is inconsistent insurance coverage and thus disparity of health care services[12]. Furthermore, a consistent breast density assessment is beneficial for both recognising patients individual breast cancer risk[13–15], as well as identifying dense breast patients who would possibly benefit from supplemental screening methods[12]. Therefore, it is timely to establish inter rater variability, internationally.

Ciatto et al., (2005)[6] has suggested the use of a two-scale breast density category (non-dense and dense) that would increase agreement levels between radiologists. This two-scale would aid in identifying women with medium to high risk of cancer being obscured by dense tissue[6]. Additionally, it provides the benefit of improving image readers underlying interpretation and performance, with establishing the image readers robustness[16]. Currently there are a number of fully automated software in use[17–19], however, despite extensive research these methods are not fully adopted in practice.

In the USA, 27 states have introduced legislative requirements not only to report density but also to notify patients of their density category[20,21]. There are no such requirements in Europe at present[22]. Literature states, that the most widely used breast density classification is Breast Imaging Reporting and Data system (BI-RADS)[23], which was initially developed by American College of Radiology (ACR)[24], to standardise reporting and minimize the uncertainty in the interpretations and management of recommendations[17]. BI-RADS classification has four categories based on the overall estimation of the percentage of fibrograndular tissue within the breast. In 2013, the BI-RADS 5th edition was released, which became more subjective in design with four-categories of breast composition of fibrograndular tissue[17]. BI-RADS is used both in the USA and Europe[23].

Multiple previous studies have shown a range of inter rater variability in categorising breast density[6,7,25]. However, these studies did not include radiologists from different countries and practices. To our knowledge, only one recently published study[26] has undertaken intercountry study, however, this study only used left breast images and a small number of images. Moreover, the UK image readers in Damases et al., (2017)[26], only involved radiographer image readers. Therefore this study endeavours to address this deficiency, by examining the inter rater variability in categorising breast density with a larger number of images, (including right and left breast images) and a large group of radiologists from two jurisdictions which have differing legal requirements related to breast density categorization. This could aid in enhancing the possible understanding of the causes of variation if it exists. This work will investigate if subjective BI-RADS remains a feasible way to categorise breast density or whether the two-scale category (non-dense and dense) is preferred. A standardised, reproducible breast density assessment would be beneficial in contributing to improved breast cancer risk stratification, and in customising breast screening for females with dense breasts more appropriately, worldwide.

METHODOLOGY:

The required ethical approvals were confirmed by the institutional Human Research Ethics committee. Permission was also granted by the American Board of Radiology to undertake the study with their expert examiners, and by the British Society of Breast Radiology (BSBR) for data collection at their Annual Scientific Meeting through voluntary enrolment.

250 fully anonymised digital mammographic cases, which included 180 cases, plus 70 repeated cases were used to facilitate inter and intra observer reliability analysis. Each case included four images: "Right and Left Mediolateral-Obliques and Cranial-Caudal projections". These cases were gathered with full patient consent from 18 units in a national breast screening programme as part of previous research. These cases were selected via consensus by two mammography researchers (WA / DOL) and were categorized using the Hand Delineation breast density assessment method. Hand Delineation method was developed by Byng et al., (1994)[27] and was performed in this study by a single researcher (WA) following the McCormack et al, (2007)[28] and Li et al, (2012)[29] methodology, where the interpreter can recognize the boundaries of the breast tissue and mark the threshold for dense tissue on the mammogram. The measurement of the percentage density was calculated from the values provided (dense area/total breast area)[29] and the values were converted into BI-RADS. Moreover, according to these studies[28,29] the Hand Delineation method is considered to be the gold standard or ground truth in assessing breast density[28,29].

Images were selected with the least or none of the following artefacts; distracting pathology, mal-positioning, technique factors and exposure factors errors, asymmetrical breast tissue and asymmetrical breast size between the left and right breast. Although, one mammogram per set did have one of the above to provide a challenge to test radiologists' consistency. According to Ko et al., (2014)[30], asymmetry of breast size and pathology are factors causing possible disagreement between the breast density assessment methods[22]. The

250 cases were divided into five sets, each set included randomly displayed 36 cases and 14 repeated cases for intra and inter-rater reliability analysis.

The density distribution within each set was not equal to avoid increasing the radiologists' sense of predictability, as shown in Table 1. Furthermore, 50 cases per set were deemed reasonable, time permitting, with radiologists given the option to read more than one set where possible, while ensuring the power of the study would still exceed 80%. According to literature, 30 cases and at least three radiologists are the minimum requirements for an accurate statistical analysis regarding inter-observer agreement level[7,31,32]. Study power was calculated using R Package 'KappaSize'.

Table 1: The distribution of BI-RADS categories for the repeated images and within each set of images.

	Image Sets					
Breast Density	**A**	**B**	**C**	**D**	**E**	**Repeated Images**
BI-RADS 1	28%	26%	22%	22%	34%	29%
BI-RADS 2	22%	20%	26%	26%	22%	35%
BI-RADS 3	32%	42%	36%	30%	30%	29%
BI-RADS 4	18%	12%	16%	22%	14%	7%

In each location participants were recruited via local advertising at an ABR examination event, Kentucky, USA and at the BSBR conference, London UK. Both USA and UK participants' years of experience reporting breast images were recorded. The UK participants were asked whether they were breast radiologists or mammographers, however, all the participants were radiologists.

For both cohorts, images were displayed using Ziltron software (Ziltron Ltd., Dublin)[33], which facilitated pan/zoom as well as rapid image brightness/contrast alteration. Furthermore, both were given an instruction sheet containing study information and details on the use of Ziltron software.

In the USA, radiologists reviewed images on two computer screens, ViewSonic ViewPanel (Viewsonic Corporation, Brea, CA), VP201mb with 1200*1600 pixel resolution, each with 20" full viewable diagonal area, oriented in portrait position. For the UK radiologists, images were presented on a single monitor, 23” TFT TOBII eye-tracker computer screen (TOBII Technology, Stockholm), 1920*1080 pixels resolution. As observers were not seeking pathology per se and only categorizing overall density, screen resolution was deemed acceptable for this purpose[34]. Quality assurance testing was performed using the Digital Imaging and Communications in Medicine (DICOM). Part 14: Grayscale Standard Display Function (GSDF) using VeriLUM calibration software and luminance pod (IMAGE Smiths Inc., Germantown, Maryland), to ensure all screens met the standard range[35,36].

Statistical analysis:

Weighted Kappa (κw), (95% confidence interval)[37], was performed to assess the inter-rater reliability between radiologists’ assessment of breast density category and for each BI-RADS category, within each set. Prior to performing this test, the mode was calculated for each cohort (the majority reported by radiologists), and in cases where there was no majority report of breast density category the answer was rounded off to the next BI-RADS category[7], creating two groups (USA, UK), for comparison. Further work was completed, by discriminating the importance of the ratings (BI-RADS category) given by each radiologist, by using their experience as the weight of ratings when computing the median for each image for UK radiologists. However, as the experience for the USA radiologists are the same, there was no need to weight.

Fleiss' Kappa (95% confidence interval) was used to assess the level of agreement within each cohort, individually[38]. The Intra-class Correlation Coefficient (ICC) was calculated to determine intra-rater reliability[39,40]. κw was used to assess the level of agreement between radiologists for the two-grade scale, BI-RADS 1 and 2 as (low-density) and BI-RADS 3 and 4 as (high-density).

The interpretation of the Kappa agreement levels for categorical data, <0 Poor, 0.01–0.20 Slight, 0.21–0.40 Fair, 0.41–0.60 Moderate, 0.61–0.80 Substantial, 0.81–1.00 Almost perfect agreement[41].

RESULTS:

A total of 49 radiologists participated, 25 USA breast radiologists all with more than 10 years of breast imaging reporting experience, while of the 24 UK breast radiologists, 29% had three years or less experience, 33% from four to nine years and only 38% of the cohort had more than 10 years' experience. The power of the sample size for 25 USA radiologists and 24 UK radiologists reviewing 180 images was calculated to be in excess of 97%. The percentage of radiologists from both cohorts that reviewed more than one set were USA 84% and UK 8%, as per Figure 1.

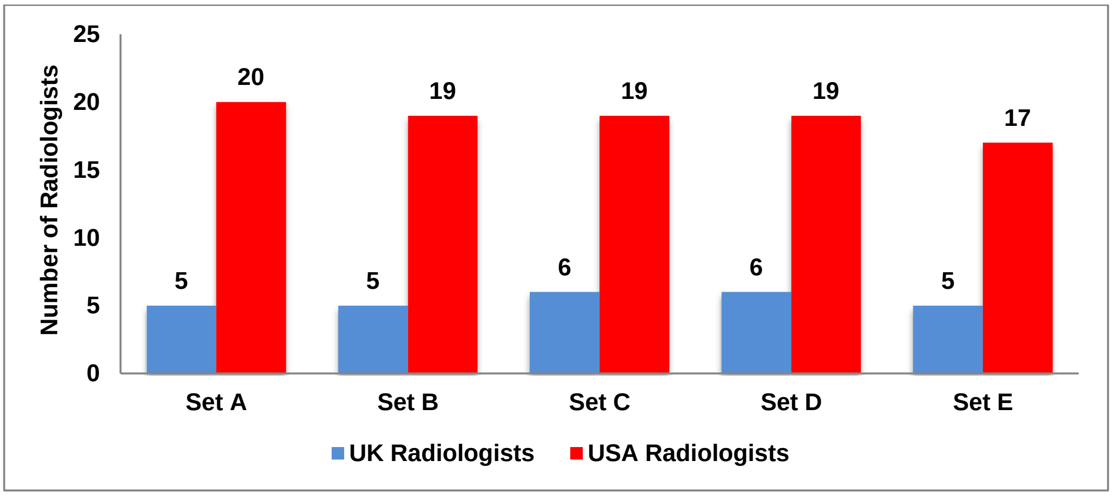


Figure 1. Number of radiologists who assessed each set of images.

Viewing conditions were monitored and luminance levels for the USA and UK study screens' are presented as per Table 2 below.

Table 2: Screen luminance levels.

Luminance levels (candela per square metre [cd/m^2])	USA		UK
	Screen 1	Screen 2	Screen
Maximum	164.2	175.5	300.0
Minimum	0.35	0.31	0.67

Mammographic density subjective assessments:

Overall agreement for all sets was substantial (κw=0.760), between USA and UK radiologists. When data were split into sets, agreement varied from substantial to high agreement with significant p values for each (p<0.001), Set A (κw=0.831), B (κw=0.819), C (κw=0.685), D (κw=0.771) and E (κw=0.696).

When the BI-RADS were weighted according to the radiologists experience, the overall agreement for all sets was substantial (κw=0.747), between USA and UK radiologists. When data were split into image sets, agreement varied from substantial to high agreement with significant p values for each (p<0.001), Set A (κw=0.831), B (κw=0.803), C (κw=0.760), D (κw=0.724) and E (κw=0.611).

The agreement level between the USA and UK radiologists when the data were split into BI-RADS categories was statistically significant and varied from fair to substantial agreement, BI-RADS 1 (κw=0.352), BI-RADS 2 (κw=0.327), BI-RADS 3 (κw=0.715) and BI-RADS 4 (κw=0.681).

The level of agreement between the USA and UK radiologists for the weighted results according to the radiologists experience when the data were split into BI-RADS categories was statistically significant and varied from fair to substantial agreement, BI-RADS 1 (κw=0.374), BI-RADS 2 (κw=0.501), BI-RADS 3 (κw=0.684) and BI-RADS 4 (κw=0.635).

The distribution of BI-RADS scores resulting from USA and UK radiologists, across the five sets of images identified USA radiologists as categorising fewer images as mostly fatty BI-RADS 1 compared to UK radiologists. In summary the USA radiologists classified a greater number of images in the higher categories, in particular in BI-RADS 3 category (heterogeneously dense), as per Figure 2.

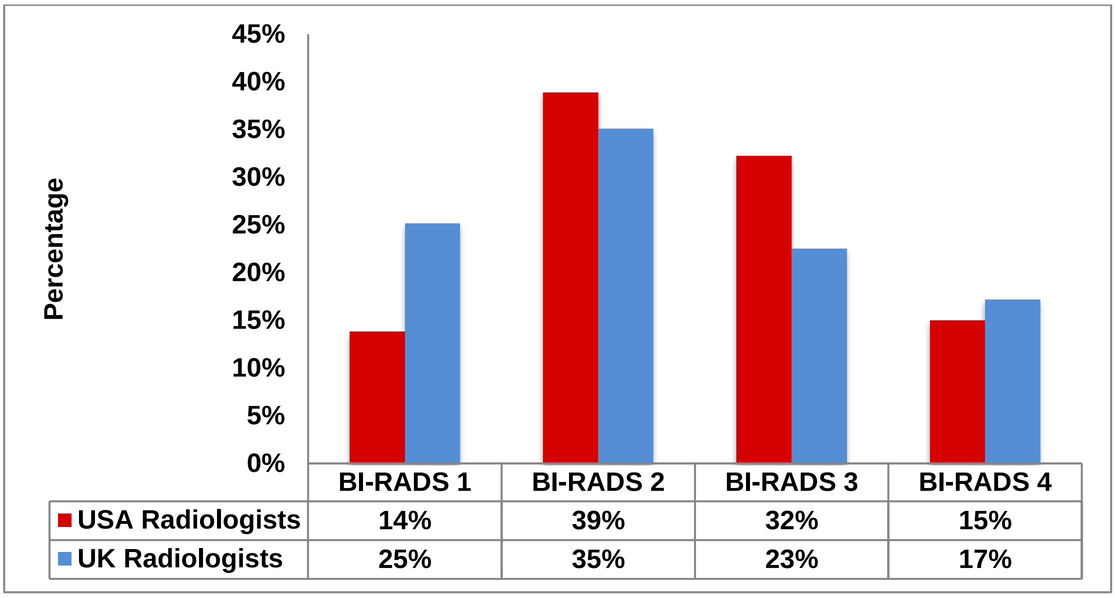


Figure 2. Distribution of BI-RADS scoring among USA, UK radiologists.

The overall agreement level within each cohort, for the USA radiologists was found to be substantial, and for the UK radiologists the agreement level was moderate. When the data were split into BI-RADS categories, the USA radiologists' agreement level increases with the higher categories. Meanwhile, the UK radiologists level of agreement was less on the middle BI-RADS (2 and 3), as per Table 3.

Table 3: Fleiss' Kappa results showing agreement level within each cohort individually, overall agreement and level of agreement when data were split into BI-RADS categories.

ACR Categories	USA Radiologists	UK Radiologists
BI-RADS 1	0.480*	0.563*
BI-RADS 2	0.552*	0.429*
BI-RADS 3	0.682*	0.416*
BI-RADS 4	0.850*	0.684*
All	0.629*	0.502*

* Statistical significance ($p<0.001$)

The intra-class correlation coefficient agreement for intra-rater reliability for the radiologists in both countries on the repeated images within all 5 sets was high (ICC >0.9), being 0.973 for USA radiologists (CI: 0.966 to 0.978 ($F(219,876)=36.974$, $p<0.001$) and 0.927 for UK radiologists (CI: 0.821 to 0.975 ($F(13,26)=15.194$, $p<0.001$).

The level of agreement between both cohorts for the two-grade scale, the κw agreement achieved almost perfect agreement (0.845, $p<0.001$).

The median of all image readers from USA and UK was compared to the used ground truth (Hand Delineation), the overall agreement was substantial and significant (0.680, $p<0.001$).

DISCUSSION:

Many studies have been undertaken to test inter-observer variability in assessing breast density. They vary in the number of radiologists included, the methodology employed, as well as the results[6,25,26,42–44]. However, this study sought to explore the establishment of the variation in two countries, USA and UK, using a greater number of expert radiologists reviewing a larger number of images compared to previous studies. This methodology allowed for a contrast of jurisdictions, with USA radiologists working under breast density legislation[20,45], while UK radiologists have no legal requirement to report breast density[22]. Furthermore, UK radiologists use a three-point scale, (fatty, mixed and dense)[46], in comparison to the four scale BI-RADS used by USA radiologists. While the UK radiologists were not given formal training in BI-RADS for this research, they were actively participating in an educational breast imaging event at the time of their participation with multiple sessions on breast density (ACR BI-RADS) so this difference in categorisation was clear. On the other hand, USA radiologists were specialised breast imaging examiners at an ABR exam sitting.

This study confirms that radiology inter-rater variability in categorising breast density using the universal BI-RADS system exists across geographic regions, as the overall results indicated a substantial agreement, which ranged from substantial to high agreement, even when the data was weighted according to the radiologists' experience, similar results were found. This could be due to variable perception on the part of radiologists, and can be improved by training based on standards and reference images[3]. However, this study's agreement levels are higher than previously reported studies which did not incorporate observer training, either oral or Atlas BI-RADS instruction[6,26,42–44] but, similar to studies that included clear BI-RADS instructions for participants[7]. This could be due to enrolment bias, as at the time of the research activity both radiology cohorts were involved in specialised mammography focussed activity and breast density was considered an important topical issue.

When the data were divided into categories, surprisingly, it was found that the least agreement after BI-RADS 2 (0.327), is in BI-RADS 1 (0.352). For BI-RADS 1, this was also confirmed when the data was weighted according to the radiologists experience. However, BI-RADS 2 agreement increased to moderate agreement. BI-RADS 1 results were not expected, as it's considered to be a straightforward categorisation due to the presence of mostly fatty breast tissue with minimal density. According to previous studies the agreement level for BI-RADS 1 ranged from moderate to substantial agreement (0.51, 0.54 and

0.76)[6,7,42]. This variation was also noticed on the BI-RADS scoring distribution between the two cohorts, as UK radiologists classified breasts as mostly fatty in almost twice the number of mammograms as compared to USA radiologists. On the other hand, USA radiologists placed the majority of the images in BI-RADS 2 and 3 classifications. This dissimilarity and lower agreement of BI-RADS 1 between the two cohorts may be due to either the American breast density legislation or participant experience levels, given that the USA cohort were a more experienced group. Additionally, this variation could be due to the distribution of the breast density BI-RADS within the sets of images, which are not necessarily representative of the radiologists' home country population. While the viewing environments, in particular the background lighting, differed between cohorts, display monitors for each were comparable quality and unlikely to have resulted in these differences. Moreover, this variability might be different if the selection and categorization of the images method was different. However, when the median of all image readers from USA and UK was compared to the used ground truth the overall agreement was substantial, which support the validity of the used method. While other studies have used expert opinion, this study incorporated the acknowledged gold standard for breast density assessment, Hand Delineation[28 29]. However, Hand Delineation is not clinically suitable as its time consuming. Therefore, to further support the study results the level of agreement between the USA and UK radiologists was tested when BI-RADS were weighted according to the radiologists experience.

Gubern-Mérida et al., (2014)[47] and Sauber et al., (2013)[48] have both suggested that, according to the European standard in using BI-RADS breast density categorisation, European radiologists are underestimating breast density compared to the USA radiologists[47,48]. Our study adds to the evidence base for such differences, while using a larger number of radiologists and images compared to the Gubern-Mérida et al, (2014)[47] study, which involved only a single radiologist, and used a larger data set of mammographic images than the Sauber et al, (2013)[48] study, which consisted of a small data set containing 12 images[47,48].

According to previous studies the largest differences were noticed in BI-RADS 2 and 3[6,7]. However, in this study the level of agreement for BI-RADS 3 was higher compared to Ciatto et al, (2005)[6], Ooms et al, (2007)[7] and Berg et al, (2000)[42] studies. In the case of BI-RADS 2 and 3, where a difference occurs between the percentage agreement and Weighted Kappa (κw) test results. The percentage agreement was calculated only on images categorised under each BI-RADS, regardless of whether all radiologists were in agreement on individual images, in this particular categorisation. Additionally, κw test was calculated to determine agreement level among radiologists. This led the level of agreement to be different than the distribution of BI-RADS 2 and 3 scores. The reasoning behind this is well represented by Ko et al., (2014)[30] study which suggested that radiologists tend to give BI-RADS 3 category to the images with a high concentration of tissue in some areas of the breast even if this concentration is (<50%) of total breast volume[30]. On the other hand, if the fibroglandular tissue is uniformly distributed within the breast, the radiologist may report it as non-dense, even though it measures as (>50%) of total breast volume. Radiologists may believe that the scattered tissue would not minimize the sensitivity of mammography[30]. As a way to reduce this variation the updated 5th edition of the ACR BI-RADS atlas recommends that any high dense area within the breast that might obscure any small mass and (<50%), should be categorised as BI-RADS c or 3[24].

Finally, in this study, the overall level of agreement and agreement level for both cohorts individually on BI-RADS 4 classification demonstrated substantially strong to almost perfect agreement, as anticipated, in agreement with Sprague et al., (2016)[25]. Additionally, when the level of agreement was tested between each cohort individually, it was found that USA radiologists level of agreement was substantial, which is close to Ooms et al, (2007)[7] results, where radiologists received instructions in the form of a set of reference images of BI-RADS density categories, however, USA radiologists are practicing breast density categorisation on a daily basis. This effect may be a consequence of the introduction of breast density legislation in the USA. Radiologists may upgrade breast density and recommend further imaging to minimize any liability if a cancer was missed[49]. The findings indicate mandatory reporting of breast density is potentially impacting on clinical decision-making compared to UK radiologists. The UK radiologists' level of agreement in comparison to USA findings was moderate, which is similar to both Ciatto et al, (2005)[6] and Berg et al, (2000)[42] results, where both did not have previous long-term experience in using BI-RADS density categories, similar to UK radiologists in this study. When the data were divided into categories, least agreement was in BI-RADS 2 and 3, which is anticipated, in subjective assessment due to the subtle differences in the classification criteria between these two categories[50]. For example, any asymmetrical density found in mammography will give the impression of pathology and therefore requires further investigation[50]. In addition, it could be because UK radiologists are familiar with the three-point scale, where BI-RADS 2 and 3 appear under one category called moderate or mixed[51]. This could have affected their confidence in distinguishing between the classifications, and may have been impacted by variable training and practice with ACR BI-RADS categorisation. Therefore, individual screening pathways selected for women with dense breasts can and will vary due to the differing levels of agreement between both cohorts demonstrated here.

Previous work by Ciatto et al, (2005)[6] and Gweon et al., (2013)[2] established that intra-rater reliability was of higher concordance than inter-rater variability, whereby reduced consistency is observed among different radiologists[2,6]. Similarly, in agreement with these studies, this study has found that intra-rater reliability for both UK and USA radiologists is in almost perfect agreement (both ICC results >0.9), which suggests that individual categorization standards are robust.

Furthermore, the agreement level for two-scale grade was an almost perfect agreement between radiologists. This high agreement has also been reported by other authors[6], however, this current work has demonstrated this high agreement across two jurisdictions. As perceptual errors may occur due to the wide BI-RADS classification criteria, reducing it to only two would aid especially in reducing the variability between the middle BI-RADS (2 and 3) by assisting in eliminating any confusion between these two categories. In general, radiologists may concur in their clinical interpretations, especially in the detection of malignancy[52]. However, they may provide different further recommendations, due to a difference in their thresholds of concern[53]. In order to reduce variability and increase the predictive value, implementing a simpler two-scale categorisation system, may considerably improve such subjective ratings and readers' robustness. However, this system increases the numbers of women in the higher category, requiring further investigation, which may potentially lead to higher cancer detection rates. This fulfils the aim of screening services, as radiologists' further recommendations for additional imaging will impact upon clinical outcomes to a greater extent than a single diagnostic interpretation[53]. Alternatively, adoption of a more objective approach such as that facilitated by computer-assisted automated software techniques that are currently available, would lead to more consistent categorisation, worldwide.

There were differences in experience and country of residence between the two cohorts and this may have impacted on the result. However, these differences in experience represent the existing radiologist population that work in breast imaging centres, and which is itself a mixture of different experience. Also, because of the shortage of radiologists in breast imaging departments[54], the researcher availed of opportunities to have a large number of radiologists to participate in this study. Moreover, these differences fulfil the aim of the study to find the level of agreement between radiologists from different countries and working under different legislation. While the USA participants reviewed images in a controlled environment, the UK radiologists, reviewed the images in a moderately quiet area. These differences may have impacted the breast density categorization decisions, however, the impact will not be as marked or as noticeable as if they were asked to look for any pathology within the breast, as density depends on the general overall view of the amount of fibroglandular tissue compared to the fatty tissue.

CONCLUSION:

This research indicates that the overall inter-rater variability using the BI-RADS system is substantial between radiologists from the two participating countries. Inconsistencies exist between the two cohorts, especially when the image sets are divided into BI-RADS categories, which substantially increases the possibility of a woman receiving over or underestimated breast density category depending on which image reader reports her mammographic images. Therefore further investigation is merited especially for those with high breast density. These findings support the requirement for improved reproducibility of BI-RADS categorisation across various jurisdictions to enhance breast cancer prediction models and individualised breast cancer screening pathways for dense breast. Use of the two-scale grade has shown greater agreement between cohorts, which may provide a feasible choice for existing clinical practice. However, fully automated breast density software may further improve consistency where economically feasible.