Article

Subjective Versus Quantitative Methods of Assessing Breast Density

Wijdan Alomaim [1,*] , **Desiree O'Leary** [2] , **John Ryan** [3] , **Louise Rainford** [3] , **Michael Evanoff** [4] **and Shane Foley** [3]

1   Radiography & Medical Imaging, Fatima College of Health Sciences, Abu Dhabi, UAE
2   Radiography (Diagnostic Imaging), Keele University, Keele ST5 5BG, UK; D.s.o'leary@keele.ac.uk
3   Radiography & Diagnostic Imaging, School of Medicine, University College Dublin, 4 Dublin, Ireland;
    John.ryan@ziltron.com (J.R.); Louise.rainford@ucd.ie (L.R.); Shane.foley@ucd.ie (S.F.)
4   American Board of Radiology, Tucson, AZ 85711, USA; Mevanoff@theabr.org
*   Correspondence: Wijdan.Alomaim@fchs.ac.ae; Tel.: +9712-5078639

check for updates

Abstract: In order to find a consistent, simple and time-efficient method of assessing mammographic breast density (MBD), different methods of assessing density comparing subjective, quantitative, semi-subjective and semi-quantitative methods were investigated. Subjective MBD of anonymized mammographic cases (*n* = 250) from a national breast-screening programme was rated by 49 radiologists from two countries (UK and USA) who were voluntarily recruited. Quantitatively, three measurement methods, namely VOLPARA, Hand Delineation (HD) and ImageJ (IJ) were used to calculate breast density using the same set of cases, however, for VOLPARA only mammographic cases (*n* = 122) with full raw digital data were included. The agreement level between methods was analysed using weighted kappa test. Agreement between UK and USA radiologists and VOLPARA varied from moderate ($\kappa w = 0.589$) to substantial ($\kappa w = 0.639$), respectively. The levels of agreement between USA, UK radiologists, VOLPARA with IJ were substantial ($\kappa w = 0.752, 0.768, 0.603$), and with HD the levels of agreement varied from moderate to substantial ($\kappa w = 0.632, 0.680, 0.597$), respectively. This study found that there is variability between subjective and objective MBD assessment methods, internationally. These results will add to the evidence base, emphasising the need for consistent, simple and time-efficient MBD assessment methods. Additionally, the quickest method to assess density is the subjective assessment, followed by VOLPARA, which is compatible with a busy clinical setting. Moreover, the use of a more limited two-scale system improves agreement levels and could help minimise any potential country bias.

Keywords: breast density; breast imaging; quantitative density assessment; automated volumetric breast density measurement; VOLPARA; ImageJ; BI-RADS; American College of Radiology Breast Imaging Reporting and Data System; mammographic breast density

1. Introduction

Mammographic breast density (MBD) represents the amount of fat and fibroglandular tissue, which is a mixture of fibrous tissue, glandular tissue, collagen and epithelial cells that line the ducts of the breast [1,2]. Increasing MBD is thus associated with greater breast cancer incidence and is regarded as a risk factor for breast cancer [1,2]. Additionally, MBD has been associated with reduced sensitivity and thus decreased diagnostic accuracy in the detection of any lesions on mammographic images, which may lead to non-detectable cancers [3–6].

Several methods of breast density classification have been developed, both subjective and quantitative [1,7,8]. Studies using subjective BI-RADS assessment as used in clinical practice [9], or using Cumulus software, a visual thresholding technique [10].

Yaffe, in 2008, described several other quantitative methods to measure and classify breast density subjectively or objectively [1]. These include two-dimensional dense area measures, which do not incorporate the thickness of the dense area, Hand Delineation (HD) and semi-automated thresholding technique with ImageJ (IJ). One of these methods, HD, is still considered to be the gold standard in assessing breast density [11,12]. Li et al. developed IJ based on a public domain Java image-processing programme while imitating Cumulus HD [12]. Sovio et al., compared IJ and Cumulus, however, only MLO analogue images were used, left MLO for the Cumulus and both MLOs' for IJ [13]. The study found that IJ is a valid method as a substitute for Cumulus (gold standard), with high inter and intra reader reliability in research settings [11]. However, this is not reproducible in clinical settings [14]. These methods are challenging in practice, tedious and time-consuming [15,16]. Furthermore, any misjudgment of the grey scale would lead to inaccurate breast density classification [16]. For this reason, Cumulus is only used as a gold standard in research and to create further quantitative methods for measuring breast density.

Alternatively, fully automated methods incorporate 3D measurements of dense areas, providing a more comprehensive assessment of the volume of dense breast tissue, which has been shown to be an indicator of increased risk of breast cancer [17–19].

Currently, there are a number of fully automated methods that quantitatively and objectively measure breast density [20,21]. One of the most well-known pieces of software is VOLPARA (Volpara Solutions, New Zealand), which has received Food and Drug Administration (FDA), Health Canada, the Therapeutic Goods Administration (TGA) approval, as well as the CE mark for use in mammography practices and has been used widely in clinical practice [22].

Many studies found an inter-rater variability between the subjective assessment and fully automated method in categorizing breast density [23–26]. However, to the knowledge of the authors no study has compared a larger number of radiologists across different regions assessing the same cases (including both right and left breast images) with VOLPARA, Cumulus HD and semi-objective IJ. Therefore, this study endeavours to address this deficiency, by examining the inter-rater variability between a range of methods currently used in both research and clinical settings worldwide in categorising breast density. This study uses a larger number of mammography cases and a large group of radiologists from two jurisdictions. This could provide a clear idea of the competence of this software globally.

Furthermore, Ko et al. proposed reducing the BI-RAD'S and VOLPARA classification scale from four categories to just two, in order to minimize variations when categorizing MBD [27]. The method was tested in this study with the aim of identifying a method that would be most consistent, simple and time-efficient in discriminating mammographic density, while simplifying the creation of an individual imaging pathway for different densities.

Breast density notification legislation launched in many USA states from 2013 [28]. Additionally, many studies have proven that the inconsistency exists between radiologists globally when using subjective assessment [29,30]. This research was therefore timely in examining the efficiency of some of these objective and semi-objective methods compared to the subjective breast density evaluation, internationally.

MBD outcomes assist in defining an individualized patient imaging pathway for follow up imaging with the least amount of radiation exposure for the radio-sensitive breast tissues. Additionally, MBD is important for each patient's breast cancer risk. A consistent categorization system has the potential to help reduce potential variations in breast density assessment [31]. Such variations can lead to unnecessary imaging or biopsies and increased patient anxiety [31–33]. Therefore, the current study was conducted to compare mammographic density evaluation methods (Hand Delineation, ImageJ and VOLPARA) with the subjective assessment of USA and UK radiologists. Additionally,

to our knowledge no previous study has looked at the impact of distractors within mammographic images in order to document which distractor has the greatest impact on radiologist BI-RADS decision making. According to Ko et al., both the asymmetry of breast size and lesions may cause disagreement between the MBD assessment methods [27]. Furthermore, the timing for each assessment method will be calculated to find the most sufficient and efficient method for clinical setting.

2. Materials and Methods

An exemption from full ethical approval was granted by the University College Dublin, Human Research Ethics Committee (HREC). (Approval ID: LS-E-13-100-Alomaim-OLeary, May 15th 2013).

Permissions were granted by the American Board of Radiology (ABR), AZ, USA, for their examiners and by the British Society of Breast Radiology (BSBR), Hertfordshire, UK, for their respective member attendees to be both recruited voluntarily as participants to undertake this study. Both participants were recruited via local advertising. USA and UK participants' years of experience reporting mammographic cases were recorded. Additionally, UK participants confirmed that they all were breast radiologists.

2.1. Mammographic Cases

Fully anonymised digital mammographic cases ($n = 250$), comprised 180 cases and 70 repeated cases. These cases were collected as part of previous research, from 18 centres in a national breast screening programme, with full patient consent. Each mammography case included "Right and Left; Mediolateral-Oblique and Cranial-Caudal". The researcher (2+ years mammography experience) and a lead researcher and expert in mammography (5+ years), using consensus agreement along with the ACR-BI-RADS Atlas 4th edition selected and categorised the cases using BI-RADS. The cases were selected based on the following criteria: cases with the least artefacts, lesions causing deviation from the norm, technical errors and amount of asymmetry of breast size and MBD (left vs. right). However, one patient image per set did include a distractor for a challenge. The cases were randomly split up into five groups of 50 mammographic cases, including 36 cases and 14 repeated cases.

For the purpose of avoiding radiologists' sense of predictability, the density distribution was not equal within each set (Table 1). Additionally, 50 cases were considered sensible, time permitting, as the radiologists had the choice to review more than one set where possible. Furthermore, using R Package 'KappaSize', the power of the study was calculated to ensure it exceeds 80%. In accordance with previous studies, where a set of 30 mammographic cases reviewed with a minimum of three radiologists were defined as the minimum requirements to have a valid statistical analysis regarding inter-observer agreement level [34–36]. The sets of cases were presented using Ziltron software (Ziltron Ltd., Dublin, Ireland), which allows the cases to be viewed in high-quality, with easy navigation for users to move between cases, pan/zoom and alter the image contrast. Ziltron instruction was provided to both cohorts.

Table 1. The distribution of BI-RADS categories within each set without the repeated cases and for the repeated cases.

Breast Density	Image Sets					
	A	**B**	**C**	**D**	**E**	**Repeated Cases**
BI-RADS 1	14%	25%	22%	17%	22%	7%
BI-RADS 2	36%	17%	22%	28%	25%	57%
BI-RADS 3	36%	39%	39%	39%	36%	21%
BI-RADS 4	14%	19%	17%	17%	17%	14%

To facilitate the automated density analyser (VOLPARA, version 1.5.0, Volpara Solutions, Wellington, New Zealand) to automatically measure density, only those mammographic cases ($n = 122$) with full raw digital data were included [37]. The categorization breakdown of the dataset as reported by the researcher and VOLPARA are listed in Table 2 below. Additionally, for quality control purposes,

the researchers' preliminary BI-RADS categorisation data are tested with the three quantitative methods (VOLPARA, Hand Delineation and ImageJ), for consistency.

Table 2. The distribution of BI-RADS categories for the 122 cases according to the researcher and VOLPARA.

Breast Density	Researcher	VOLPARA
BI-RADS 1	16%	10%
BI-RADS 2	29%	31%
BI-RADS 3	41%	31%
BI-RADS 4	14%	28%

2.2. Subjective Density Assessment

The target population consisted of breast radiologists from both USA and UK who were purposively sampled at two educational events, following voluntary recruitment. Datasets were displayed on secondary reporting monitors using two computer screens, each one with 20″ full viewable diagonal area, with landscape and portrait display modes, called 'ViewSonic ViewPanel, VP201mb (Viewsonic Corporation, Brea, CA, USA) with 1200 × 1600 pixel resolution for USA participants and a TOBII 23″ TFT (TOBII Technology, Stockholm, Sweden) with 1920 × 1080 pixels screen resolution, for UK participants. The luminance level of each monitor for both cohorts was first checked using the Digital Imaging and Communications in Medicine (DICOM). Part 14: Grayscale Standard Display Function (GSDF) using VeriLUM calibration software and pod (ImageSmiths) to ensure all screens met the standard range [37,38]. The measurements for the USA first screen: 164.2 candela per square metre (cd/m^2) maximum, 0.35 cd/m^2 minimum and for the second screen: (175.5 cd/m^2 max, 0.31 cd/m^2 min), for UK screen: (300 cd/m^2 max, 0.67 cd/m^2 min).

All participants were asked to simply record their subjective MBD score using the Breast Imaging Reporting and Data System (BI-RADS) 4th edition (almost fatty, scattered fibroglandular densities, heterogeneously dense and extremely dense) [9], for each image displayed.

2.3. Quantitative Assessment

Byng et al., [39] and Boyd et al., [40], both used 6 MBD categorisation (0%, 1% to <10%, 10% to <25%, 25% to <50%, 50% to <75%, and ≥75%), but later work condensed this to four by combining the upper two categories (<10%, 10–24%, 25–49% and ≥50%.) [13], as a consequence of the small number of subjects in some mammographic density categories. Also ≥50% of mammographic density are counted as dense breast [41]. The modified four category approach was used in this study for the semi-objective (IJ) and semi-subjective (HD) assessment of the MBD.

2.3.1. Semi-Subjective Density Assessment

Using GE Centricity RA 600 (GE Healthcare, Milwaukee, WI, USA) image software. The images were displayed on two side-by-side Barco (Barco NV, Belgium) Coronis 3 Mega Pixel (MFGD-3420) screens, with luminance levels of 402.6 cd/m^2 max, 0.41 cd/m^2 min and 529.1 cd/m^2 max, 0.63 cd/m^2 min respectively. The area of the breast of each mammographic image for each projection (RCC, RMLO, LCC, LMLO) was quantified by selecting the contrast and density toolbar to adjust the contrast and density to enhance the skin edge of the breast. Then, the freehand area calculation tool was selected from the menu, to draw a semi-circular boundary around the area of the breast using a computer mouse pad and polygon generating a measurement in mm^2. The measurements of the delineation of the breast area for the MLO images were done by tracing individually around the area of the whole breast excluding the pectoral muscle, see Figure 1a [16].
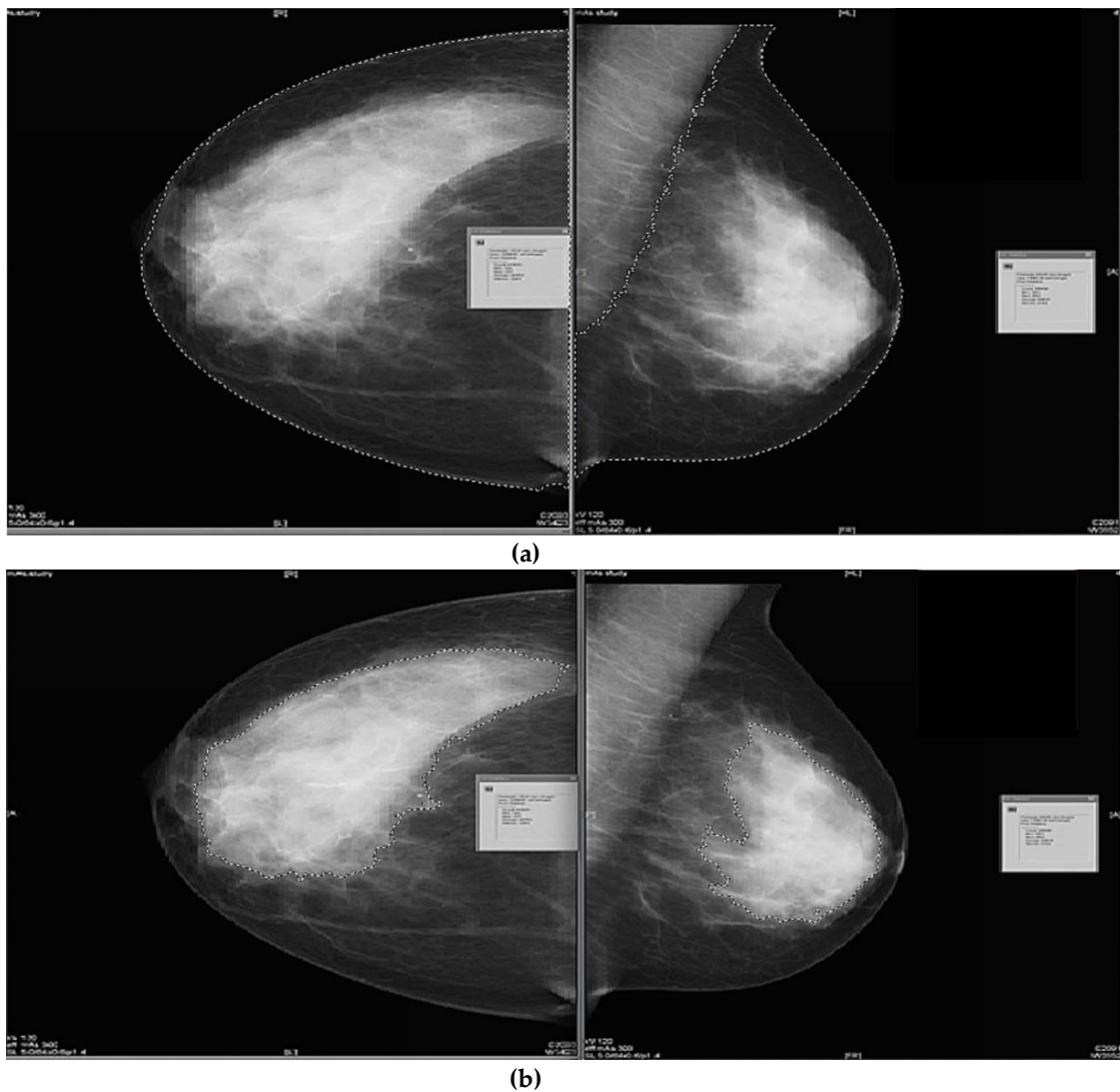
Figure 1. (**a**). Screen shot of the semi-subjective (HD) method showing the craniocaudal and mediolateral oblique breast images with boundaries drawn around the outer surface of the breast. (**b**). Screen shot of the semi-subjective (HD) software showing the craniocaudal and mediolateral oblique breast images with boundaries drawn around the fibroglandular tissue within the breast.

To measure dense breast tissue, the same steps were followed, by adjusting the contrast and density to enhance the fibroglandular tissue and using a computer mouse pad and polygon to draw around the area of interest. The measurement was generated in mm^2, see Figure 1b.

Both area measurements were converted to cm^2 then multiplied by the depth of the breast on the mammogram image in order to calculate the whole volume of breast tissue versus the volume of dense breast tissue. The ratio of the density was thus calculated in a semi-subjective manner.

2.3.2. Semi-Objective Density Assessment

This assessment method uses an automated thresholding procedure that separates the interest area (dense tissue) from the background. The dense portion of the breast was segmented via IJ freeware software for Portable Network Graphics images, which is a built-in design extended by Fiji plugin a 3rd party Java modules [42]. This software performed an area measurement, which it then represented as a percentage. This software also allows the user to alter the area selected by the software. The software performed the area measurement using the interfaces/pixel edges between the various portions of dense tissue in the breast, thus the proportion of dense breast to fatty breast tissue was calculated

(Figure 2). For more accurate measurement, the pectoral muscle was cropped manually from the MLO images [10,43].
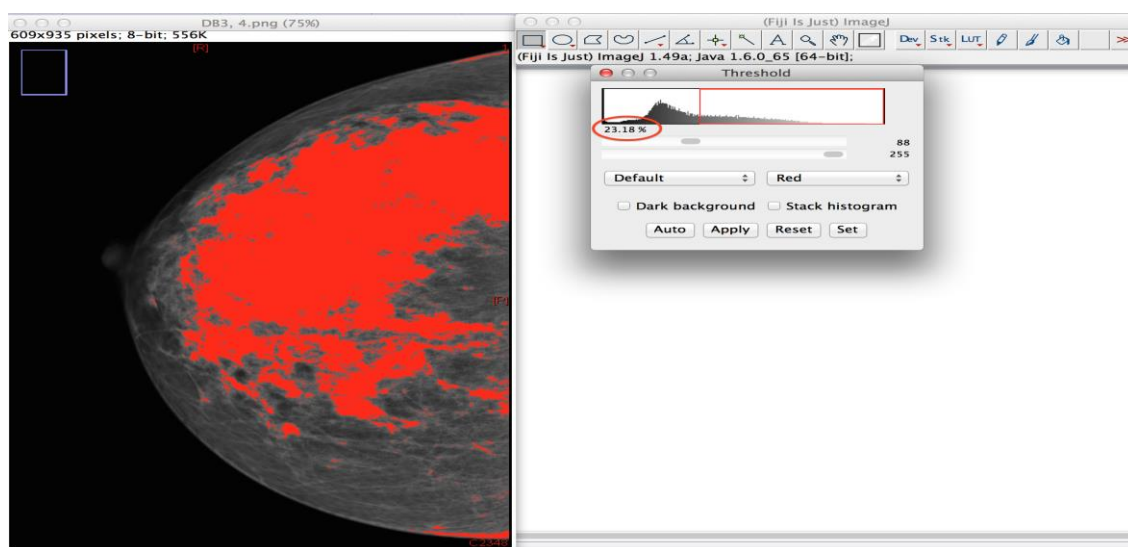

Figure 2. Screenshot of the semi-objective (ImageJ) software, showing a right Cranio-Caudal image with the density as highlighted and the red circle represents the volume density percentage by the software.

2.3.3. Automated Objective Density Assessment

The test set of patient digital mammography cases were saved as raw data DICOM images and then uploaded into a fully automated volumetric analysis, VOLPARA (version 1.5.0, Volpara Solutions, Wellington, New Zealand) density assessment software for quantitative assessment of MBD. Image analysis begins by locating a reference point in the breast, mainly close to the chest wall which contains fatty tissue. Following that, the x-ray attenuation is calculated for each pixel. A density map is created by calculating the attenuation degree in a pixel, the x-ray source and the tissue composition located between the pixel. VOLPARA computes the volume of the fibroglandular tissue and the breast (both in cm^3), volumetric MBD (%), all by analysing the values in the density map. Then the volumetric density is calculated from these data, ranging from 0% to 40%. The information from MLO and CC images are averaged, and then the density information is given per breast. Following that, the density category is provided for each patient. To obtain percentage density, total fibroglandular tissue volume is divided by total breast volume. The percentage density is graded as following: BI-RADS1 (<4.5%), BI-RADS 2 (≥4.5% and <7.5%), BI-RADS 3 (≥7.5% and <15.5%) and BI-RADS 4 (≥15.5%) [44]. The final step automatically processes the data and sends a DICOM secondary capture image to the radiologist's screen, see Figure 3 [45]. The MBD score is given according to the BI-RADS MBD classifications [46,47], which allows for comparison with other methods used in this study.

2.4. Time Difference between Methods

The average length of time for HD, IJ and VOLPARA was measured using a digital stopwatch (iPhone 6 plus, Apple Inc., California). A random stratified sample of eight cases (two cases from each BI-RADS category) was selected for both the HD and IJ methods. Meanwhile, the time taken by VOLPARA to provide the BI-RADS for nine cases saved on one compact disc (CD) was measured. Finally, the length of time for the subjective BI-RADS categorization was calculated using ZILTRON software instant reporting and feedback of the length of time taken by each radiologist to move from one image to the next image for the five sets of cases. In addition, the time was calculated individually for the cases with and without distractors, as reported by the UK radiologists.

For the HD method: the average time was calculated for all BI-RADS except for the two BI-RADS 1 cases, as they were fatty breasts. This time was calculated by adding the drawing time taken for each image for the outer surface to the time taken for outlining the dense area. Times were taken for each of the four projections for each patient, and the average time was calculated.
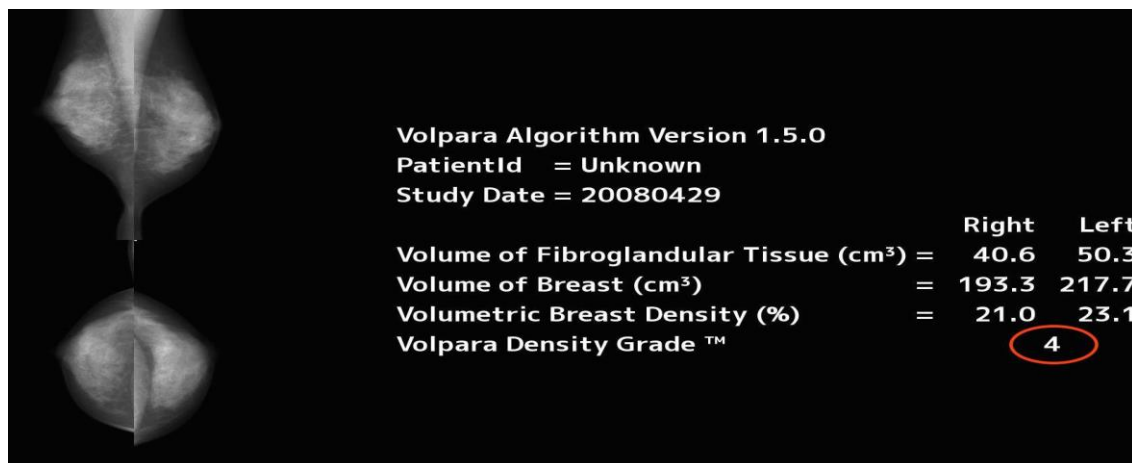

Figure 3. Screen shot of an example of the automated objective density assessment (VOLPARA) software; the red circle shows the BI-RADS as graded by the software.

2.5. Amended, BI-RADS Scale

To investigate further possible adaptations to the existing BI-RADS method, subjective BI-RADS ratings from radiologists in each country (USA and UK) and objective VOLPARA grades were additionally amended using just two categories (high vs. low). This was achieved by combining BI-RADS 1 and 2 together (Low Density), with BI-RADS 3 and 4 (High Density) as a second category.

2.6. Distractors within Cases That Impacted the Radiologists' Breast Density Decisions

Further analysis was performed on the data collected for the cases with distractors from both cohorts. However, the USA radiologists, unlike the UK radiologists, rarely used or commented on this question; this will be explored in the discussion. Each answer from UK radiologists was counted and datasets created by the number of radiologists, thereby representing opinions from 1350 reads. An example is set C, which contains 50 cases read by six radiologists, yielding 300 reads. This procedure was used with all datasets.

Additionally, further analysis was performed to explore whether cases with distractors ($n = 56$) and cases without distractors ($n = 66$) would have an influence on the level of agreement between the subjective BI-RADS decision of UK radiologists and the automated software (VOLPARA) categorisation.

2.7. Data Analysis

The levels of agreement between the radiologists from both countries (USA and UK) assessing MBD were compared separately to VOLPARA using Weighted Kappa test (κw) (95% confidence interval) for the categorical items [48]. Additionally, the levels of agreement between all the different methods were compared using a Weighted Kappa test (κw) (95% confidence interval).

Intra-class Correlation Coefficient (ICC) (95% confidence interval) [49,50] was carried out to determine the intra-rater reliability for each radiologist, who completed more than one set of cases for the repeated cases within the sets.

The Spearman Correlation Coefficient (r) was calculated following the Li et al., [12] study methodology to test for an association between the HD, and IJ MBD measurements.

Kruskal–Wallis non-parametric ANOVA (Chi-Square Test) was used to determine whether there is a significant difference between the time taken for each mammographic density assessment method, including UK radiologists' and VOLPARA [51,52].

3. Results

A total of 25 USA breast imaging expert radiologists with >8 years' experience, post radiology board certification, as well as 24 UK breast imaging experts, participated, with half of the cohort having >8 years' experience and only 12% having ≤1 years' experience post radiology board certification.

3.1. MBD Assessments

Table 3 below shows that the levels of agreement between USA and UK radiologists with the fully automated method (VOLPARA), were substantial and moderate, respectively. Moreover, the table shows that the levels of agreement between USA and UK radiologists with semi-objective (IJ) and semi-subjective (HD) methods varied from moderate to substantial agreement.

Table 3. The level of agreement using (κw) between USA and UK radiologists with Hand Delineation, ImageJ, and VOLPARA.

Subjects	VOLPARA	Hand Delineation	ImageJ
USA radiologists	0.639 *	0.632 *	0.752 *
UK radiologists	0.589 *	0.680 *	0.768 *

* Statistical significance ($p < 0.001$).

Table 4 below shows the levels of agreement of VOLPARA with semi-objective (IJ) and semi-subjective (HD) methods varied from moderate to substantial agreement.

Table 4. The level of agreement using (κw) between Hand Delineation and ImageJ, with VOLPARA.

Subjects	VOLPARA
Hand Delineation	0.597 *
ImageJ	0.603 *

* Statistical significance ($p < 0.001$).

The intra-class correlation coefficient agreement for intra-rater reliability for the radiologists in both countries on the repeated cases within all five sets was high (ICC > 0.9): the average ICC measure for USA radiologists being 0.973 with a 95% confidence interval from 0.966 to 0.978 (F (219,876) = 36.974, $p < 0.001$). For the UK radiologists the average ICC measure was 0.927 with a 95% confidence interval from 0.821 to 0.975 (F (13,26) = 15.194, $p < 0.001$).

The findings for the quality control tests are: the level of agreement between the researcher's subjective BI-RADS categorisation and VOLPARA and Hand Delineation was substantial (κw = 0.660 and 0.782, respectively, $p < 0.001$), while, with ImageJ was almost perfect agreement (κw = 0.856, $p < 0.001$).

The level of agreement using weighted Kappa between Hand Delineation and ImageJ, is substantial with κw = 0.699, ($p < 0.001$).

When the correlation was tested between ImageJ and Hand Delineation assessment for breast density, a strong, positive association between the two subjects ($r = 0.809$, $p < 0.001$, 2-sided), was found.

Using the modified two grade scale the agreement levels increased to substantial agreement between VOLPARA and USA radiologists (0.702, $p < 0.001$) and UK radiologists (0.630, $p < 0.001$).

3.2. Time Analysis

Examination of the time difference between each method showed that there was a significant difference between each method ($p < 0.001$), with visual assessment of the MBD taking an average of 14 s (range: 7–33 s), VOLPARA takes approximately 57 sec for each patient, with HD and IJ methods taking an average time of 320 sec (range: 198–453 s) and 131 sec (range: 83–168 s), respectively.

3.3. Image Distractors Influencing MBD Assessment

Of the original total of 250 mammographic cases, 92 mammographic cases with distractors were included. However, results show that different decisions were taken by each radiologist ($n = 24$), according to the number of reads. Figure 4 below shows that the highest percentage of distractors identified by radiologists while assessing breast density was lesions.
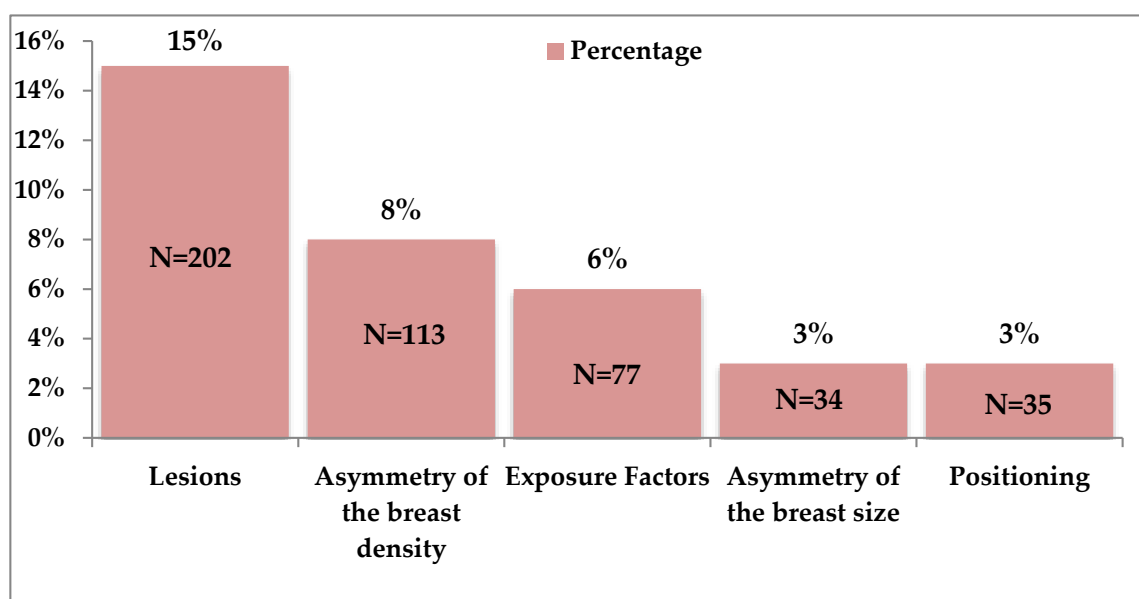

Figure 4. Percentage of the distractors within total cases according to the number of reads, as decided by the UK radiologists.

When the agreement level for the BI-RADS categorization between UK radiologists and VOLPARA for cases with distractors was calculated, the agreement was substantial ($\kappa w = 0.656$, $p < 0.001$), meanwhile, the agreement level for the cases without any distractors was moderate ($\kappa w = 0.518$, $p < 0.001$). The average time for the subjective categorization for cases with distractors was 12 s (range: 7–23), and 16 s (range: 9–33 s) for those without any distractors, however, the time difference was not significant ($p > 0.05$). Furthermore, VOLPARA takes approximately 57 s for each patient, whether the image is with or without any distractors.

Figure 5 shows that for cases with no distractors, VOLPARA placed less cases in BI-RADS 1, while UK radiologists placed less cases in the BI-RADS 4. For the cases with distractors, VOLPARA placed most of the cases on the higher categories. However, UK radiologists placed cases more evenly across the categories.
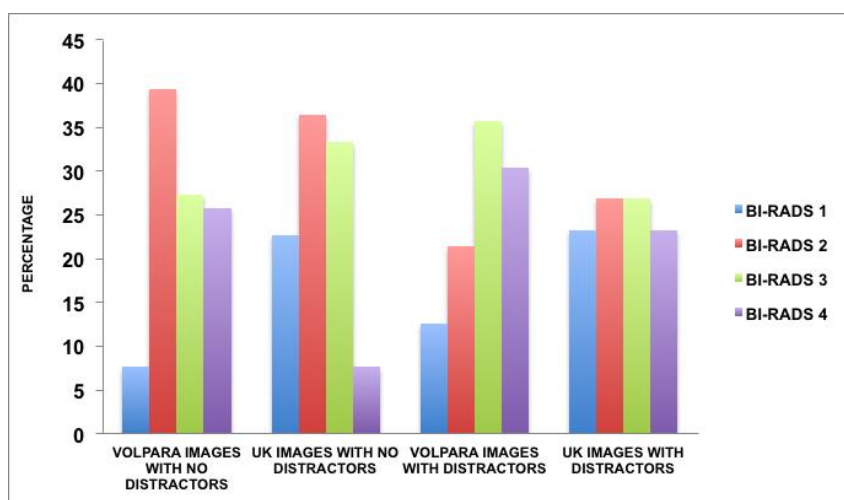
Figure 5. Distribution of the BI-RADS scoring amongst cases with and without distractors for VOLPARA and UK radiologists.

4. Discussion

As MBD categorisation is subjective, this may lead to under- or overestimates of the percentage of fibroglandular tissue within the mammographic cases. This finding has been confirmed and it has been proposed that MBD should be assessed using automated software with a view to unifying MBD judgment worldwide [29,31,32,53], as the significance of proper MBD categorization is well documented [17–19].

When the radiology cohort findings were compared with the fully automated objective method (VOLPARA), it was found that this method has a substantial agreement with the American radiologists. However, a moderate agreement was found with the British radiologists. The higher agreement between USA radiologists and VOLPARA might be expected, as VOLPARA software thresholds were determined by a USA expert group of radiologists [54]. This difference between both cohorts and VOLPARA raises some concerns, as it may impact negatively upon patient imaging and follow up pathway [31–33,55].

Lee et al., [24], a Korean-based study, reported a substantial level of agreement (0.799) between a single radiologist and VOLPARA. However, another Korean-based study, found moderate agreement (0.54) between two radiologists and VOLPARA [17], while an Indian-based study comparing two radiologists' BI-RADS assessment results with VOLPARA found fair agreement (0.398 and 0.388) [45]. Varying levels of agreement have thus been found internationally, especially where small cohorts of radiologists were involved. By contrast, in this work a large cohort from two countries participated, with the results varying again (from 0.589 to 0.639). This could be driven by the difference in experience levels, clinical practice and legal requirement to report MBD between the two cohorts. This difference may also be because UK radiologists are using a three-category system, meanwhile, the USA uses the four BI-RADS. Moreover, the viewing environments differed between the two cohorts, specifically the setting lighting and the display monitors [29]. However, both monitors were similar in quality and it was unlikely to influence these results [29].

Sartor et al., [56] in his study has explained a few reasons for the low agreement between VOLPARA and subjective BI-RADS categorization: BI-RADS are judged according to processed images, while, VOLPARA only analyses raw data which contains information about the pixel intensities and the X-ray attenuations of the fatty tissues versus the fibroglandular tissues, which calculate the depth of the fibroglandular tissues [57]. Moreover, as outlined, VOLPARA's assessment scale is a continuous one, while, BI-RADS categories are an estimation that are divided into four groups, for this reason the MBD that is near the limits in the different VOLPARA's assessment scales could be classified into the upper or the lower adjacent BI-RADS category, since the radiologists would not be able to detect small

differences in mammographic density. Similarly, van der Waal et al. [58] state, following the direction of the ACR, that radiologists tend to rate at the maximum level, in contrast to VOLPARA, which is designed to estimate density using the average of multiple views.

Subjective variations noted worldwide could lead to increases in both radiologist and patient uncertainty. Radiologists may fear over or underestimating MBD, while patients may fear the possibility of a cancer being missed. Such variation could also create difficulties in following a consolidated MBD categorization, thus making it more difficult to develop a specific imaging pathway, which is dependent on MBD, and unifying patient care [32,55,59–61]. Additionally, it increases the woman's concern of going through demanding procedures, the pain involved with extra biopsies and imaging and, more importantly, questions the reliability of mammography [31–33,55]. However, it is important to consider that the results of this study were collected without VOLPARA guidance to assist radiologists to categorize MBD. Therefore, the use of VOLPARA guidance is recommended. An American-based study of Schilling et al. [25] compared eight radiologists before and after using the guidance of VOLPARA. This showed that the level of agreement before the guidance of VOLPARA was moderate (0.566), but, following the aid of VOLPARA to re-categorize the same cases, the level of agreement became substantial (0.626). This is similar to the level of agreement between the American radiologists and VOLPARA in this study (0.639), where radiologists were not provided with VOLPARA's decisions in advance, however they were ABR examiners, which means that they had more experience and training when compared to others. Meanwhile, the initial result before the aid of VOLPARA is similar to the level of agreement of the UK radiologists in the current study, where the radiologists possibly had less experience with the ACR BI-RADS and were not practicing MBD assessment in their routine practice. Moreover, Jeffers et al.'s [62] study found that the level of agreement between the subjective single image reader and VOLPARA was fair, however, this low level of agreement may be because the image reader only had two years of experience and categorised without the aid of VOLPARA. To elaborate further on this point using quality control test findings, the researcher's level of agreement with VOLPARA was tested, and it was found to be substantial (0.660, $p < 0.001$), which was similar to the USA radiologists. This may be because the researcher has been specifically trained to use BI-RADS, and categorised the cases with the aid of BI-RADS Atlas 4th edition. These results add to the evidence supporting the fact that, without the assistance of VOLPARA, the image readers will have a lower agreement. This emphasises the value of incorporating VOLPARA or other such similar automated systems into MBD assessment as they may positively impact on the radiologists' agreement, as also proven by Schilling et al. [25]. Thus, this current study recommends a further investigation between the radiologists and VOLPARA agreement levels when VOLPARA is used as a reference.

A further part of the study tested the agreement between the subjective analysis and VOLPARA on a more limited version of the current BI-RAD'S classification by reducing the scale to just two separate categories, to see whether differences in MBD interpretation could be reduced. The findings demonstrate that the level of agreement between the two cohorts' subjective evaluation and VOLPARA showed more substantial agreement. Thus, the scale of two separate categories—low density and high density—is reliable and robust, and therefore could lead to reducing patient uncertainty, as well as potentially minimizing the costs associated with unnecessary imaging and additional procedures, and help in unifying the imaging pathway for patient cohorts.

Moving toward the area-based methods (HD and IJ), which assess MBD by segmenting two-dimensional mammographic images into dense and non-dense areas, and have been only partially automated [10,63], when the correlation between these two methods was tested, a high correlation was observed ($r = 0.809$). This was expected and is in agreement with previous studies of Li et al., [12] and Couwenberg et al., [64], as IJ was developed mimicking the HD method. Furthermore, using quality control findings, the level of agreement between HD and IJ and the researcher's categorization according to the BI-RADS 4th ed, varied from substantial to almost perfect (0.782 and 0.856). Both methods were performed by the researcher. Meanwhile, the level of agreement between the two cohorts with HD and IJ was lower than for the researcher. These results support the assumption that these two methods

are dependent on visual decisions to determine the breast area and the MBD edges. Furthermore, the intra-rater reliability for both UK and USA radiologists are in almost perfect agreement, which supports the possibility that individual categorization standards for each radiologist is strong and is one of the limitations that might affect final outcome. However, the researcher acknowledges that, in other studies, HD and IJ methods were performed by radiologists, rather than the researcher.

The second limitation, as stated by Byng et al. [10], is that the HD takes a greater amount of time, even when carried out by skilled and professional radiologists. Byng et al., [10] reported that the time to conduct the HD method is less than four minutes to evaluate the four mammographic images for each patient [10,16,65]. Meanwhile, in this study the HD took almost 5 min, and IJ almost 2 min to evaluate the four mammographic images, which, for HD, is greater than the time reported by Byng et al., [10], proving that this method is largely dependent on the subjective assessment of the image readers, as well as the skill level and training in the use of the method.

On the other hand, when the level of agreements between HD and IJ MBD assessment methods with VOLPARA were tested, moderate agreement was found between VOLPARA and HD (0.597) and IJ (0.603). The variation in agreement between VOLPARA and both HD and IJ is due to the fact that VOLPARA is a fully volumetric measurement, while both HD and IJ are area-based methods with visual computer assistance. The differences in the methodologies may have an impact on the final results, as area-based methods are highly user-dependent. The two-dimensional image of the compressed breast is segmented and the threshold is defined by the user. This increases the possibility of inter-observer variability [10,16,63,65]. In addition, according to van Engeland et al. [66] and McCormack et al.'s [67] studies, the potential source of variation between both methods is that, as a quantitative measure of MBD volume, fully volumetric methods are a more suitable approach than a projected area.

When agreements between HD and IJ breast density assessment methods with the BI-RADS subjective assessment were tested, the level of agreement between these two methods and the two cohorts was substantial. In general, this is higher than the agreement level between VOLPARA and the two cohorts. Possible explanations for this higher agreement include the fact that, as in volumetric methods, it is difficult to locate a reference point in the breast that only contains fatty tissues, either within an extremely dense breast or in a very small breast [66]. This software is also affected by asymmetry or asymmetric breast thickness as the breast thickness influences both breast and density volume; when the thickness increases, both volumes will increase [54]. Moreover, as the x-ray device nearly always underestimates breast thickness, VOLPARA's assessment will increase, therefore, women will be placed in a higher BI-RADS category [54].

British radiologists when asked regarding potential distractors, reported lesions causing deviation from the norm as the greatest distractor, followed by the asymmetry of the MBD. Lee et al., [24] found that the asymmetry of bilateral MBD had the biggest impact on the disagreement between the BI-RADS assessment by the radiologists and VOLPARA. The level of agreement between the radiologists and VOLPARA for the cases with distractors, in this study, showed that the agreement level (0.451, $p < 0.001$) was higher than the agreement for the cases with no distractors (0.293, $p < 0.001$), which was not expected, and this supports the need for further studies to explore visual processing reasoning. Furthermore, as shown in Figure 5, the distribution of the BI-RADS differs between VOLPARA and UK radiologists. This may be the reason for the low agreement level. Additionally, high body mass index (BMI) is significantly related to volumetric-assessed MBD [68]. Another explanation is that, when the average viewing time was calculated for the radiologists viewing the cases with distractors compared to the cases without distractors, it was found that they spent less time on the cases with distractors than the cases without any distractors. This was not expected as, logically, the cases with possible distractors need more time to be explored, because the radiologists will have to focus on the general image plus the distractor. As has been stated by Eagleman et al. [69], perception time is still shrouded in mystery. As stated by Nodine et al. [70], the image reporter's confidence level decreases when the interpretation time increases, which increases the likelihood of errors. In view of this research, further studies should be carried out to compare the level of agreement between

radiologists with the time spent viewing cases, both with and without distractors. This could further test the hypothesis that radiologists viewing cases for longer times tended to both incorrectly diagnose and to misinterpret MBD.

The authors acknowledge that this work has some limitations: the two-dimensional nature of IJ and HD methods does not take into consideration the thickness of dense tissue within the breast [37,71]. This method is also a user-assisted (semi-automated) method, so it is only for research settings and not suitable for use in clinical settings [71]. The radiologists from UK and USA were asked to give their opinion as to whether there were any distractors within the image that might affect their assessment of MBD. Unlike their UK colleagues, the USA radiologists rarely used or commented on this question. These differences in response could be due to the time constraints, also this question was mandatory for the UK radiologists, as explained previously. Additionally, the use of repeated cases could be a limitation, by creating a memory effect in the rating of MBD by radiologists.The viewing environments and the monitors used differed for both cohorts, which was not ideal [29]. However, as observers were not seeking any abnormality per se and only categorizing overall density, these factors are thought not to be as marked or noticeable as if they were asked to look for any lesions within the breast, as density depends on the general overall view of the amount of fibroglandular tissue compared to the fatty tissue.

5. Conclusions

There is variability between breast density assessment methods, and that all methods are not in complete or even high agreement with each other. The level of agreement varied from moderate to substantial between VOLPARA with both cohorts. Despite the high agreement between HD, IJ and both cohorts, HD, IJ are user dependent, requires skills, training and are time consuming. Therefore, these are costly in relation to radiologists' time and are impractical in a clinical setting. Moreover, subjective assessment is the quickest method to assess density, followed by VOLPARA, and both are appropriate in a busy clinical setting. In addition, the use of the limited two-scale system improves agreement levels, and thus could help in minimizing any potential country bias while providing a clear imaging pathway for patients. To date, there is no universal agreement as to which methods are best-suited to creating individual imaging pathways, therefore, further investigation is strongly recommended.

References

1. Yaffe, M.J. Mammographic density. Measurement of mammographic density. *Breast Cancer Res.* **2008**, *10*, 209. [CrossRef] [PubMed]
2. Ziv, E.; Shepherd, J.; Smith-Bindman, R.; Kerlikowske, K. Mammographic breast density and family history of breast cancer. *J. Natl. Cancer Inst.* **2003**, *95*, 556–558. [CrossRef] [PubMed]
3. Carney, P.A.; Miglioretti, D.L.; Yankaskas, B.C.; Kerlikowske, K.; Rosenberg, R.; Rutter, C.M.; Geller, B.M.; Abraham, L.A.; Taplin, S.H.; Dignan, M.; et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann. Intern. Med.* **2003**, *138*, 168–175. [CrossRef] [PubMed]

4.　Persson, I.; Thurfjell, E.; Holmberg, L. Effect of estrogen and estrogen-progestin replacement regimens on mammographic breast parenchymal density. *J. Clin. Oncol.* **1997**, *15*, 3201–3207. [CrossRef]

5.　Yankaskas, B.C.; Cleveland, R.J.; Schell, M.J.; Kozar, R. Association of Recall Rates with Sensitivity and Positive Predictive Values of Screening Mammography. *Am. J. Roentgenol.* **2001**, *177*, 543–549. [CrossRef]

6.　Thurfjell, E. Breast density and the risk of breast cancer. *New Engl. J. Med.* **2002**, *347*, 866. [CrossRef]

7.　Petroudi, S.; Kadir, T.; Brady, M. Automatic classification of mammographic parenchymal patterns: A statistical approach. In Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439), Cancun, Mexico, 17–21 September 2003; Available online: http://www.robots.ox.ac.uk/~{}timork/Mammo/embs511.pdf (accessed on 18 June 2017).

8.　Tagliafico, A.; Tagliafico, G.; Tosto, S.; Chiesa, F.; Martinoli, C.; Derchi, L.E.; Calabrese, M. Mammographic density estimation: Comparison among BI-RADS categories, a semi-automated software and a fully automated one. *Breast* **2009**, *18*, 35–40. [CrossRef]

9.　Sickles, E.; D'Orsi, C.; Bassett, L.; Al, E. ACR BI-RADS ® Mammography. In *ACR BI-RADS ® Atlas, Breast Imaging Reporting and Data System*; Reston, V.A., Ed.; American College of Radiology: Reston, VA, USA, 2013.

10.　Byng, J.W.; Boyd, N.F.; Fishell, E.; A Jong, R.; Yaffe, M.J. The quantitative analysis of mammographic densities. *Phys. Med. Boil.* **1994**, *39*, 1629–1638. [CrossRef]

11.　McCormack, V.; Highnam, R.; Perry, N.; Dos-Santos-Silva, I. 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*, 1148–1154. [CrossRef]

12.　Li, J.; Szekely, L.; Eriksson, L.; Heddson, B.; Sundbom, A.; Czene, K.; Hall, P.; Humphreys, K. High-throughput mammographic-density measurement: A tool for risk prediction of breast cancer. *Breast Cancer Res.* **2012**, *14*, R114. [CrossRef]

13.　Sovio, U.; Li, J.; Aitken, Z.; Humphreys, K.; Czene, K.; Moss, S.; Hall, P.; McCormack, V.; Dos-Santos-Silva, I. Comparison of fully and semi-automated area-based methods for measuring mammographic density and predicting breast cancer risk. *Br. J. Cancer* **2014**, *110*, 1908–1916. [CrossRef]

14.　Eng, A.; Gallant, Z.; Shepherd, J.A.; McCormack, V.; Li, J.; Dowsett, M.; Vinnicombe, S.; Allen, S.; Dos-Santos-Silva, I. Digital mammographic density and breast cancer risk: A case-control study of six alternative density assessment methods. *Breast Cancer Res.* **2014**, *16*, 439. [CrossRef] [PubMed]

15.　Kerlikowske, K.; Ichikawa, L.; Miglioretti, D.L.; Buist, D.S.M.; Vacek, P.M.; Smith-Bindman, R.; Yankaskas, B.; Carney, P.A.; Ballard-Barbash, R. Longitudinal Measurement of Clinical Mammographic Breast Density to Improve Estimation of Breast Cancer Risk. *J. Natl. Cancer Inst.* **2007**, *99*, 386–395. [CrossRef] [PubMed]

16.　Martin, K.E.; Helvie, M.A.; Zhou, C.; Roubidoux, M.A.; Bailey, J.E.; Paramagul, C.; Blane, C.E.; Klein, K.A.; Sonnad, S.S.; Chan, H.-P. Mammographic Density Measured with Quantitative Computer-aided Method: Comparison with Radiologists' Estimates and BI-RADS Categories. *Radiology* **2006**, *240*, 656–665. [CrossRef] [PubMed]

17.　Vachon, C.M.; Van Gils, C.H.; A Sellers, T.; Ghosh, K.; Pruthi, S.; Brandt, K.R.; Pankratz, V.S. Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Res.* **2007**, *9*, 217. [CrossRef] [PubMed]

18.　Shepherd, J.A.; Kerlikowske, K.; Ma, L.; Duewer, F.; Fan, B.; Wang, J.; Malkov, S.; Vittinghoff, E.; Cummings, S.R.; Malkov, S. Volume of mammographic density and risk of breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **2011**, *20*, 1473–1482. [CrossRef]

19.　Ng, K.-H.; Yip, C.-H.; Taib, N.A.M. Standardisation of clinical breast-density measurement. *Lancet Oncol.* **2012**, *13*, 334–336. [CrossRef]

20.　Alonzo-Proulx, O.; Mawdsley, G.; Patrie, J.T.; Yaffe, M.J.; A Harvey, J. Reliability of Automated Breast Density Measurements. *Radiology* **2015**, *275*, 366–376. [CrossRef]

21.　Youk, J.H.; Gweon, H.M.; Son, E.J.; Kim, J.-A. Automated Volumetric Breast Density Measurements in the Era of the BI-RADS Fifth Edition: A Comparison With Visual Assessment. *Am. J. Roentgenol.* **2016**, *206*, 1–7. [CrossRef]

22.　Volpara Solutions Receives FDA Clearance for Volpara Density Maps. Volpara Solutions. 2016. Available online: https://volparasolutions.com/about-volpara/news/volpara-solutions-receives-fda-clearance-for-volpara-density-maps-designed-to-help-identify-the-risk-of-breast-density-masking-a-potential-cancer/ (accessed on 18 June 2017).

23.　Gweon, H.M.; Youk, J.H.; Kim, J.-A.; Son, E.J. Radiologist Assessment of Breast Density by BI-RADS Categories Versus Fully Automated Volumetric Assessment. *Am. J. Roentgenol.* **2013**, *201*, 692–697. [CrossRef]

24. Na Lee, H.; Sohn, Y.M.; Han, K. Comparison of mammographic density estimation by Volpara software with radiologists' visual assessment: Analysis of clinical–radiologic factors affecting discrepancy between them. *Acta Radiol.* **2015**, *56*, 1061–1068. [CrossRef] [PubMed]

25. Schilling, K.; The, J.; Griff, J.; Oliver, L.; Mahal, R.; Saady, M.; Velasquez, M.V. Impact of quantitative breast density on experienced radiologists' assessment of mammographic breast density Aims and objectives. In *Electronic Presentation Online System*; European Congress of Radiology: Vienna, Austria, 2015; Available online: http://www.bocaradiology.com/data1/BreastDensityCAD.pdf (accessed on 18 June 2017).

26. Damases, C.N.; Brennan, P.C.; Mello-Thoms, C.; McEntee, M.F. Mammographic Breast Density Assessment Using Automated Volumetric Software and Breast Imaging Reporting and Data System (BIRADS) Categorization by Expert Radiologists. *Acad. Radiol.* **2016**, *23*, 70–77. [CrossRef] [PubMed]

27. Ko, S.Y.; Kim, E.-K.; Kim, M.J.; Moon, H.J. Mammographic Density Estimation with Automated Volumetric Breast Density Measurement. *Korean J. Radiol.* **2014**, *15*, 313–321. [CrossRef] [PubMed]

28. Durning, M. Breast Density Notification Laws by State—Interactive Map. Diagnostic Imaging. 2016. Available online: http://www.diagnosticimaging.com/breast-imaging/breast-density-notification-laws-state-interactive-map (accessed on 3 June 2017).

29. Alomaim, W.; O'Leary, D.; Ryan, J.; Rainford, L.; Evanoff, M.; Foley, S. Variability of Breast Density Classification Between US and UK Radiologists. *J. Med Imaging Radiat. Sci.* **2019**, *50*, 53–61. [CrossRef] [PubMed]

30. Damases, C.N.; Hogg, P.; McEntee, M.F. Inter country analysis of breast density classification using visual grading. *Br. J. Radiol.* **2017**, *90*, 20170064. [CrossRef]

31. Redondo, A.; Comas, M.; Macià, F.; Ferrer, F.; Murta-Nascimento, C.; Maristany, M.T.; Molins, E.; Sala, M.; Castells, X. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *Br. J. Radiol.* **2012**, *85*, 1465–1470. [CrossRef]

32. Marcus, E.N.; Yepes, M. The conundrum of explaining breast density to patients. *Clevel. Clin. J. Med.* **2013**, *80*, 761–765. [CrossRef]

33. McCann, J.; Stockton, D.; Godward, S. Impact of false-positive mammography on subsequent screening attendance and risk of cancer. *Breast Cancer Res.* **2002**, *4*, R11. [CrossRef]

34. Giraudeau, B.; Mary, J.Y. Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stat. Med.* **2001**, *20*, 3205–3214. [CrossRef]

35. Kundel, H.L.; Polansky, M. Measurement of Observer Agreement. *Radiology* **2003**, *228*, 303–308. [CrossRef]

36. Ooms, E.; Zonderland, H.; Eijkemans, M.; Kriege, M.; Delavary, B.M.; Burger, C.; Ansink, A. Mammography: Interobserver variability in breast density assessment. *Breast* **2007**, *16*, 568–576. [CrossRef]

37. Gubern-Merida, A.; Kallenberg, M.; Platel, B.; Mann, R.M.; Martí, R.; Karssemeijer, N. Volumetric Breast Density Estimation from Full-Field Digital Mammograms: A Validation Study. *PLoS ONE* **2014**, *9*, e85952. [CrossRef]

38. NEMA. *Digital Imaging and Communications in Medicine (DICOM) Part 14: Greyscale Standard Display Function*; (NEMA) National Electrical Manufacturers Association: Rosslyn, VA, USA, 2004.

39. Byng, J.W.; Yaffe, M.J.; A Jong, R.; Shumak, R.S.; A Lockwood, G.; Tritchler, D.L.; Boyd, N.F. Analysis of mammographic density and breast cancer risk from digitized mammograms. *Radiographics* **1998**, *18*, 1587–1598. [CrossRef] [PubMed]

40. Boyd, N.; Guo, H.; Martin, L.J.; Sun, L.; Stone, J.; Fishell, E.; Jong, R.A.; Hislop, G.; Chiarelli, A.; Minkin, S.; et al. Mammographic Density and the Risk and Detection of Breast Cancer. *New Engl. J. Med.* **2007**, *356*, 227–236. [CrossRef] [PubMed]

41. Boyd, N.F.; Byng, J.W.; Jong, R.A.; Fishell, E.K.; Little, L.E.; Lockwood, G.A.; Tritchler, D.L.; Yaffe, M.J.; Miller, A.B. Quantitative Classification of Mammographic Densities and Breast Cancer Risk: Results From the Canadian National Breast Screening Study. *J. Natl. Cancer Inst.* **1995**, *87*, 670–675. [CrossRef]

42. Schindelin, J.; Arganda-Carreras, I.; Frise, E.; Kaynig, V.; Longair, M.; Pietzsch, T.; Preibisch, S.; Rueden, C.; Saalfeld, S.; Schmid, B.; et al. Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **2012**, *9*, 676–682. [CrossRef] [PubMed]

43. Boone, J.M.; Lindfors, K.K.; Beatty, C.S.; Seibert, J.A. A breast density index for digital mammograms based on radiologists' randing. *J. Digit. Imaging* **1998**, *11*, 101–115. [CrossRef] [PubMed]

44. Highnam, R.; Sauber, N.; Destounis, S.; Harvey, J.; McDonald, D. Breast Density into Clinical Practice. In *Breast Imaging IWDM 2012 Lecture Notes in Computer Science*; Maidment, A.D.A., Bakic, P.R., Gavenonis, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7361, pp. 466–473. [CrossRef]

45. Singh, T.; Sharma, M.; Singla, V.; Khandelwal, N. Breast Density Estimation with Fully Automated Volumetric Method. *Acad. Radiol.* **2016**, *23*, 78–83. [CrossRef]

46. Brand, J.S.; Czene, K.; Shepherd, J.A.; Leifland, K.; Heddson, B.; Sundbom, A.; Eriksson, M.; Li, J.; Humphreys, K.; Hall, P. Automated Measurement of Volumetric Mammographic Density: A Tool for Widespread Breast Cancer Risk Assessment. *Cancer Epidemiol. Biomarkers Prev.* **2014**, *23*, 1764–1772. [CrossRef]

47. Holland, K.; Kallenberg, M.; Mann, R.; van Gils, C.; Karssemeijer, N. Stability of Volumetric Tissue Composition Measured in Serial Screening Mammograms. In *International Workshop on Digital Mammography*; Fujita, H., Hara, T., Muramatsu, C., Eds.; Springer: Cham, Switzerland, 2014; Volume 8539, pp. 239–244. [CrossRef]

48. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef]

49. McGraw, K.O.; Wong, S.P. Forming inferences about some intraclass correlation coefficients. *Psychological Methods.* **1996**, *1*, 30–46. [CrossRef]

50. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef] [PubMed]

51. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]

52. Gosall, N.K.; Gosall, G.S. *The Doctor's Guide to Critical Appraisal*, 3rd ed.; PasTest: Knutsford, UK, 2012.

53. Sauber, N.; Chan, A.; Highnam, R. BI-RADS breast density classification-an international standard? In *European Congress of Radiology*; Electronic Presentation Online System: Vienna, Austria, 2013.

54. Highnam, R.; Brady, S.M.; Yaffe, M.J.; Karssemeijer, N.; Harvey, J. Robust Breast Composition Measurement-VolparaTM. In *10th International Conference on Digital Mammography*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 342–349. [CrossRef]

55. Barton, M.B.; Moore, S.; Polk, S.; Shtatland, E.; Elmore, J.G.; Fletcher, S.W. Increased patient concern after false-positive mammograms: Clinician documentation and subsequent ambulatory visits. *J. Gen. Intern. Med.* **2001**, *16*, 150–156. [CrossRef] [PubMed]

56. Sartor, H.; Lång, K.; Rosso, A.; Borgquist, S.; Zackrisson, S.; Timberg, P. Measuring mammographic density: Comparing a fully automated volumetric assessment versus European radiologists' qualitative classification. *Eur. Radiol.* **2016**, *26*, 4354–4360. [CrossRef]

57. Destounis, S.; Arieno, A.; Morgan, R.; Roberts, C.; Chan, A. Qualitative Versus Quantitative Mammographic Breast Density Assessment: Applications for the US and Abroad. *Diagnostics* **2017**, *7*, 30. [CrossRef]

58. Van Der Waal, D.; Heeten, G.J.D.; Pijnappel, R.M.; Schuur, K.H.; Timmers, J.M.H.; Verbeek, A.L.M.; Broeders, M.J.M. 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. [CrossRef]

59. Smith, M.L. The Density Conundrum: Does Legislation Help or Hurt? *J. Am. Coll. Radiol.* **2013**, *10*, 909–912. [CrossRef]

60. Bruno, M.A.; Walker, E.; Abujudeh, H.H. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *Radiographics* **2015**, *35*, 1668–1676. [CrossRef]

61. Winkler, N.S.; Raza, S.; Mackesy, M.; Birdwell, R.L. Breast Density: Clinical Implications and Assessment Methods. *Radiographics* **2015**, *35*, 316–324. [CrossRef]

62. Jeffers, A.M.; Sieh, W.; Lipson, J.A.; Rothstein, J.H.; McGuire, V.; Whittemore, A.S.; Rubin, D.L. Breast Cancer Risk and Mammographic Density Assessed with Semiautomated and Fully Automated Methods and BI-RADS. *Radiology* **2016**, *282*, 348–355. [CrossRef] [PubMed]

63. Sivaramakrishna, R.; Obuchowski, N.A.; Chilcote, W.A.; Powell, K.A. Automatic Segmentation of Mammographic Density. *Acad. Radiol.* **2001**, *8*, 250–256. [CrossRef]

64. Couwenberg, A.M.; Verkooijen, H.M.; Li, J.; Pijnappel, R.M.; Charaghvandi, K.R.; Hartman, M.; Van Gils, C.H. Assessment of a fully automated, high-throughput mammographic density measurement tool for use with processed digital mammograms. *Cancer Causes Control.* **2014**, *25*, 1037–1043. [CrossRef] [PubMed]

65. Prevrhal, S.; Shepherd, J.A.; Smith-Bindman, R.; Cummings, S.; Kerlikowske, K. Accuracy of mammographic breast density analysis: Results of formal operator training. *Cancer Epidemiol. Biomarkers Prev.* **2002**, *11*, 1389–1393. [PubMed]

66. Van Engeland, S.; Snoeren, P.; Huisman, H.; Boetes, C.; Karssemeijer, N. Volumetric breast density estimation from full-field digital mammograms. *IEEE Trans. Med Imaging* **2006**, *25*, 273–282. [CrossRef]

67. McCormack, V.; Dos-Santos-Silva, I. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **2006**, *15*, 1159–1169. [CrossRef]

68. Lokate, M.; Karssemeijer, N.; Kallenberg, M.G.; Bosch, M.A.A.J.V.D.; Peeters, P.H.M.; Van Gils, C.H. Volumetric Breast Density from Full-Field Digital Mammograms and Its Association with Breast Cancer Risk Factors: A Comparison with a Threshold Method. *Cancer Epidemiol. Biomarkers Prev.* **2010**, *19*, 3096–3105. [CrossRef]

69. Eagleman, D.M.; Tse, P.U.; Buonomano, D.; Janssen, P.; Nobre, A.C.; Holcombe, A.O. Time and the Brain: How Subjective Time Relates to Neural Time. *J. Neurosci.* **2005**, *25*, 10369–10371. [CrossRef]

70. Nodine, C.F.; Mello-Thoms, C.; Kundel, H.L.; Weinstein, S.P. Time Course of Perception and Decision Making During Mammographic Interpretation. *Am. J. Roentgenol.* **2002**, *179*, 917–923. [CrossRef]

71. Cheddad, A.; Czene, K.; Eriksson, M.; Li, J.; Easton, D.; Hall, P.; Humphreys, K. Area and Volumetric Density Estimation in Processed Full-Field Digital Mammograms for Risk Assessment of Breast Cancer. *PLoS ONE* **2014**, *9*, e110690. [CrossRef]