Original Investigation

# Mammographic Breast Density Assessment Using Automated Volumetric Software and Breast Imaging Reporting and Data System (BIRADS) Categorization by Expert Radiologists

Christine N. Damases, MTech Radiography, Patrick C. Brennan, PhD, Claudia Mello-Thoms, PhD, Mark F. McEntee, BSc (Hons) Radiography, PhD

**Rationale and Objectives:** To investigate agreement on mammographic breast density (MD) assessment between automated volumetric software and Breast Imaging Reporting and Data System (BIRADS) categorization by expert radiologists.

**Materials and Methods:** Forty cases of left craniocaudal and mediolateral oblique mammograms from 20 women were used. All images had their volumetric density classified using Volpara density grade (VDG) and average volumetric breast density percentage. The same images were then classified into BIRADS categories (I–IV) by 20 American Board of Radiology examiners.

**Results:** The results demonstrated a moderate agreement ($\kappa = 0.537$; 95% CI = 0.234–0.699) between VDG classification and radiologists' BIRADS density assessment. Interreader agreement using BIRADS also demonstrated moderate agreement ($\kappa = 0.565$; 95% CI = 0.519–0.610) ranging from 0.328 to 0.669. Radiologists' average BIRADS was lower than average VDG scores by 0.33, with their mean being 2.13, whereas the mean VDG was 2.48 ($U = -3.742$; $P < 0.001$). VDG and BIRADS showed a very strong positive correlation ($\rho = 0.91$; $P < 0.001$) as did BIRADS and average volumetric breast density percentage ($\rho = 0.94$; $P < 0.001$).

**Conclusions:** Automated volumetric breast density assessment shows moderate agreement and very strong correlation with BIRADS; interreader variations still exist within BIRADS. Because of the increasing importance of MD measurement in clinical management of patients, widely accepted, reproducible, and accurate measures of MD are required.

**Key Words:** Breast density; Mammographic; Automated density assessment; Volpara; BIRADS categories.



## INTRODUCTION

Women with dense breasts have a two- to sixfold increased risk of breast cancer compared to fatty breasts (1). Mammographic density (MD), the most common measure of breast density, is defined by the relative amount of fat and fibroglandular tissue in the breast as seen on a mammogram. This is usually expressed as a percentage, where MD is the proportion of the breast area on a mammogram that is radiodense or opaque (2). However, area-based, two-dimensional measures of MD such as semiautomated Cumulus do not take the volume of density into account. It has been proposed that MD might be used to stratify women into different screening regimes, such as increasing the frequency of screening or using adjunctive imaging modalities for women with dense breasts (3,4). However, this would rely on reproducible and accurate measurement of MD, which to date has proven to be troublesome (5). Measurement reproducibility is important if MD is to be incorporated into breast screening imaging pathways and cancer risk predictive models (6).

There is no gold standard method for measuring actual breast density, although attempts to quantify this feature using mammography are well reported. The most common (conventional)

**Acad Radiol 2016;** 23:70–77

From the Faculty of Health Sciences, Discipline of Medical Radiation Sciences and Brain and Mind Research Institute, M205, Cumberland Campus, The University of Sydney, 75 East St, Room M205, Lidcombe, Sydney, NSW 2141, Australia (C.N.D., P.C.B., C.M-T., M.F.M.E.); Faculty of Health Sciences, Department of Radiography, University of Namibia, M-Block, Room M-105, Mandume Ndemufayo Avenue, Private Bag 13310, Windhoek 9000, Namibia (C.N.D.). Received June 16, 2015; revised August 31, 2015; accepted September 16, 2015. **Address correspondence to:** C.N.D. e-mail: cdam2504@uni.sydney.edu.au


http://dx.doi.org/10.1016/j.acra.2015.09.011

method in practice involves visual (qualitative) assessment by the radiologist, for example, the fourth edition of the Breast Imaging Reporting and Data System (BIRADS) (7,8), which defines MD from I (describing an entirely fatty breast) to IV (representing an extremely dense breast) (9). The recent update in the BIRADS standard changed numbered categories to letters and removed percentage descriptors from the four categories (7–9).

BIRADS assessment of MD has formed the basis for the majority of studies evaluating the importance of MD on both mammographic sensitivity and breast cancer risk (10–12). However, this method is subjective and not without its disadvantages; having low reproducibility as reported by Zhou et al. (13), being less reproducible than other methods (14,15), and demonstrating wide intra- and interreader disagreement, particularly in less-experienced readers (5,15,16).

Because of these difficulties with BIRADS, computer-assisted solutions for quantifying density have been developed (15,17). Computer-assisted quantification can either have some level of user involvement or the process can be fully automated. Fully automated systems (18,19), such as VolparaDensity (Volpara Solutions Limited, New Zealand) and Quantra (Hologic Inc., Bedford, MA), calculate the volume of dense tissue based on measurement and modeling, do not require manual intervention, and have been shown to be more reproducible than BIRADS (20). Recently, Alonzo-Proulx et al. presented data from a comparison of several fully automated volumetric methods, and showed that Volpara was the most reliable (20).

Volpara automatically measures volumetric MD and then converts these values to a Volpara density grade (VDG), a grading system that correlates with the American College of Radiology (ACR) BIRADS density grading classifications. The "For Processing" data are processed by Volpara using a standard mammography form algorithm to obtain an objective VDG score. VDG scores differ from the ACR BIRADS system in that Volpara uses quantitative volumetric measurements of fibroglandular and fatty tissue to calculate volumetric percent density rather than a qualitative visual assessment.

To date, data available comparing volumetric MD estimates to BIRADS assessment have used either a single radiologist (21) or less than five radiologists (22,23). Given the reported intra- and interreader variation, a larger sample of readers is required to determine the range of agreement that can be expected between Volpara scores and expert BIRADS assessment. The current work will address this deficiency, by comparing the two methods in terms of the density scores produced and the range of expert reader agreement. Previous work by our group has investigated the impact of differing mammographic imaging systems on MD assessment (24).

It has been acknowledge that the data set used in this paper has previously been analyzed to evaluate the impact of different mammographic imaging systems on breast density assessment (24) and to compare mammographic breast density assessment using automated volumetric software with BIRADS categorization by radiologists (25). The results of these studies have previously been reported (24,25). The results of previous work (24) demonstrated that the effect of using two mammographic imaging systems on volumetric MD measurement was negligible. The focus of that article was to investigate the impact of mammography systems on MD assessment, whereas the current article is aimed at investigating agreement between automated volumetric software and BIRADS categorization by expert radiologists in mammographic breast density assessment.

## MATERIALS AND METHODS

Institutional review board approval was provided for the study (IRB 2013/448). Twenty American Board of Radiology examiners volunteered to participate in this study.

### Selection of Images

A test set of 40 cases was obtained from 20 women between the ages of 42 and 89 years. Previous studies on breast density assessment reported typical age range between 30 and 86 years (23,26). This age range is appropriate as it reflects the whole spectrum of age ranges normally encountered in mammography. Mammograms were acquired 1 year apart on a GE Healthcare Senographe Essential (or DS) and a Hologic Lorad Selenia. The 40 cases each contained three images: a left craniocaudal (LCC), a left mediolateral oblique (LMLO), and a combined image of the LCC and the LMLO. To ensure observers could evaluate the images in an acceptable time frame, only the left breast images were used for this study. For each breast density assessment method, the ratings were calculated using the mean for both the GE system and the Hologic system. The mean BIRADS is the mean rating given by the radiologists for the GE and the Hologic systems combined. Likewise, the mean VDG and average volumetric breast density percentage (AvBD%) is the mean rating given by Volpara over all images.

### Image Display and MD Quantification using BIRADS

Images were displayed on a single five-megapixel diagnostic quality monitor (EIZO, Japan) using ViewDEX 2.0 (27). During evaluation of the images, the ambient lighting was kept between 25 and 35 lux as confirmed by a calibrated photometer (model 07–621, Nuclear Associates) (28). Radiologists had the ability to adjust the window width and level, as well as pan and zoom the image. For each case, there were three images: first the LCC, followed by the LMLO, and finally the combined LCC and LMLO presented together. A score on the fourth edition ACR BIRADS density scale of I–IV was recorded with (I) The breast is almost entirely fat (<25% glandular); (II) There are scattered fibroglandular densities (approximately 25%–50% glandular); (III) The breast tissue is heterogeneously dense, which could obscure detection of small masses (approximately 51%–75% glandular); and (IV) The breast

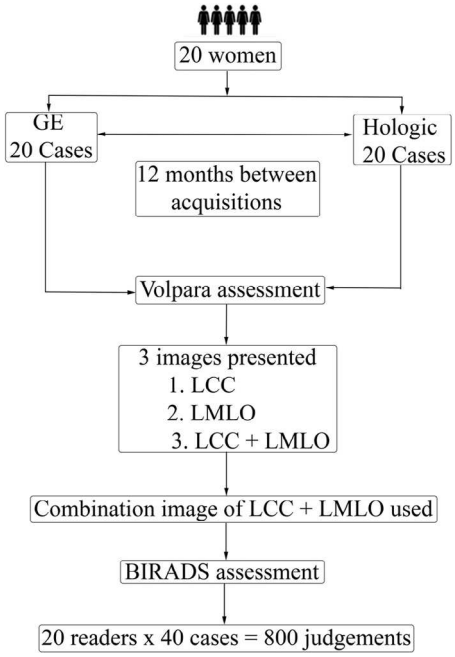


**Figure 1.** Schematic overview of image selection.

tissue is extremely dense. Denser breasts may lower the sensitivity of mammography (7–9). To replicate clinical practice, radiologists made one overall judgment from the combined image of LCC and LMLO for each of the 40 cases. A total of 800 MD judgments were made (20 readers × 40 cases) (Fig 1). The breast densities were then grouped as low and high or binary classification; low including BIRADS I and II and high including BIRADS III and IV.

### Breast Density Quantification Using Volpara Automated Software

The combined images of the left breast were categorized using Volpara imaging software version 1.4.3 into VDG categories 1–4. The software also generates automatic measurement of volumetric MD values reported as the AvBD%. Using the following preset thresholds, the density percentages are classified as: VDG 1 = 0–4.5%, VDG 2 = 4.5–7.5%, VDG 3 = 7.5–15.5%, and VDG 4 = over 15.5% (29). Of the 40 cases of MD measured using Volpara, nine were in VDG 1, 13 in VDG 2, eight in VDG 3, and 10 in VDG 4.

### Agreement Between BIRADS and Volpara (VDG and AvBD%)

VDG thresholds were established based on the BIRADS density assessment of one radiologist (30). Although using a single expert has advantages such as no interreader variability, the findings of a single expert may not represent the population of radiologists. Therefore, data from the current study were reanalyzed to examine the effect of changing the preset Volpara thresholds on agreement between BIRADS and Volpara.

### Statistics

Statistical analyses were performed using SPSS 21.0 (SPSS, Chicago, IL). Wilcoxon signed ranks test was used to compare BIRADS to Volpara (VDG). Spearman rank coefficient of correlation ($\rho$) was used to examine the relationship between BIRADS and VDG and between BIRADS and AvBD%. The agreement on BIRADS categorization between the 20 readers was compared and expressed as Cohen's kappa ($\kappa$) with the use of a $20 \times 20$ matrix. Agreement was examined using the four-point scale and binary classifications, respectively. Results were considered to be statistically significant at $P < 0.05$. One-way analysis of variance with a Bonferroni post hoc test was used to determine whether there were any significant differences in reading times spent on cases in each of the four BIRADS categories.

## RESULTS

All 20 participants were specialized in breast imaging, and had a mean 24.9 years (sd = 8.3) specialization, ranging from 11 to 42 years. The mean annual number of mammograms read was 7107 (sd = 5,308) with a range of 840–25,000.

The percentage of images in each BIRADS and VDG category are shown in Figure 2. Radiologists allocated the images to each of the four BIRADS classifications as follows: BIRADS I = 34%, BIRADS II = 26%, BIRADS III = 31%, BIRADS IV = 9%, whereas VDG allocated the images as VDG 1 = 22.5%, VDG 2 = 32.5%, VDG 3 = 20%, and VDG 4 = 25%. Radiologists ranked density using BIRADS significantly lower than VDG with a mean of 2.13 versus 2.48 ($Z = -3.742$; $P < 0.001$).

### Correlation Between BIRADS and Volpara (VDG and AvBD%)

Very strong positive correlations were found between four-point scale BIRADS with VDG ($\rho = 0.91$; $P < 0.001$) and AvBD% ($\rho = 0.94$; $P < 0.001$), as well as between binary scale BIRADS and VDG ($\rho = 0.90$; $P < 0.001$). No significant correlation was demonstrated between years of experience ($\rho = -0.265$; $P < 0.260$) and number of cases read ($\rho = -0.122$; $P < 0.608$) with the radiologists' interreader agreement.

### Agreement Between BIRADS and Volpara (VDG and AvBD%)

There was moderate agreement between four-point scale BIRADS and VDG with a kappa ($\kappa$) of 0.537 (95% CI = 0.234–0.699) as shown in Table 1, whereas binary scale BIRADS and VDG demonstrated an almost perfect agreement ($\kappa = 0.898$; 95% CI = 0.701–1.000).

Agreement between four-point scale BIRADS and VDG increased from moderate ($\kappa = 0.537$) to substantial ($\kappa = 0.659$) (31) by changing Volpara preset thresholds for VDG 1 from 0%–4.5% to 0%–4.9%; for VDG 2 from 4.5%–7.5% to

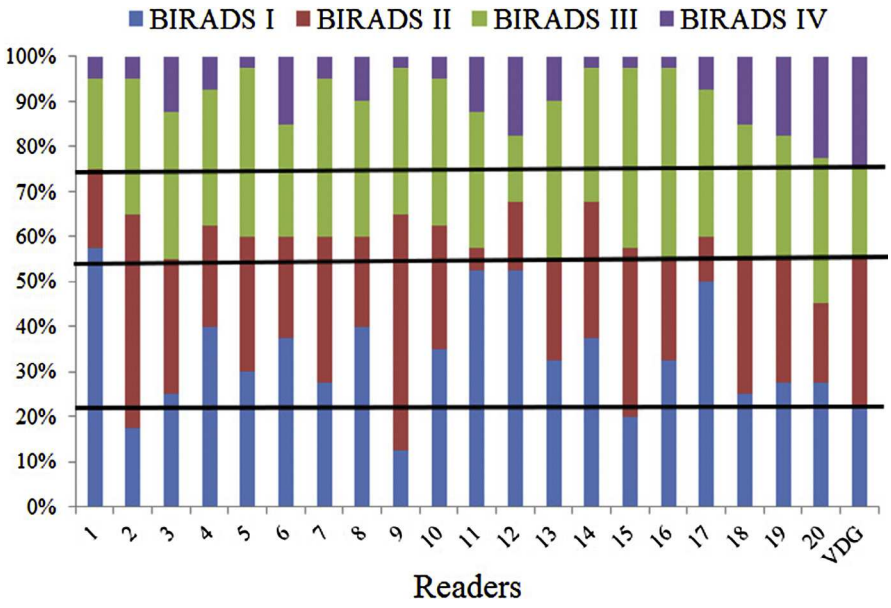


**Figure 2.** The percentage MD categorization by each radiologist. Each reader's images were judged by Volpara as: VDG 1 = 22.5%, VDG 2 = 32.5%, VDG 3 = 20%, and VDG 4 = 25%. (Color version of figure available online).

**TABLE 1. Mean Difference in Time Spent on Each BIRADS Category Assessment by Radiologists**

| | | | | | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| BIRADS | BIRADS | Mean Difference | Std. Error | *P* Value | Lower Bound | Upper Bound |
| I | II | –0.55317 | .51747 | 1.000 | –1.9218 | .8155 |
| | III | –2.54092* | .49325 | .000 | –3.8455 | –1.2363 |
| | IV | –3.96609* | .74457 | .000 | –5.9354 | –1.9968 |
| II | III | –1.98775* | .52820 | .001 | –3.3848 | –.5907 |
| | IV | –3.41293* | .76817 | .000 | –5.4446 | –1.3812 |
| III | IV | –1.42518 | .75207 | .351 | –3.4143 | .5640 |

* The mean difference is significant at the 0.05 level.

4.9%–8.7%; for VDG 3 from 7.5%–15.5% to 8.7%–17.1%; and for VDG 4 from >15.5% to >17.1%.

### Interreader Agreement on BIRADS Categorization

The interreader agreement for radiologists using four-point scale BIRADS had an average kappa of 0.565 (95% CI = 0.519–0.610), and ranged between 0.328 and 0.669. Figure 3 demonstrates a large variation in the distribution of radiologists' BIRADS categories. The interreader agreement for radiologists using binary scale BIRADS had an average kappa of 0.855 (95% CI = 0.824–0.866), and ranged between 0.656 and 0.901. BIRADS categories I and II took significantly less time for radiologists to rate than BIRADS III and IV at $F = 15.517$; $P < 0.001$ as shown in Table 1.

Although there was agreement between BIRADS and VDG in 26 of 40 cases, the current study demonstrates that VDG generally provides a higher assessment of MD than BIRADS. For instance, there were five cases rated as BIRADS I that were given a VDG score of 2 by Volpara, whereas there were seven cases that were rated as BIRADS III that were given a VDG 4 by Volpara (Fig 4). Overall, Volpara VDG rated

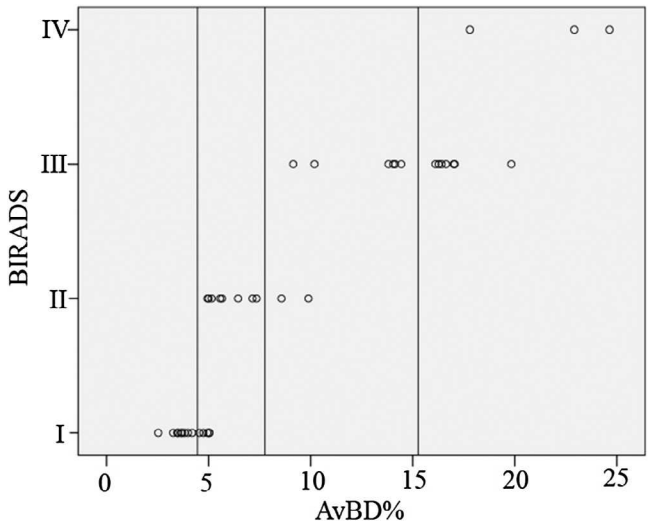


**Figure 3.** Scatter plot demonstrating radiologists' BIRADS classifications compared to AvBD%.

the MD higher in 14 cases, whereas BIRADS rated the MD as lower than VDG in 14 cases. For the crucial II or III classifications, there were differences in only two of 40 classifications (Table 2).

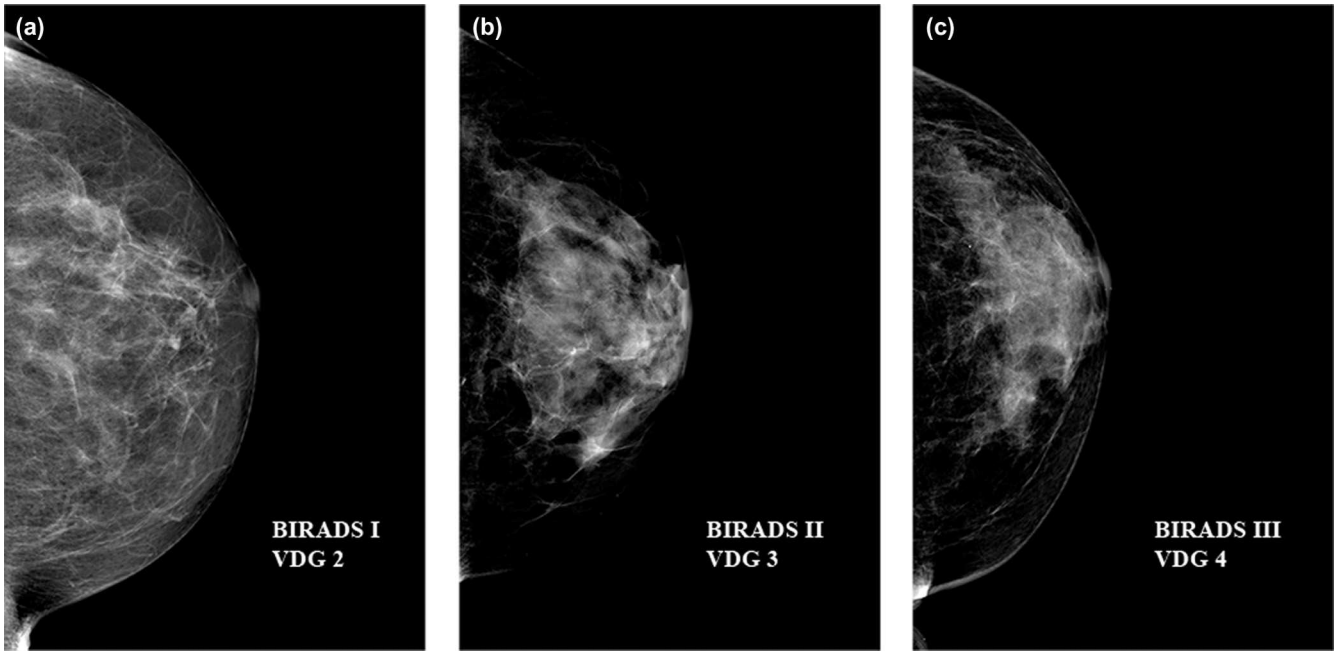


**Figure 4.** Examples of mammograms showing discrepancy between BIRADS and VDG. These mammograms demonstrate the variations in breast density between BIRADS and Volpara. (**a**) Majority of the radiologists rated this case as BIRADS I, whereas Volpara rated the same case as VDG 2. (**b**) Majority of the radiologists rated this case as BIRADS II, whereas Volpara rated the same case as VDG 3. (**c**) Majority of the radiologists rated this case as BIRADS III, whereas Volpara rated the same case as VDG 4.

**TABLE 2. The Number of Images Allocated to BIRADS Categories I, II, II, and IV in Our Research Setting and VDG Categories 1, 2, 3, and 4 as well as the Number of Images on which BIRADS and VGD Agreed for Each Category (Shown in Squares)**

| | | VDG | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | Total |
| BIRADS | I | 9 | 5 | 0 | 0 | 14 |
| | II | 0 | 8 | 2 | 0 | 10 |
| | III | 0 | 0 | 6 | 7 | 13 |
| | IV | 0 | 0 | 0 | 3 | 3 |
| Total | | 9 | 13 | 8 | 10 | 40 |

## DISCUSSION

The current work found moderate agreement between the BIRADS MD assessments of 20 radiologists and Volpara. Interreader variation continues to be an issue for BIRADS, as a wide range of interreader agreement is seen on a four-point scale ($\kappa = 0.33$–$0.67$). On a binary scale, BIRADS expert radiologists had an almost perfect average interreader agreement with a range of $\kappa = 0.656$–$0.901$. Martin et al. also previously demonstrated that there is a range of interreader agreement for MD assessment, although their range was smaller at 0.61–0.76 (32). Ciatto et al. examined interreader variability during MD assessment using BIRADS and demonstrated a wide range of kappa values, with interobserver variations ranging from $\kappa = 0.02$ to $\kappa = 0.77$ (33). Even though BIRADS is the most commonly used method for MD assessment, it is subjective and results in only moderate reproducibility. Automated methods, such as Volpara, have no interreader variability (23,30,32,34). Therefore, there is strong evidence to conclude that large interreader variations will continue to be an issue with BIRADS assessment, and that volumetric MD assessments such as Volpara are inherently more reproducible (20).

### Comparison Between BIRADS and Volpara MD Assessments

Wide ranges of AvBD% for each BIRADS category demonstrated that a single BIRADS category is likely to contain AvBD% that extend beyond the thresholds assigned by Volpara (Fig 3). For example, BIRADS I had AvBD% ranging from 2.5% to 5%. Volpara indicates that a BIRADS I should range from 0% to 4.5%; therefore, the range of AvBD% seen for the radiologists' BIRADS assessments exceeds this range. The same observation was made for BIRADS II, which ranged from 4.9% to 9.9%, again outside the Volpara AvBD% range of 4.5–7.5%. BIRADS III exceeded the AvBD% ranges of 7.5–15.5% as set by Volpara, ranging from 9.2% to 19.8%, whereas BIRADS IV had a range of AvBD% from 17.8% to 24.6%. Radiologists' range for BIRADS IV did not violate Volpara's recommended threshold for VDG 4 category as set by Volpara to be >15% (Fig 3). This demonstrates that there are multiple AvBD% (as classified by the software) in every category of BIRADS, as classified by the experts. For the same amount or volume of dense tissue in the breast, radiologists

are assigning a range of BIRADS scores. Although there was 65% agreement between BIRADS and VDG in assessment of 40 cases, the current study demonstrates that VDG generally provides a higher assessment of MD than BIRADS (Fig 4).

Martin et al. reported similar findings in their study comparing BIRADS MD categories and computer-aided MD estimates using Mammographic Density Estimation (MDEST) (32). They demonstrated that BIRADS I had a percentage density range of 1–11%; BIRADS II had a range from 2% to 45%; BIRADS III had the largest range from 8% to 60%; and BIRADS IV had a range from 20% to 82%. Although MDEST is a different method of MD assessment and uses different percentage categorizations, it demonstrates a similar pattern of large variations by radiologists across percentages indicated by automated MD categorization systems.

The pattern of a wide range of quantitative percentage densities in each BIRADS category was also demonstrated by Jeffreys et al. (35). The authors compared BIRADS MD classifications of one expert radiologist with the MD classifications of Cumulus (36–38) and Volpara. They demonstrated that BIRADS I had a percentage density range of 2.5–7.0%, BIRADS II from 3% to 13%, BIRADS III from 4% to 22%, and BIRADS IV from 8% to 31% with Volpara. The results of Jeffreys et al.'s study also indicated that Volpara AvBD% correlated well with the Cumulus MD% ($r = 0.85$). Thus the literature supports our finding that a single BIRADS category is likely to contain AvBD% that extends beyond the thresholds assigned by Volpara and that there are wide interreader variations in the assessment of BIRADS MD.

Variability in assessment of MD can potentially influence the individualized breast cancer screening pathway of a woman. This variability becomes crucial when it affects the classifications that separate "low" from "high" density, that is, those that could fall in either BIRADS I and II or BIRADS III and IV. Women with high MD are more likely to be referred for ultrasound or magnetic resonance imaging than those with low density, particularly in the United States, where density notification legislation has been passed in many states. Nicholson et al. examined the binary classification of MD using two readers' consensus on BIRADS MD classification. The authors reported a median reader-assigned mammographic percentage density of 6.0% with a range from 0.5% to 19.2% for fatty breasts (BIRADS I), 1.2%–52.7% for scattered densities (BIRADS II), 15.9%–82.2% for heterogeneously dense (BIRADS III), and 60.1%–87.9% for extremely dense breasts (BIRADS IV) (12). The authors reported that the reader-assigned percent MD ranges for fatty (BIRADS I) ($r = 2.4$–13.1) and extremely dense breasts (BIRADS IV) ($r = 70.5$–87.2) correlated well with BIRADS definitions, whereas the ranges of densities in the scattered (BIRADS II) ($r = 5.8$–32.2) and heterogeneously dense (BIRADS III) ($r = 31.0$–67.7) categories were considerably wider. This means that the differentiation between BIRADS II and III was generally poor, creating a problem for the classification of dense breast. Radiologists are more reliable in assigning BIRADS IV to dense tissue but BIRADS II and III have variations sufficient to cause considerable doubt about classifying the breast as dense. When the agreement between BIRADS and VDG on a binary scale was assessed, they demonstrated an almost perfect agreement with a kappa of 0.898 (95% CI = 0.701–1.000).

The moderate agreement between BIRADS and VDG and the apparent higher density ratings with VDG might be addressed through adjusting the thresholds used to classify VDG categories. Improving agreement between BIRADS and VDG through adjusting Volpara preset thresholds based on feedback from a larger number of expert readers is a novel and important finding of this work. Improved agreement is likely to result in improved adoption of automated volumetric systems. When Volpara preset thresholds were changed, the overall agreement between BIRADS and VDG improved from moderate ($\kappa = 0.537$) to substantial ($\kappa = 0.659$). This improved agreement is supported by Morris et al. who demonstrated "scores with the area-based density given by Quantra yielded a low correlation ($r = 0.55$, $P < .001$). Correlations of observer's scores with the volumetric density results gave $r$ values of 0.60 ($P < .001$) and 0.63 ($P < .001$) for Quantra and Volpara, respectively," and they suggested that this could be because radiologists use a semivolumetric approach to measuring visual density (39). Modeling the effect of changes in thresholds on the agreement between BIRADS and VDG may lead to greater improvements; however, these improvements may be population and reader specific, and may need to be adjusted for the population of radiologists and breast densities in each country.

Our previous work indicated that the change in MD takes many years to manifest; however, for women going through menopause or taking hormonal therapy, the difference in readings between the images taken 1 year apart could be substantial. The women in our study showed no such large change in breast density over 1 year period. As the images were collected retrospectively, we do not know the menopausal status for the selected cases. A study by Holland et al. (40) demonstrated high reproducibility of MD measurement with Volpara compared to BIRADS over a mean screening interval period of 22.65 months between their two studies. The authors reported that in 89.7% of the cases, MD remained in the same category. In 3.2% of the pairs, an increase in percentage density was actually reported, resulting in a change from nondense to dense category. The authors report that this effect may have been due to differences in the breast thickness measurement in the Digital Imaging and Communications in Medicine (DICOM) headers. Our previous work demonstrated that MD did not change over a period of one year as the correlation in AvBD% between year 1 and 2 was $r = 0.997$ ($P \leq 0.001$) and no large differences in VDG were noted.

The findings of the current work are based on cases taken from US women assessed by American radiologists. Future work might ascertain the need for further assessment as an outcome rather than a density measurement itself; for example, does the patient need ultrasound or further imaging as a result of MD. Knowledge of a woman's MD can be used to predict her risk to cancer and personalize her imaging pathway.

However, measurement of MD has proven to be troublesome with wide variations in density recorded by radiologists. Having an accurate global standard for MD measurement will eliminate unnecessary differences in clinical decision-making for women with dense breast; thus, further work will investigate whether the geographic location of the readers impacts their perception of density. Further work will also investigate whether for the same breast test set the range of breast density assigned by experts varies by country and whether the scale of the variation differs from country to country.

Previous studies on MD assessment have used smaller numbers of radiologists, typically 10 or less (41,42). Thus, a strength of the current work is the consensus result of 20 radiologists. This study has several limitations. Firstly, there is no widely accepted ground truth for MD measurement and thus a reference standard to evaluate MD does not exist. Secondly, Volpara reference thresholds have been calibrated based on one radiologist estimation, which might have affected the outcome of the results. Thirdly, only the left breast was used for BIRADS estimation, and it is not known whether including the right breast would have increased or decreased the correlation between BIRADS and Volpara. A power calculation was performed and indicated that a sample size of 79 images was required for power of 0.80 at an alpha of 0.05. As the sample size for the current study was 40 images, this results in a lower power of 0.51 at an alpha of 0.05. Finally, the radiologists may not have been familiar with the presentation state of the images and may be used to a different presentation, and this may have affected their conclusions on density.

## CONCLUSIONS

This work demonstrates that expert radiologists differ in their MD assessment on a four-point scale and less so on a binary scale. Reduction in the variability of MD assessment may require a widely acceptable, reproducible, and accurate method of assessment. MD measurements with automated methods eliminate subjectivity and are more reproducible than BIRADS. Because of the increasing importance of MD measurement in clinical management of patients, automated methods are recommended to avoid unnecessary variations.

## ACKNOWLEDGMENTS

The authors acknowledge Dr. Ralph Highnam and Dr. Ariane Chan of Volpara Solutions Ltd for their assistance with data analysis and statistical support for this study. The authors thank Professor Steve Hillis, University of Iowa, Patrick Kasi, University of Western Sydney, and Ziba Gandomkar, University of Sydney for their assistance with statistical support.

## REFERENCES

1. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2006; 15:1159–1169.
2. Ursin GQS. Mammographic density—a useful biomarker for the breast cancer risk in epidemiologic studies. Norsk Epidermiologi 2009; 19:59–68.
3. Yaffe MJ. Mammographic density. Measurement of mammographic density. Breast Cancer Res 2008; 10:209.
4. Harvey JA, Bovbjerg VE. Quantitative assessment of mammographic breast density: relationship with breast cancer risk. Radiology 2004; 230:29–41.
5. Ng KH, Yip CH, Taib NAM. Standardisation of clinical breast-density measurement. Lancet Oncol 2012; 13:334–336.
6. Shepherd JA, Kerlikowske K, Ma L, et al. Volume of mammographic density and risk of breast cancer. Cancer Epidemiol Biomarkers Prev 2011; 20:1473–1482.
7. D'Orsi CJ, Bassett LW, Berg WA, et al. BI-RADS mammography. In: D'Orsi CJ, Mendelson EB, Ikeda DM, et al., eds. Breast imaging reporting and data system: ACR BI-RADS—breast imaging atlas. 4th ed. Reston, VA: American College of Radiology, 2003:61–128.
8. Radiology ACo. The American College of radiology BIRADS ATLAS and MQSA: frequently asked questions. Available at: http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/BIRADS/BIRADSFAQs.pdf. Accessed September 25, 2012.
9. Radiology ACo. BI-RADS mammography 2013-ACR BI-RADS atlas. 5th ed. Available at: http://www.acr.org/Quality-Safety/Resources/BIRADS. Accessed March 17, 2014.
10. Andersson I, Ikeda DM, Zackrisson S, et al. Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. Eur Radiol 2008; 18:2817–2825.
11. Duffy SW, Nagtegaal ID, Astley SM, et al. Visually assessed breast density, breast cancer risk and the importance of the craniocaudal view. Breast Cancer Res 2008; 10:R64.
12. Nicholson BT, LoRusso AP, Smolkin M, et al. Accuracy of assigned BI-RADS breast density category definitions. Acad Radiol 2006; 13:1143–1149.
13. Zhou C, Chan HP, Petrick N, et al. Computerized image analysis: estimation of breast density on mammograms. Med Phys 2001; 28:1056–1069.
14. McCormack VA, Highnam R, Perry N, et al. 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.
15. Ciatto S, Bernardi D, Calabrese M, et al. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. Breast 2012; 21:503–506.
16. Ooms EA, Zonderland HM, Eijkemans MJC, et al. Mammography: interobserver variability in breast density assessment. Breast 2007; 16:568–576.
17. Highnam R, Brady M, Yaffe MJ, et al. Robust breast composition measurement: Volpara. In: Marti J, Oliver A, Freixenet J, et al., eds. Digital mammography. 2010; 342–349.
18. Tagliafico A, Tagliafico G, Tosto S, et al. Mammographic density estimation: comparison among BI-RADS categories, a semi-automated software and a fully automated one. Breast 2009; 18:35–40.
19. Byng JW, Boyd NF, Fishell E, et al. Automated analysis of mammographic densities. Phys Med Biol 1996; 41:909–923.
20. Alonzo-Proulx O, Mawdsley GE, Patrie JT, et al. Reliability of automated breast density measurement. Radiology 2015; 275:366–376.
21. Sauber N, Chan A, Highnam R. BI-RADS breast density classification—an international standard. ECR; 2013.
22. Wang K, Chan A, Highnam R. Robustness of automated volumetric breast density estimation for assessing temporal changes in breast density. ECR; 2015.
23. Gweon HM, Youk JH, Kim JA, et al. Radiologist assessment of breast density by BI-RADS categories versus fully automated volumetric assessment. Am J Roentgenol 2013; 201:692–697.
24. Damases CN, Brennan PC, McEntee MF. Mammographic density measurements are not affected by mammography system. J Med Imag 2015; 2:15501–15505.
25. McEntee MF, Damases CN. Mammographic density measurement: a comparison of automated volumetric density measurement to BIRADS. Med Imag 2014: Image Percept, Obs Perform, Technol Assess 2014; 9037:8.

26. Wanders JO, Holland K, Veldhuis WB, et al. Effect of volumetric mammographic density on performance of a breast cancer screening program using full-field digital mammography. ECR 2015.
27. Börjesson S, Håkansson M, Båth M, et al. A software tool for increased efficiency in observer performance studies in radiology. Radiat Prot Dosimetry 2005; 114:45–52.
28. Brennan PC, McEntee M, Evanoff M, et al. Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist. Am J Roentgenol 2007; 188:W177–W180.
29. Volpara Solutions. Volpara clinical breast density and its implications for your patients. Available at: http://www.volparadensity.com/clinicians/. Accessed March 27, 2013.
30. Berg WA, Campassi C, Langenberg P, et al. Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. Am J Roentgenol 2000; 174:1769–1777.
31. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005; 37:360–363.
32. Martin KE, Helvie MA, Zhou C, et al. Mammographic density measured with quantitative computer-aided method: comparison with radiologists' estimates and BI-RADS categories. Radiology 2006; 240:656–665.
33. Ciatto S, Houssami N, Apruzzese A, et al. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. Breast 2005; 14:269–275.
34. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998; 90:1801–1809.
35. Jeffreys M, Harvey J, Highnam R. Comparing a new volumetric breast density method (Volpara™) to Cumulus. In: Marti J, Oliver A, Freixenet J, et al., eds. Digital mammography. Berlin: Springer-Verlag Berlin, 2010; 408–413.
36. Byng JW, Boyd NF, Fishell E, et al. The quantitative analysis of mammographic densities. Phys Med Biol 1994; 39:1629–1638.
37. Byng JW, Yaffe MJ, Jong RA, et al. Analysis of mammographic density and breast cancer risk from digitized mammograms. Radiographics 1998; 18:1587–1598.
38. Boyd NF, Byng JW, Jong RA, et al. 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.
39. Morris OWE, Tucker L, Black R, et al. Mammographic breast density: comparison of methods for qualitative evaluation. Radiology 2015; 275:356–365.
40. Holland K, Kallenberg M, Mann R, et al. Stability of volumetric tissue composition measured in serial screening mammograms. In: Fujita H, Hara T, Muramatsu C, eds. Breast imaging. Switzerland: Springer International Publishing, 2014; 239–244.
41. Schilling K, The J, Griff S, et al. Impact of quantitative breast density on experienced radiologists' assessment of mammographic breast density. Vienna, Austria: European Congress of Radiology, 2015;C-1281.
42. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. NEJM 1994; 331:1493–1499.