Mammographic Breast Density: Comparison of Methods for Quantitative Evaluation

Oliver W. E. Morrish, MSc
Lorraine Tucker, DCR
Richard Black, MSc
Paula Willsher, DCR
Stephen W. Duffy, MSc
Fiona J. Gilbert, MD

Purpose:
To evaluate the results from two software tools for measurement of mammographic breast density and compare them with observer-based scores in a large cohort of women.

Materials and Methods:
Following written informed consent, a data set of 36,281 mammograms from 8,867 women were collected from six United Kingdom centers in an ethically approved trial. Breast density was assessed by one of 26 readers on a visual analog scale and with two automated density tools. Mean differences were calculated as the mean of all the individual percentage differences between each measurement for each case (woman). Agreement in total breast volume, fibroglandular volume, and percentage density was assessed with the Bland-Altman method. Association with observer’s scores was calculated by using the Pearson correlation coefficient (r).

Results:
Correlation between the Quantra and Volpara outputs for total breast volume was $r = 0.97$ ($P < .001$), with a mean difference of 43.5 cm$^3$ for all cases representing 5.0% of the mean total breast volume. Correlation of the two measures was lower for fibroglandular volume ($r = 0.86$, $P < .001$). The mean difference was 30.3 cm$^3$ for all cases representing 21.2% of the mean fibroglandular tissue volume result. Quantra gave the larger value and the difference tended to increase with volume. For the two measures of percentage volume density, the mean difference was 1.61 percentage points ($r = 0.78$, $P < .001$). Comparison of observer’s 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.

Conclusion:
Automated techniques for measuring breast density show good correlation, but these are poorly correlated with observer’s scores. However, automated techniques do give different results that should be considered when informing patient personalized imaging.

© RSNA, 2015

Clinical trial registration no. ISRCTN 73467396
The estimation and measurement of breast density has been reported in many studies during the past 2 decades (1–9). This has been partly driven by the role of breast density as a strong, independent, and modifiable risk factor of breast cancer (3–5). Most of the work has used mammograms for the evaluation of density where the appearance of the breast reflects variations in the relative amounts of fat, connective tissue, and epithelial tissue and their different x-ray attenuation characteristics (10). Breast density is expressed as a percentage of the mammogram occupied by the fibroglandular and stromal tissue (9).

Assessment of breast density is traditionally performed by a reader on the basis of a visual assessment of the standard two-view mammogram. Consistency of this measure requires an experienced observer to be able to correctly assess the relative proportions of glandular and fatty tissue while accounting for variations in breast shape, size, fibroglandular pattern, and the various radiographic factors used. Density scores are then given either on a continuous scale as a percentage (11) or within discrete ranges, such as the four-point Breast Imaging Reporting and Data System, BI-RADS (12), scale or the Boyd five-point scale (5). Studies suggest that training and experience are essential in ensuring that the scores are accurate and reproducible (12,13).

The introduction of full-field digital mammography technologies has provided an opportunity to implement automated breast density measurement algorithms, which had been initially developed for digitized analog mammograms (11,14). These algorithms work by applying thresholds to the pixel values within the digital image to identify the area of the image that contains the breast and to then determine the proportion of that breast which contains fibroglandular tissue. For example, the pixel values with the highest signal (radiation dose detected by the pixel) can identify the areas of the image where no breast tissue has attenuated the primary x-ray beam. The areas of lowest signal, on the other hand, represent areas where the x-rays have passed through a section of tissue that is relatively most attenuating (15).

Later developments have led to software that estimates the volume of dense fibroglandular tissue rather than just the area represented on the mammogram. By using the image pixel data in combination with information about the acquisition parameters (eg, compression paddle height, x-ray tube potential, target, and filter), more recent algorithms are capable of providing estimates of the relevant tissue volumes by means of derivation of the tissue composition represented by each pixel (16–18).

The researchers in one small study (19) have compared a range of volumetric software tools and found that the density measurements were in substantial agreement with data from breast magnetic resonance (MR) imaging. The aim of this work was to evaluate the results from two software tools for the measurement of mammographic breast density and compare them with observer-based scores in a large cohort of women.

### Materials and Methods

Support for this study was given by two companies—Hologic (Bedford, Mass) and Matakanaka Technology (Wellington, New Zealand). Both provided automated breast density software and technical advice. The authors had full access to all data in the study and the information submitted for publication.

### Study Data

Standard two-dimensional digital mammograms (n = 36281) were obtained from the TOMosynthesis with digital...
MammographY, or TOMMY, trial in the United Kingdom National Health Service Breast Screening Program (20) (clinical trial registration ISRCTN 73467396) by using the full-field digital mammography system (Selenia Dimensions; Hologic) installed in six centers in the United Kingdom. All the systems underwent regular quality control testing to demonstrate ongoing compliance with expected system performance (21). The trial was approved by the Scotland A Research Ethics Committee, and participants gave written informed consent.

Study Cohort

The study cohort included women aged 47–73 years who were recalled to an assessment clinic following abnormal screening mammography results and women aged 40–49 years with a family history of breast cancer who were attending annual screening. The imaging protocol for the trial included acquisition of standard two-view mammograms (cranial caudal [CC] and mediolateral oblique [MLO]) of both breasts, and these images were used for the density assessment. A complete set of images for a total of 8867 individual women were available for analysis.

Reader Assessment

In the study, 26 readers were asked to assess breast density on a visual analog scale, giving a score ranging from 0% to 100%. For each subject, a single score assigned by a single reader was obtained at review of the mammograms, except for subjects with a family history of breast cancer, in whom two readers assigned scores. These readers, who were not authors of this article, consisted of 21 radiologists, two breast physicians, and three advanced practitioner radiographers, each with a minimum of 2 years of experience reading at least 5000 mammograms per year. The visual analog scale is a 10-cm line on a paper form on which readers indicate their density score by making an appropriate mark, with the left end of the line representing 0% and the right end representing 100%. Each center then digitized and processed the forms to extract density scores for each subject on the basis of the position of the reader’s mark (11). In four centers, this resulted in a scale of scores given to the nearest whole percentage; however, one center rounded scores to the nearest 5%, and another center rounded scores to the nearest 10%. When the readers were scoring the images, they were advised not to alter the window levels of images.

Density Assessment with Software

Two software packages were used to assess the breast density on each mammogram (Quantra version 2.0, Hologic; and Volpara version 1.4.2, Matakana Technology) (17,22). The output from each program consisted of values for the absolute volume of fibroglandular tissue and overall breast volume, as well as the volumetric breast density on a per-image basis. In addition, with Quantra, an area-based score was determined.

To obtain the overall score for each examination, the absolute values of total breast volume and fibroglandular tissue volume for each of the four views (left and right, CC and MLO) were examined. For cases (women) in which no cancer was assessed as being present, the largest total breast volume and the largest fibroglandular tissue volume for each breast (either from the CC or MLO view) were separately determined, and the average tissue volumes of the two breasts were calculated. The same logic was applied to the scores determined with the Quantra system for area breast density (which should nominally be comparable with the observer scores).

For cases in which cancer was confirmed with histopathologic examination, results were used from the contralateral breast. We believe this procedure to mirror the behavior of observers. Also, the rejection of data from the affected breast reflects methods used in cancer cohort studies in which mammographic density is assessed from the contralateral mammogram (5,9). If no contralateral data were available, results from the affected breast were used. Volumetric density was calculated from the ratio of the fibroglandular tissue volume to the overall breast volume.

Statistical Analysis

Agreement in the measurements of absolute tissue volumes and density measurements was assessed by using the Bland-Altman method. The mean difference between each measurement was derived by subtracting the relevant Volpara value from the Quantra value and then calculating the mean value of this difference for all cases. The mean percentage difference was derived by expressing the difference for each case as a percentage of the mean of the two measurements and then calculating the mean percentage for all cases. The relationship between the various measurements was assessed by calculating the Pearson correlation coefficient (r). Logarithmic transformation of the data was performed where examination of residual plots indicated that the data would behave more linearly and variance would be more stable when compared with the untransformed data. These analyses were further performed on the fibroglandular tissue volumes above and below 300 cm² (mean volume) to assess the software’s performance at greater fibroglandular volumes.

For the subset of women with a family history of breast cancer, breast density was assessed by two observers at each center. In these cases, agreement between the observers’ scores was analyzed by examination of the absolute difference between the scores.

Results

A complete set of standard mammographic images (right MLO, left MLO, right CC, left CC) from 8867 individual women were available for analysis. However, the software was unable to produce scores for every image analyzed. Reasons for this inability were not available for the Quantra algorithm; Volpara supplied error messages for each image without a score. Reasons for algorithm failures were varied and are summarized in Table 1. Most of the unexpected errors appeared to be related to the incorrect use of certain exposure settings on the acquisition workstation, leading to inconsistent values being used in the
DICOM image header. Also, as the data set was very large, consisting of 36281 images, we were unable to manually exclude cases with additional views (eg, magnification or mosaic views).

The logical combination of maximal tissue volumes from each view resulted in scores being used from the MLO view in most of the cases. From the Quantra results, where there were 16957 cases of paired CC and MLO scores, the MLO view gave the largest total breast volume in 13953 cases and the largest fibroglandular tissue volume in 12347 cases. From the Volpara results where there were 16991 cases of paired CC and MLO scores, the MLO view gave the largest total breast volume in 14469 cases and the largest fibroglandular tissue volume in 9200 cases. When we combined these volumes from both breasts for each woman, there were therefore cases where the CC view from one breast was averaged with the MLO view from the other. For the Quantra results, this averaging occurred in 1471 cases for total breast volume and in 2495 cases for fibroglandular tissue volume. For the Volpara results, this averaging occurred in 1337 cases for total breast volume and in 2544 cases for fibroglandular tissue volume.

The summary of results for the study cohort is shown in Table 2. An overall density score was assigned by observers in 8391 cases. Quantra scores were available for 8512 women, and 8532 women had scores calculated from Volpara software. Observer scores were therefore unavailable for 5.4% (476 of 8391) of the cohort, and Quantra and Volpara scores were not generated for 4.0% (355 of 8867) and 3.8% (335 of 8867) of the cohort, respectively.

Figure 1 illustrates the comparison of total breast volume, as measured with the Quantra and Volpara software, throughout the study population. The Pearson coefficient for correlation between Quantra breast volume and Volpara breast volume was $r = 0.97$ ($P < .001$). The mean difference between the values calculated for each case was 43.5 cm$^3$, and the mean percentage difference between the values was 5.0%, suggesting good agreement between the two systems.

![Figure 1](image1.png)

**Table 1**

<table>
<thead>
<tr>
<th>Error*</th>
<th>No. of Images</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast implant present: yes</td>
<td>32</td>
<td>On review, not all of these images had implants present; however, all were stated as such in the DICOM image header, suggesting occasional errors in the use of the ‘implant present’ mode at the acquisition workstation</td>
</tr>
<tr>
<td>Invalid quantitative values—not a standard mammogram?</td>
<td>9</td>
<td>Images that were in this category were ones with implants present but not indicated in the DICOM header and spot compression images; a mammogram where dense tissue was found largely just behind the nipple also appeared in this category</td>
</tr>
<tr>
<td>Magnification factor too large (&gt; 1.1)</td>
<td>770</td>
<td>These were all magnification images that were not removed from the image data set</td>
</tr>
<tr>
<td>Filter material invalid/unknown</td>
<td>5</td>
<td>These were specimen images identified as such at acquisition; this leads to there being no filter specified in the DICOM header</td>
</tr>
<tr>
<td>Possible mosaic</td>
<td>292</td>
<td>Many of these were mosaic view where there was only partial coverage of the breast; some of the images in this category were spot compression views</td>
</tr>
<tr>
<td>No background—not a standard mammogram?</td>
<td>379</td>
<td>All of these images turned out to be specimen images where the specimen view was not selected at acquisition</td>
</tr>
<tr>
<td>No energy—all fat found</td>
<td>1</td>
<td>This was an implant image not indicated at acquisition</td>
</tr>
<tr>
<td>Recorded breast thickness invalid</td>
<td>3</td>
<td>Two of these were specimen images where a breast thickness of 0 mm is used in the DICOM header; the other image appeared to be a magnification view where the x-ray system has not recognized it as such (ie, broad focus and grid are indicated with a thickness of 317 mm)</td>
</tr>
</tbody>
</table>

Note.—DICOM = Digital Imaging and Communications in Medicine.

* The error messages are as transcribed from the software.

**Table 2**

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Observers</th>
<th>Quantra</th>
<th>Volpara</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of cases*</td>
<td>8391</td>
<td>8512</td>
<td>8532</td>
</tr>
<tr>
<td>Breast volume (cm$^3$)</td>
<td>. . .</td>
<td>953.5 (73.0–4986.5)</td>
<td>921.4 (33.4–5009.3)</td>
</tr>
<tr>
<td>Fibroglandular tissue volume (cm$^3$)</td>
<td>. . .</td>
<td>93.0 (4.0–1024.0)</td>
<td>71.6 (6.8–628.5)</td>
</tr>
<tr>
<td>Area breast density (%)</td>
<td>37.0 (0, 100)</td>
<td>14.8 (0–76.5)</td>
<td>. . .</td>
</tr>
<tr>
<td>Volumetric breast density (%)</td>
<td>. . .</td>
<td>9.5 (1.4–56.2)</td>
<td>7.7 (2.5–54.2)</td>
</tr>
</tbody>
</table>

Note.—Unless otherwise indicated, numbers are median values, and numbers in parentheses are the minimum and maximum values, respectively, in the ranges.

* Cases = women.

Figures 2 and 3 show the distribution and comparison of the fibroglandular tissue volumes within the breast reported by the two algorithms. The Pearson coefficient for correlation between the natural logarithm
of Quantra fibroglandular volume and the natural logarithm of Volpara fibroglandular volume was $r = 0.86 \ (P < .001)$. The mean difference between the values calculated for each case was 30.3 cm$^3$, and the mean percentage difference between the values was 21.2%, with Quantra giving the larger of the two values. The differences between the two measurements increase as the amount of fibroglandular tissue increases. This is particularly noticeable for data above 300 cm$^3$, where the mean difference between cases is 226.4 cm$^3$ and the correlation coefficient reduces to 0.15 ($P = .06$) if only those data are considered. The remaining data below 300 cm$^3$ have a mean difference between cases of 26.4 cm$^3$ and a correlation coefficient of 0.85 ($P < .001$). In addition, the range of results for fibroglandular volume is greater for the Quantra system, with most values lying between 0 and 400 cm$^3$, as opposed to results with Volpara, with values lying between 0 and 300 cm$^3$.

Figures 4 and 5 show the results for volumetric breast density. The Pearson coefficient for correlation between the natural logarithm of Quantra volumetric density and the natural logarithm of Volpara volumetric density was $r = 0.78 \ (P < .001)$. The mean difference between the values calculated for each case was 1.61 percentage points, and the mean percentage difference between the values was 16.3%.

Figures 6 and 7 show the results for breast density estimated from the projected mammogram, the area-based breast density score assigned by the observers, and the Quantra software. The Volpara software did not give an area-based density result. The Pearson coefficient for correlation between the natural logarithm of the observer density scores and the natural logarithm of the Quantra area-based density was $r = 0.55 \ (P < .001)$.

Figure 8 shows a comparison of the area-based breast density assigned a score by the observers and the volumetric density measurement from each program. The Pearson coefficient for correlation between the observer density scores and the natural logarithms of the volumetric densities was $r = 0.60 \ (P < .001)$ for Quantra and $r = 0.63 \ (P < .001)$ for Volpara, although there does appear to be a better correlation in breasts with lower density. The
large numbers of values given at each 5% mark on the observer histogram in Figures 7 and 8 are a consequence of the way that the visual analog scale was processed in two centers, with one rounding results to within the nearest 5% and the other rounding results to the nearest 10%. Figure 9 shows how the Quantra outputs of area-based and volumetric density compare.

For the cases in women with a family history of breast cancer, density was assessed by two observers at each of the centers, giving us 638 cases with two scores for comparison. In 70% (449 of 638) of these cases, the score agreed to within 10%; however, in 8% of cases, the score disagreed by more than 20%. The 54 cases with more than 20% disagreement were found across the whole range of densities, with 33 of them in the density range of 50%–75%. In 13% (85 of 638) of cases, one reader assigned a score to the density above 50%, with the other reader assigning a score to the density below 50%.

**Discussion**

The two automated density assessment techniques have relatively good agreement in the evaluation of the overall breast volume; there was less agreement in the assessment of the fibroglandular volumes. We presume from the agreement in total volume that there are similarities in the way each algorithm identifies the breast against the background, while each applies their own corrections to account for compression paddle height and tilt and to estimate the volume at the edge of the breast where the paddle is not in contact with the breast (17,19,23).

Although there was relatively good agreement at lower fibroglandular volumes, it became poorer as those volumes increased, particularly above 300 cm³. The reasons for this discrepancy are most likely due to the differing ratios of fibroglandular, adipose, and skin tissues allocated to each pixel, on the basis of their relative x-ray attenuation with reference to pixels in the image defined as pure adipose or fatty tissue (19,22). As density increases, it becomes harder to identify these reference areas, and each manufacturer’s solution to this problem is likely to result in differences in the final...
volumetric density results (17). Further differences might be the way that skin and fibroglandular tissue are differentiated in the assessment of dense tissue (18,24,25).

While the specific corrections that each algorithm makes to account for paddle tilt are unknown because of their proprietary nature, Kallenberg et al (26) have shown that such corrections significantly influence fibroglandular tissue volumes and volumetric densities. In larger fatty breasts, the paddle tilt generates an inhomogeneity in the mammogram, which can cause dense tissue volumes to be overestimated, especially when intentionally flexible paddles are used (23). Conversely, in dense breasts, the increased height at the chest wall edge of the mammogram can lead to an underestimation of density (26).

There is low correlation between observer scores of breast density and automated analysis scores. Observer measurement of breast density has been shown in other studies to be affected by interobserver variability (12,25,27,28). When the histograms of the area-based measurements from both human observers and software analysis are examined, there is clearly a difference in the distribution of scores. This may in part be caused by the observers’ application of a semivolumetric approach to the assessment rather than a purely area-based approach.

---

**Figure 5**: Bland-Altman plot of volumetric breast density reported by Volpara subtracted from that reported by Quantra compared with the mean of the two results. Middle dashed line = mean difference and top and bottom dashed lines = 95% limits of agreement (± 2 standard deviations).

**Figure 6**: Histograms show distribution of area-based breast density across the study population from the observers (left) and the Quantra software (right).
One possible technical reason for the difference between observer-based density scores and automated scores could be the processing of the displayed image. The software analyzes the raw digital data, whereas the observers make their estimation of density on processed images optimized for display at the workstation and have the ability to alter the window width and level of the gray scale applied to the image’s pixels. This adjustment can substantially alter the image presentation. In this study, readers were advised not to alter the window levels.

In the current study, it is not possible to know what the ground truth is with regard to breast density measurement. Both of the automated density measurement tools used in this study have been shown to correlate with measurements made with MR imaging (17,19,24), with methods using a calibration phantom (21) and with visual assessment using Breast Imaging Reporting and Data System categories (17,30). Correlation, however, does not necessarily mean agreement between methods.

In recent work, Wang et al (19) retrospectively analyzed the breast density methods (25,29), it is the software that produces the more comparable results.
of 99 women and compared the performance of Quantra and Volpara with that of a fuzzy-clustering segmentation method on MR images. Total breast volume with both mammographic measurement tools was highly correlated with MR imaging results ($r = 0.95$), although these researchers reported a greater difference between Quantra and Volpara than we have found. One possible explanation may be their use of the CC image and our use of the higher score from either the MLO or the CC view. This is also likely to have an effect on the volumetric density measurements that were reported to be a median of 22.0% and 13.3% for Quantra and Volpara, respectively, which are larger than those found in our work. For fibroglandular volume, the difference between the reported median values was 36.2 cm$^3$, with Quantra giving the larger score, which is similar to that found in our study.

One of the limitations of this study was that, despite the same cases being submitted to each of the three density measurement methods, the failure in some cases to provide a score means that the population used for comparing each of the three methods is different. Density is not routinely assessed in the National Health Service Breast Screening Programme, and so, it was sometimes overlooked during image reading. However, because each comparison of any two methods uses the results where the populations coincide, and are therefore identical, the effect of this difference is likely to be small. Also, the visual analog scale data were rounded to 5% in some cases and 10% in others, which presents another limitation of the study, although, again, the effect of this difference is unlikely to be large. Finally, to obtain percentage density scores and absolute volumes, we combined the density scores from each image given by the two software tools using the same logic for both. In their derivation of a category-based score for the subject, the software tools may combine the results using a different method than that used in the current study. The fact that the women selected for this study are from two different groups, those with abnormal mammographic results and those with a family history of breast cancer, is not thought to be a confounding factor.

Breast density has been correlated with risk of developing breast cancer ($3\text{--}5,14,31$). Because cancers originate in the fibroglandular tissue, it is proposed that measurements that describe its volume would be a better predictor of risk than those based on projected area ($1,6,31$). If volumetric density is to be used to estimate breast cancer risk, it is important that the measurements are reliable. Technical differences in the way in which each software package determines the fibroglandular tissue volume, and therefore the density, produce different values. This factor needs to be considered when density is used to inform patient personalized imaging.


Disclosures of Conflicts of Interest: O.W.E.M. financial activities related to the present article: institution received a grant from National Institute for Health Research’s Health Technology Assessment Programme; institution received payment for provision of Quantra software by Hitachi Technology. Financial activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. R.B. financial activities related to the present article: institution received a consulting fee or honorarium for data processing and computer services from University of Cambridge, Department of Radiology. Financial activities not related to the present article: disclosed no relevant relationships. S.W.D. disclosed no relevant relationships. L.T. disclosed no relevant relationships. P.W. disclosed no relevant relationships. S.W.D. disclosed no relevant relationships. E.J.G. financial activities related to the present article: institution received a grant from National Institute for Health Research’s Health Technology Assessment Programme; institution received payment for loaning of Quantra software to enable density measurements of the
mammograms by Hologic and for density measurements of mammograms using Volpara software by Mammotakina Technology. Financial activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

References


