Volumetric Breast Density Assessment: Reproducibility in Serial Examinations and Comparison with Visual Assessment

Volumetric Breast Density Assessment: Reproducibility in Serial Examinations and Comparison with Visual Assessment

Authors

J. M. Singh1, E. M. Fallenberg1, F. Diekmann2, D. M. Renz2, R. Witlandt1, U. Bick1, F. Engelken1

Affiliations

1 Department of Radiology, Charité – Universitätsmedizin Berlin, Campus Mitte, Berlin
2 Department of Radiology, Charité – Universitätsmedizin Berlin, Campus Virchow, Berlin

Key words

- mammography
- breast cancer
- breast density
- volumetric measurement

Zusammenfassung


Ergebnisse: Interexamination-Reproduzierbarkeit für die volumetrische Bestimmung des Brustparenchymanteils betrug 0,91 (ICC; 95 % CI 0,87–0,93). Es bestand kein Unterschied im Ausmaß der Reproduzierbarkeit zwischen Patientinnen mit starker vs. geringer Abweichung der aufgebrachten Kompressionskraft für die unterschiedlichen Untersuchungen. Intra- und Interrater-Variabilität reichten von 0,81–0,84 und 0,71–0,77. Interexamination-Reproduzierbarkeit des visuellen Assessments betrug 0,75–0,81. Übereinstimmung von visuell-erhobener und volumetrischer Dichtebestimmung war vergleichbar mit der Übereinstimmung der Reader untereinander.

Schlussfolgerung: Unsere Ergebnisse zeigen, dass die volumetrische Brustdichtebestimmung eine höhere Reproduzierbarkeit für serielle Untersuchungen aufweist als die optische Bestimmung. Die volumetrische Bestimmung könnte daher im

Abstract

Background and Purpose: Mammographic breast density is the strongest known marker of breast cancer risk. Visual breast density assessment is subject to significant intra- and inter-rater variability. The aim of the present study was to test the reproducibility of automatic breast density assessment and to compare the results to the visual assessment.

Patients and Methods: Serial mammograms of 141 patients were retrospectively reviewed. Breast density was assessed both visually using a BI-RADS four-category breast density scale and with a software tool for volumetric breast density measurement.

Results: The intra- and inter-rater reproducibility as well as inter-examination reproducibility were assessed for both techniques by calculating the intraclass correlation coefficient (ICC). The inter-examination reproducibility of the volumetric measurement of breast percent density was 0.91 (ICC; 95 % CI 0.87–0.93). There was no difference in the strength of the correlation between patients with a large vs. small difference in compression force. The intra- and inter-rater reproducibility ranged from 0.81–0.84 and 0.71–0.77, respectively. The inter-examination reproducibility of visual assessment was 0.75–0.81. The agreement of visual assessment with volumetric measurement was similar to the agreement among readers.

Conclusion: Our results indicate that volumetric breast density measurement provides higher reproducibility in serial examinations than visual assessment and may thus be preferable in the longitudinal assessment of breast density and in the measurement of breast density for risk stratification.

Citation Format:

Materials and Methods

Patients
We searched our records from June 2002 to December 2006 for patients satisfying the following inclusion criteria: two consecutive examinations performed on the same mammography unit no more than 24 months apart, raw image data stored in the picture archiving and communication system (PACS), unremarkable mammography reports for at least one breast, and minimum of 18 months of normal follow-up of the eligible breast(s). The exclusion criteria were: previous surgery on the eligible breast(s), change in hormone status such as starting or stopping hormone-replacement therapy or menopause, and technical deficits of the mammogram such as inadequate positioning or presence of large skin folds.

A total of 170 patients were identified. Raw image data of the two consecutive mammography examinations were sent to an R2 Cenova server for analysis by the R2 Quantra breast density assessment algorithm. In 29 patients, the algorithm failed to produce results for one or both examinations. These patients were excluded from the analysis. Therefore, 141 patients were included in the study. In 21 patients, the algorithm produced results but marked the results as potentially inaccurate. This occurs when there is a discrepancy between the measurements in the CC and MLO projections. This was recorded for subgroup analysis of the reproducibility.

Only one breast per patient was chosen for analysis to avoid linkage of data points. If both breasts were eligible for analysis, one side was chosen at random. Institutional review board was obtained.

Image acquisition
All patients underwent digital mammography using the same full-field digital mammography system with a flat-panel detector and a cesium iodide absorber, field size 19 × 23 cm, pixel size 100 μm, image matrix size 1914 × 2294 (Senographe 2000 D, General Electric Healthcare, Chalfont St. Giles). All mammograms were acquired in standard cranio-caudal and mediolateral oblique projections using automatic optimization of acquisition parameters and standard supplier presets.

Image analysis
For all patients included in the study, breast density was assessed both visually and with the automatic software tool. Visually, breast density was assessed by three independent, board-certified radiologists of our hospital using the BI-RADS lexicon. Reading was performed on a diagnostic mammography workstation (syngo MammoReport, Siemens Medical, Erlangen, Germany) in a blinded manner without knowledge of the woman’s age, the original mammography interpretation, and risk profile for breast cancer. The three observers independently assessed the mammograms for breast density, assigning one of the BI-RADS breast density categories on a standardized form. The first mammogram of each patient was read first, followed by another reading session for the second mammogram after an interval of 4 weeks or more. In a third reading session, again after an interval of at least 4 weeks, the first set of mammograms was read a second time to estimate the intra-rater reproducibility.

The BI-RADS scheme of breast densities, developed by the American College of Radiology (ACR) is intended to provide a standardized classification system for mammographic studies. The ACR classification identifies four categories of breast composition: (1) the breast is almost entirely fat (<25% glandular); (2) there are scattered fibroglandular densities (25–50% glandular); (3) the breast is almost entirely dense; and (4) the breast is almost entirely fatty. This classification is widely used in clinical practice and provides a standardized method for reporting mammographic findings.

The BI-RADS scheme of breast densities, developed by the American College of Radiology (ACR) is intended to provide a standardized classification system for mammographic studies. The ACR classification identifies four categories of breast composition: (1) the breast is almost entirely fat (<25% glandular); (2) there are scattered fibroglandular densities (25–50% glandular); (3) the breast is almost entirely dense; and (4) the breast is almost entirely fatty. This classification is widely used in clinical practice and provides a standardized method for reporting mammographic findings.
breast tissue is heterogeneously dense (approximately 51–75% glandular); and (4) the breast tissue is extremely dense (>75% glandular).

For the software-based analysis, raw image data were sent to a dedicated server running the R2 Quantra software. Briefly, R2 Quantra™ is a software tool for automatically calculating volumetric breast density from the ratio of fibro glandular tissue to the estimated total breast volume. The algorithm uses a physical model of the imaging process to deduce the density and composition of breast tissue from the degree of X-ray attenuation on mammograms. To achieve this, the algorithm estimates the amount of fibro glandular tissue an X-ray beam must have passed to deposit the amount of energy measured at the detector. Images are processed within minutes. The output of the R2 Quantra software includes the estimated total breast volume and fibro glandular tissue volume in ml (cm³) and the calculated breast PD (Fig. 1).

Statistical analysis

Data analysis was performed using statistical software packages (SPSS, version 18.0; SPSS Chicago, Illinois; MedCalc 12.3.0). The intra- and inter-rater reproducibility as well as the inter-examination reproducibility of the visual and software-based analysis were assessed by calculating the intraclass correlation coefficient (ICC). For comparison with other studies of visual density assessment, quadratic-weighted kappa values were also calculated for the intra- and inter-rater reproducibility. For the correlation of categorical BI-RADS density levels of examinations 1 and 2 versus ordinal volumetric breast density values, BI-RADS classes 1 – 4 were replaced with the mean PD value of the respective category (1 = 12.5%; 2 = 37.5%; 3 = 62.5%; 4 = 87.5%), and the ICC was calculated.

To investigate the effects of different compression forces on breast density estimates by volumetric assessment, we assigned the patients to one of four subgroups based on the magnitude of the difference in compression force applied for the first and the second mammogram in each patient. For each subgroup, the inter-examination agreement of the measured breast density was determined. Differences in correlation coefficients were tested for statistical significance using the Fisher r-to-z transformation.

Results

The patients had a mean age of 62 years (range, 45–78 years). 61 patients underwent mammography in the setting of surveillance after breast surgery and had one unaffected breast. The remaining 80 patients had workup of a palpable lump or unclear ultrasound findings. The median interval between the first and the second examination was 13.2 months with a range of 9–24 months. 29 patients were premenopausal, 112 patients were postmenopausal. Of the premenopausal patients, 6 patients took oral contraceptive agents. Of the postmenopausal patients, 12 received hormone replacement therapy and 17 received antihormonal therapy.

The results for inter-rater agreement in visual breast density assessment between pairs of observers for both examinations, 1 and 2, are summarized in Table 1. The inter-rater agreement ranged from 0.71 – 0.77 (ICC). Table 2 summarizes the results for intra-rater agreement for examination 1, the inter-examination variability for raters and volumetric measurements, as well as the comparison between visual breast density assessment and volumetric analysis. The inter-rater agreement ranged from 0.81 – 0.84 (ICC). The inter-examination agreement of examinations 1 and 2 for individual readers varied from 0.75 – 0.81 versus 0.91 for volumetric analysis. The difference in the strength of correlation between volumetric and visual assessment was statistically significant for all readers and constellations (p<0.01). In patients where breast density was marked as potentially inaccurate by the R2 Quantra software, the inter-examination agreement was 0.90 (95% confidence interval 0.77–0.96). Table 3 shows the inter-examination correlation of the volumetric analysis of the whole group and for the four subgroups based on magnitude of difference in compression forces. The inter-examination correlation of the volumetric analysis was sim-

![Fig. 1](image)

**Fig. 1** Representative mammogram of the right breast in craniocaudal and mediolateral oblique projection and the corresponding datasheet provided by the QuantraR2 software.

<table>
<thead>
<tr>
<th>Factor</th>
<th>rater A vs. B</th>
<th>rater A vs. C</th>
<th>rater B vs. C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>0.71 (0.62–0.78)</td>
<td>0.77 (0.71–0.86)</td>
<td>0.77 (0.69–0.83)</td>
</tr>
<tr>
<td>κ</td>
<td>0.69 (0.61–0.77)</td>
<td>0.76 (0.79–0.82)</td>
<td>0.69 (0.58–0.78)</td>
</tr>
</tbody>
</table>

**Table 1** Inter-rater variability. Intraclass correlation coefficients (ICC) and quadratic-weighted kappa values were calculated. Numbers in parentheses represent 95% confidence intervals.

**Table 2** Interrater-Variabilität. Es wurden die Intraklassen-Korrelationskoeffizienten (ICC) und quadratisch-gewichteten Kappa-Koeffizienten (κ) bestimmt. Zahlen in Klammern stellen die jeweiligen 95% Konfidenzintervalle dar.
Table 2  Intrarater agreement, agreement of R2 Quantra and visual assessment, and inter-examination agreement for visual and software-based breast density assessment. Numbers in parentheses represent 95 % confidence intervals.

<table>
<thead>
<tr>
<th></th>
<th>rater A</th>
<th>rater B</th>
<th>rater C</th>
<th>quantra PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>intrain-ration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>0.83 (0.78 – 0.88)</td>
<td>0.81 (0.74 – 0.86)</td>
<td>0.84 (0.77 – 0.88)</td>
<td>/</td>
</tr>
<tr>
<td>K</td>
<td>0.81 (0.75 – 0.87)</td>
<td>0.80 (0.72 – 0.87)</td>
<td>0.82 (0.75 – 0.89)</td>
<td>/</td>
</tr>
<tr>
<td>agreement Quantra vs. visual assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>examination 1</td>
<td>0.68 (0.58 – 0.76)</td>
<td>0.68 (0.58 – 0.76)</td>
<td>0.65 (0.55 – 0.74)</td>
<td>/</td>
</tr>
<tr>
<td>examination 2</td>
<td>0.69 (0.59 – 0.77)</td>
<td>0.63 (0.51 – 0.83)</td>
<td>0.73 (0.64 – 0.80)</td>
<td>/</td>
</tr>
<tr>
<td>inter-examination agreement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>0.75 (0.67 – 0.83)</td>
<td>0.81 (0.74 – 0.86)</td>
<td>0.76 (0.67 – 0.84)</td>
<td>0.91* (0.87 – 0.93)</td>
</tr>
</tbody>
</table>

* indicates statistical significance of the difference in ICC compared with all other ICC values (p<0.01).

Tab. 3 Interexamination-Reproduzierbarkeit der Software-basierten Analyse in Abhängigkeit vom Ausmaß der Kompressionskraftschwankungen zwischen den zwei Mammografieuntersuchungen. Zahlen in Klammern repräsentieren die 95 % Konfidenzintervalle.

<table>
<thead>
<tr>
<th>difference in compression force</th>
<th>N (%)</th>
<th>inter-examination reproducibility (ICC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 39 N</td>
<td>49 (35 %)</td>
<td>0.89 (0.82 – 0.94)</td>
</tr>
<tr>
<td>40 – 79 N</td>
<td>51 (36 %)</td>
<td>0.92 (0.86 – 0.95)</td>
</tr>
<tr>
<td>80 – 119 N</td>
<td>27 (19 %)</td>
<td>0.92 (0.83 – 0.96)</td>
</tr>
<tr>
<td>≥ 120 N</td>
<td>14 (10 %)</td>
<td>0.91 (0.73 – 0.97)</td>
</tr>
<tr>
<td>total</td>
<td>141 (100 %)</td>
<td>0.90 (0.87 – 0.93)</td>
</tr>
</tbody>
</table>

Breast density has been shown to be the strongest known risk factor for breast cancer [1 – 4, 17]. There is some evidence that breast density may reflect changes in breast cancer risk associated with interventions such as tamoxifen treatment [18]. From a clinical perspective, breast density has a strong effect on mammographic sensitivity [19, 20]. Future breast cancer screening programs may employ individualized screening regimens for women according to their personal breast cancer risk as well as their chance of benefiting from additional procedures like breast ultrasound or digital breast tomosynthesis [21, 22]. Therefore, accurate and reproducible measurement of breast density is very desirable both in the clinical and research setting. The results of our study show that volumetric analysis provides highly reproducible measurements of breast density in consecutive examinations and clearly exceeds the performance of human readers. The method appears to be robust with respect to differences in breast compression as well as the small differences in breast orientation and projection angle, which may occur in consecutive examinations. Volumetric analysis is therefore preferable to visual assessment in the setting of longitudinal studies of breast density. Most studies investigating the reproducibility of breast density assessment have looked at intra- and inter-rater reproducibility. Software-based volumetric analysis always yields the same result when confronted with the same mammogram, thereby eliminating intra- and interobserver variability. As immediate acquisition of a second mammogram after a satisfactory mammogram has been obtained is not possible for ethical reasons, we used serial mammograms for estimating the reproducibility of the method. The reproducibility of visual breast density assessment has been shown to be substantial but not perfect [9 – 11]. Interactive thresholding in one study of digitized film mammograms improved both the inter- and intra-rater reproducibility, with an increase in the intraclass coefficients to 0.84 – 0.94 and 0.93 – 0.99, respectively.[12] Another study showed better correlation of the Cumulus method with another automated density assessment algorithm than with the four-category BI-RADS scale on digitized mammograms [13]. However, ours is the first study to investigate the reproducibility of breast density assessment in serial examinations. Three-dimensional imaging techniques, such as MR volumetry and digital breast tomosynthesis, may yield similar information and potentially provide more accurate volume measurements. However, the strength of quantifying breast tissue density from...
digital mammograms is that these are inexpensive and widely available. A current limitation of the software is the failure rate of around 8.5% observed in this study, which may be improved with future developments.

The results of our study are relevant both to the use of this method in longitudinal studies and to the comparison of results obtained in different imaging centers, where variations in imaging technique cannot be fully avoided. The lack of reader interaction and the avoidance of intra-rater variability represent notable advantages over alternative breast density assessment approaches. It should be noted that the high reproducibility (precision) of this method does not allow assumptions about its accuracy, i.e. the closeness of the software result to the true breast composition. While a highly accurate measurement would be highly reproducible, high reproducibility does not prove high accuracy. However, the high reproducibility of this algorithm means that changes in breast density over time will be detected with much higher precision by volumetric assessment than by visual assessment.

The major limitation of our study is the long interval between consecutive mammography examinations in the same patients. While 1–2 years is the minimum interval for performing serial mammography after an initial unremarkable mammogram, this is long enough for changes in weight to occur and changes in hormone levels to manifest. The reproducibility found in this study, therefore, very likely represents an underestimate.

In conclusion, volumetric breast density measurement is highly reproducible in serial mammograms in a routine clinical setting. The performance significantly exceeds the reproducibility of visual assessment by human readers. The method appears robust with respect to variations in breast compression. Given the lack of reader interaction and the avoidance of intra- and inter-rater variability, this method is a useful tool for longitudinal studies of breast density and for the quantification of breast density for breast cancer risk stratification.

Acknowledgement

This manuscript is dedicated to Professor Bernd Hamm for his 60th birthday.

References