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

Volumetrische Bestimmung der Brustdichte: Reproduzierbarkeit in seriellen Mammografie-Untersuchungen und Vergleich mit visueller Dichtebestimmung

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-Varibility 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 Brustdichtemessung eine höhere Reproduzierbarkeit für serielle Untersuchungen aufweist als die optische Bestimmung. Die volumetrische Bestimmung könnte daher im...
longitudinalen Assessment der Brustdichte sowie für die Dich-
temessung des Brustgewebes zur Risikostratifizierung der opti-
schen Dichtebestimmung vorzuziehen sein.

Introduction

Mammographic breast density has been shown to be one of the strongest known markers of breast cancer risk and has been pro-
posed as a variable for individual risk assessment [1–4]. Some in-
vestigators have used breast density as an intermediate end point
for interventional studies [5, 6]. An assessment of radiographic
breast density is required in every mammography report and is
an important variable in research studies. Breast density may in
the future become a factor for individualizing breast cancer
screening regimens according to each woman’s risk profile and
the expected sensitivity of mammography given her individual
breast density [7, 8]. A number of different reporting schemes
have been developed, with the American College of Radiologists
Breast Imaging Reporting and Data System (BI-RADS) perhaps
being the most widely used system. However, visual assessment
of breast density has limited intra- and interobserver reproduc-
bility [9–11]. A variety of approaches have been tested to object-
tify breast density assessment [12–15]. A drawback of most
approaches is the demand on reader time, which limits their use
in the clinical setting and population studies alike. Given the im-
portance of breast density for risk stratification, an accurate, fast,
and reproducible method for assessing breast density is needed
[4, 16]. Volumetric breast density measurement provides an esti-
mate of breast percent density (PD) without reader interaction.
The method uses a model of the imaging chain to estimate total
breast volume and the amount of glandular tissue present. An
important advantage of this approach is that it avoids subjectiv-
ity, which is introduced whenever different readers rate breast
density on the same study, as this software-based method always
produces the same result when presented with identical image
input. However, while the algorithm has been calibrated to vol-
ume measurements of sample breasts, it is not clear to what ex-
tent the produced result deviates from reality. Also, there is no
study data on the reproducibility of measurements when there is
variation in data input, for example in repeated examinations of
the same patient, when differences in projection angle, breast
compression and image acquisition parameters may affect the
apparent breast density. Given that a possible application of this
algorithm is its use in longitudinal studies of breast density, this
application requires testing in a sample of consecutive mammom-
grams. This information is necessary for estimating the magni-
tude of error resulting from variations in the imaging chain and
provides a measure of the reproducibility of the process as a
whole. The aim of this study was to assess the reproducibility of
breast density assessment using the R2 Quantra software in seri-
al mammography examinations and to compare its performance
with that of human readers.

Materials and Methods

Patients

We searched our records from June 2002 to December 2006 for
patients satisfying the following inclusion criteria: two consecu-
tive examinations performed on the same mammography unit
no more than 24 months apart, raw image data stored in the pic-
ture 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 ex-
clusion criteria were: previous surgery on the eligible breast(s),
change in hormone status such as starting or stopping hor-
mone-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
CenovaTM server for analysis by the R2 QuantraTM breast density
assessment algorithm. In 29 patients, the algorithm failed to pro-
duce results for one or both examinations. These patients were
excluded from the analysis. Therefore, 141 patients were inclu-
ded 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 link-
age of data points. If both breasts were eligible for analysis, one
side was chosen at random. Institutional review board was ob-
tained.

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 craniocaudal and mediolateral oblique pro-
jections 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 wom-
án’s age, the original mammography interpretation, and risk
profile for breast cancer. The three observers independently as-
sessed 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 an-
other reading session for the second mammogram after an inter-
val 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 Amer-
ican College of Radiology (ACR) is intended to provide a standard-
ized 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

Singh JM et al. Volumetric Breast Density... Fortschr Röntgenstr 2013; 185: 844–848
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 fibroglandular 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 fibroglandular 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 fibroglandular 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 1 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 intra-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 intervals, 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-
The aim of our study was to assess the reproducibility of breast density measurement in consecutive examinations using volumetric breast density analysis software and to compare the results with the performance of human readers. We found substantial, but not excellent, intra- and interobserver reproducibility of the visual density classification, comparable to the results reported by other studies. The inter-examination reproducibility of visual assessment was equal to or slightly less than the intra-examination reproducibility, depending on the reader. In comparison, breast density measurement by volumetric analysis showed an excellent inter-examination reproducibility, which was significantly higher than that of human readers. There was good agreement of the readers’ results with the volumetric analysis. We found no influence of differences in breast compression on the reproducibility of volumetric breast density analysis. Results that were marked as discrepant in CC and MLO views and therefore potentially inaccurate by the software were as reproducible as results that were not marked as potentially inaccurate.

Discussion

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 algorithm than with the four-category BI-RADS scale on digitized mammograms 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

Table 2 Intra-rater 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>intra-rater agreement</th>
<th>rater A</th>
<th>rater B</th>
<th>rater C</th>
<th>quantra PD</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

Table 3 Inter-examination reproducibility of software-based analysis by magnitude of difference in compression forces between the two mammography examinations. Numbers in parentheses represent 95% confidence intervals.

<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>
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