Inter-observer reproducibility according to three methods of evaluating mammographic density and parenchymal pattern: Impact on risk prediction

Poster No.: C-1037
Congress: ECR 2015
Type: Scientific Exhibit
Authors: R. R. Winkel, M. von Euler-Chelpin, M. Nielsen, M. Bachmann Nielsen, W. Uldall, P. Diao, I. Vejborg; Copenhagen/DK
Keywords: Breast, Mammography, Observer performance, Comparative studies, Cancer, Epidemiology, Image registration
DOI: 10.1594/ecr2015/C-1037

Any information contained in this pdf file is automatically generated from digital material submitted to EPOS by third parties in the form of scientific presentations. References to any names, marks, products, or services of third parties or hypertext links to third-party sites or information are provided solely as a convenience to you and do not in any way constitute or imply ECR's endorsement, sponsorship or recommendation of the third party, information, product or service. ECR is not responsible for the content of these pages and does not make any representations regarding the content or accuracy of material in this file.
As per copyright regulations, any unauthorised use of the material or parts thereof as well as commercial reproduction or multiple distribution by any traditional or electronically based reproduction/publication method is strictly prohibited.
You agree to defend, indemnify, and hold ECR harmless from and against any and all claims, damages, costs, and expenses, including attorneys' fees, arising from or related to your use of these pages.
Please note: Links to movies, ppt slideshows and any other multimedia files are not available in the pdf version of presentations.
www.myESR.org
Aims and objectives

*Breast density* or *parenchymal pattern* of the breast are major risk factors for breast cancer with a four to six-fold increased risk of breast cancer for high-density women [1], [2]. This increased risk so-exists with a reduction in mammographic sensitivity—the so-called masking effect—besides density being an independent marker of risk.

In recent years personalized screening strategies based on a woman’s risk profile (density profile) have been in focus. However, a great number of methodologies for measuring density or parenchymal pattern on mammograms exist. These methodologies include different visual classifications [3]-[5] as well as several newer semi or fully-automated techniques [6]-[8]. The manual (visual and semi-automated) breast density methods do not require very advanced or expensive technical equipment but are greatly influenced by subjectivity.

We aimed to report inter-observer agreement on three different manual ways of assessing mammographic density and parenchymal pattern, and to examine what impact inter-observer variability has on breast cancer risk prediction in terms of odds ratios.
Methods and materials

This retrospective nested case-control study included 122 cases and 262 age-matched controls based on all 14,736 women with negative screening mammograms from a population-based screening service in Copenhagen in 2007 (follow-up period 2007-2011). Use of screening data and tumour-related information was approved by the Danish Data Inspection Agency (2013-41-1604).

Film-based craniocaudal (CC) and mediolateral oblique (MLO) images (right and left) from the screening in 2007 were digitized. The randomized mammograms were classified independently by two MDs (a senior breast radiologist and a resident in radiology) according to two visual classifications—the BI-RADS density classification (4th edition, 2003; Figure 1) [5] and the Tabár classification on parenchymal patterns (Figure 2) [4], [9]—as well as a computerized interactive threshold technique measuring area-based percent mammographic density (denoted PMD; Figure 3) [10]. Readings were done without knowledge of the woman's age at screening or (later) cancer status. The right and the left breasts were read independently and considered independent measurements. Readings by the three different methodologies were completed at different stages in a MatLab scoring-database, where the readers were blinded from evaluations by the other classifications in order to reduce artificial agreement between the methods.

Statistics:

Inter-observer consistency was investigated on both a multiple-category scale and on a high/low-risk scale based on all independently scored right and left mammograms. High-risk density was defined as: BI-RADS: D3 and D4, Tabár: PIV and PV and the upper two quartiles of PMD (within density range corresponding to the BI-RADS classification). Kappa statistics and the Intraclass Correlation Coefficient (ICC) (which is also equivalent to the weighted kappa) were used to evaluate inter-observer agreement.

The association between mammographic density/parenchymal pattern and breast cancer risk was estimated using logistic regression to calculate odds ratios (OR) adjusted for age. The higher density groups were compared individually with the lowest density group for each method (reference categories: BIRADS: D1; Tabár: PII; PMD: the lowest quartile). In addition, dichotomized high and low-risk groups were compared.

Exact two-sided P-values and 95% confidence intervals (95% CI) have been listed and results were considered statistically significant with P-values # 0.05.
Fig. 1: The BIRADS density classification illustrated on MLO views. A) D1: Fatty (<25% fibro-glandular tissue), B) D2: Scattered fibro-glandular densities (25-50%), C) D3: Heterogeneously dense (51-75%), D) D4: Extremely dense (>75%)

© Department of Radiology, University Hospital Copenhagen, Rigshospitalet, Copenhagen, Denmark 2015

Fig. 2: The Tabár classification on parenchymal patterns illustrated on MLO images. A) PI: Scalloped contours with oval-shaped lucencies and evenly scattered 1-2 mm nodular densities. B) PII: Almost complete fatty replacement. C) PIII: Like PII with a retroareolar prominent duct pattern. D) PIV: Dominated by extensive nodular and linear densities with nodular densities larger than normal lobules. E) PV: Dominated by homogeneous, ground glass like and structure-less densities.

© Department of Radiology, University Hospital Copenhagen, Rigshospitalet, Copenhagen, Denmark 2015
**Fig. 3:** The interactive threshold technique. In a MatLab programme the readers marked the nipple (blue arrow) and outlined the breast boundary (green arrow) and the pectoral muscle (yellow arrow). Secondly, the readers chose the most optimal threshold separating the dense tissue (red color) from the non-dense tissue by sliding the threshold bar (red arrow). Accordingly, the pixels in the image were identified as either dense or non-dense tissue. PMD was computed by dividing the total number of dense pixels by the total number of pixels within the breast area.

© Biomediq, Copenhagen, Denmark 2015
Results

The women were aged between 50 and 69 years (mean age of cases 57.8 (SEM 0.49) and controls 58.1 (SEM 0.34)). Invasive cancer was diagnosed in 110 women and ductal carcinoma in situ (DCIS) in 12 women.

Substantial to almost perfect inter-observer reproducibility was seen for all three methods both on a multiple and a dichotomous scale (Table 1).

Risk estimates are shown in table 2. A comparable stepwise increase in risk with increasing density measured by BI-RADS and PMD was seen for both readers. However, risk estimates for the different Tabár categories varied among the two readers. Both readers demonstrated the highest age-adjusted ORs for Tabár's pattern IV, though, 4.14 (2.26-7.61) and 7.69 (3.49-16.91), respectively. No significant difference in ORs between readers was seen for any of the methodologies on a high/low-risk basis.
### Table 1: Inter-observer agreement according to three methods (BIRADS, Tabár and PMD) based on 765 independent breast measurements in 384 women.

<table>
<thead>
<tr>
<th>Method</th>
<th>Agreement (%)</th>
<th>Kappa (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BIRADS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low/high-risk</td>
<td>88.9</td>
<td>0.74</td>
<td>(0.69-0.79)</td>
</tr>
<tr>
<td>5-categories</td>
<td>74.5</td>
<td>0.64</td>
<td>(0.60-0.68)</td>
</tr>
<tr>
<td><strong>PMD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low/high-risk**</td>
<td>92.7</td>
<td>0.78</td>
<td>(0.72-0.83)</td>
</tr>
</tbody>
</table>
**Table 2:** Association between density/parenchymal pattern (assessed by two readers using three different scoring methods) and breast cancer risk in terms of odds ratios.

© - Copenhagen/DK
Conclusion

This is to our knowledge the first study to report inter-observer agreement on the Tabár classification. Whereas the BI-RADS density classification and the PMD measurements are based on simple quantitative assessment of density, the Tabár classification is far more intuitive. However, we found inter-observer reproducibility on the Tabár classification to be highly comparable with the two other methods. The Inter-observer concordances we demonstrated are also comparable with previous inter-observer studies reporting kappa values ranging from the extremes of 0.02-0.87 [11]-[15] regarding the BIRADS classification, and ICC values of 0.94 on CC views [1] and 0.91 on MLO views [16] regarding the interactive threshold technique.

Despite substantial to almost perfect inter-observer reproducibility for all three methods, different impact on breast cancer risk prediction on a multiple-category scale was observed depending on the density scale used. This indicates that the overall concordance is not as important as the specific type of "misclassification" when estimating risk, as has also previously been discussed by Grove et al [17]. However, no difference in OR risk estimates between readers was seen after categorising into only two risk-groups.

Only a few studies have investigated the association between the Tabár classification and the risk of breast cancer. In line with Jakes and colleagues we also found the correlation of parenchymal pattern and breast cancer risk to be specifically associated with PIV [18].

The study have some limitations to address: In this retrospective study we have not been able to control for other breast cancer risk variables other than age, such as e.g. BMI and reproductive variables. Moreover, we did not differentiate between interval cancers (defined as cancers diagnosed between two screenings) and screen-detected cancers. These limitations are important to bear in mind when interpreting the risk estimates. The lack of BMI adjustment has probably led to some underestimation of risk. On the other hand, we might have included some "excess" cancers which may have been initially undetected (masked at the negative screening in 2007), leading to an overestimation of risk.

Additionally, readings were done on analogue digitized mammograms reducing the quality of the images. Finally, it would have strengthened our study methodologically to have had more readers.

In conclusion, parenchymal pattern as well as density may play a role in a future individualized screening setting; however, automated computerized techniques are needed to fully overcome the impact of subjectivity.
References


