The Breast 29 (2016) 49–54

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Original article

Consistency of breast density categories in serial screening mammograms: A comparison between automated and human assessment

Katharina Holland [a,*], Jan van Zelst [a], Gerard J. den Heeten [b,c], Mechli Imhof-Tas [a], Ritse M. Mann [a], Carla H. van Gils [d], Nico Karssemeijer [a]

[a] Department of Radiology and Nuclear Medicine, Radboud University Medical Center, PO Box 9101, 6500 HB Nijmegen, The Netherlands
[b] LRCB – Dutch Reference Center for Screening, PO Box 6873, 6503 GJ Nijmegen, The Netherlands
[c] Department of Radiology/Biomedical Engineering and Physics, Academic Medical Center Amsterdam, PO Box 22660, 1100 DD Amsterdam, The Netherlands
[d] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form
23 June 2016
Accepted 23 June 2016

Keywords:
Breast
Screening
Mammography
Computer-assisted
Mammographic density

ABSTRACT

Reliable breast density measurement is needed to personalize screening by using density as a risk factor and offering supplemental screening to women with dense breasts. We investigated the categorization of pairs of subsequent screening mammograms into density classes by human readers and by an automated system.

With software (VDG) and by four readers, including three specialized breast radiologists, 1000 mammograms belonging to 500 pairs of subsequent screening exams were categorized into either two or four density classes. We calculated percent agreement and the percentage of women that changed from dense to non-dense and vice versa. Inter-exam agreement (IEA) was calculated with kappa statistics. Results were computed for each reader individually and for the case that each mammogram was classified by one of the four readers by random assignment (group reading).

Higher percent agreement was found with VDG (90.4%, CI 87.9–92.9%) than with readers (86.2–89.2%), while less plausible changes from non-dense to dense occur less often with VDG (2.8%, CI 1.4–4.2%) than with group reading (4.2%, CI 2.4–6.0%). We found an IEA of 0.68–0.77 for the readers using two classes and an IEA of 0.76–0.82 using four classes. IEA is significantly higher with VDG compared to group reading.

The categorization of serial mammograms in density classes is more consistent with automated software than with a mixed group of human readers. When using breast density to personalize screening protocols, assessment with software may be preferred over assessment by radiologists.

Introduction

The association between breast density and breast cancer risk is well established. Several studies show that the risk of developing breast cancer is two to six times higher for women with dense breasts than for women in the lowest density category [1–4]. Though studies suggest that with the introduction of digital mammography differences in sensitivity across density categories disappear [5], sensitivity of mammography is still impaired by density, because dense tissue can mask cancers [6,7]. Therefore, personalized breast cancer screening protocols involving ultrasound and MRI are developed taking into account breast density [8].

The most common breast density reporting method is the Breast Imaging Reporting and Data System (BI-RADS) [9] which uses four categories. Studies have shown a considerable inter- and intra-reader variability when using BI-RADS [10–13].

To overcome these variabilities, semi and fully automatic methods were developed to quantify breast density. A first

* Corresponding author.
E-mail addresses: Katharina.Holland@radboudumc.nl (K. Holland), Jan.vanZelst@radboudumc.nl (J. van Zelst), G.denHeeten@lrcb.nl (G.J. den Heeten), Mechli.Imhof-Tas@radboudumc.nl (M. Imhof-Tas), Ritse.Mann@radboudumc.nl (R.M. Mann), C.vanGils@umcutrecht.nl (C.H. van Gils), Nico.Karssemeijer@radboudumc.nl (N. Karssemeijer).

http://dx.doi.org/10.1016/j.breast.2016.06.020
0960-9776/© 2016 Elsevier Ltd. All rights reserved.

approach was the area based method Cumulus [14]. With Cumulus, the radiologist has to set a threshold to distinguish fibroglandular tissue from fatty tissue. Subsequently, the proportion fibroglandular tissue is calculated with respect to the breast area. BI-RADS and Cumulus are limited by the fact that they are based on the two-dimensional projection of fibroglandular tissue. This projection varies with the projection angle and threshold settings and ignores the three-dimensional anatomical breast structure. To overcome these limitations, quantitative image analysis methods were developed based on imaging physics [15–19]. These methods take the thickness of the compressed breast and imaging parameters into account to measure the absolute (cubic centimetres) and relative (percentage of the breast volume) amount of fibroglandular tissue.

Development of automated breast density assessment methods is an important step towards the introduction of personalized screening protocols adjusted to the need of individual women. This includes supplemental screening to mammography or the replacement of mammography with MRI or ultrasound. To be accepted in practice, it is important to have a consistent, objective and reproducible measurement of breast density to stratify women unambiguously in non-dense and dense categories. With a poor density measurement clinicians and women may lose confidence in the stratification process. Therefore, the temporal aspect of density measurements is very important, as it may be hard to explain why supplemental screening is offered in an irregular pattern. This is acknowledged in a recent review paper, where concerns are raised that radiologists' variability of BI-RADS density assessments over time may lead to inconsistent information in mandated communications about elevated breast cancer risk and supplemental screening [20].

Changes in density classes over time can be caused by changes in hormonal status or a change of BMI. It is known that density usually decreases gradually with lifetime [21], so a change to a lower category is expected for some women. The reproducibility of automated volumetric breast density measurements was studied with repeated exams [22,23]. In these studies no significant differences in density measurements were observed. Furthermore, several studies found good correlations between automated and human density assessment [24–26].

The purpose of this study is to investigate the consistency of density classifications in serial screening mammograms with fully automated volumetric density measurements and to compare these results to classifications of human readers, operating individually or as a group with mammograms distributed randomly over the readers. The latter does better reflect screening practice as serial mammograms are usually not read by the same radiologist.

Methods

Material

Digital Mammograms from the Dutch breast cancer screening program were used which were acquired in a population of 56,000 women between 2003 and 2012. In this program, women aged 50–75 receive a biennial invitation for breast cancer screening. All mammograms were recorded with Lorad Selenia systems (Hologic, Bedford, USA). Consecutive exam pairs were selected in which we call the oldest exam the prior and the more recent exam the current. All mammograms were processed with Volpara (v1.5.0, Volpara Health Technologies, Wellington, New Zealand) to obtain breast density scores. For this purpose we used the ‘for processing’ (raw) data. In total, there were 67,260 pairs and for 64,308 pairs density computation was successful. Missing values were due to breast implants (1.3%) and software failures (3.1%). For this study, we randomly selected 500 women, where for every woman one pair with a prior and current exam was selected at random.

The average screening interval in the 500 pairs was 30 months and is more than 24 months, because sometimes women skip screening. A screening interval of 26 months was measured most frequently, which corresponds to the median screening interval. The mean age was 58.8 ± 6.7 years at the prior screening.

Not all mammograms in our study had four views, because until recently four-view mammography was not standard in the Dutch screening program. Instead, four views were taken in the first screening round and in subsequent rounds only medial lateral oblique (MLO) images were acquired unless there was an indication for additional cranial caudal (CC) images, like high breast density or a possible abnormality judged by the radiographer. Of the 1000 exams used, 415 exams had MLO views only, while 585 exams had MLO and CC views. For 473 exams, only ‘for processing’ images were available. To enable density assessment by the radiologists these exams were converted to ‘for presentation’ format using dedicated software. It was verified that the presentation quality of these images was appropriate for density assessment.

Experimental design

For all images, volumetric percent density was calculated with Volpara by dividing fibroglandular tissue volume by breast volume. Volpara uses quantitative image analysis algorithms based on physical models [16–18]. We averaged all available percent density estimations of an exam. Using the averaged percent density estimate, we categorized all exams using the Volpara density grade (VDG) [27], which is a four point scale matched to the BI-RADS categories. Additionally, we categorized studies with a VDG of one or two as non-dense while we labelled studies with VDG of three or four as dense.

Three radiologists (R1, R2 and R3) with more than eight years of experience in breast imaging and a PhD student (R4) with a medical degree and two years of experience with breast imaging assigned BI-RADS scores (4th edition) individually to each exam. The radiologists were familiar with the density categories, as these are routinely assessed in clinical practice. We categorized studies with a score of one or two as non-dense while studies with a score of three or four were categorized as dense.

Each reader performed the BI-RADS scoring in two reading sessions with at least one week between the sessions. In each of the sessions 500 exams were scored, including either the prior or the current mammogram of a pair. Each of these sessions contained 250 prior and 250 current exams. In screening practice current and prior mammograms are often read by different radiologists. Therefore, we also constructed a ‘group reading’, by assigning the score of a randomly chosen reader to each exam. The group reader is abbreviated with RG.

Statistical methods

The percentage of women categorized in the same class for the prior and current exam and the percentage of women that change from the non-dense to the dense category and vice versa were calculated using the results of the two reading sessions. We used bootstrapping to calculate the 95% confidence interval (CI) of these percentages and to determine if there are significant differences between VDG and results of each of the readers.

The inter-exam agreement was calculated with Cohen's weighted kappa, using either two or four density classes. Moreover, we determined the average kappa value of the readers. To compare the kappa value of the group reader and VDG we used bootstrapping [28]. We also calculated the agreement for a subset of