Visual assessment of breast density using visual analogue scales: observer variability, reader attributes and reading time

DOI:
10.1117/12.2253797

Document Version
Final published version

Published in:
Proc. SPIE 10136, Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment

Visual assessment of breast density using Visual Analogue Scales: observer variability, reader attributes and reading time

Teri Ang[a], Elaine F Harkness[b,c], Anthony J Maxwell[b,c,d], Yit Y Lim[b,c], Richard Emsley[e], Anthony Howell[b,d], D Gareth Evans[b,d], Susan Astley*[b,c,d], Soujanya Gadde[b]

[a]Manchester Medical School, University of Manchester, Oxford Road, Manchester M13 9PT, UK; [b]Nightingale Centre, University Hospital of South Manchester, Manchester M23 9LT, UK; [c]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Oxford Road, Manchester M13 9PT, UK; [d]Manchester Breast Centre, Manchester Cancer Research Centre, University of Manchester, Christie Hospital, Withington, Manchester, M20 4QL UK; [e]Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, University of Manchester, Oxford Road, Manchester M13 9PT, UK

*sue.astley@manchester.ac.uk

ABSTRACT

Breast density is a strong risk factor for breast cancer and has potential use in breast cancer risk prediction, with subjective methods of density assessment providing a strong relationship with the development of breast cancer. This study aims to assess intra- and inter-observer variability in visual density assessment recorded on Visual Analogue Scales (VAS) among trained readers, and examine whether reader age, gender and experience are associated with assessed density. Eleven readers estimated the breast density of 120 mammograms on two occasions 3 years apart using VAS. Intra- and inter-observer agreement was assessed with Intraclass Correlation Coefficient (ICC) and variation between readers visualised on Bland-Altman plots. The mean scores of all mammograms per reader were used to analyse the effect of reader attributes on assessed density. Excellent intra-observer agreement (ICC>0.80) was found in the majority of the readers. All but one reader had a mean difference of <10 percentage points from the first to the second reading. Inter-observer agreement was excellent for consistency (ICC 0.82) and substantial for absolute agreement (ICC 0.69). However, the 95% limits of agreement for pairwise differences were -6.8 to 15.7 at the narrowest and 0.8 to 62.3 at the widest. No significant association was found between assessed density and reader age, experience or gender, or with reading time. Overall, the readers were consistent in their scores, although some large variations were observed. Reader evaluation and targeted training may alleviate this problem.

Keywords: breast density, visual assessment, visual analogue scale, mammogram, observer variability, observer study

1. INTRODUCTION

Breast density is the relative proportion of radio-opaque fibroglandular tissue on a mammogram, and is recognised as a strong risk factor for breast cancer.[1] Increased breast density is also associated with reduced sensitivity in detecting breast cancer lesions in mammograms.[2] The use of breast density measurements may lead to more accurate predictions of breast cancer risk at screening, and hence it can play a role in tailoring breast screening according to individual cancer risk.

The Predicting Risk Of Cancer At Screening (PROCAS) study in Greater Manchester, UK[3] aimed to determine the feasibility of introducing personalised breast cancer risk prediction into the National Health Service (NHS) Breast Screening Programme by collecting data on breast cancer risk factors including breast density measurement to predict individual women's risk. One of the methods used to quantify breast density in the PROCAS study is assessment of percentage density by expert readers recorded on Visual Analogue Scales (VAS).

Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment
edited by Matthew A. Kupinski, Robert M. Nishikawa, Proc. of SPIE Vol. 10136, 1013608
 · doi: 10.1117/12.2253797

Visual assessment of breast density using VAS is an area-based two-dimensional quantitative method which involves the estimation of the percentage breast density on a continuous scale by expert readers. Although it has the advantage of simplicity, it is subjective and gives rise to both intra- and inter-observer variability.[4] Despite the availability of automated volumetric methods of assessing breast density, visual assessments are still useful for longitudinal analyses, or in cases where the raw 'for processing' mammograms are unavailable. However, stability over time and inter-reader variability are important factors in such situations. Visually assessed breast density also adds useful information to breast cancer risk prediction models[5] and shows a stronger relationship with breast cancer risk than automated volumetric methods.[6]

This study aims to assess the extent of intra- and inter-observer variability in assessing breast density using VAS among a group of readers in the Nightingale and Prevent Breast Cancer Centre at the University Hospital of South Manchester, UK. It addresses the issue of reproducibility over time, and also aimed to identify whether reader attributes or the time taken to assess images influenced density assessment.

2. METHODS

2.1 Study design

The PROCAS study recruited women between the ages of 47 to 73 who were attending routine NHS breast screening, and collected data on a number of breast cancer risk factors including breast density at the time of screening. 120 mammograms from PROCAS were used to assess intra- and inter-observer variability in visual assessment of breast density over a three-year interval. Eleven readers scored the percent density of the mammograms on VAS on two occasions. Mammograms were assessed in four batches of 30 and the time taken to assess each batch was recorded.

The VAS scores were used to assess intra- and inter-observer agreement. For the inter-observer component, the most recent scores were used to assess the level of agreement between readers. Mean density scores were compared with age, gender and experience of the readers, and time spent assessing the mammograms.

2.2 Mammogram selection

Mammograms were selected on the basis of the average visually assessed percentage density recorded on VAS during PROCAS, to provide examples distributed evenly between deciles of density. All images had been acquired on GE Senographe Essential mammography systems. Mammograms of subjects who had a breast cancer diagnosis in the past or at time of entry to PROCAS, had breast implants, had undergone a mastectomy or self-reported a previous breast biopsy were excluded from this analysis.

2.3 Breast density assessment with VAS

Four mammographic projections were available: the cranial-caudal (CC) and mediolateral-oblique (MLO) views for both left (L) and right (R) breasts; hence there were four scores per reader per case: LCC, RCC, LMLO and RMLO. Breast density for each woman was recorded on paper score sheets with four 10cm long VAS marked 0% at one end and 100% at the other, one for each available view. The score sheets were scanned using custom software which converted the marks to percentages. The exercise was undertaken in 2012 and repeated in 2015.

2.3 Participating readers

The readers were a group of eight consultant radiologists, two advanced practitioner radiographers and one breast physician, with varying levels of experience in reading mammograms ranging from 4 to 28 years. Readers were blinded to scores undertaken on entry to PROCAS, their own scores in 2012 and scores of other readers. Reader demographics are shown in Table 1.

2.4 Statistical analysis

For the purpose of analysis, the average of the four readings (LCC, RCC, LMLO, RMLO) for each mammogram was used. Intra- and inter-observer agreement for VAS readings was performed by calculating the Intraclass Correlation Coefficient (ICC). ICC is used to estimate the reliability of quantitative and continuous data and lies between 0 and 1. A higher ICC corresponds to higher level of agreement and reliability. The interpretation of the ICCs obtained is based on

the work of Landis and Koch[7] who defined the strength of agreement into five categories: slight 0.00-0.20; fair 0.21-0.40; moderate 0.41-0.60; substantial 0.61-0.80; excellent 0.81-1.00.

Table 1. Reader demographics

Reader	Age range	Gender	Years of experience
A	51-55	Male	22
B	56-60	Male	27
C	51-55	Female	22
D	46-50	Female	7
E	41-45	Female	7
F	56-60	Female	25
G	46-50	Female	9
H	35-40	Female	7
I	61-65	Female	28
J	51-55	Female	4
K	41-45	Male	8

The mean difference between the initial (2012) and subsequent (2015) scores for each reader was computed to examine the general trend in scoring over the two time points. To aid the analysis of inter-observer variability, a scatterplot matrix was used to demonstrate the relationship between the scores of any two readers. Bland-Altman plots[8,9] were constructed to assess the 95% limits of agreement (LOA) of VAS readings between two readers. The maximum absolute difference was also computed; this is the maximum difference between any two readings for the same mammogram.

The mean score of the 120 mammograms for each reader was computed. Readers were grouped into two categories for each of the following attributes: age (below or above 50 years); gender (female or male); and experience (less than or greater than 20 years). The mean scores of the readers for each category were calculated and group means were compared using a one-way ANOVA to investigate possible associations. Linear regression was used to examine the relationship between time taken to read the set of mammograms and mean VAS scores.

3. RESULTS

3.1 Intra-observer variability

The ICCs for the readers ranged from 0.52 to 0.94 (Table 2). Eight of the 11 readers had ICCs above 0.80 which corresponds to an excellent level of agreement between their VAS readings at the two time points. Nine readers had negative mean differences in their scores, indicating that these readers had, on average, a decline in their scores from the initial to subsequent readings (Figure 1). Only one reader (Reader H) had a mean difference above 10 percentage points of 14.8. The maximum absolute difference ranged from 20.1 to 71.0 (Table 2).

3.2 Inter-observer variability

Inter-observer agreement was found to be excellent for consistency (ICC 0.82) and substantial for absolute agreement (ICC 0.69), meaning the VAS scores are very consistent for each reader but to a lesser extent for absolute agreement, which takes into account the systematic differences between readers.

In the Bland-Altman analysis the 95% limits of agreement (LOA) were narrowest for Readers G and K (-15.7 to 6.8 to 15.7) and widest for Readers C and J (0.8 to 62.3) (Figure 2).

Figure 3 shows the scatter plot matrix of all pairwise scores among the 11 readers. The line of equality, where the corresponding scores from the two readers are equal, acts as a reference line for perfect pairwise concordance. The large proportion of points deviating from the lines of equality demonstrate the extent of the variability in VAS scores between readers (Figure 3). This is particularly evident for Reader J who had scores significantly lower than other readers.

Table 2. ICC for each reader and the difference between their mean VAS scores in 2015 and at the initial assessment in 2012

Reader	ICC	Density difference[a] (subsequent-initial)		
		Mean	SD	Max abs[b]
A	0.88	- 4.8	9.2	30.0
B	0.82	5.0	6.5	23.8
C	0.86	- 5.0	11.5	44.0
D	0.85	- 3.8	11.2	40.3
E	0.57	- 9.8	7.7	24.5
F	0.83	6.4	9.8	33.6
G	0.90	- 6.7	5.9	32.0
H	0.52	- 14.8	9.3	37.8
I	0.90	- 1.2	6.7	20.5
J	0.76	- 8.0	12.2	71.0
K	0.94	- 1.5	6.3	25.8

a The units of the differences are given in percentage points.
b Max abs denotes the maximum absolute difference.

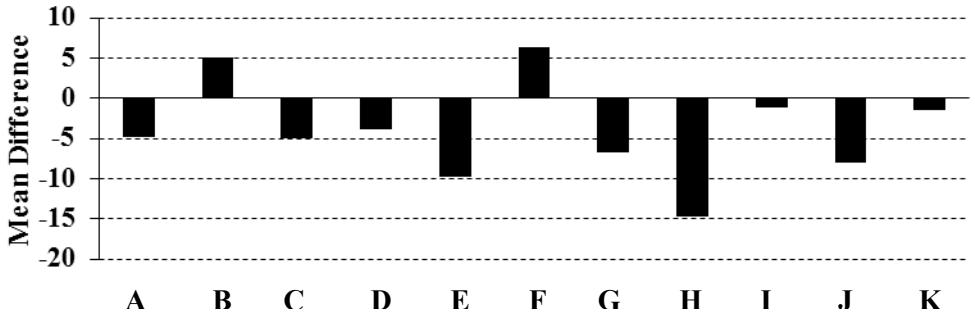

Figure 1. Mean differences density scores (subsequent assessment-initial assessment) by reader

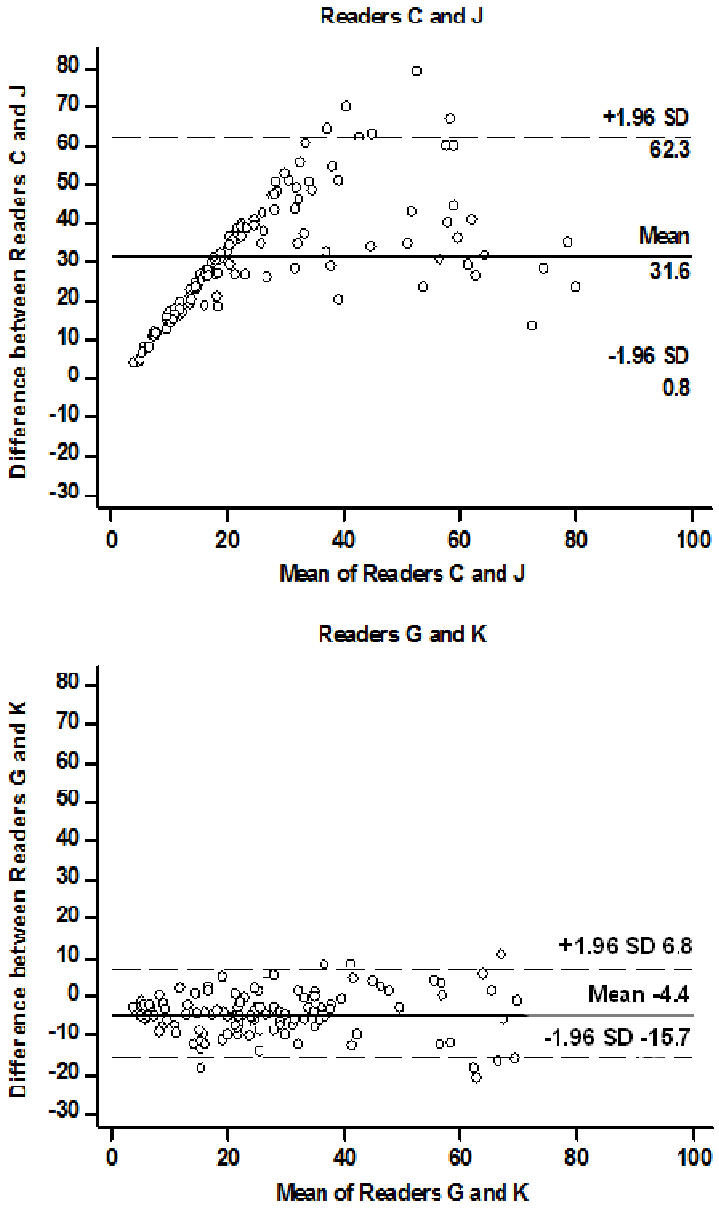

Figure 2. Bland-Altman plots for two pairs of readers. Differences in scores (subsequent – initial) are plotted against the mean scores of the two readers. The solid horizontal lines are the mean differences and the dashed horizontal lines are the 95% limits of agreement (LOA) between which 95% of differences are expected to lie, and which are defined as the mean difference plus and minus 1.96 times the standard deviation of the mean difference.

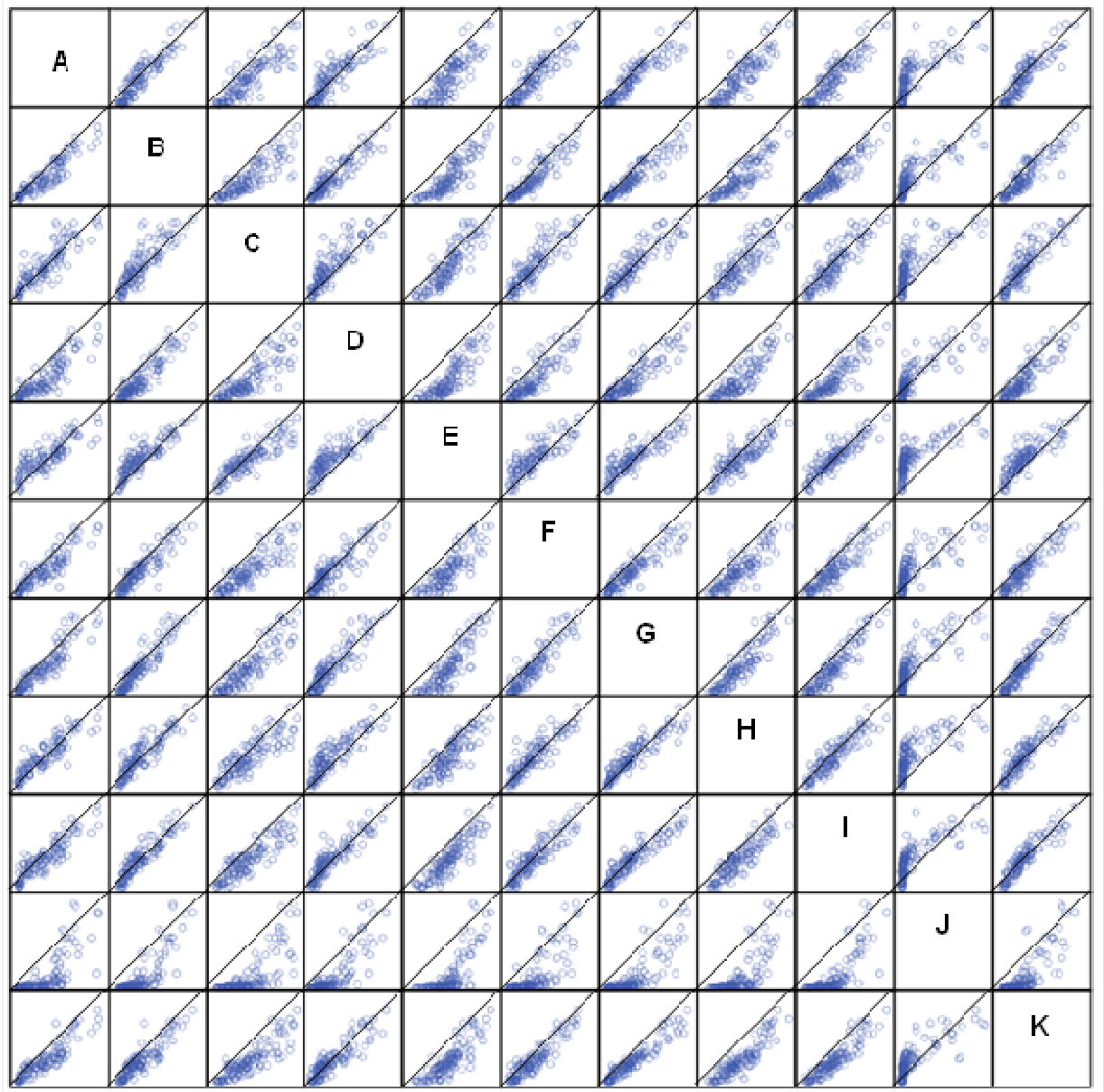

Figure 3. VAS scores for all pairs of readers (A to K). All axes range from 0 – 100%. The line of equality in each plot is shown.

3.2 Reader attributes

3.2.1 Age, gender and experience

Table 3 shows the mean density scores in each category and Figure 4 displays the mean scores and 95% confidence intervals (CI). No association was found between age, gender or experience and assessed density.

Table 3. Mean VAS density score by reader age, gender and experience.

Reader attribute		N	Mean score	SD
Age	<50	6	29.1	3.1
	>50	5	28.9	10.4
Gender	Female	8	28.7	8.8
	Male	3	29.7	4.0
Experience	<20 years	5	26.2	7.8
	>20 years	6	32.4	6.5

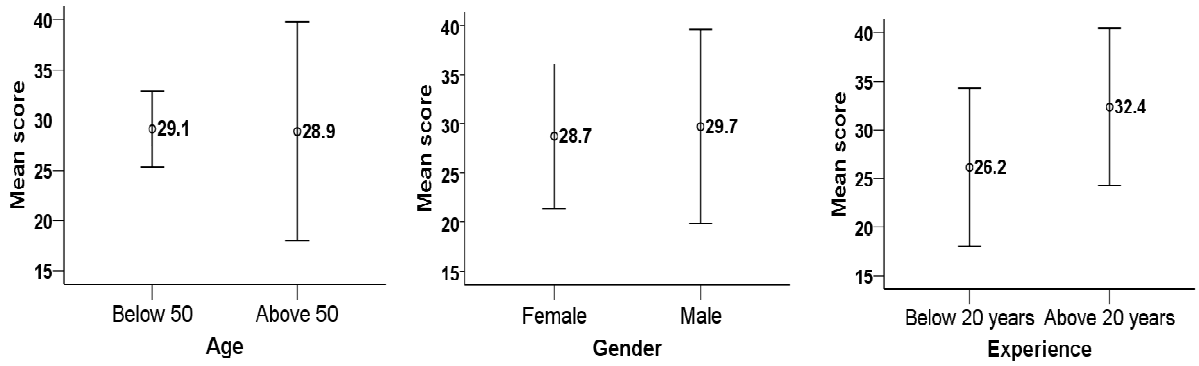

Figure 4. Mean VAS by reader factor with 95% CIs.

3.2.2 Time taken to read the mammograms

The total time taken to read the set of mammograms had no significant association with mean VAS (P = 0.08) (Figure 5).

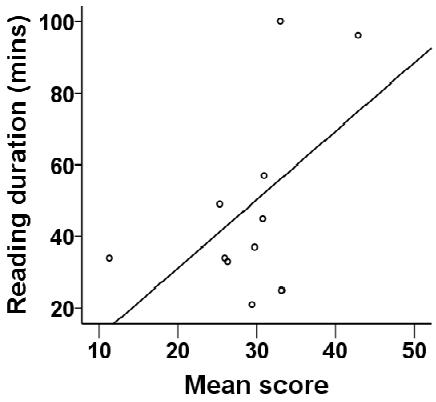

Figure 5. Total time taken to read the mammograms in minutes against mean scores of readers. The linear regression line is included.

4. DISCUSSION AND CONCLUSIONS

We studied intra- and inter-observer agreement in assessment of mammographic density using VAS among 11 readers at the Nightingale Centre, University Hospital of South Manchester. The readers assessed the same set of 120 mammograms on two occasions, three years apart. A strength of this study was that mammograms were selected to

represent the full range of densities encountered in a screening population; whilst the readers were informed that this was the case, they were used to assessing a screening mix clinically. This could have resulted in underestimating density in some cases; it is apparent in figure 3 that higher densities are under-represented. The readers in this study did not have specific training in density assessment, since at the time of the PROCAS study it was not known in detail how visual assessment recorded in this way related to risk. With a large body of data now available, training could be implemented to share best practice and improve risk prediction across the population of readers.

Overall, intra-observer agreement of VAS scoring was excellent for the majority of the readers. Whilst inter-observer agreement was substantial, there were significant variations in scores between readers for some mammograms. Bland-Altman analysis illustrated wide limits of agreement between any two readers, with the widest at (0.8 to 62.3) and narrowest at (-6.8 to 15.7). These results suggest that the use of VAS scores by a pool of readers might be problematic when assessing density longitudinally, for example in assessing whether tamoxifen has resulted in a reduction in mammographic density and cancer risk. In the IBIS study[10] density reductions of 10% or more were associated with a reduction in risk. If VAS is to be used for a similar purpose, based on our analysis we would recommend that the same reader assesses mammograms at both time points. For breast cancer risk stratification, a potential approach is to correct scores to adjust for inter-rater bias.[11] A recent publication from the PROCAS study has demonstrated that VAS risk estimates were attenuated but still statistically significantly associated with increased breast cancer risk after adjusting for observer variability[12].

Reader attributes were also examined in relation to density assessment, however there was no significant association between age, gender, experience of reading mammograms or time taken to read the batches of images in this study. However, a limitation of this study is the small number (11) of participating readers. There were too few readers in each group to enable detailed examination of any association between reader attributes and density assessment, in particular there were only three male readers available, and whilst they all had several years' experience reading mammograms, their experience of density assessment varied widely with the number of mammograms read in PROCAS ranging from 1653 to 16443 (median = 7763).

5. ACKNOWLEDGEMENTS

We acknowledge the support of the National Institute for Health Research (NIHR) and the Prevent Breast Cancer Appeal. We would like to thank the study radiologists and advanced practitioner radiographers for VAS reading. We would also like to thank the many radiographers in the screening programme, and the study centre staff for recruitment and data collection. This article represents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (RP-PG-0707-10031): "Improvement in risk prediction, early detection and prevention of breast cancer") with additional funding from the Prevent Breast Cancer Appeal. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

REFERENCES

[1] Machida, Y., Tozaki, M., Shimauchi, A., and Yoshida, T., "Breast density: the trend in breast cancer screening," Breast Cancer 22(3), 253-256 (2015)

[2] Boyd, N.F., Guo, H., Martin, L.J., Sun, L., Stone, J., Fishell, E., et al. "Mammographic density and the risk and detection of breast cancer," N Eng J Med. 356(3),227-236 (2007)

[3] Evans, D.G.R., Warwick, J., Astley, S.M., Stavrinos, P., Sahin, S., Ingham, S., McBurney, H., Eckersley, B., Harvie, M., Wilson, M., Beetles, U., Warren, R., Hufton, A., Sergeant, J.C., Newman, W.G., Buchan, I., Cuzick, J., and Howell, A., "Assessing individual breast cancer risk within the UK National Health Service Breast Screening Program: a new paradigm for cancer prevention," Cancer Prevention Research 5(7), 943-951 (2012)

[4] Sergeant, J.C., Walshaw, L., Wilson, M., Seed, S., Barr, N., Beetles, U., Boggis, C., Bundred, S., Gadde, S., Lim, Y.Y., Whiteside, S., Evans, D.G.,Howell, A., and Astley, S.M., "Same task, same observers, different values: the problem with visual assessment of breast density," SPIE 8673, Medical Imaging (2013)

[5] Brentnall, A.R., Harkness, E.F., Astley, S.M., Donnelly, L.S., Stavrinos, P., Sampson, S., Fox, L., Sergeant, J.C., Harvie, M., Wilson, M., Beetles, U., Gadde, S., Lim, Y.Y., Jain, A., Bundred, S., Barr, N., Reece, V., Howell, A., Cuzick J., and Evans, D.G., "Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort," Breast Cancer Research 17(1),147 (2015)

[6] Evans, D.G.R., Warwick, J., Astley, S.M., Stavrinos, P., Sahin, S., Ingham, S., McBurney, H., Eckersley, B., Harvie, M., Wilson, M., Beetles, U., Warren, R., Hufton, A., Sergeant, J.C., Newman, W.G., Buchan, I., Cuzick, J., and Howell, A., "Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study," Programme Grants for Applied Research, 4(11), (2016)

[7] Landis, J.R., and Koch, G,G., "The Measurement of Observer Agreement for Categorical Data," Biometrics 33(1), 159-1741(1977)

[8] Bland, J.M. and Altman, D.G., "Statistical methods for assessing agreement between two methods of clinical measurement," The Lancet 327(8476), 307-310 (1986)

[9] Bland, J.M. and Altman, D.G., "Measuring agreement in method comparison studies," Stat. Methods Med. Res. 8(2), 135-160 (1999).

[10] Cuzick, J., Warwick, J., Pinney, E., Duffy, S.W., Cawthorn, S., Howell, A., Forbes, J.F., and Warren, R.M., "Tamoxifen-induced reduction in mammographic density and breast cancer risk reduction: a nested case–control study," Journal of the National Cancer Institute 103(9), 744-752 (2011)

[11] Sperrin, M., Bardwell, L., Sergeant, J.C., Astley, S., Buchan, I., "Correcting for rater bias in scores on a continuous scale, with application to breast density" Stat Med. 32(26), 4666-4678 (2013)

[12] Harkness, E.F., Sergeant, J., Wilson, M., Beetles, U., Gadde, S., Lim, Y.Y., Howell, A., Evans, D.G., and Astley, S.M., "Should We Adjust Visually Assessed Mammographic Density for Observer Variability?" In: Breast Imaging. Lecture Notes in Computer Science 9699, 540-547 (2016)