How quickly do breast screeners learn their skills?

Hossein Nevisi, Leng Dong, Yan Chen*, Alastair G. Gale
Applied Vision Research Centre, Loughborough University, UK

ABSTRACT

The UK's Breast Screening Programme is 27 years old and many experienced breast radiologists are now retiring, coupled with an influx of new screening personnel. It is important to the ongoing Programme that new mammography readers are quickly up to the skill level of experienced readers. This raises the question of how quickly the necessary cancer detection skills are learnt. All breast screening radiologists in the UK read educational training sets of challenging FFDM images (the PERFORMS® scheme) yearly to maintain and improve their performance in real life screening. Data were examined from the PERFORMS® annual scheme for 54 new screeners, 55 screeners who have been screening for one year and also for more experienced screeners (597 screeners). Not surprisingly, significant differences in cancer detection rate were found between new readers and both of the other groups. Additionally, the performance of 48 new readers who have now been screening for about a year and have taken part twice in the PERFORMS® scheme were further examined where again a significant difference in cancer detection was found. These data imply that cancer detection skills are learnt quickly in the first year of screening. Information was also examined concerning the volume of cases participants read and other factors.

Keywords: Breast screening, Screeners, radiologists, PERFORMS®, training, FFDM images

1. INTRODUCTION

The UK's Breast Screening Programme screens all women aged 50-70 every three years. In England alone, the latest data show that each year more than two million women undergo breast cancer screening[1]. Nationally the number of radiologists and advanced practitioners (technologists) who undertake screening is over 700. The Royal College of Radiologists have recommended that a radiologist must read at least 5,000 cases a year in order to participate in the national screening programme[2]. Despite this large number of cases, an individual radiologist can possibly only expect on average to see a malignant case about once or twice in a working week when they are operating as a first screen reader (the UK employs a double reading policy in screening).

Therefore, it is extremely important for radiologists to work to the best of their abilities and to maintain vigilance in reading mammograms. To this aim, all breast screening radiologists in the UK read self-assessment sets of challenging FFDM digital breast screening cases (the PERFORMS® scheme) yearly to maintain and improve their performance in real life screening[3]. New case sets are regularly carefully constructed. Examining carefully selected sets of difficult screening cases, coupled with immediate confidential feedback on how mammography readers identified key mammographic signs of abnormality, and how their screening decisions agreed with an expert panel of radiologists as well as large numbers of their radiological colleagues, is key to understanding their skills and improving their cancer detection - especially for new screening personnel[4].

The UK Screening Programme is 27 years old and has a deserved international reputation for its high quality. When screening was first introduced many radiologists went into this domain and have now become experts, several are also considered international experts. However, many of these radiologists are now reaching retirement age and new screening personnel are coming into the Programme. The question is - will the overall quality of the Programme diminish with the loss of this expert body of personnel? To examine this issue, data from the PERFORMS® scheme were examined for different groups of participants with different levels of experience of breast screening.

*y.chen@lboro.ac.uk

Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment
edited by Matthew A. Kupinski, Robert M. Nishikawa, Proc. of SPIE Vol. 10136, 101360D

2. METHODS

In order to see how new breast screening readers perform in comparison with more experienced readers, data from 706 individuals who have recently participated in the scheme were examined in three groups: Group A consisted of 54 individuals who participated in the PERFORMS® scheme for the first time; Group B consisted of 55 individuals who had participated in the PERFORMS® scheme for the second time, and Group C consisted of 597 individuals who had completed the scheme more than twice. The results of these three groups were analysed and compared on different factors. These included: the mean values of CD (cancer detection) as identified based on the PERFORMS® case pathology reports; the mean values of CR (correct recall), and CS (correct return to screen) decisions as judged against the radiological decisions of a panel of expert UK breast screening radiologists and ROC measures of Area under the Curve (AUC). The number and type of false negative responses and the ability to correctly identify abnormality sites were also examined. Additionally, the volume of cases participants read in real-life was also studied.

Furthermore, we analysed the performance of 48 members of Group A who had subsequently participated in the latest round of the PERFORMS® scheme, and compared that with their data from the previous scheme which they had completed previously.

3. RESULTS

3.1 Performance data analyses

Data for the three groups were analysed in terms of; mean values of cancer detection (CD), correct recall (CR), correct return to screen (CS), negative predictive value (NPV), and positive predictive value (PPV) (see Figure 1). Data were similarly examined in terms of area under the ROC curve (AUC) (see Figure 2). Not surprisingly, the mean values for cancer detection rate and correct recall rate improved as individuals' experience in screening grew, as did NPV values, although CS fell and PPV remained at the same value.

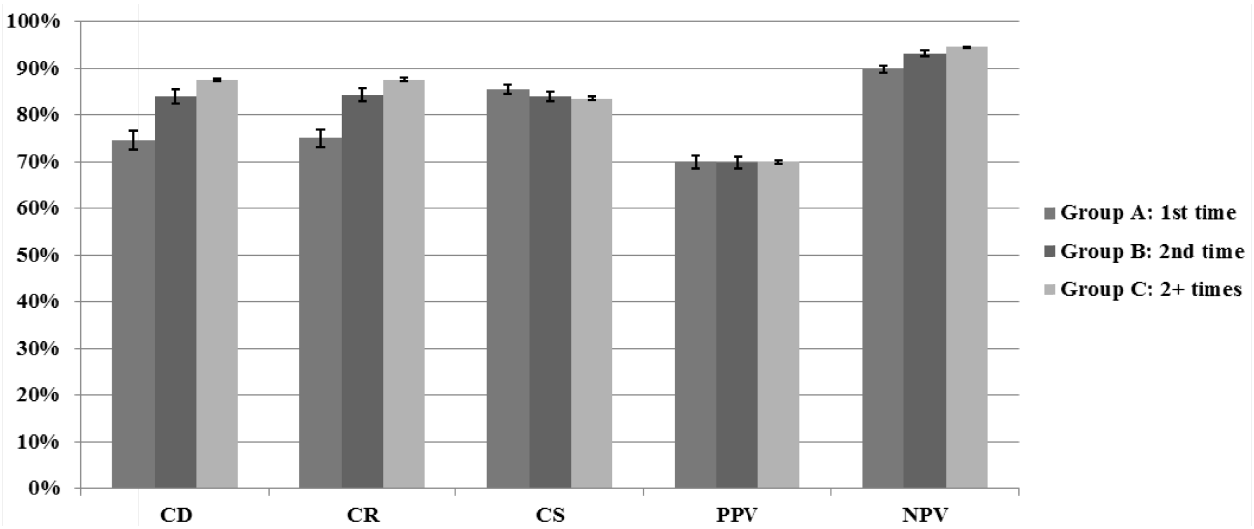

Figure 1. Comparison of the mean values of CD, CR, CS, PPV and NPV in groups A, B and C.

For CD, CR, AUC and NPV, there were statistically significant differences between groups as determined by one-way ANOVA ((CD: $F(2,703) = 44.819$, $p<0.01$; CR: $F(2,703) = 45.334$, $p <0.01$; AUC: $F(2,703) = 36.634$, $p <0.01$; NPV: $F(2,703) = 42.217$, $p <0.01$). Tukey post hoc tests revealed that CD, CR, AUC and NPV were significantly higher for second-time participants: Group B (CD: M= 83.92%, $p<0.01$; CR: M= 84.26 %, $p<0.01$; AUC:M=0.8844, $p<0.01$; NPV: M= 93.16%, $p<0.01$, CS: M=83.94%, $p>0.05$; PPV=69.76%, $p>0.05$) and '2+ time' participants: Group C (CD: M= 87.48%, $p<0.01$; CR: M= 87.55%, $p<0.01$; AUC: M=0.9026, $p<0.01$; NPV: M= 94.47%, $p <0.01$; CS: M=83.50, p= n.s., PPV: M=69.90%, p=n.s.) compared to first-time participants: Group A (CD: M= 74.65%; CR: M= 74.97%; NPV:

M=89.78%; AUC: M=0.8554; CS: M=85.45%; PPV: M=69.93%). Also it was found that CD, CR, AUC and NPV were significantly higher for '2+ time' participants (Group C) compared to second-time participants (Group B) with CD: $p<0.05$; CR: $p<0.05$; AUC: $p<0.01$; NPV: $p<0.05$; CS: p=n.s.; PPV: p=n.s.

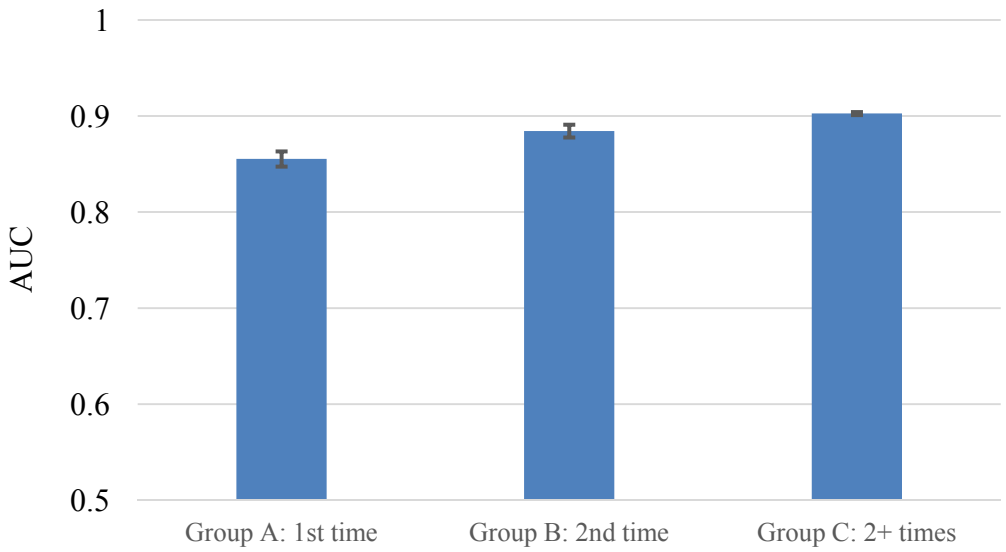

Figure 2. Comparison of mean values of AUC between group A, B and C

3.2 False negative responses

To understand what may underlie the key performance differences between the groups the number of false negative responses were investigated, based on the various abnormalities' feature types (Table 1). These instances involve cases where participants have either identified a feature but then not interpreted that case as a recall case (misinterpreted the feature information) or they have not identified any feature and have returned that case to normal screen (undetected the feature). In general, the ability to identify types of abnormal features correctly increased as the number of participations in PERFORMS® schemes increased. This did not hold true for well-defined masses.

Table 1. False negative responses by feature types.

	Mean % False Negative Responses		
Feature Type	**Group A**	**Group B**	**Group C**
Architectural Distortion	16.14	11.17	8.21
Calcification	16.14	10.13	6.03
Ill Defined Mass	28.52	18.18	14.47
Spiculate Mass	32.04	19.27	16.06
Well Defined Mass	44.44	27.27	30.82

For Architectural Distortion, Calcification, Ill Defined Mass and Spiculate Mass, there were statistically significant differences between groups as determined by one-way ANOVA (Architectural Distortion: $F(2,703) = 9.438$, $p<0.01$; Calcification: $F(2,703) = 17.755$, $p<0.01$; Ill Defined Mass: $F(2,703) = 25.711$, $p<0.01$; Spiculate Mass: $F(2,703)=35.894$, $p<0.01$).

For Well Defined Mass, there was no significant difference between groups ($F(2,703)=2.389$, $p=0.092$). Tukey post hoc tests revealed that the mean false negative rates for Architectural Distortion, Calcification, Ill Defined Mass and Spiculate Mass were significantly higher for first time participants: Group A (Architectural Distortion: M= 16.14%, $p<0.01$; Calcification: M= 16.14%, $p<0.01$; Ill Defined Mass: M=28.52%, $p<0.01$; Spiculate Mass: M=32.04%, $p<0.01$; Well Defined Mass: M=44.44%, $p=0.098$) compared to 2+ time participants: Group C (Architectural Distortion: M= 8.21%; Calcification: M= 6.03%; Ill Defined Mass: M=14.47%; Spiculate Mass: M=16.06%; Well Defined Mass: M=30.82%).

The mean false negative rates for Calcification, Ill Defined Mass and Spiculate Mass were significantly lower for 2nd time participants: Group B (Calcification: M=10.13%, $p=0.034$; Ill Defined Mass: M=18.18%, $p<0.01$; Spiculate Mass: M=19.27%, $p<0.01$; Architechtural Distortion: M=11.17%, $p=0.129$; Well Defined Mass: M=27.27%, $p=0.131$) compared to 1st time participants: Group A. No significant difference between Group B and Group C was found for all feature types ($p>0.05$).

3.3 Examples of difficult cases

To further illustrate differences between the three groups, data from five very difficult malignant cases (based on Group A responses) and which had been reported as false negative cases by the Group A are shown in Table 2. This further indicates that, in general, the more experienced individuals performed better on the difficult malignant cases, making fewer false negative responses.

Table 2. False negative responses for these five difficult cases.

	Mean % False Negative Responses		
Malignant Case	**Group A**	**Group B**	**Group C**
1	51.85	29.09	15.24
2	50	34.55	29.15
3	48.15	20	20.10
4	44.44	27.27	30.82
5	40.74	36.36	27.30

3.4 Detecting suspicious areas

Data of the three groups were also analysed in terms of correctly identifying the areas of interest (AOIs), which had been pre-determined by the expert panel of radiologists, to ascertain their abilities in detecting these suspicious areas. Overall, the mean value of detecting at least one AOI on an abnormal breast for Group A, Group B and Group C were approximately 77%, 84% and 86% respectively.

For example, for the case shown in Figure 3, there are three abnormality areas on the two mammograms. The abnormalities are defined in these images by the AOIs constructed around them by the expert panel and the participants' responses are shown by the individual points. Some 27 individuals (50%) from Group A could not detect any of the AOIs (one is an Architectural Distortion and the other is a Spiculate Mass and these are both highly suspicious of breast cancer); however, this rate for Group B and Group C are 29% and 16% respectively.

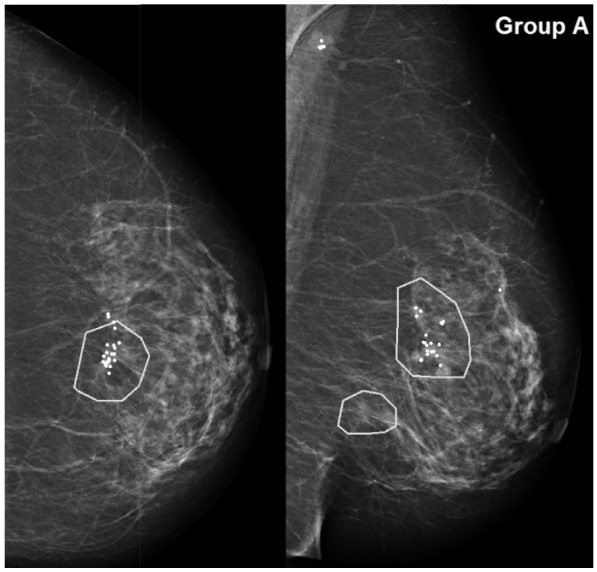

Figure 3a. Responses for Group A

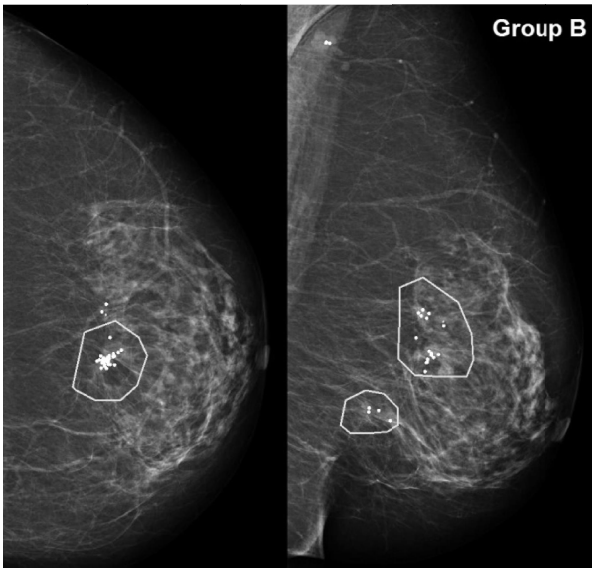

Figure 3b. Responses for Group B

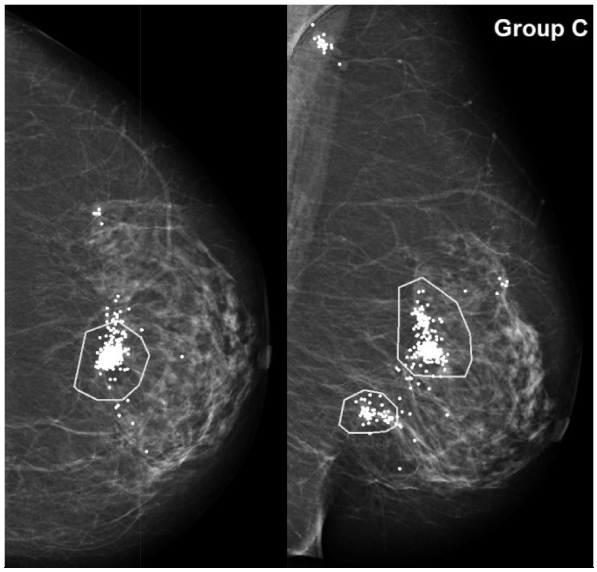

Figure 3c. Responses for Group C

Figure 3. The red dots show the locations that the individuals determined as an abnormal feature. The yellow polygons show the areas of interest around the respective abnormalities.

3.5 New users in their second participation

Forty-eight members of Group A, since the initial analyses, have also taken part in the next round of the scheme. To make it easier to reference, here the first scheme is termed Scheme 1, and the second more recent scheme, Scheme 2. The two schemes use different difficult cases. We compared the results of these 48 participants in these two schemes (Figure 4).

In terms of cancer detection rate the mean value for Scheme 1 was 74.93% which increased to 81.56% in Scheme 2, this being a significant increase (two-tailed t-test, $t(47)=-3.40$, $p<0.01$). Similarly, this improvement in cancer detection skills can also be shown by the sensitivity measure (correct recall) ($t(47)=-4.64$, $p<0.01$). There are no significant differences found in correct return to screening (CS), PPV and NPV.

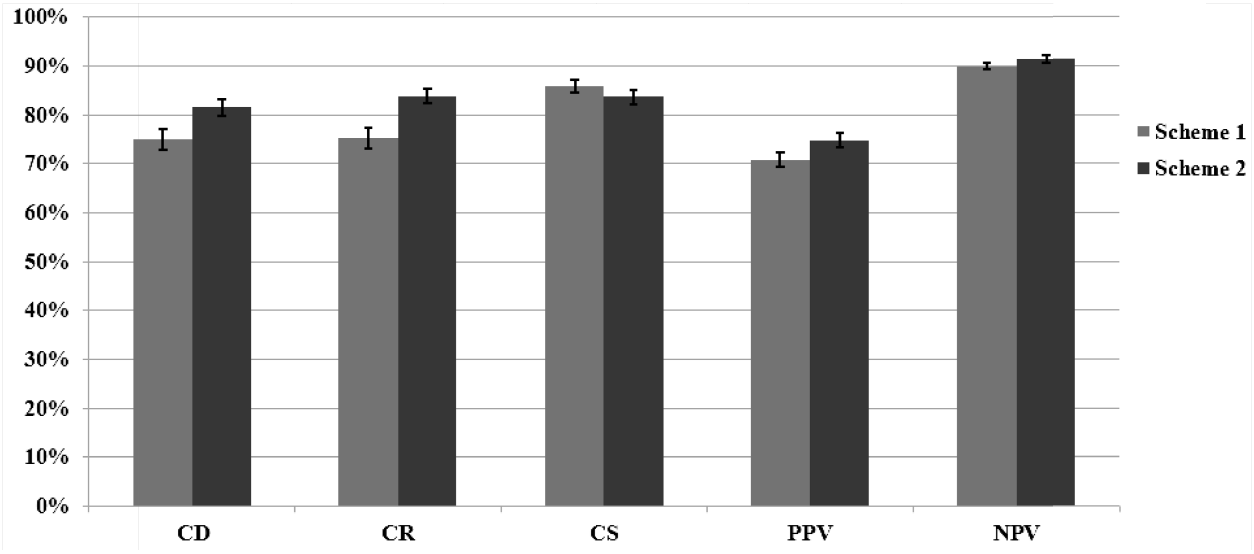

Figure 4. Comparison of Group A results in Scheme 1 with their results in Scheme 2.

3.6 Volume of cases participants read in real-life

When participants take part in PERFORMS® they also report their real-life reading volume of breast screening cases during the previous year. Data were examined for 343 individuals and their real-life reported screening case volume and their PERFORMS® AUC measures. A Pearson product-moment correlation coefficient demonstrated that there was a positive correlation between the two variables ($r = 0.154$, $n = 342$, $p < 0.01$) - increases in the number of cases read per year were correlated with increases in AUC scores from the PERFORMS® scheme.

4. DISCUSSION

The data show that on several performance measures, new breast screening readers performed significantly worse than more experienced readers. This may partly be due to them being new to taking part in the PERFORMS® scheme but that cannot fully explain the very significant differences found. The purpose for the PERFORMS® scheme is to highlight such performance differences and to then help new, and poor performers, improve and maintain their performance so that the quality of the National Screening Programmed remains high despite any workforce changes.

Based on our analysis on the performance of PERFORMS® new readers after a year, it is evident that the skills of newcomers increases rapidly during the first year. We have elsewhere reported[5] on the importance of maintaining a high reading volume of cases as a factor in developing breast cancer screening skills and these data again contribute to that argument.

5. CONCLUSIONS

Breast cancer identification skills appear to be largely quickly learnt in the UK Screening Programme within the first year of participation which is reassuring for the Programme with many personnel changes taking place.

6. ACKNOWLEDGEMENTS

We acknowledge the support of Public Health England.

REFERENCES

[1] http://content.digital.nhs.uk/catalogue/PUB20018/bres-scre-prog-eng-2014-15-rep.pdf .

[2] Royal College of Radiologists: Quality Assurance Guidelines for Radiologists, (1990)

[3] Gale A.G., "PERFORMS – a self-assessment scheme for radiologists in breast screening," Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills, 6(3), 148-152, (2003)

[4] Gale A.G., "Maintaining quality in the UK breast screening program", In D.J. Manning & C. Abbey (Eds.) Proc. SPIE Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment. 7627, 1-11 (2010).

[5] Scott H.J., Gale A.G., & Wooding D.S., "European Breast Screening Performance: does case volume matter?" In: Image Perception, Observer Performance, and Technology Assessment, D.P. Chakraborty & M.P. Eckstein (eds.) Proceedings of SPIE Vol. 5372, (2004).