A potential method to identify poor breast screening performance

Article *in* Proceedings of SPIE - The International Society for Optical Engineering · February 2012

DOI: 10.1117/12.913610

4 authors, including:

Leng Dong
Loughborough University
12 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Alastair Gale
Loughborough University
193 PUBLICATIONS **1,620** CITATIONS
SEE PROFILE

Dev P Chakraborty
ExpertCAD Analytics, LLC
172 PUBLICATIONS **4,360** CITATIONS
SEE PROFILE

A potential method to identify poor breast screening performance

Leng Dong*[a], Yan Chen[a], Alastair G. Gale[a], Dev P. Chakraborty[b]
[a]Applied Vision Research Centre, Loughborough University, Loughborough, UK;
[b]University of Pittsburgh, Pittsburgh, USA

ABSTRACT

In the UK all breast screeners undertake the PERFORMS scheme where they annually read case sets of challenging cases. From the subsequent data it is possible to identify any individual who is performing significantly lower than their peers. This can then facilitate them being offered further targeted training to improve performance. However, currently this under-performance can only be calculated once all screeners have taken part, which means the feedback can potentially take several months. To determine whether such performance outliers could usefully be identified approximately much earlier the data from the last round of the scheme were re-analysed. From the information of 283 participants, 1,000 groups of them were selected randomly for fixed group sizes varying from four to 50 individuals. After applying bootstrapping on 1,000 groups, a distribution of low performance threshold values was constructed. Then the accuracy of estimation was determined by calculating the median value and standard error of this distribution as compared with the known actual results. Data indicate that increasing sample sizes improved the estimation of the median and decreased the standard error. Using information from as few as 25 individuals allowed an approximation of the known outlier cut off value and this improved with larger sample sizes. This approach is now implemented in the PERFORMS scheme to enable individuals who have difficulties, as compared to their peers, to be identified very early after taking part which can then help them to improve their performance.

Keywords: performance, bootstrapping, ROC analysis, breast screening

1. INTRODUCTION

All UK breast screeners undertake the PERFORMS self-assessment scheme where they examine sets of challenging recent screening cases and receive immediate feedback on how they have performed[1,2]. Once all individuals have participated, which currently takes circa five months, then their anonymous data are calculated; both on how well they have fared in terms of correctly identifying actual early signs of malignancy and also on how each person has performed as compared to their peers. The scheme facilitates teasing out, not only a person's cancer identification abilities, but also individuals' agreements on how particular cases should be clinically treated[3,4]. Namely, if this woman presented anywhere in the UK would she always be recalled or judged to be normal/benign and so returned to the next screening round?

One aspect of this scheme is that because all screeners read the same set of test cases under broadly similar reporting conditions then this allows those individuals who have performed much worse on the scheme than their peers to be identified easily. The underlying reasons for any such poor performance can be identified by examination of a person's raw data which facilitates examination of factors such as; the time of day when the scheme was undertaken[5], how long it took to complete the test set, how long it took to read each case, how many rest breaks were taken etc.. If someone is deemed to have under-performed sufficiently then an agreed process with the Royal College of Radiologists in the UK allows suitable follow up actions to be deployed should these be deemed appropriate.

Whilst the scheme uses recent difficult screening cases, taking part in PERFORMS is distinct from typical screening. Of necessity the various PERFORMS case sets are loaded with interesting and challenging examples of difficult normal, benign and malignant appearances. Additionally in reporting the cases the participants are asked to identify a range of features and their locations, rate breast density, and rate each breast in terms of malignancy and other factors. Thus reading a PERFORMS case set, whilst essentially equivalent to reading the same cases in a screening environment, requires different behaviour from the participants as they have to make many more decisions on every case. In the UK, to read a set of 60 screening cases may take a typical radiologist about an hour, to read the same number of PERFORMS

*l.dong@lboro.ac.uk

Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment,
edited by Craig K. Abbey, Claudia R. Mello-Thoms, Proc. of SPIE Vol. 8318, 831819

cases will take circa two hours or even more. Therefore we are always at pains to draw a distinction between performance as measured on the PERFORMS scheme and typical routine screening performance. That said; it must not be forgotten that these cases have all been originally seen in routine clinical screening.

Thus, whilst differences are acknowledged between the scheme and screening there are many similarities between the two. For instance, it is most unlikely that an individual performs poorly on the scheme and yet performs very well in real life screening. Various studies have examined scheme and real screening performances and attested to similarities between how people perform on the scheme and in real life[6,7].

Consequently, finding that someone is under-performing on the scheme cannot simply be regarded as an interesting experimental finding. Whilst it may not fully reflect their real life behavior it can be taken at least as a potential indicator that something may be awry and may require following up. Poor performing individuals (statistical outliers) can then be offered further training, if necessary, which can be specifically targeted for them[8,9,10,11,12]. Available data from such individuals on subsequent rounds of the PERFORMS scheme show that they do not remain as poor performers but instead improve.

This process of following up outliers works very well overall, although such poor performers can logically only be identified once all, or nearly all, other UK screeners have participated and read the same case set. Of necessity this process then causes a delay in providing useful feedback to these individuals. If it was possible to provide a more rapid feedback informing them that they may not have performed on the scheme as well as their colleagues then this is thought to be more practically useful. Therefore, a way of potentially identifying such poor performers much earlier than is currently possible was investigated so that these individuals would receive feedback quicker and so be encouraged to undertake further training earlier if necessary.

2. METHOD

The PERFORMS case sets originate as carefully selected examples of challenging cases from breast screening centres across the UK. These are all recent Full Field digital Mammographic (FFDM) images from different vendors which are then prepared for examination in the scheme both as mammographic film (by processing and printing out digital laser films) and as FFDM images suitable for viewing on any vendors' mammographic workstations (again, by suitable processing).

In the last national round of the PERFORMS scheme some 404 screeners read the case set as mammographic film and 283 read the set as FFDM soft copy images. As our interest primarily lies in digital mammography then the data of those 283 who had read the case set on their workstations were used here. For this group of participants various performance measures had been compiled on various measures including correct recall, correct return to screen decisions, positive predictive value (PPV), negative predictive value (NPV), cancer detection rate and Receiver Operator Characteristic (ROC) measures of performance as judged against other participants as well as as judged against known pathology. For all these, the mean, inter-quartile range, and upper and lower bounds were known. In addition, for the ROC measures the inner and outer statistical fences had already been compiled. The key interest was in these fence values which determine the cut off limits for ascribing either mild (inner fence) or severe (outer fence) under-performance. For these individuals the inner fence value was 0.916, below which was judged severe outlier performance, and 0.947, below which was judged mild outlier performance. Note that both these values are very high. Typically we would find fence values of circa 0.7 and 0.8 respectively. These high values here solely represent extremely good performance for these participants in reading this particular case set.

Our task was to then determine whether we could arrive at these same fence values. To do this the data of randomly selected small groups of these participants were repeatedly bootstrapped with the aim of artificially determining equivalent thresholds of such mild and severe under-performance. Consequently, varying numbers of participants, from four to 50, were used in each group. For each group size then 1,000 randomly selected samples were constructed. After bootstrapping each group, a distribution of 1,000 thresholds of low performance was constructed and the mean values and standard errors of this distribution calculated to determine how the number of participants affected the mean estimation accuracy.

3. RESULTS

As would be expected, the standard error of the estimated inner and outer fence thresholds reduced as group size increased, indicating better estimation accuracy. Figure 1 plots the standard error values against the increasing number of participants in the groups. Note that the standard error of the outer fence is always higher than that of the inner fence, which means that the estimation of the outer fence does not perform as well as the estimation of the inner fence. The standard errors begin to plateau as group size increases beyond 25.

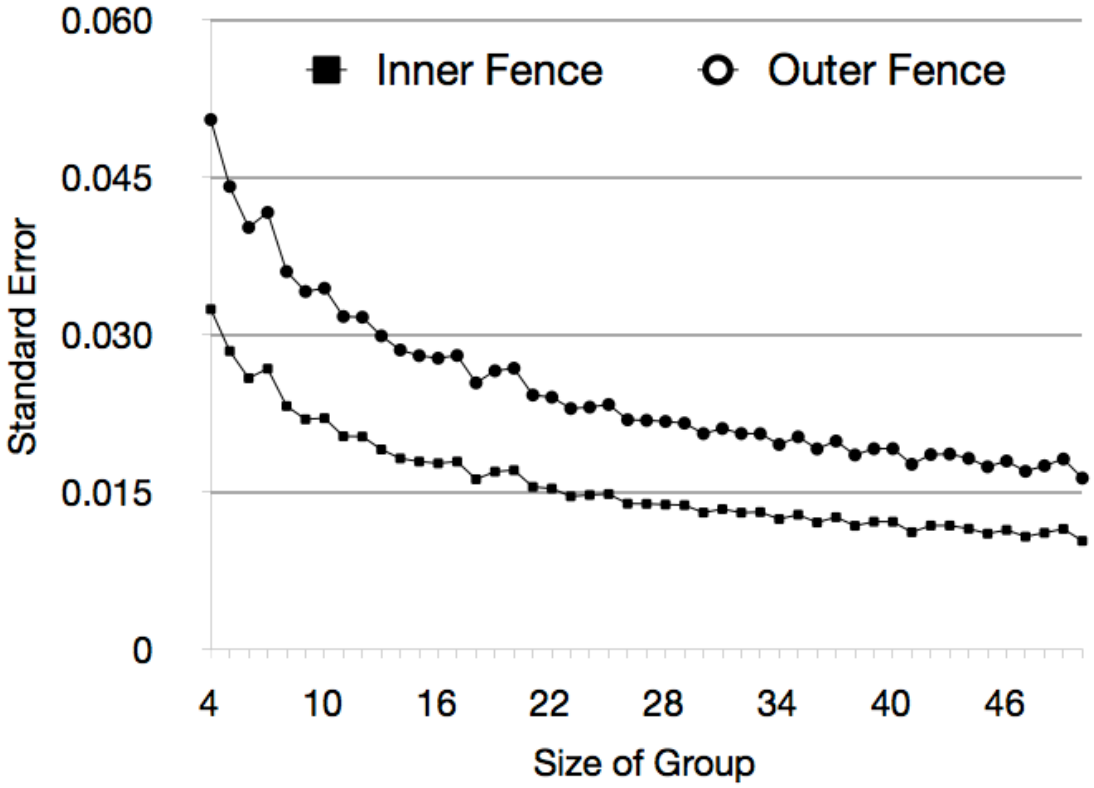

Figure 1. Standard Errors of the estimated Inner Fence and Outer Fence values

With increasing group sizes the mean values of the estimated inner and outer fences approached (figures 2 and 3) the actual known values (shown as dotted horizontal lines in these figures). The y axis is the fence value and the x axis is the increasing number of participants in the groups.

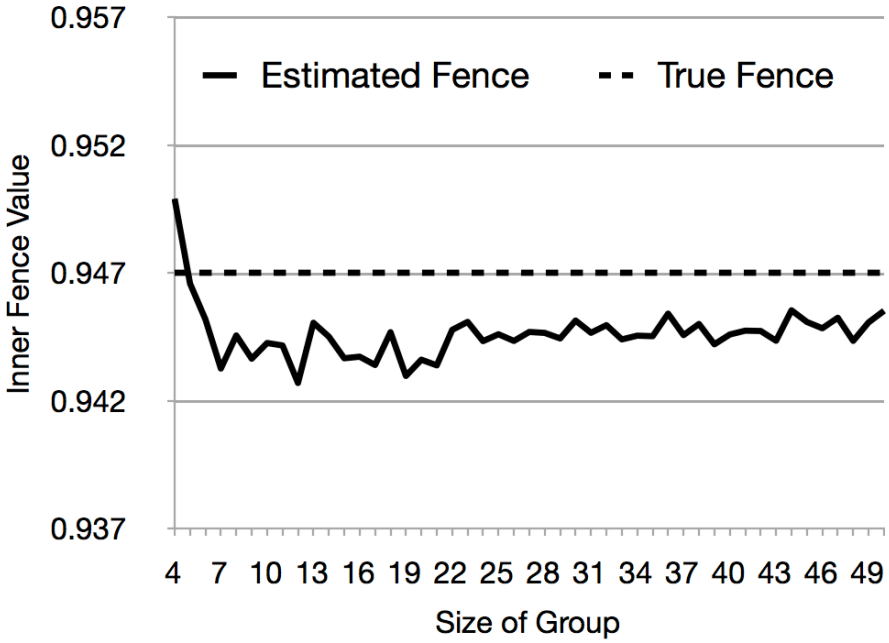

Figure 2. Mean Values of Estimated Inner Fence

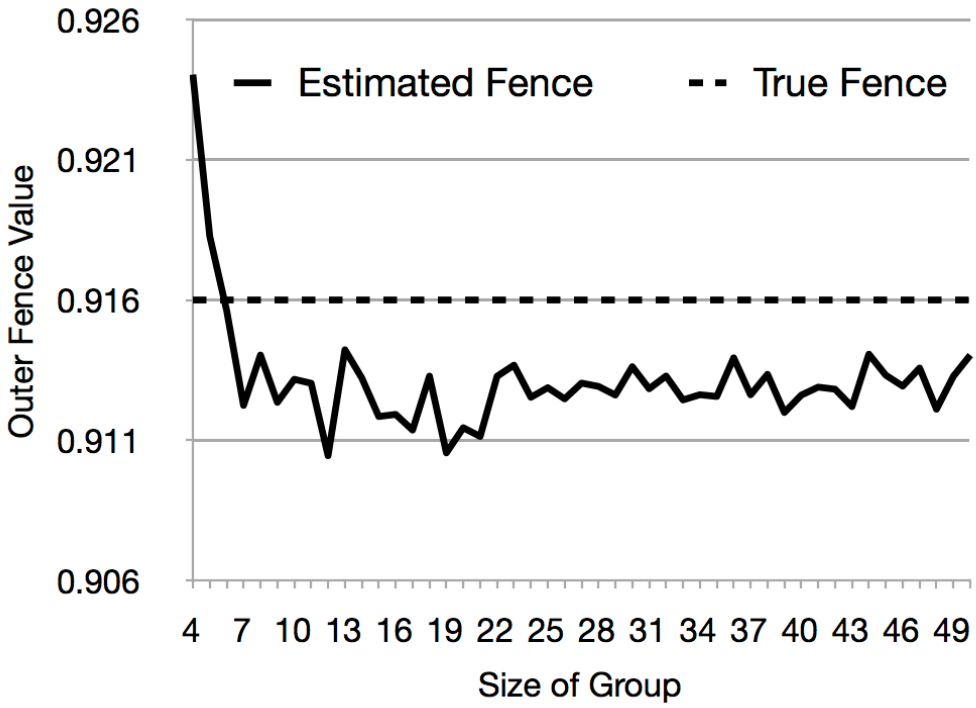

Figure 3. Mean Values of Estimated Outer Fence

The estimated threshold distributions are plotted in figures 4 to 7 for different size groups, together with the true inner and outer fence values (shown as vertical lines to the right and left of each figure respectively). With as few as four people the inner fence approximation is evident (figure 4) and is much clearer with 10 people (figure 5). As group size increases (25 to 50 people – figures 6 and 7 respectively) then a very good approximation of the true values is achieved.

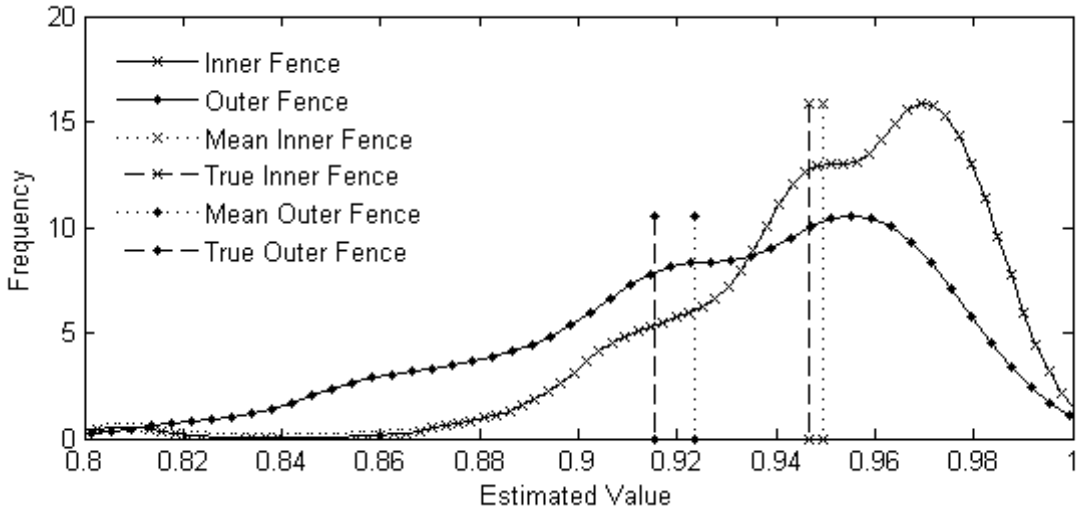

Figure 4. Thresholds Distribution of Groups with Size 4

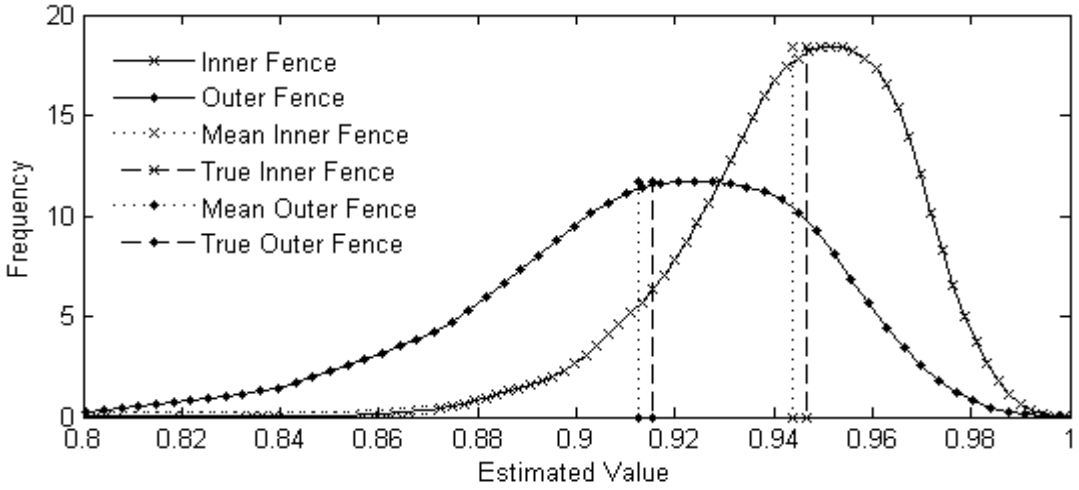

Figure 5. Thresholds Distribution of Groups with Size 10

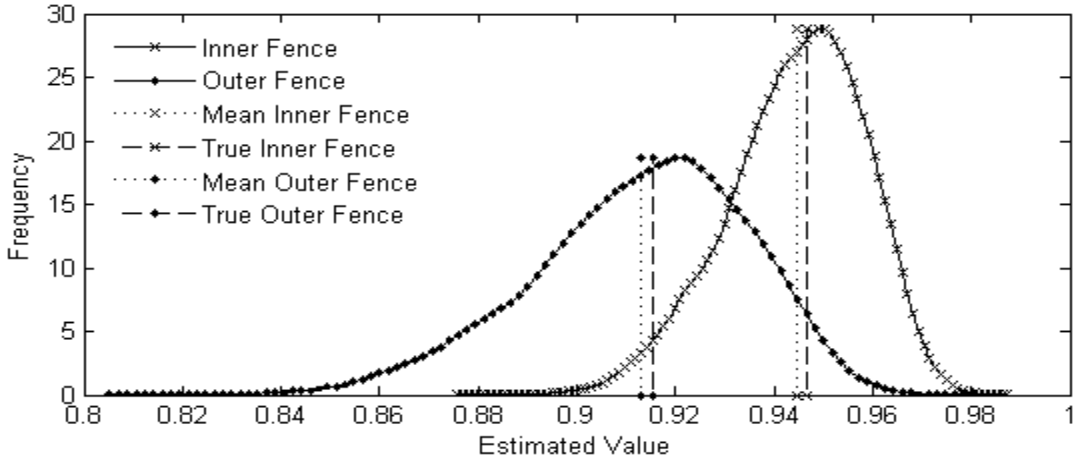

Figure 6. Thresholds Distribution of Groups with Size 25

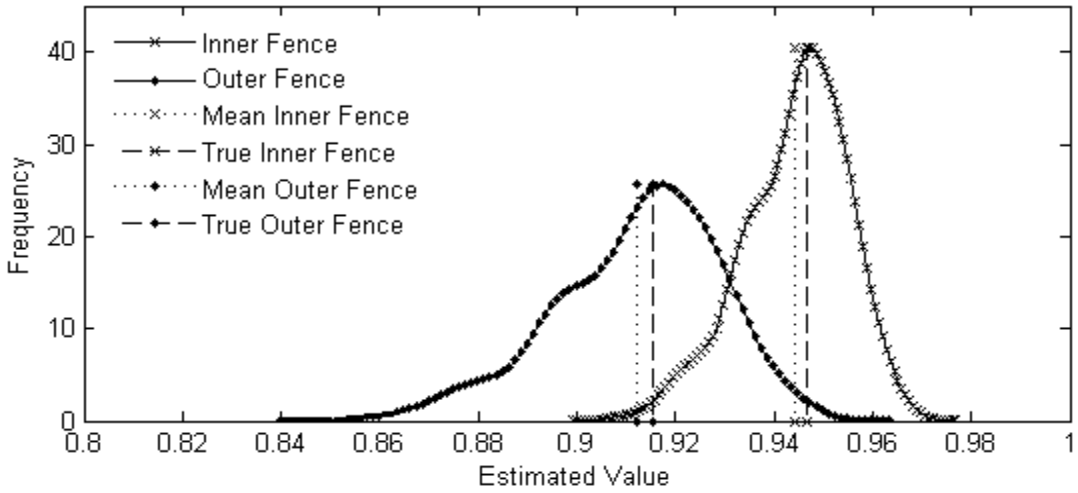

Figure 7. Thresholds Distribution of Groups with Size 50

In order to quantify the estimation accuracy for the plots in the figures above, three error offsets (0.01, 0.02 and 0.05) have been set up as reference standards. For instance, when we have an already known low performance threshold inner fence value of 0.947, then an error-offset value 'e' is used to construct an error interval which is 0.947±e. Applying this approach to the data underlying figures 4 to figure 7 then the area under the inner fence curve between this error interval (0.947±e) has been calculated. Then this value divided by the area under the whole inner fence curve yields a percentage value that is equivalent to the accuracy achieved in estimating the empirical fence value.

Implementing this approach for the inner (Table 1) and outer fence (Table 2) then for both measures the estimation accuracy increased as the size of groups used increased. By comparing Table 1 with Table 2, we can see that the estimation accuracy of the outer fence did not perform as well as the estimation accuracy of the inner fence. Furthermore, from Table 1, when the error offset used is 0.05, then the estimation accuracy reached 100% with a group size of 50 and 99.6% with a group size of 25 which is also very high.

Group size	*0.01*	*0.02*	*0.05*
4	23.4%	44.3%	93.9%
10	33.4%	63.4%	96.6%
25	54.3%	86.8%	99.6%
50	68.7%	92.7%	100%

Table 1. Inner fence accuracy rate of groups with different numbers of screeners under the error interval of 0.01, 0.02 and 0.05

Group size	*0.01*	*0.02*	*0.05*
4	16%	29.3%	73.1%
10	22.7%	43.3%	87.8%
25	37.5%	64.2%	95.8%
50	48.9%	78.3%	98.5%

Table 2. Outer fence accuracy rate of groups with different numbers of screeners under the error interval of 0.01, 0.02 and 0.05

4. DISCUSSION & CONCLUSION

Errors occur in any situation for numerous reasons ranging from system failures, design errors and human error of various kinds. In breast screening when an error occurs it can be extremely traumatic and affect several women. In the UK in 2002 a case happened where cancers in 11 women had been missed and in 2007 seven women were similarly given the all clear. In 2009 some 14 women attending one centre were found to have cancer which had been missed and in 2010 a screening centre stopped screening because of concerns. In each instance the reasons for such oversights have been investigated and reported on by the Department of Health. Whilst the reason for problems in screening can be very varied it is important to minimize the potential for any type of error. Human oversight in identifying visible early signs of abnormality is key and it is argued here that the PERFORMS scheme can help in this process of minimizing errors. The Burns report[13] (2011) into the errors at East Lancashire in 2009 recommended that the UK screening programme 'should mandate that all clinical staff involved in reading mammograms participate in the external PERFORMS QA process'.

Notwithstanding any discussions about real life screening performance and performance on the PERFORMS scheme it is argued that early identification on PERFORMS of someone who does significantly less well than their colleagues is important. This may be an indicator of something going wrong with their real life screening behaviour or may simply be something reflecting how they undertook the scheme on that particular day. Either way, by identifying potential outliers as early as possible then this allows such individuals the opportunity to reflect on how such low scores may have come about and facilitates them undertaking further training if necessary as early as possible. PERFORMS allows calculation of outlying poor performances based on examining the data of all participants, circa 700, in the scheme. Here we have presented an approach which can give an indication that someone may be performing poorly simply once over 25 people have taken part. This suggests that as the scheme is deployed then an ongoing process can be run in parallel which identifies possible outliers and feeds such information back to them. Once the full scheme has been completed then such potential underperformers can be confirmed, or otherwise, by calculating actual outlier values. Further information about the scheme is available at www.performs.org.uk .

ACKNOWLEDGEMENTS

This work is partly supported by the UK National Health Service Breast Screening Programme.

PERFORMS is a registered trade mark.

REFERENCES

[1] 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)

[2] Gale A.G. & Scott H., "Measuring Radiology Performance in Breast Screening," M. Michell (ed.) Contemporary Issues in Cancer Imaging - Breast Cancer, Cambridge UP, Cambridge,(2010).

[3] Scott H.J., Gale A.G. & Hill S., "How are false negative cases perceived by mammographers? Which abnormalities are misinterpreted and which go undetected?" D. Manning and B Sahiner (Eds.). Image Perception, Observer Performance, and Technology Assessment. Proceedings of SPIE 6917:13, 1-11(2008).

[4] Gale, A.G. and Cowley, H.C., "Analysis of breast screening results," Doi, M L Giger, R M Nishikawa & R.A. Schmidt (Eds.), Digital Mammography '96. K Elsevier, Amsterdam, (1996).

[5] Cowley H.C. & Gale A.G., "Time of Day Effects on Mammographic Film Reading Performance," Kundel H. (Ed.) Medical Imaging 1997: Image Perception. SPIE Vol. 3036, (1997).

[6] Cowley, H C and Gale, A G., "Breast cancer screening: comparison of radiologists performance in a self-assessment scheme and in actual breast screening," In Medical Imaging 1999: Image and Performance, E.A. Krupinski (ed), Proceedings of SPIE Vol. 3663(1999).

[7] Scott H.J., Evans A., Gale A.G., Murphy A. & Reed J., "The relationship between real life breast screening and an annual self-assessment scheme," B. Sahiner & D.J. Manning (Eds.) SPIE Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment, Vol 7263, ppE1-E9(2009).

[8] Chen, Y.; Gale A.G.; Scott HJ., "Mammographic interpretation training in the UK: current difficulties and future outlook," B. Sahiner & D.J. Manning (Eds.) SPIE Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment, Vol 7263, pp C1 - C10(2009).

[9] Yap M.H. & Gale A.G.," Individualised Grid-enabled Mammographic Training System," K.M. Siddiqui & B.J. Liu (Eds.); Medical Imaging 2009: Advanced PACS-based Imaging Informatics and Therapeutic Applications, 72640V(2009).

[10] Chen Y, Gale A, Scott H, Evans A, James J., "Computer-Based Learning to Improve Breast Cancer Detection Skills," Jacko J.A. (Ed.). Proceedings of the 13th International Conference on Human-Computer Interaction. Part IV: Interacting in Various Application Domains; Springer; pp 49-57(2009).

[11] Yap M, Gale A.G. & Scott HJ., "Generic Infrastructure for Medical Informatics (GIMI): the Development of a Mammographic Training System," Krupinski E. (ed.) Digital Mammography, Springer, Berlin, (2008).

[12] Chen Y., Gale A.G., "Intelligent Computing Applications based on Eye Gaze: their role in Medical Image Interpretation," Edited proceedings of the Sixth International Conference on Intelligent Computing (Springer), Changsha, China.

[13] http://www.elht.nhs.uk/pdf/Burnsreport_Breastscreening_ELHTFINALVERSION.pdf