Title: PERFORMS – a self assessment scheme for radiologists in breast screening

Author: Alastair G Gale PhD

From: The Institute of Behavioural Sciences, University of Derby, UK

Supported by: the NHS Breast Screening Programme

Mailing address:-

Prof A.G. Gale

Institute of Behavioural Sciences

University of Derby

Kingsway House

Kingsway

Derby

DE22 3HL

UK

Email: a.g.gale@derby.ac.uk

A difficulty for any radiologist involved in breast screening is to know how well he or she is performing, given the low incidence of the disease. The latest available UK data indicate a breast cancer rate of circa 6.8% per 1,000, within the screened population of 50-64 year old women[1]. Some 15 years ago when the UK Breast Screening Programme was established it was recommended that a radiologist should read 6,000 cases a year[2]. The Royal College of Radiologists subsequently amended this figure to 5,000 cases[3] although several radiologists read many more cases than this – exceeding 20,000 annually and in rare cases up to 30,000. A similar recommendation of 5,000 cases per annum exists in other European countries[4] although in some countries, such as the USA, the number is much lower, circa 480[5].

Even when screening 5,000 cases a year a radiologist can possibly only expect on average to see a malignant case less than once in a working week. Reading a high volume of cases then is an important practice as this increases experience of the wide variety of normal mammographic appearances thereby enabling a radiologist to develop those particular skills that help identify abnormalities. Originally in the UK radiologists were recommended to read some 60 cases an hour. This rate is achievable as eye movement studies of experienced radiologists when screening demonstrate that they actually only spend a several seconds (sometimes less than 15s) examining the case itself, with the rest of the time devoted to recording their decisions[6,7]. Experience increases the speed of dealing with each case, both in terms of examining the mammograms themselves and in recording screening decisions.

Specific expertise in identifying early abnormal appearances is also related to the individual's experience and their skill in recognising key mammographic features. Theoretical models of how radiologists examine radiographic images and arrive at a decision emphasise the role of appropriately attending to such features[8]. These approaches also elucidate how errors are made due to; not visually searching the image appropriately, failing to detect information, or detecting information appropriately but then not utilising this information[9].

For each case screened the radiologist must decide either to recall the woman for further assessment or return her to routine screen where in the UK she would be screened again in three years. Feedback on whether a particular screening decision is a correct detection of malignancy or a false positive report is confirmed by subsequent follow-up or at biopsy. However, feedback on whether a case is truly negative or a false negative report is a more difficult issue. Typically the radiologist would have to wait until that woman presents herself again in the next screening round in order to confirm a decision of normality. Alternatively, a false negative report may result in the woman presenting symptomatically in the interim period. Such potential misses by an individual reader can be much reduced with double reading[10] of every screened case, an approach widely implemented across the UK.

The PERFORMS (PERsonal PerFORmance in Mammographic Screening) self assessment scheme[11,12,13] is an educational exercise which was established in 1991 as a partial response to the difficulty which an individual has regarding slow feedback on their screening performance and partially as a development of earlier research on developing a computerised decision aid, based on radiologists correctly identifying particular

mammographic features on an image[14]. Although the perceived relevant importance of some of the features originally used in this approach has since changed, the technique emphasised the importance of accurate feature identification[15]. The PERFORMS scheme is funded by the National Health Service Breast Screening Programme and it reports annually both to this programme and to the National Co-ordinating Committee for Quality Assurance Radiologists in Breast Screening.

In the UK all practising breast screening radiologists, and other suitably qualified individuals (e.g. breast clinicians/physicians and specially trained technologists) involved in reading breast screening cases, are offered the opportunity to participate annually in this free and confidential self-assessment film-reading scheme. In this system the individual reads a number of difficult recent screening cases each year. The purpose is to increase the participants' experience of a range of abnormal appearances within a short time frame. This is coupled with immediate feedback on their performance and subsequent detailed feedback where an individual's decisions on each case are judged both against any known case pathology and also against the opinions of their peers on the radiological appearance. Although PERFORMS is a voluntary scheme, the majority of film-readers involved in the UK Breast Screening Programme elect to participate and for which they receive CME credits.

The scheme functions as follows. Each year examples of interesting and difficult screening cases are collected from UK breast screening centres. These two view mammographic cases are digitised and then returned to the originating centres. In digitising, all patient and any breast centre identity information are removed and then the images are processed and stored. Any pathology related to a benign or malignant case

is recorded. A number of copies of these cases are then laser printed and sent to a small panel of experienced radiologists. Working individually, panel members identify for each case: which key mammographic features are present and indicate the location of these features on breast diagrams; categorise each breast in terms of normal, benign or malignant appearance and whether, in typical screening, that breast image should be recalled for assessment based solely upon the radiological appearance. The technical quality of each mammogram and the overall case difficulty are also rated. From these individual opinions a set of 120 cases is derived which comprise a range of normal, benign and malignant appearances, together with an agreed standard radiological opinion for each one. Importantly, this opinion is not based on the original mammograms but upon the digitised copies that are identical to the images subsequently examined by participants. For the normal cases a three-year follow up confirming normality is required, the benign and malignant cases will have been confirmed by pathology.

A number of copies of this full case set are then printed. The case set is then split into two sets of 60 cases, which are randomised with constraints, so as to provide a suitable mix of the film types. These sets are then circulated around the UK breast screening centres, which takes approximately a full calendar year for all participants to complete. As participants receive detailed feedback about each case examined then each year a new set of cases is employed.

For each screening centre radiologists and others who perform screening indicate if they wish to participate in the scheme and dates are then agreed when the PERFORMS cases are couriered to them and are then returned one or two weeks later. Each participant is given a unique user name and password. All data collection and analyses

are then related to this user name and not to the named individual themselves in order to comply with data protection legislation.   A participant typically takes some 120-150 minutes to read and review each set of 60 cases. By giving participants two sets of 60 cases to read, at approximately a six monthly interval, then a fair estimate of their overall skill is obtained as this reduces the likelihood of extraneous variables affecting their performance on each session.

At the breast screening centre the films are mounted on a multiviewer and participants can then examine them at their convenience. Originally, each individual did this by completing a simple paper-based form to record his or her decisions about a case. These forms were then returned to us where this information was transposed into a computer for subsequent detailed analyses.

Currently we use a small PDA (Personal Digital Assistant) coupled with a bar code reader and associated bar-coded reporting form - although we are updating this process. Each mammographic film is also bar coded. The radiologist first identifies which case they are examining by scanning the bar code on any of the individual films and then records decisions about the case by utilising the bar coded form. Any recording errors can be corrected and changes in decisions made. The process is simple and user friendly, as a result of extensive usability testing, although the procedure is somewhat different from recording radiological decisions in typical UK screening practice. Importantly, for cases considered normal or benign the participant is also required to indicate if there are any key mammographic features present.

For each case the individual has to:

- identify and rate their confidence in whether key mammographic features are present

- specify the location of these features using a bar coded diagram of the mammographic views.

- classify each breast, both as to whether it would merit recall or not, and whether it is of normal, benign or malignant appearance.

PERFORMS predates the BI-RADS[16] reporting approach although the classification of the cases is somewhat similar. There is no equivalent to the BI-RADS '0' classification as all cases here are prejudged of diagnostic quality by the panel of experienced radiologists.

The key mammographic features used are:
- predominantly well-defined mass

- predominately ill-defined mass

- spiculate mass

- architectural distortion

- asymmetry

- calcification

- other - this allows a keyboard entry description

Once a participant has read all 60 cases then they are given immediate feedback on three aspects of their performance. This comprises: the number of malignancies they correctly identified as judged against the known pathology of these cases; the percentage of cases which they correctly recalled, and the percentage of cases correctly

returned to screen. These latter decisions are judged against the standard radiological

opinion for each case.   The radiologist can then review as many of the cases as he or

she wishes and where known pathology exists for particular cases then they can also

view this. Once they have reviewed the cases they complete a computerised

questionnaire covering details of their individual real-life screening practices - such as

weekly volume of cases read and years of screening experience - as well as feedback

information concerning their experience of the PERFORMS scheme. When all

participants at a screening centre have taken part then the PDA and mammographic

cases are returned and the recorded data downloaded into a database for analyses. The

mammographic films are checked for potential damage and then randomised prior to

sending out again to another screening centre.

There are several facets to an individual's performance on the scheme. The number of

cancers identified is a robust indicant as measured against the case pathology.

However, judgements concerning whether the radiological appearance of a case merits

its recall need to be made against peer radiologist opinion rather than against the

opinions of a small number of highly experienced radiologists. Consequently, after all

participants have completed the scheme then a 'mean peer radiological opinion' about

each case is determined, based solely upon its radiological appearance. Each

participant's various decisions concerning every case are then re-calculated against this

mean peer opinion. This comparison then provides a fairer estimate of this aspect of an

individual's actual performance.

Screening in the UK is organised within a number of health regions and so from sets of

individuals' data it is possible to produce useful anonymous regional group data as well

as national information. Consequently, all PERFORMS participants within a health region receive an annual report which details their personal performance that year as against the anonymised individual data of their regional peers. The report covers; the percentage of correct recall decisions, percentage of correct return to screen decisions, ROC performance measures (e.g. d', A_z), and the percentage of cancers identified. Such data also include regional and national mean performance information for comparison. Furthermore, the report details how the individual fared as compared to all of the UK participants by using distribution curves for each of the variables. From these distributions a participant can gauge whether they are performing above or below the mean of other participants' performance. Information can also be provided on any differences between the mammographic features identified by an individual and by their peers, together with the feature locations, which can potentially indicate where someone has a specific weakness. Individuals can then request dedicated training sets of cases which are targeted at specific mammographic features and utilised in a similar fashion to the main scheme[17,18].

Additionally it is possible to determine whether any individual who has performed poorly on some measure is simply at the bottom of the relevant data distribution curve or whether their performance merits them being classed statistically as an outlier. A formal procedure has been agreed with the National Co-ordinating Committee of QA Radiologists concerning any individuals who are found to be outliers to ensure that any reasons for poor performance are documented and additional training undertaken if necessary.

The main purpose of the scheme is to provide individuals with feedback concerning the specific screening cases that they examine so that this will aid them in interpreting future screening cases. It is not the purpose of the PERFORMS scheme to act directly as an external quality assurance device. However, an annual report is produced for the National Co-ordinating Committee for QA Radiologists using anonymised regional information. From this, the committee can study any variations across the health regions when the same set of cases has been read by virtually all UK screeners.

When individuals participate in the scheme they are presented with a case set containing many more malignancies than would be expected in typical screening practice; additionally the cases are difficult exemplars. Therefore a participants' approach in undertaking the scheme may well differ to that adopted in routine screening. Consequently, although recent screening cases are actually examined in the scheme, the process of reading these cases is not fully equivalent to the real-life situation. Some caution must then be exercised in any extrapolations from PERFORMS data to real-life screening. Nonetheless, comparisons have been made between both real-life screening and symptomatic data for radiologists in one screening centre and their data from the scheme which have demonstrated interesting correlations[19].

Of necessity, participants read the self-assessment cases at different times of the day as taking part has to fit in with their everyday duties. All responses made by participants are time-logged and so factors such as the effect of time of day on film reading performance can be examined[20]. It is also possible to compare anonymous data across different groups of readers such as radiologists and technologists[21]. This is particularly apt within the UK where there is some growing shortage of experienced radiologists in breast

screening. Such comparisons demonstrate that technologists who have been specially

trained to read screening cases perform as well as consultant radiologists on these

cases. Variations of the scheme have also been implemented as trials elsewhere, such

as in California and Germany. Comparisons have been made[22] between UK screening

radiologists and groups of Californian radiologists, split according to real-life case reading

volume, demonstrated that on this particular set of cases then those who read more

cases in real life did perform better on this case set. Whilst it must be noted that there

are various differences between the countries in their approaches to screening, this does

lend support to the argument that case volume increases skill in reading mammograms,

although case volume by itself is not the sole factor[23,24].

REFERENCES

1.   Patnick J (ed.): Annual Review 2003: Saving women for 15 years, NHS Cancer

        Screening Programmes, London, Citigate Communications, 2003

2.   Pritchard J: Quality assurance guidelines for mammography. NHS Breast

        Screening Programme, London 1989.

3.   Royal College of Radiologists: Quality Assurance Guidelines for Radiologists,

        1990

4.   Warren Burhenne LJ, Smith RA, Tabar L et al: Mammographic screening: international perspective. Seminars in Roentgenology, 36(3),187-194, 2001

5.   Pub L No. 102-539

6.   Gale AG, Walker GE: Visual search in breast cancer screening. In Brogan D, Gale, AG and Carr K (eds.) Visual Search II. London, Taylor and Francis 1993, pp 231-238.

7.   Mugglestone MD, Gale AG: CAD implementation: implications from real life visual search of screening cases. In Karssemeijer N, Thijssen M, Hendriks J. and van Erning L (eds.) Digital mammography 1998, London, Kluwer, 1998, pp375-382.

8.   Gale AG: Human response to visual stimuli. In Hendee W & Wells P. (Eds.) Perception of Visual Information, second edition. New York, Springer Verlag, 1997, pp127-147.

9.   Kundel HL, Nodine CF, Carmody D: Visual scanning, pattern recognition, and decision making in pulmonary nodule detection. Invest. Radiol. 13:175-181,1978

10.   Dinnes J, Moss S, Melia J, et al: Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. The Breast 10:455-463, 2001.

11.   Gale AG, Walker GE: Design for performance: quality assessment in a national breast screening programme.   In Lovesay E (ed.) Ergonomics - design for performance, London, Taylor & Francis, 1991, pp197-202.

12.   Savage CJ, Gale AG, Pawley EF, Wilson ARM: To err is human, to compute divine? In Gale AG, Astley SM, Dance DR et al (eds.) Digital Mammography, Elsevier, Amsterdam 1994, pp405-414.

13.   Cowley HC, Gale AG: Minimising human error in the detection of breast cancer. In: Robertson SA (ed).   Contemporary Ergonomics 1996, Taylor and Francis, London, 1996, pp379-384.

14.   Gale AG, Roebuck EJ, Riley P et al: Computer aids to mammographic diagnosis. British Journal of Radiology 60: 887-891,1987.

15.   Whatmough P, Gale AG, and Wilson ARM: Do radiologists agree on the importance of mammographic features? In K Doi K, Giger ML, Nishikawa RM et al (eds.), Digital Mammography 1996, Amsterdam, Elsevier, 1996, pp111-116.

16.   American College of Radiology, Breast Imaging And Reporting System (BI-RADS) 3rd edition. Reston, Va; American College of Radiology 1998

17   Cowley HC, Gale AG: Approaches to computer assisted training in mammographic screening. In Doi K, Giger ML, Nishikawa RM et al (ed) Digital Mammography Amsterdam, Elsevier, 1996, pp117-122

18.   Cowley HC, Gale AG, Wilson ARM.: Mammographic training sets for improving
       breast cancer detection. In Kundel HL (ed.) medical Imaging 1996: Image
       Perception Proc. SPIE 1996, pp102-112

19.   Cowley HC, Gale AG: Breast cancer screening: comparison of radiologists'
       performance in a self-assessment scheme and in actual breast screening. In
       Krupinski EA (Ed.) Medical imaging 1999: image perception and performance.
       Proceedings of SPIE, Vol 3663, 1999, pp157-168

20.   Cowley H.C. and Gale A.G. Time of day effects on mammographic film reading
       performance. In Kundel H.L. (Ed.) medcial Imaging 1997: image perception.
       Proceedings of SPIE ,1997, Vol. 3036, pp212-221

21.   Scott HJ, Gale AG, Wooding D.S. Breast Screening Technologists: does real-life
       case volume affect performance? In Chakraborty DP, Eckstein MP (Eds.) Image
       perception, observer performance and technology assessment. Proceedings of
       SPIE Vol 5372, in press.

22.   Esserman L, Cowley H, Eberle C et al: Improving the Accuracy of
       Mammography: Volume and Outcome Relationships J Natl Cancer Inst., 94: 369-
       375, 2002

23.   Smith-Bindman R, Chu PW, Miglioretti DL, et al: Comparison of screening

        mammography in the United States and the United Kingdom. Journal of the

        American Medical Association, October 2003, Vol 290, pp2129-2137.

24.   Mugglestone MD, Gale AG, Wilson ARM: Perceptual processes involved in

        mammographic film interpretation. In Kundel HL (Ed.), Medical Imaging 1997:

        Image perception, Proceedings of SPIE, Vol 3036, pp212-221

25.   Beam CA, Conant EF, Sickles EA: Association of volume and volume-

        independent factors with accuracy in screening mammogram interpretation. J

        National cancer Institute, 2003, 95, 282-290.