The relationship between real life breast screening and an annual self assessment scheme

The relationship between real life breast screening and an annual self assessment scheme.

Hazel J. Scott[1], Andrew Evans[2], Alastair G. Gale[1], Alison Murphy[2] & Jacquie Reed[2]

[1]Applied Vision Research Centre, Loughborough University, Loughborough UK, LE11 3TU
[2] Nottingham International Breast Education Centre, Nottingham City Hospital, Nottingham UK, NG5 1PB

ABSTRACT

Incidence of cancer in the UK NHS Breast Screening Programme (NHSBSP) is relatively low (approximately 7% per 1,000 cases screened). As such, feedback from cancers missed or interval cancers can be a relatively lengthy process (whereby a woman will not present for corroborating imaging for a further three years). Therefore in order to monitor their radiological skill, all breast screening radiologists and technologists read a self-assessed, standard set of challenging mammographic images bi-yearly. This scheme, 'PERFORMS' (Personal Performance in Mammographic Screening) has been running since near the inception of the NHSBSP in 1991. Although PERFORMS has functioned as an educational tool for film-readers on the UKBSP for decades, its relation to real life screening in past years has proven to be somewhat equivocal (Cowley & Gale, 1999). The present study investigated the relationship between performance measures in real life and their equivalent on the PERFORMS self assessment scheme namely: Miss Rate (FN), Cases Arbitrated and Returned to Routine screening and Incorrect recall (FP), Specificity (TN) and Cancer Detection (TP). Over 40 individuals from one NHS region in the UK submitted their real life data for comparison with PERFORMS results from the same time frame. Data from this initial study were taken from the year 2005-2006 and compared with the relevant PERFORMS set of cases. Results indicated a significant positive correlation between PERFORMS performance measures and performance measures for real life. These results are discussed in the light of the legitimacy of self-assessment comparative to film-reading skill (during real life clinical practice).

Keywords: Observer Performance Evaluation, Image Perception, PERFORMS, Breast Screening, Mammographic skill.

1. INTRODUCTION

The UK Breast screening programme recommends that film-readers read approximately 5,000 screening cases per annum. Given the low incidence of breast cancer, approximately 7 per 1,000 any given film-reader can expect to see few malignant cases per week. Feedback on those cases that have been recalled is relatively rapid (pending further imaging or biopsy) but information on false negatives can be extremely lengthy, given that a woman may not return for a future screen for two to three years hence. PERFORMS (PERsonal PerFORmance in Mammographic Screening) est. 1991, functions as a free and anonymous self-assessment scheme whereby film readers complete 60x2 difficult cases annually and enables all participants to receive both immediate and confidential feedback on their performance (Gale and Walker, 1991). These film-sets are sourced nationally (from breast screening centres throughout the UK) and are annually renewed with current breast screening cases. The 'gold standard' for these cases are initially the majority expert decisions of five experienced radiologists and latterly the majority decision of all participating film readers (plus pathology). Typically circa 90% of breast screening film-readers complete the PERFORMS set.

However, although previous studies have examined the relationship between performance on PERFORMS and performance in real life (Cowley, Gale and Wilson, 1996 & Cowley and Gale, 1999), results have been inconclusive. Cowley and Gale (1999) concluded that similar portions of features (taken from the interval cancer database) were missed in real life *and* on PEFORMS schemes. The interval cancer database is a record of those cases that were 'missed' by radiologists in the regular screening round but presented symptomatically before the next screening round; therefore as these were cases incorrectly reported in real life they are akin to those missed on self-assessment. Analysis

Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment,
edited by Berkman Sahiner, David J. Manning, Proc. of SPIE Vol. 7263, 72631E

revealed that there were significant correlations for the proportion of specific features missed in real life and those missed on the PERFORMS set. Further detailed examination of individual radiologist's data revealed that PERFORMS data in some instances (for certain PERFORMS 'rounds') mirrored real life performance. Specifically, PERFORMS measures of 'correct recall'(a measure of sensitivity) and 'correct return to screen' (a measure of specificity) were correlated with radiologists real life 'recall rates'. Real life sensitivity measures for the years 1996 and 1997 were also significantly correlated with the amount of missed cancers on the PERFORMS set. It was concluded that these results, although inconclusive, were encouraging due to a well reported variability in radiologists' performance. These results were somewhat indicative of a symbiotic relationship between PERFORMS (an artificial self-assessment task) and actual film reading in the UK National Health Service Breast Screening Programme (NHSBSP). Messick (1989) when addressing the validity of educational measurement admits "that test responses are a function not only of items, tasks, or stimulus conditions but of the persons responding and the context of measurement" (p14). Such persons, as in this instance, can produce the aforementioned variations in performance. One way of counteracting these individual differences is in the comparisons of as many individuals as possible comparing similar or the same measurements. Such an attempt was made in a further examination of the results by Cowley and Gale (unpublished Departmental Report) where real life measures and those on PERFORMS were subjected to the same measurement formula (such as 'Miss Rate' see next section). Therefore for the current analysis we aim to reproduce these methods, comparing wherever possible 'like with like'.

As PERFORMS aims to function as a learning/skill improvement exercise for film readers in the NHS, and all breast screening units in the UK allocate a small but significant proportion of time for its execution (twice yearly) a pertinent question is one of its legitimacy. PERFORMS with its rapid confidential feedback serves as an early warning for those who under-perform and anecdotal reports suggest that such individuals consequently undertake further training (a further commitment of time). PERFORMS also has a mechanism whereby those individuals who significantly under-perform (outliers) are subject to further communication and in some cases action by the relevant radiological body. It would therefore seem critical to examine in greater detail if *what* the scheme measures is objectively useful and therefore valid (the previously cited variability in radiologists' performance not withstanding).

Although in the measurement of test validity three key elements are commonly cited (newer approaches (Messick (1989) expand this three tier classification to six); validity related to content (measuring the correct skills), criterion (the degree to which one can infer the test results are related to another criteria representing achievement i.e. compared to real life behaviour), and construct (if the items in the test measure a uniform set of behaviours). PERFORMS could be arguably strong in two of the aforementioned, as PERFORMS self assessment tests use the same cases radiologists read in real life and logically must tap into the same skill base (content-validity) and in addition each case derives from a source with the same spread of radiological features (conceivably test items i.e. construct-validity). However, the criterion-validity requires matching PEFORMS 'test' results with that of real life - which has not been examined for recent PERFORMS sets. One could also consider ecological validity and the differences between something which is 'naturalistic' (taking place in natural environment with familiar items – which PERFORMS does) compared to something which is 'realistic' in that comparisons can be made to real life. Although it might also be proposed that PERFORMS does not adhere strictly to 'Mundane Realism' (Carlsmith et al, 1976) as the sets are manipulated in order to contain a high percentage of abnormalities as well as a far higher than normal range of radiological appearances.

In this study we aimed to re-examine this relationship for more current PERFORMS film sets with real life data from the same time frame. Previous studies (Cowley & Gale 1999) have shown a tentative relationship between performance on PERFORMS and real life screening. This research aims to update and expand on these previous works with a view to validating more fully PERFORMS self assessed 'tests' as an educational tool within the Breast Screening Programme.

2. METHODOLOGY

2.1 Design

Results, from 48 individuals from one region in the UK, including breast-screening radiologists, technologists (specially trained in mammographic film-reading) and other health professionals, for PERFORMS sets (in the year 2005-2006), were compared to real life performance from the same time period. PERFORMS results were extracted from the main PERFORMS database and real life data was extracted from the NBSS NHS database. Nine participants were not

included in the final analysis as they did not fulfil the inclusion criterion for the study, whereby each participant had to have completed at least 2 rounds of PERFORMS sets in the 12 month period of the study and must have read over 2,000 real life cases (as a first reader). A within-subjects design was employed with one group of circa 50 participants from several Breast Screening Units (in East Midlands). Participants performance in real life screening was compared/correlated with similar performance measures on the PERFORMS film-set.

2.2 Participants

Inclusion criteria for all participants were as follows: at least two rounds of PERFORMS completed within 12 months of the screening data period and at least 2,000 screening cases read as first reader with results entered onto NBSS. Participants included were radiologists, technologists, breast physicians and registrars. From a cohort of 48 screening readers in 05'-06', 9 individuals could not be included in the comparison because they did not meet the inclusion criterion of these:-

- 8 individuals had not read 2,000 screening cases as first reader of these:
    - 2 had not completed at least two recent rounds of PERFORMS
    - 1 was not present on the PERFORMS database
- 1 participant had not completed two recent rounds of PERFORMS

In addition one other participant did not quite complete 2,000 cases as a first reader but was very close to this number (by 12 cases), this individual is also an expert radiologist so was included in the study.
All performs measures were calculated using the case pathology and the National Radiological Opinion (gleaned from the national average as well as pathology).

2.3 Materials

Real life data was extracted by crystal report from the NBSS database for the years 2005 and 2006 - these were approximately comparable to the appropriate PERFORMS set (SA07(1)).

2.4 Procedure

Initially all film-readers were contacted for their permission to have their data included in the study. Following this, a standard crystal report was written to extract the relevant data from the NBSS database (this crystal report was made available to all breast screening units using NBSS). PERFORMS data from the last three rounds was extracted from the PERFORMS database detailing results on all measures as compared with a 'National Opinion' (based on pathology and the majority decisions of all PERFORMS film readers for that set).
Data extracted for comparison included:-

1. PPV and PPV on PERFORMS (TP Measure)
2. Missed Cancers on PERFORMS and Incorrectly Returned Cases in routine screening i.e. missed cancers in breast-screening (FN measure).
3. Miss Rate on PERFORMS and Real Life (FN Measure).
4. Percentage of Cases Arbitrated and Returned to Routine Screening and Incorrect Recall Percentage on PERFORMS (FP).
5. Specificity (in real life) and Correct Return to Screen Percentage on PERFORMS (TN).
6. Correct Recall Percentage (in real life) and Correct Recall Percentage on PERFORMS (TP).
7. Cancer Detection and PERFORMS Malignancies Detected (TP).

3. RESULTS

We compared the following measures in real life and on the PERFORMS self assessment scheme, specifically we looked to see if there were significant correlations between PERFORMS and real life results for all FP,FN,TP,TN measures as well as PPV. Results were, where possible, fractionated for reading as a first reader and reading as a second reader.

3.1 PPV and PPV on PERFORMS (TP Measure)

Positive Predictive Value on PERFORMS and real life were compared. Pearson product moment correlation revealed that there was a significant correlation (one-tailed) between PPV in real life and PPV in PERFORMS ($r(39)= 0.407$; $p<.01$. $r^2 = 016$) and PPV by first ($r(39)= 0.381$; $p<.01$. $r^2 = 0.14$) and second reader($r(39)= 0.371$; $p<.01$. $r^2 = 0.137$).

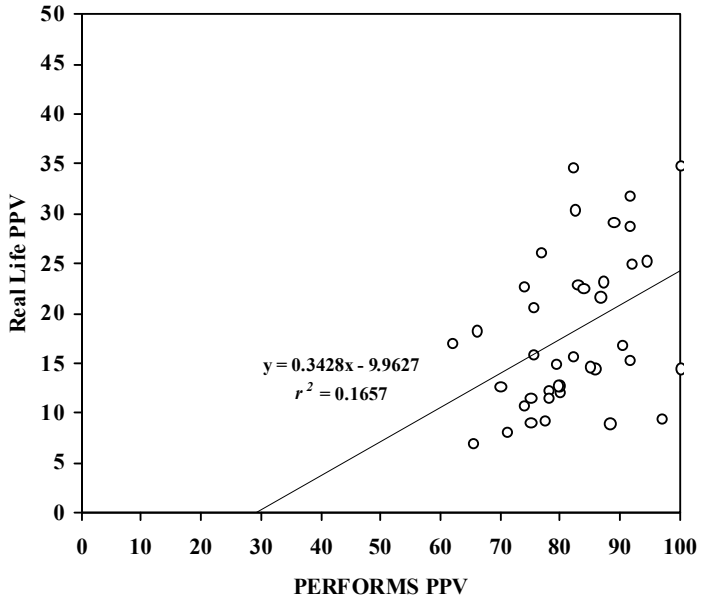

Fig 1. Overall PPV in PERFORMS and Real Life

3.2 Missed Cancers on PERFORMS and Incorrectly Returned Cases in routine screening i.e. missed cancers in breast-screening (FN measure).

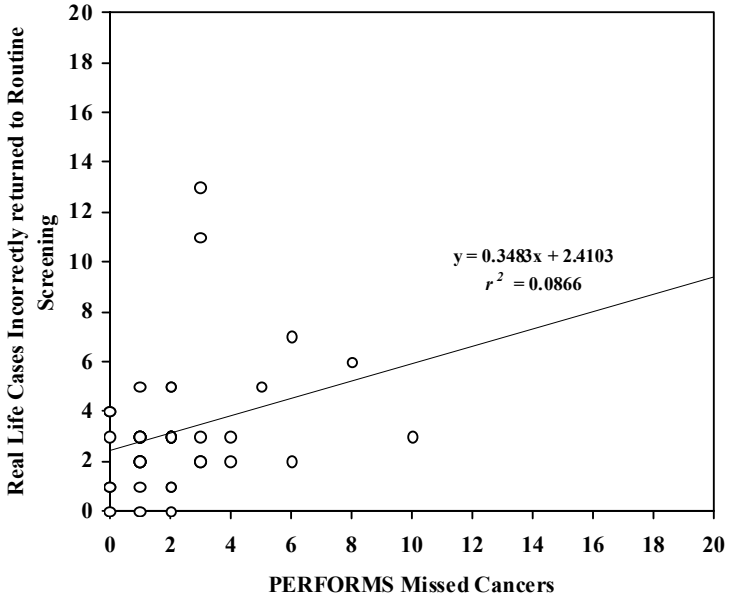

Fig 2. Missed Cancers on PERFORMS and Real Life Cases that were Incorrectly Returned to Screen

The number of cancers missed on PERFORMS were compared to the number of cases incorrectly returned to routine screening in the same time frame. There were significant positive correlations (one-tailed) for overall cases incorrectly returned to routine screening and overall missed cancers on PERFORMS ($r(39)= 0.294$; $p<.05$. $r^2 = 0.08$) as well as for cases incorrectly returned to screen as a first reader ($r(39)= 0.355$; $p<.05$. $r^2 = 0.12$). There were no significant correlations for PERFORMS missed cancers and overall cases incorrectly return to screen as a second reader.

3.3 Miss Rate on PERFORMS and Real Life (FN Measure)

As a more direct comparison of FN measures, miss rate was calculated from real life data as well as for PERFORMS data using the following formula:

$$\textbf{Miss rate(\%)} = \frac{\textbf{Number of cancers missed}}{\textbf{Number of cancers detected + missed}} \times 100$$

There were significant correlations between overall miss rate in real life and miss rate in PERFORMS $r(39)= 0.321$; $p<.05$. $r^2 = 0.10$).

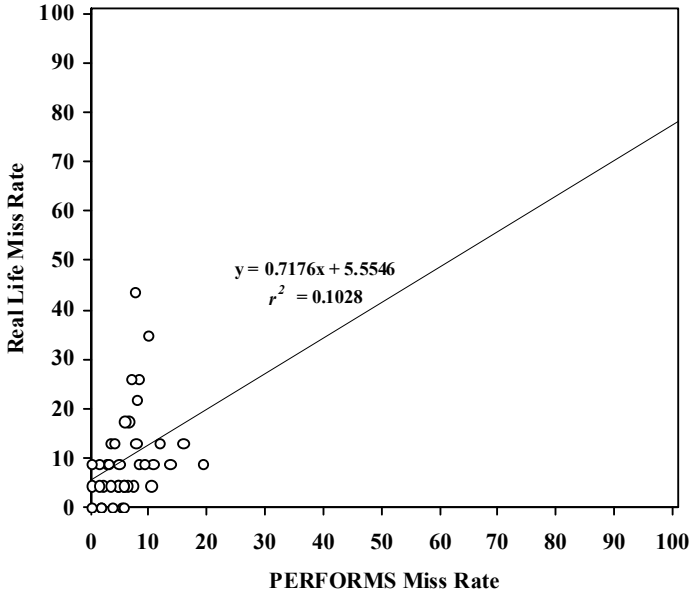

Fig 3. Overall Miss Rate (real life) and Miss Rate on PERFORMS

There were also significant correlations between PERFORMS miss rate and miss rate as a first reader ($r(39)= 0.409$; $p<.005$. $r^2 = 0.17$) - although miss rate as a second reader comparisons failed to reach significance (p=n.s).

3.4 Percentage of Cases Arbitrated and Returned to Routine Screening and Incorrect Recall Percentage on PERFORMS (FP)

The percentage of cases arbitrated and returned to routine screening were compared to percentage of incorrect recall on PERFORMS. A Person correlation (one-tailed) revealed that there was a significant association between these measures ($r(39)= 0.372$; $p<.01$. $r^2 = 0.13$)

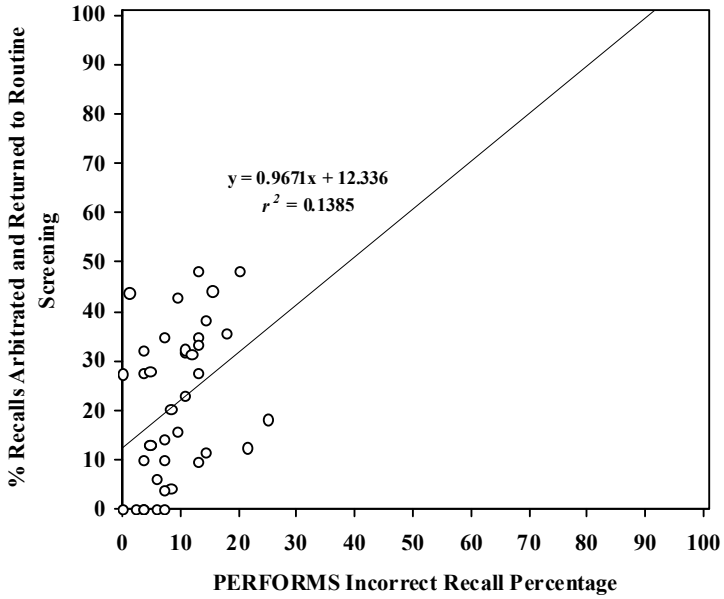

Fig 4. Overall % Recalls Arbitrated and Returned to Routine Screening (real life) and Incorrect Recall Percentage on PERFORMS

3.5 Specificity (in real life) and Correct Return to Screen Percentage on PERFORMS (TN)
In order to approximate a True Negative Score from the NBSS data the following formula was used to calculate Specificity.

No. of Normal Films Read = (No. of Films read – Cases Incorrectly Returned to Routine Screening) – (No. of Cases Recalled – No. of Recalls Arbitrated and returned to routine screening)
No. of Normal Films Correct = No. of Normal Films Read – Number of Recalls Arbitrated and returned to routine screening.

$$\textbf{Specificity \%} = \frac{\textbf{No. of Normal Films Correct}}{\textbf{No. of Normal Films Read}} \times 100$$

There was a significant correlation overall for Specificity (in real life) and Correct Return to Screen Percentage on PERFORMS (($r(39) = 0.349$; $p < .05$. $r^2 = 0.12$).

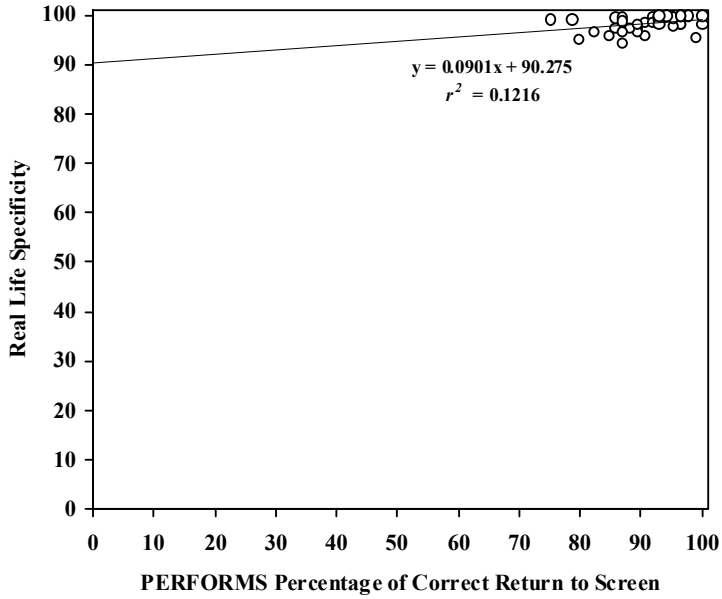

Fig 5. Specificity(real life) and in PERFORMS (Correct Return to Screen)

3.6 Correct Recall Percentage (in real life) and Correct Recall Percentage on PERFORMS (TP)
Correct Recall Percentage was calculated from real life data using the following formulae:-

No. Total Recalls = No. of Women Recalled + Cases Incorrectly Returned to Routine Screening
No. of Correct Recalls = No. of Women Recalled – Number of Recalls Arbitrated and Returned to Routine Screening

%Correct Recalls = **No. of Correct Recalls** x 100
No. of Possible Recalls

Correlations between Correct Recall Percentage (real life) and Correct Recall Percentage (PERFORMS) failed to reach significance for any of the measures.

3.7 Cancer Detection and PERFORMS Malignancies Detected (TP)

Percentage of cancers detected to film read overall (in real life) was compared to percentage of Malignancies Detected on PERFORMS. There was a significant (one-tailed) person correlation between PERFORMS malignancies detected and overall percentage of cancers detected ($r(39)= 0.301$; $p<.05$. $r^2 = 0.09$). The correlation for cancers detected to films as first reader and PERFORMS Malignancies Detected was approaching significance (p=.06).

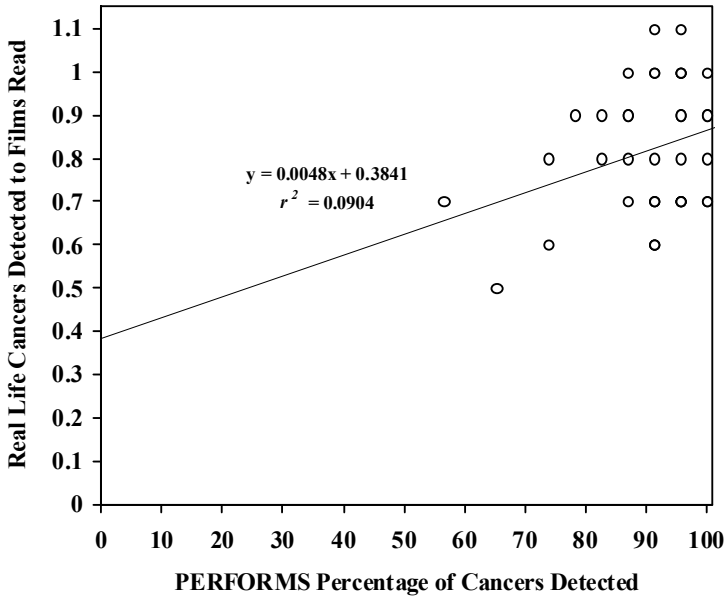

Fig 6. Overall Malignancies Detected to Films Read (Real Life) and Correct Recall Percentage on PERFORMS

4. DISCUSSION

Significant correlations were found between most real life measures and PERFORMS data. Positive correlations were found for overall measures of PPV, Cases Incorrectly Returned (FN), Miss Rate (FN), Cases Arbitrated and Returned to Routine Screening (FP),Specificity Measures (TN) and Cancers Detected (TP). The only overall correlations that failed to reach significance were for Correct Recall Percentages (TP). Overall, real life performance measures and PERFORMS measures, as well as other measures calculated for both data sets, reveal an almost consistently positive (albeit very weak in some instances) relationship.

These results compare favourably to previous work in this area (looking at the relationship between PERFORMS and real life at the inception of the scheme) and illustrates a stronger relationship between real life and PERFORMS performance than was found previously. Perhaps this was due to the inclusion of calculated measures for Specificity, Miss Rate and Correct Recall as well as increased sample size.

Real life comparisons (when available) for first reader were positively related to PERFORMS for PPV, Missed Cancers, Miss Rate and were approaching significance for malignancies detected. Save PPV, there were no significant positive correlations for reading as a second reader and PERFORMS data. This result was not unexpected in that reading as a first reader (where there are no previous radiological opinions to aid judgement) is akin to reading a PERFORMS set which is completed individually.

There were some methodological issues with the data set, which although larger than previously, was relatively small ($n<50$) compared to the number of film readers who read PERFORMS (n=500+), sample size was further reduced by study criterion. The correlations although significant showed small r^2 values indicating a weak relationship for many of the measures which may be due to this still limited sample size. Therefore, widening the study to at least double the inclusion numbers will be the focus of future work.

These data indicate that there is a definite relationship between how one reads PERFORMS films in a somewhat artificial way (although as naturalistic as is logistically possible) and how film-readers assess films in real life clinical practice. These results go someway to providing solid criterion-validation for the PERFORMS scheme, showing that PERFORMS measures are directly comparable to observable achievement for the same skill set outside of the self

assessed 'test' environment. Further work is needed to strengthen this validation, as due to some of the weak correlations, PERFORMS cannot yet be said to the 'litmus test of legitimacy' for predicting real life performance.

These data also suggest that individuals' performance is a more stable factor than previously reported, as participants in this study showed similar results across real life and self-assessment domains. This may be due simply to a larger sample size (than previously) but may also be due to an increase in skill level over years of screening experience (Scott, Gale & Wooding 2004, Scott and Gale 2007) as the initial study in 1999 was completed with data from the beginning of the breast screening programme. The current study includes a cohort with considerably larger years of experience in not only mammography but in reading PERFORMS sets as well. Previous papers have outlined that those who have read PERFORMS for the first time or for the first year tend to read the set less well than more experienced mammographers. In addition, first time film-readers are common in the outlying group but research shows that they improve significantly on future sets (Gale and Scott, 2008).

Correlations between PERFORMS measures (TP,TN,FP,FN) are *generally* correlated with most breast screening data. Skill in PERFORMS generally reflects how an individual is performing in breast screening, but is not always wholly representative of every aspect of an individual's real life performance – a strong indicator rather than a measure.

5. CONCLUSIONS

Work is presented with a view to ascertaining the legitimacy of self-assessed performance as an indicator of real life practice. It was concluded that self-assessment on PERFORMS, although dissimilar to breast screening practice, broadly reflected breast-screening performance for the same time interval.

ACKNOWLEDGEMENTS

We would like to thank and gratefully acknowledge all those who allowed their data to be included for this study, in particular we would like to thank those involved in compiling the real-life data for these analyses.

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

REFERENCES

[1] Gale A.G. & Walker G.E., "Design for performance: quality assessment in a national breast screening programme." in *Ergonomics - design for performance 1991*, edited by E. Lovesay, Taylor & Francis, London.
[2]Cowley, H., Gale A. & Wilson, R. "Mammographic training sets for improving breast cancer detection." in *Medical Imaging 1996: Image Perception and Performance* edited by Miguel P. Eckstein & D.P. Chakraborty, Proceedings of SPIE Vol. 2712, pp. 102-112.
[3]Cowley, H. & Gale A., "Breast Cancer Screening: Comparison of radiologists' performance in a self-assessment scheme and in actual breast screening. " in *Medical Imaging 1999: Image Perception and Performance* edited by Elizabeth A. Krupinski, Proceedings of SPIE Vol. 3663, pp. 157-168.
[4] Messick, S., [Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103)]. New York: Macmillan (1989).
[5] Carlsmith, J., Ellesworth, P. & Arnson, E., [Methods of research in social psychology.] Reading, Mass.: Addison-Wesley (1976).
[6] Scott, H. J., Gale, A. G., Wooding, D. S., "Breast screening technologists: Does real-life case volume affect performance?" in *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*, edited by Dev P. Chakraborty, Miguel P. Eckstein, Proceedings of SPIE Vol. 5372 (SPIE, Bellingham, WA 2004) pp. 399-406.
[7] Scott, H. J., Gale, A. G., "How much is enough? Factors affecting the optimal interpretation of breast screening mammograms" in *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*, edited by Yulei Jiang, Berkman Sahiner, Proceedings of SPIE Vol. 6515 (SPIE, Bellingham, WA 2007) 65150F.
[8] Gale, A. G. & Scott, H.J., "Patient Safety in Radiology: An Example from Breast Screening" in *Proceedings of Improving Patient Safety 2008 Conference*, edited by Sue Hignett, Beverley Norris, Ken Catchpole, Allen Hutchinson & Sarah Tapley, (The Ergonomics Society) pp. 345-349.