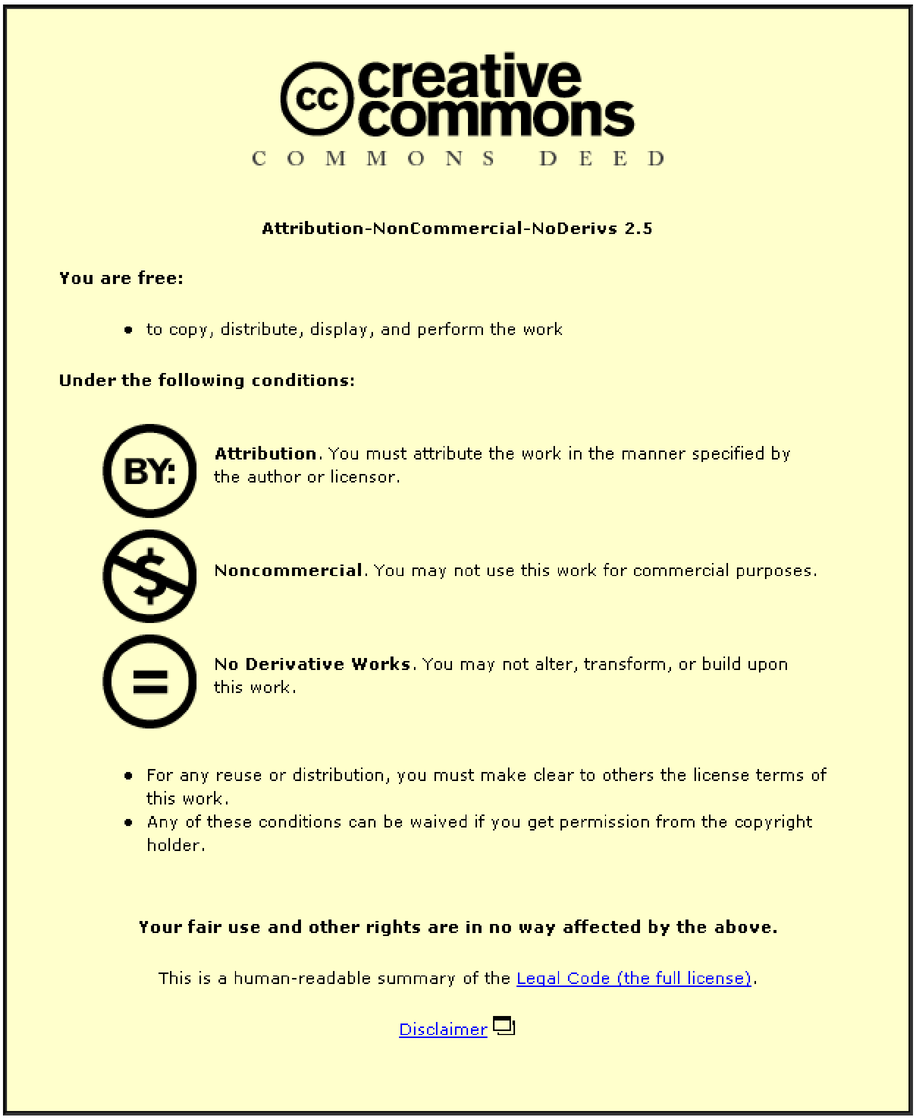

Does mammographic practice affect film reading style – Breast Screening vs. Symptomatic Radiologists?

Hazel J. Scott and Alastair G. Gale*
Applied Vision Research Centre, Loughborough University, Garendon Wing, Holywell Way,
Loughborough, LE11 3TU, UK

ABSTRACT

In the UK there are two groups of radiologists who routinely read mammographic cases: Symptomatic and Screening Radiologists. We examined the performance of these two film-reading populations, Breast Screening Radiologists and Symptomatic Radiologists, to evaluate if there were group differences in their 'style' of reading the same set of cases. Specifically we looked at each group's sensitivity and specificity measures. In addition we investigated if there were any individual group differences apparent in the cases which they found challenging and what (if any) were the characteristics of those cases. Data from 66 Breast Screening Radiologists and a matched group of 66 Symptomatic Radiologists were compared over a number of years (360 cases). Results are presented which demonstrate that whilst the two groups show overall similarities in performance there exist subtle underlying differences which we attribute to the differences in their everyday experience of the types of cases that they read. In conclusion, we argue that these differences are related to the volume of cases which UK Screening Radiologists read in order to maintain skill level.

Keywords: PERFORMS, Breast Screening, Performance, Reading Styles, Radiologist, Symptomatic, Sensitivity.

1. INTRODUCTION

In the UK there are two groups of Radiologists who routinely read mammographic cases: these are Symptomatic and Screening Radiologists. A Screening Radiologist is an individual who is employed within the UK National Breast Screening Programme and consequently must read a minimum of 5,000 screening cases per annum. Given the low incidence of breast cancer (circa 6.8 per 1,000 screened women in the age range 50-64 years in the UK (1)) and a typical screening volume of 20,000 cases per annum then a typical Screening Radiologist can expect to examine a great number of normal cases, some benign cases and a few malignant cases per month. Even those individuals who read close to the minimum number of 5,000 cases a year will potentially, on average, only interpret very few malignant cases a month. In marked contrast, Symptomatic Radiologists in general will read far fewer cases per annum and generally only read mammograms where there is a potential abnormality (e.g. a palpable lump), which has been previously detected either by a General Practitioner, by the patient themselves, or where some other breast symptom (e.g. pain) has been reported to the woman's General Practitioner. Therefore Symptomatic Radiologists will examine a far higher ratio of potential abnormal cases, often where they additionally know the potential location of a possible abnormality, as compared to Screening Radiologists.

Virtually all UK Breast Screening Radiologists take part in a self-assessment scheme ('PERFORMS') where they read a difficult set of recent screening cases each year (2). In this scheme, participants interpret two sets of 60 cases per annum and receive detailed feedback on their skills as compared to other participants. Additionally, some UK Symptomatic Radiologists also opt to voluntarily read these case sets. This then offers the opportunity to anonymously compare how these two groups of radiologists fare when they examine these sets of difficult cases.

The interest in undertaking this comparison lies in the theoretical approach which we adopt in considering that for Screening Radiologists - their skill in interpreting such screening mammograms lies in them examining a large number of normal cases. This then allows them to build up a lexicon of what constitutes variable normal mammographic appearances and that this implicit knowledge then allows them to readily identify when a case contains some feature or

*a.g.gale@lboro.ac.uk; phone, 0044 1509635701; fax, 0044 1509635709; www.appliedvision.org

Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment, edited by Yulei Jiang,
Miguel P. Eckstein, Proc. of SPIE Vol. 6146, 61460X, (2006) · 1605-7422/06/$15 · doi: 10.1117/12.651705

features which indicate the presence of a possible abnormality. We have previously argued that reading a high volume of screening cases is an important factor in developing and maintaining screening performance (3).

In contrast, for the Symptomatic Radiologists, the reverse argument would be that they amass a knowledge of abnormal appearances (which includes both malignant and benign appearances) but do not build up such an extended knowledge of the range of possible normal appearances. An earlier study has shown (7) that performance in symptomatic radiology is not as good as that in a screening environment, however this may be, in part, due (the authors argue) to the different types of cases (as symptomatic reading is likely to involve entirely different patient demographic consisting of younger women with denser breast parenchyma) as well as marked differences in the diagnostic practice. We suggest that it may also be due to differing volume of cases in both practices. However, other work has shown (8) that there is no difference in accuracy (in terms of FP and FN) between breast screening units which do both breast screening and symptomatic examinations and those that do symptomatic examinations only. Some Screening Radiologists also undertake symptomatic examinations and we have previously examined the detailed performance of a small group of radiologists on some PERFORMS case sets and their real life screening and symptomatic performances. Some correlations were found between pooled ROC performances on the PERFORMS case sets and both symptomatic and screening performances (2).

We therefore examined if there were any differences in film reading performance and reporting style on these case sets of mammograms of known outcome between these two groups, and if so, whether such variability could be attributable to the differences in their everyday practice.

2. METHODOLOGY

PERFORMS self assessment cases are volunteered by over 35 breast screening units throughout the UK. Units submit cases that are deemed to be difficult exemplars of particular mammographic features and/or classification. These cases subsequently undergo a rigorous examination by a panel of over five highly experienced radiologists who rate each case in terms of classification, feature type, difficulty, case appearance and case suitability.

From these opinions (which make up our initial ‘gold standard’ for the cases) we can extract the most relevant cases, in terms of those that will be the most discerning, for each PERFORMS set (a case set is made up of lots of 60 cases per annum). As such we arrive at a set of cases with a range of difficulty, mammographic appearance and classification type (malignant, benign and normal cases). After all participating individuals have completed a PERFORMS set the initial radiological gold standard is discarded in favour of a more impartial standard that is based on the radiological opinion of all participating film readers (as well as case pathology). This National Radiological Opinion (NO) is used in preference to the initial Radiological Opinion (RO) for all comparisons with participants’ data. Details of the scheme have been described previously (4,5,6).

For this study we looked at the performance on the PERFORMS self-assessment scheme of a group of Symptomatic readers who read PERFORMS on a semi-regular basis compared with a group of Screening Radiologists who take part in the scheme.

We examined their accuracy on the cases in terms of sensitivity and specificity, namely; Correct Recall (CR) which was the percentage of cases where the individual correctly recalled the case (a measure of True Positive decisions) and Correct Return to Screen (CS), defined as the percentage of cases where the individual correctly returned a case to normal screen (a measure of True Negative decisions). We also examined measures which assessed true positive and true negative scores in terms of pathology alone (the correct percentage of Malignancies Detected). The results were calculated against the National Radiological Opinion.

As part of participation in the scheme, information on real life factors was collected by questionnaire and included the aforementioned information on volume of cases read per week (hence annual mammographic case volume) as well as years of mammographic radiological experience. However, completion of the questionnaire is not mandatory for participants.

Over 66 Symptomatic film-readers were matched as closely as possible over three PERFORMS case sets with the same number of Screening Radiologists (out of a total of over 300 Screening Radiologists). They were matched on two key real-life factors; the volume of cases read per week and years of mammographic experience. Their performances were then compared over three rounds of the PERFORMS scheme, i.e. on 360 difficult case exemplars of normal, benign and malignant appearances.

3. RESULTS

3.1 A Matched Design

A matched design reduces the variation due to apparent differences between the groups. Key factors which may affect participant performance were controlled for by matching the occupational groups on real-life factors (volume of cases read and years of mammographic experience) which our previous research (10) has shown to be important. For this analysis each individual Symptomatic Radiologist was closely matched, on these factors, with a Screening Radiologist. These data were gleaned from voluntary questionnaire data. Not every Symptomatic Radiologist completed this questionnaire and consequently there was a large number of missing data (21 out of 66 for years of experience and 34 out of 66 for volume of cases read). Therefore, the mean values for that year's half set was taken and this was used to form the subject match on empty data cells.

A one-way ANOVA (Analysis of Variance) was carried out which revealed no differences in participant group in terms of years of experience and volume of cases read (p = n.s.) per each half of the PERFORMS set, over a three consecutive sets (six half sets), the mean scores for these measures for each PERFORMS set were very similar (see Figure 1a and 1b).

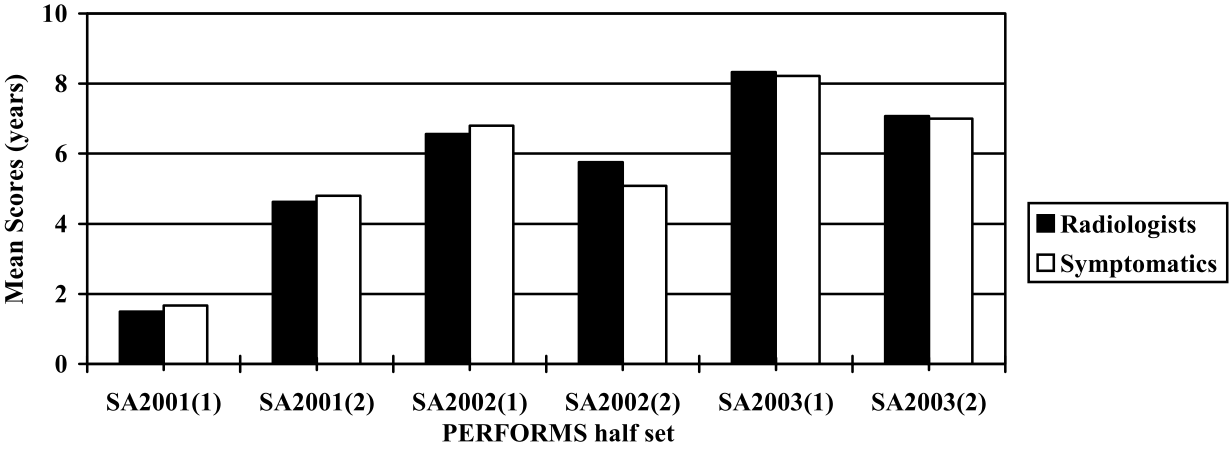

Figure 1a: Matched Participant Groups- Years of Experience

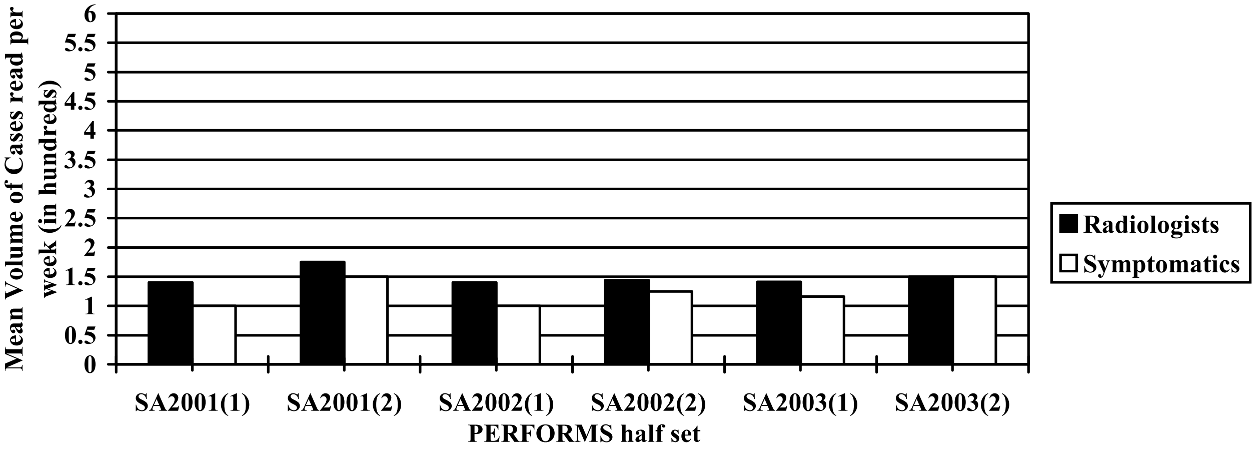

Figure 1b: Matched Participant Groups – Volume of cases read per week

3.2 Sensitivity Measures

Over the three audited rounds (termed SA2001, SA2002 and SA2003) of the scheme considered here, data concerning 360 difficult recent screening cases were examined. In each year both groups of individuals (Radiologists and Symptomatic Radiologists) participated on two occasions, reading 60 cases each time. Initially the results were analysed in terms of sensitivity measures, namely; number of malignancies detected, and correct recall (CR). For the purposes of this analysis 'correct recall' includes all classification of cases, malignant, normal and benign. This was assessed for each individual according to the scheme's 'gold standard' - a National Radiological Opinion (NO) which scores each individual against both the case pathology and the amalgamated scores of over 400 peer readers.

3.2.1 Correct Recall Comparison

A 2x2x3 uni-variate ANOVA was performed on these data with one DV (Correct Recall percentages), and two IV, occupational group (Symptomatic Radiologist and Screening Radiologist) and type of test set, split by PERFORMS set (2001-2003) and PERFORMS half case set (first or second half). Results showed that there was a significant difference of PERFORMS year [F(2,120) = 18.67, p<.001] and post hoc Student Newman Keuls (SNK) tests revealed that the percentage of correct recall was significantly higher for 2002 than for 2003 or 2001. The SA2003 was the least well performed (SA2001 mean = 85.41%, SA2002 mean = 89.54%, SA2003 mean = 79.03%). This result reflects the reported difficulty of the sets with SA2002 being reported as the easiest (measured by anecdotal comments and the highest national mean score) and SA2003 as the most difficult PERFORMS set to date.

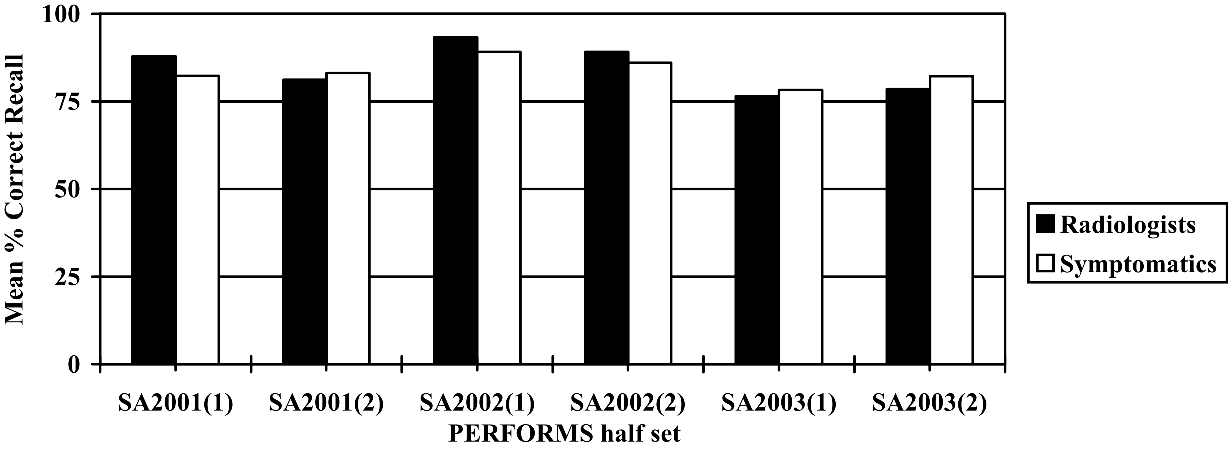

Figure 2: Percentage of Correct Recall by Occupational Group

There were no significant differences found for PERFORMS year half or occupation, nor where there any significant interactions between PERFORMS year, occupation and PERFORMS half (p=n.s) see Figure 2. Therefore there were no significant differences between the Symptomatic and Screening Radiologists for the percentage of correct recall across the three years and all 360 cases.

3.2.2. Detecting Malignant Cases

The results for the second measure of sensitivity is perhaps a purer measure of true sensitivity as it relies on the TP and TN scores from case pathology. Consequently, all benign cases are omitted in these figures. A 2x2x3 uni-variate ANOVA was performed on these data with one DV (correct malignancy detected percentages), and two IV, occupation (Symptomatic Radiologists and Screening Radiologists) and type of test set, split by PERFORMS set (SA2001, SA2002 and SA2003) and PERFORMS half case set. Results showed that there was a significant difference of PERFORMS year [F(2,120) = 31.443, p< .001] and post hoc Student Newman Keuls (SNK) tests revealed that the percentage of malignancies detected was significantly higher for 2002 than for 2003 or 2001 with 2003 being the least well performed of the three years (SA2001 mean = 80.03%, SA2002 mean = 87.54%, SA2003 mean = 93.17%).

Again, this tallies with our previous research which has examined the difficulty of particular case sets. As shown in Figure 3, there was no main effect of occupation or of year half (p=.n.s.).

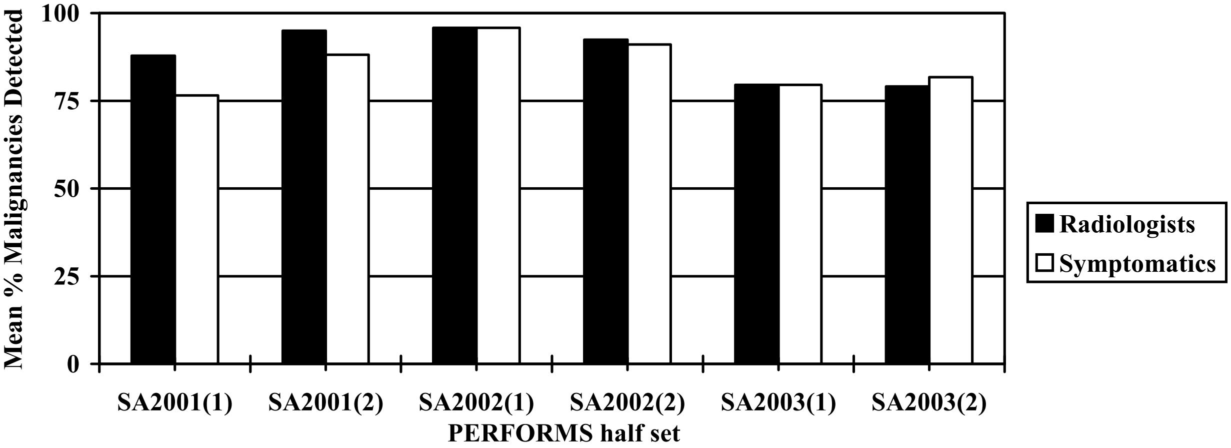

Figure 3: Percentage of Malignancies detected by Occupational Group

There was, however, a significant interaction found for PERFORMS set and PERFORMS half [F(2,120) = 5.24, p<.05], whereby participants scored higher in the first half of 2001 (p<.05) and the second half of 2002 (p<.05) but performed equally well, regardless of PERFORMS half set in SA2003, see Table 1.

Table1: Mean Percentage of Malignancies detected by Occupational Group, PERFORMS year and PERFORMS half set

	PERFORMS year by PERFORMS half set											
	2001 (1st half) n=12		**2001 (2nd half) n=16**		**2002 (1st half) n=18**		**2002 (2nd half) n=34**		**2003 (1st half) n=24**		**2003 (2nd half) n=28**	
Screening Radiologists ***Total n=66***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
CR	87.87	*5.07*	88.21	*5.59*	93.20	*5.39*	89.10	*9.63*	76.57	*12.3*	78.57	*12.9*
MAL	87.87	*7.95*	95.00	*6.54*	95.83	*4.65*	92.47	*8.10*	79.58	*9.15*	79.14	*12.3*
Symptomatic Radiologists ***Total n=66***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
CR	82.29	*8.76*	83.12	*8.93*	93.39	*7.65*	86.01	*9.07*	78.28	*10.10*	82.23	*7.67*
MAL	76.51	*10.91*	88.12	*8.42*	95.83	*5.89*	91.05	*7.42*	79.58	*9.64*	81.71	*8.41*

There was also a significant interaction between PERFORMS set and occupation [F(2,120) = 3.38, p<.05]. Post hoc SNK tests revealed that for Screening Radiologists the later PERFORMS set (SA2003) was significantly (p<.05) less well performed than the previous two. For Symptomatics, the pattern of results was quite different, with the scores on SA2003 and SA2001 higher than SA2002 (p< .05).

Sensitivity measures revealed significant within group differences over the three PERFORMS sets for Correct Recall scores. When the measure was refined to 'true' TP and TN scores however these differences revealed significantly different patterns of results for each occupational group, whereby the later PERFORMS sets were most problematic for the Screening Radiologists. There were no differences however, between the later (2003) and the first (2001) PERFORMS case sets for the Symptomatics, whose performance was significantly higher for 2002 (p<.05).

3.3 Specificity Measures – Correct Return to Screen (CS)

Specificity was measured by 'Correct Return to Screen' - the percentage of times, according to both pathology and National Radiological Opinion that each case was correctly assigned the classification of a 'normal' case and the case was 'returned' to normal screening practice (which in reality means that the woman is not screened again for a further three years). The 2x2x3 uni-variate ANOVA performed on these data with one DV (Correct Return to Screen percentages), and two IV, occupation (Symptomatic Radiologists and Screening Radiologists) and type of test set, split by PERFORMS case set (SA2001,SA2002 & SA2003) and PERFORMS half set showed no significant main effects. There was also no interaction between any of the factors. Therefore, the Symptomatics' specificity was similar to that of the Screening Radiologists as shown in Figure 4.

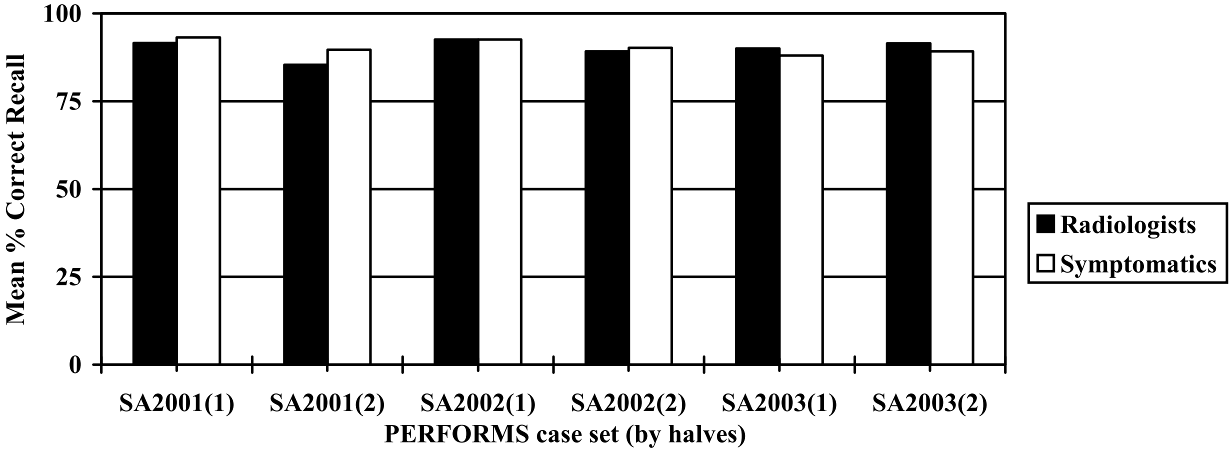

Figure 4: Correct Return to Screen Percentages

3.4 Differences in mean values – a comparison of specificity and sensitivity measures over the film sets.

In order to examine those differences in trends of the PERFORMS results over the case sets more closely we took the difference between the Screening Radiologists' and Symptomatics' (this is simply the Screening Radiologists' score minus Symptomatics' score). This allowed us to build up a profile of the Symptomatics' pattern of results as directly compared to the performance of a matched group of Screening Radiologists over the three PERFORMS sets, as well as allowing the simultaneous comparison between specificity and sensitivity.

Figure 5 shows the results for group differences for sensitivity (CR) and specificity measures (CS). Where the score is negative this indicates that the Symptomatic group scored below the Screening Radiologists, whereas where the score is positive this indicates where the Symptomatics' score is above. Figure 5 shows a clear increase in the performance of the Symptomatics' CR scores over the three PERFORMS case sets PERFORMS, however a decline (albeit less sharp) of the CS performance is also evident. A uni-variate ANOVA revealed that there were no significant differences apparent for either measure (p=n.s). This supports the previous findings and indicates that although there are group differences evident in the pattern of results, these differences did not reach significance.

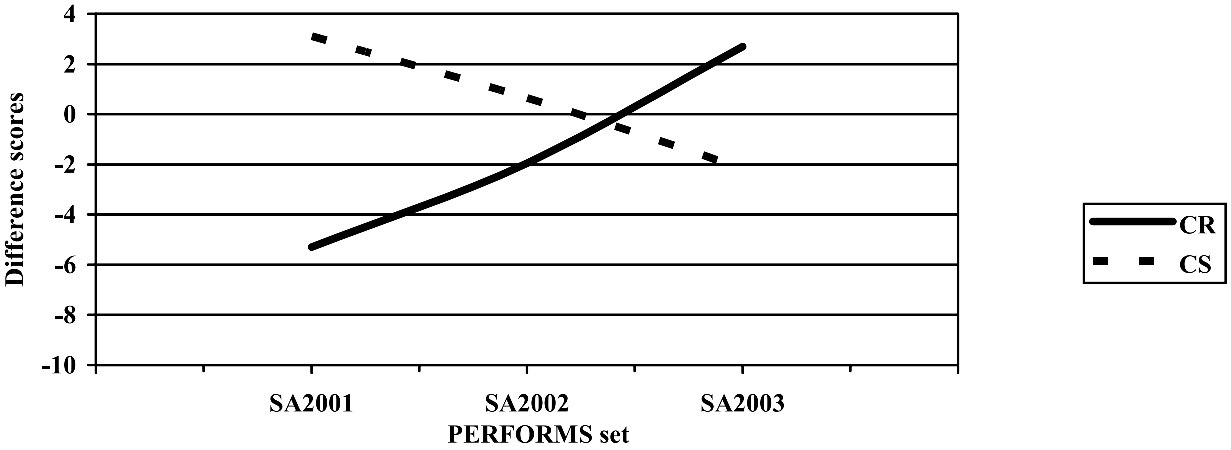

Figure 5: Radiologists/Symptomatic score differences for sensitivity (CR) and specificity (CS)

However, as a significant interaction was found between occupational group and PERFORMS set for Malignancies Detected, we also looked at the comparison with specificity difference scores for this purer measure of sensitivity. Again there is a double dissociation apparent in the results as shown in Figure 6. A univariate ANOVA revealed that there was a significant main effect of PERFORMS set (p<.05) and post hoc tests revealed that there were significant differences between case set SA2001 and SA2003 difference scores (p<.05), although no differences were found for either set compared with SA2002. Therefore, Symptomatic Radiologists' performance, when calculated as a function of a difference between Screening Radiologists' 'baseline' score and the Symptomatics' score, shows significantly different patterns of performance for cancer detection. Comparatively, Symptomatic Radiologists' performance increases over the case sets.

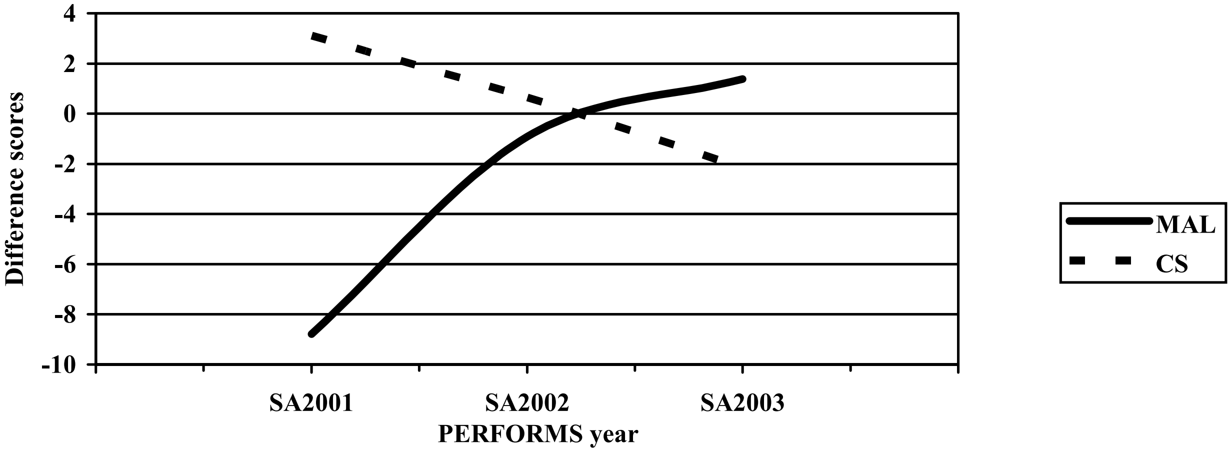

Figure 6: Radiologists/Symptomatic score differences for Malignancies detected (MAL) and Correct Return to Screen (CS)

Although there were no significant differences over the case sets for specificity measures there was a significant negative correlation between malignancy detected difference scores and correct return to screen differences scores (r= -.434, n=66, p<.0005, two-tailed). Similar results were also found for CR scores. This suggests that Symptomatics show a tendency to over-read these PERFORMS case sets whereas Screening Radiologists show the opposite tendency to under-read.

Looking at the data pertaining to participant matches however, it can be observed that although the volume of cases read per week stays relatively stable over time, with a mean range of just under 100 and just over 175 cases per week for all three PERFORMS sets, not surprisingly the years of experience in mammographic reading increases over time with a mean range of 1.5 to 8.3 years. An ANOVA revealled that these differences are significant over the three PERFORMS sets [F(2,105) = 7.63, p<.001]. SNK post hoc analysis shows that years of experience was significantly higher for 2002 and 2003 compared with 2001 (2001 mean years =3.41, 2002 mean years=5.86, 2003 mean years=7.61). This may suggest that as years of real life experience in interpreting mammographic cases increases then the reading style of the two groups becomes markedly different which may be due to the real life differences in clinical practice.

4. DISCUSSION

In this study we compared the performance of Screening and Symptomatic Radiologists on an annual self assessment scheme in order to determine if differences in real-life practice can affect overall skill level. Symptomatic Radiologists generally read a far lower volume of cases per week compared with Screening Radiologists and so this study also explored the consequences (if any) of reading a lower volume of cases per annum.

The data were analysed firstly in terms of sensitivity measures. For Correct Recall, overall both groups performed equally well at cancer detection. For three rounds of the PERFORMS scheme, for a sample that was matched for real life factors, the Symptomatic Radiologists performed as well as the Screening Radiologists. Both groups found the case set SA2003 the most difficult of the three. However, when the data were analysed in terms of TP for pathology, looking at Malignancies Detected, although there was no main effect of occupation there was a significant interaction between PERFORMS case set and occupation. Post hoc tests revealed significant within group differences for the particular pattern of cancer detection over the three PERFORMS rounds. Symptomatics performed better over the later PERFORMS case set compared to the Screening Radiologists.

For specificity measures the groups also performed equally well, with the Symptomatics detecting as many normal cases overall as compared to the Screening Radiologists.

Overall, the performance of the Symptomatic Radiologists, when matched with Screening Radiologists (so as to minimise variance due to volume of cases read and years of mammographic experience), proved to be equivocal for both sensitivity and specificity measures. The groups were matched as well as possible on the volume of cases read. However, the PERFORMS' participant questionnaire was designed primarily for Screening Radiologists and so the lowest number of cases per week that participants could select was 'below 100 cases per week'. This number was fixed at 100 because this is the lowest number a Screening Radiologist should read to keep their annual volume of cases above the recommended 5,000 cases). Consequently we were not able to take into account how much lower than 100 cases the Symptomatic Radiologists may actually read. Further investigation is required to determine this, which may be fewer than the data here suggests.

We are therefore unable to comment accurately on the effect of a lower volume of cases. However, we did find a significant difference in the pattern of results for Malignancies Detected when we looked at mean group difference scores (Symptomatic Radiologists' score minus Screening Radiologists' score) over the three rounds, where Symptomatics' sensitivity score increased over the three case sets. Initially their score was below that of the Screening Radiologists but by SA2003 their scores were higher. There was also a difference in the mean years of experience for these three sets which rose from 3.4 years in 2001 to 7.61 years in 2003. Although the mean was similar for the Screening Radiologists, their pattern of cancer detection was very different for those three rounds. For specificity scores there was an opposite trend for the Symptomatic Radiologists (although this did not reach statistical significance) as years of experience increased across the PERFORMS sets, specificity showed a slight decrease. This may be due to differences in their real life practice where Screening Radiologists learn to identify suspicious abnormal appearances from a large cluster of normal cases but Symptomatic Radiologists identify from a large group of abnormal cases where an abnormal appearance is suspicious or normal. Therefore, Symptomatics, examining what could be described as a potentially larger concentration of malignant cases per week, have a higher sensitivity for those malignant cases. The data suggest that over time Symptomatic Radiologists over-read (as their specificity subsequently goes down) and Screening Radiologists (comparatively) under-read (as their specificity goes up).

These data support the findings of Barlow et al. (9) who report the increased sensitivity of symptomatic diagnositic practice. The fact that both groups exhibit a different 'style' of reading which may develop over time has important consequences for the PERFORMS as it may necessitate the scheme to tailor case sets (and other specific training sets that we also provide) to the specific requirements of participants. For instance, it may be that Symptomatic Radiologists might benefit from a case set with a higher proportion of difficult normal cases. Further work on the exact nature of a

Symptomatic reading style is demanded with a detailed analysis of the types of features and classifications misinterpreted.

In summary, the data demonstrate that whilst both groups perform well at detecting early signs of cancer there are subtle differences in their underlying skills which we attribute to their expertise brought about by the types of cases which they routinely read.

5. CONCLUSIONS

Detecting early signs of breast cancer correctly is a difficult task which is aided by Screening Radiologists interpreting a large number of normal cases per year. Symptomatic Radiologists demonstrate a subtly different style of image interpretation.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Patnick, J. (Ed.). Annual review 2003: Saving women for 15 years, NHS Cancer Screening Programmes, 2003, Citigate Communications, London
2. 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 perception and performance. E.A. Krupinski, (Ed.). Proceedings of SPIE Vol. 3663, 157-168.
3. Esserman, L., Cowley, H., Eberle, C., Kirkpatrick, A., Chang, S., Berbaum, K., & Gale, A.G.: Improving the Accuracy of Mammography: Volume and Outcome Relationships. Journal of the National Cancer Institute, 2002,Vol. 94, No. 5, 369-375
4. Gale, A.G. & Walker, G.E. Design for performance: quality assessment in a national breast screening programme. In E. Lovesay (Ed.) Ergonomics - design for performance 1991, Taylor & Francis, London
5. Cowley, H. and Gale, A.G.. Minimising human error in the detection of breast cancer. In: S. A. Robertson (Ed). Contemporary Ergonomics 1996, Taylor and Francis, London.
6. Gale, A.G.: 2003 PERFORMS – a self assessment scheme for radiologists in breast screening. In Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills, 6(3), 148-152
7. Goodard, C.C., Gilbert, F.J., Needham, G. and Deans, H.E.: Routine receiver operating charateritic analysi in mammography as a measure of radiologists' performance. 1998, The British Journal of Radiology, 1998, Vol.71, 1012-1017
8. Mills, P.,Foord, K. and Trevethick, P.:Clinical audit and standard setting for symptomatic breast imagines in South Thames region. Clinical Radiology, 1997, Jan, Vol. 52(1), 55-58.
9. Barlow, W.E., Lehman, C.D., Zeng, Y., Ballard-Barbash, R., Yankaskas, B.C., Cutter, G.R., Carney, P.A., Geller, B.M., Rosenberg, R., Kerlikowske, Weaver, D.L., Taplin, S.H.: Performance of Diagnostic Mammography for Women With Signs or Symptoms of Breast Cancer, 2002, Journal of the National Cancer Institute, Vol. 94, No. 15, August 7, 2002.
10. 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 and performance. Miguel P. Eckstein & D.P. Chakraborty (Ed.). Proceedings of SPIE Vol. 5372, 399-406.