Mammographic feature type and reader variability by occupation –an ROC study

Hazel J. Scott*, Alastair G. Gale
Applied Vision Research Centre, Loughborough University, Loughborough UK, LE11 3TU

ABSTRACT

Previous work has outlined that certain mammographic appearances feature more prominently in reader's false negative responses on a self-assessment scheme. Bi-annually 600 breast-screening film-readers complete at least one round of the Personal Performance in Mammographic Screening (PERFORMS) self-assessment scheme in the UK. The main occupational groups in UK Breast Screening can be categorised thus, Radiologist, Technologists and Symptomatics. Previous work has shown that these groups can vary in their reading 'style' and accuracy on self-assessed cases. These groups could be said to contain individuals each with (arguably) pronounced differences in their real life reading experience, symptomatic readers routinely read a large number of cases with abnormal appearances and Technologists (specially trained to read films) do not have the same medical background as breast-screening Radiologists. We aimed to examine overall (national) and group (occupational) differences in terms of ROC analysis on those mammographic cases with different mammographic appearance (feature type). Several main feature types were identified namely; Well Defined Mass (WDM), Ill Defined Mass (IDM), Spiculate Mass (SPIC), Architectural Distortions (AD), Asymmetry (ASYM) and Calcification (CALC). Results are discussed in light of differences in real-life practice for each of the occupational groups and how this may impact on accuracy over certain mammographic appearances.

Keywords: Observer Performance Evaluation, Image Perception, PERFORMS, Breast Screening, Mammographic Feature, Occupation

1. INTRODUCTION

Breast-screening in the UK is no longer dominated by the Consultant Radiologist. The NHS Breast Screening Programme (NHSBSP) promotes the benefits of a multi-disciplinary four-tier (replacing the old two-tier system) team as promoted by the NHS cancer plan a decade ago (Department of Health, 2000). A range of film-readers now populate the UK NHS Breast Screening Programme (NHSBSP), the House of Commons Committee of Public Accounts reported, as early as 2006, that multi-disciplinary teams were now 'well embedded' in the NHS infrastructure. The number of personnel, particularly Advanced Practitioners (Technologists trained in mammographic interpretation) have increased two-fold.

All breast-screening film-readers on the NHSBSP are obliged to sit a bi-annual self-assessment weighted with difficult examples from recent breast screening (PERFORMS- Personal Performance in Mammographic Screening). Circa 90% of all breast screening mammographers in the UK choose to complete the PERFORMS self-assessment task (Gale and Walker, 1991). The scheme, which has been running since 1991 and commissioned by the Royal College of Radiologists, gives mammographers the chance to examine their radiological skill confidentially, biannually and with both immediate and detailed feedback. PERFORMS is not real life screening but has been reported to be a good indicator of real-life practices (Scott, Evans, Gale, Murphy & Reed, 2009, Cowley, Gale and Wilson, 1996 & Cowley and Gale, 1999).

Previous studies, regarding Technologists verses Radiologists on the PERFORMS self-assessment scheme has shown that occupational group is not significantly related to accuracy when those groups are matched on real life factors such

*h.scott@lboro.ac.uk phone, 0044 1509635733; fax, 0044 1509635736; http://www.lboro.ac.uk/research/applied-vision/people/h-scott.htm

Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment,
edited by David J. Manning, Craig K. Abbey, Proc. of SPIE Vol. 7627, 762715

as years of screening experience and volume of cases read (Scott, Gale & Griffiths, 2006), however overall differences in performance, as measured by percentage of false negative error have been identified (Scott and Gale, 2006). Non-Radiologists (Technologists), although they perform a different clinical role in the multi-disciplinary team, perform equally well on self-assessment (Scott, Gale and Wooding, 2004; Scott & Gale, 2006).
Screening Radiologists and Technologists are required to achieve a target, 5,000 screening cases yearly, and the volume of real life cases is self-reported to be far higher (Scott & Gale, 2007). Symptomatic Radiologists are not obliged to read such a high number. Screening film-readers may read a larger number of normal cases (where the incidence of abnormality is circa 7 per 1,000), however, Symptomatic Readers, although they read far less overall cases, may actually examine a greater percentage (per yearly volume) of abnormal appearances. Symptomatic Readers will report on referred patients presenting with symptoms (e.g. palpable lump, discharge, pain) which may also have the advantage of greater knowledge of physical location of a potential abnormality in many instances. Our research has shown that these differences in clinical practice do not generally manifest themselves when the groups are compared on self-assessment (Scott and Gale, 2006) as both groups perform well on cancer detection and specificity measures, although, akin to other studies (Barlow et al; 2002), over three consecutive PERFORMS years, Symptomatic Radiologists had a greater tendency to over-read compared to screening-readers. Occupational reader variability, therefore, may account for differing reading behaviours.

There are however a larger range of occupational groups who routinely complete PERFORMS self-assessment and who work as part of the multidisciplinary team, these include specialists such as Breast Surgeons, Breast Physicians, Breast Clinicians, Associate Specialists etc, whose performance we have yet to examine comparative to Radiologists, Technologists and Symptomatic Readers, this we aim to address in the current study.

Studies have identified the possibility that certain readers may exhibit a particular approach to mammographic reading (dubbed 'reading-style'), which reveals itself in different individuals having different strengths. One reader may be better at identifying certain feature types (subtle calcification for example) where another picks up asymmetry more easily. Studies have aimed to identify these sensitivities with a view to possible training needs (Scott and Gale, 2006) but results could also be interpreted to support the current practice of double-reading and the use of multi-disciplinary teams whereby differing strengths are an asset.

With the widespread rollout of digital mammography in the UK, technology has to be adaptable to the individual or perhaps group needs of all film-readers and an effective reader/technology interface could be most efficient if geared towards these possible differences.

The purpose of this study is to examine if certain occupational groups (non-matched, and a matched-size sample) are sensitive to error for certain mammographic appearances. We aimed to expand our previous research (Scott & Gale, 2006) on reading style (where it was shown that symptomatic readers show subtle differences compared to Radiologists in their sensitivity over a three year period) with a larger cohort of occupational groups (with the inclusion of another occupational group encompassing a range of medical specialists) in order to examine not only overall accuracy but overall performance, using ROC analysis for all prominent mammographic feature types.

Work is presented with a view to understanding both the uk national strengths and weakness in radiological skill as well as the possible identification of radiological style, perhaps mediated by clinical practice.

2. METHODOLOGY

Results, from the most recently completed PERFORMS case set (SA09) from a cohort of over 650 UK breast-screening Radiologists, Technologists (specially trained in mammographic film-reading), Symptomatic Readers and Other Medical Specialists were analysed by occupation and feature type in order to identify reader variability.

PERFORMS responses, taken from each individual's decisions about feature and classification which, after being entered into a tablet PC, were subsequently recorded on the PERFORMS database. Each individual participant received confidential feedback (via the tablet PC) on all performance measures and radiological feature/lesion type compared with pathology and a radiological 'gold standard'. All participants received additional 'national' feedback in the form of regional/national reports, on completion of a further 60 cases (120 in total). These reports detailed their performance

measures with that of their peers against a national radiological opinion (based on the majority decisions of all film-readers and pathology).

The initial radiological or 'gold' standard was gleaned from the majority decisions of an experienced panel of over five Radiologists who provided information on feature type and malignancy likelihood (on a 6 point scale ranging from normal =1 to malignant=6) in accordance with original case pathology (or 'truth'). For the purposes of this study we selected results from those cases with particular mammographic features (well defined, ill defined and spiculate mass, architectural distortions, asymmetries and calcifications) for between group analysis and based all results on pathology as a measure of absolute truth rather than utilising the radiological opinion as a measure of suspicion. As such all benign cases were removed from the analysis, as these cases have differing 'truths' one according to pathology (normal) and another according to radiological rating (on how suspicious the case is for recall).

In the PERFORMS self-assessment task,whereby the process of reporting is designed to mimic real life as far as possible, for all feature/lesion types participants' rate case classification on a 6 point scale. As such, all cases with double-lesions were omitted from this data set (n=2) as case classification for a case with one lesion could be said to be akin to lesion classification. Those cases with double lesions were problematic in that one cannot be certain that case classification accounts for the malignancy rating of both lesions.

ROC curves and Az values for all PERFORMS results were produced for feature type using ROCKIT, and PLOTROC.

3. RESULTS

3.1 Reader variation – overall analysis.

Data concerning difficult recent screening cases were examined for the most recent PERFORMS scheme (SA09) n=60. Results were analysed by reader occupation, these were as follows: Radiologists, Technologists, Symptomatic Readers and Other Medical Specialists. Due to the fact that these groups are not evenly present within the NHSBSP, reader group numbers were not equal, with the Radiologists and Technologists forming the largest groups- see Table 1. for participant characteristics.

Table 1: Participant Characteristics

Occupational Group	Number of Participants
Radiologist	***n***= 337
Technologist	***n***= 222
Other Medical Specalists	***n***= 52
Symptomatic Readers	***n***= 42
Total	***n***= 653

Initially, the results were analysed in terms of overall performance for each group over this (circa) six month period. Group performance was also measured by mammographic feature types on the PERFORMS scheme namely; Well Defined Masses, Ill Defined Masses, Spiculate Masses, Asymmetries and Suspicious Calcifications. Results on these cases, each with a single feature type one side and an adjacent normal side provided the case sample for this study cases n=36.

Az performance measures were obtained using ROCKIT and subsequently plotted by reader group using PLOTROC. A one-way ANOVA with one IV (occupational group) and one DV (Az scores) showed no significant group differences overall for Az scores [F(3,23)=.363, p=n.s.] indicating that group performance was equivocal for this scheme. Figure 1. shows the mean performance for all groups showing a descriptive trend for Technologists and Radiologists to slightly outperform the other two groups.

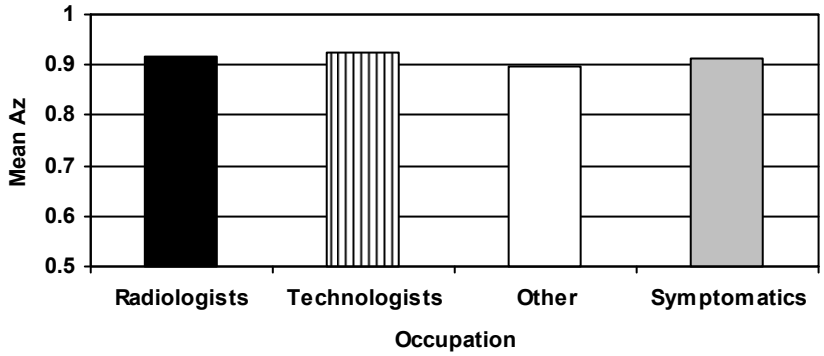

Figure 1. Group Az scores by Occupation

Participants' performance was also examined by feature type and a one-way ANOVA with one IV (feature type) and one DV (Az scores) showed a significant main effect of feature types [$F(5,23)=22.167$, $p<.0001$]. Post-hoc tests (Student Newman-keuls) revealed that cases with calcifications were less well identified than all other feature types ($p<.05$) and that Architectural Distortion and IDM were less well reported ($p<.05$) than Spiculate Masses, Well Defined Masses and Asymmetries.

ROC curves for feature type by group are shown in Figures 2a-2f. Although there were no significant group differences in this instance, several descriptive trends can be observed. For all feature types Radiologists and Technologists performance curves are very similar, with both groups showing greater AUC than the Symptomatic Readers - notably for WDM (Fig. 2a) , Spiculate Masses (Fig. 2c) and Calcification (Fig. 2f). Other Medical Specialists generally perform well but performance dips below all other groups (but not significantly so) for Spiculate Masses.

3.2 Reader variation – matched-size groups

In order to combat the effect of group size in the previous analysis (and possible Type II error) all participant groups were matched on sample size as well as inclusion on NBSS data systems (meaning that all readers' data was counted in their real life audit results). In addition, all participants had completed at least one half of the PERFORMS latest round. Participants who fulfilled the inclusion criteria were randomly selected and allocated into equal groups for analysis ($n=42$) in accordance with the smallest group in the previous overall analysis (Symptomatic Readers).

A one-way ANOVA with one IV (occupational group) and one DV (Az scores) showed no significant group differences overall for Az scores [$F(3,23)=.296$, $p=$n.s.].

Feature analysis revealed similar results, a one-way ANOVA revealed significant differences in feature types [$F(5,23)=10.974$, $p<.0001$.]. Although, post-hoc showed a different pattern of results, with Architectural Distortions and Ill Defined Masses less well reported than Spiculate Masses, Asymmetries and Calcifications ($p<.05$).

A further analysis was carried out using ROCKIT's bivariate chi-square test, on all occupational group/feature type comparisons (see Figures 3a-3f). Results revealed significant differences between Radiologists and Other Medical Specialists for Spiculate masses whereby Radiologists performance was significantly higher for this feature type. Bivariate chi-square tests were approaching significance ($p=.06$) for Technologists/Other Medical Specialist comparisons (with Technologists showing higher Az scores) - see Figure. 3c. For this particular sample, there was a surprising difference in the performance of Other Medical Specialists for calcification where they showed significantly larger AUC in comparisons with Radiologists ($p<.01$), Technologists ($p<.01$) and Symptomatic Readers ($p<.05$).
Overall, results generally show no significant differences between occupational groups, although when analysed by feature type several possible trends are discernable.

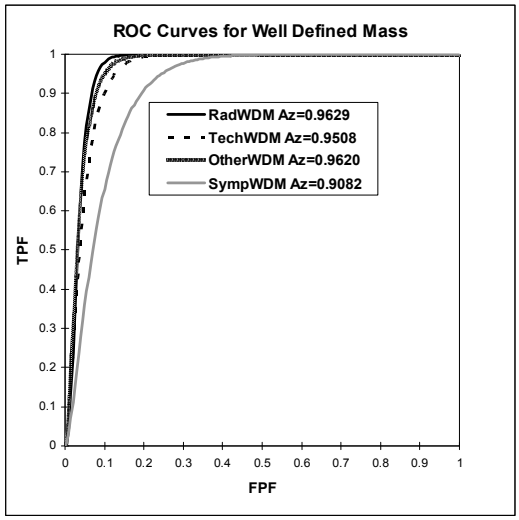

Fig. 2a. WDM ROC curves by Occupation

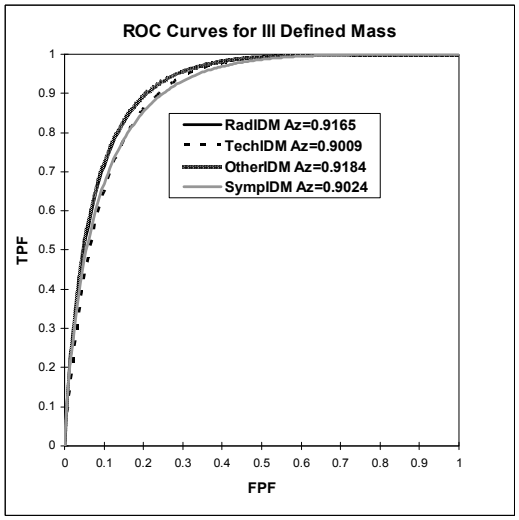

Fig. 2b. IDM ROC curves by Occupation

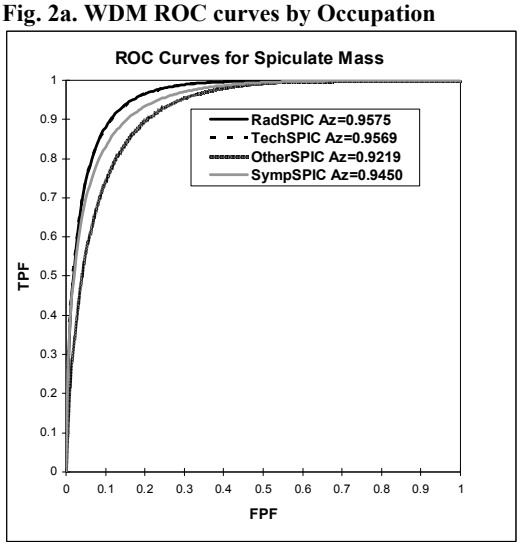

Fig. 2c. Spiculate Mass ROC curves by Occupation

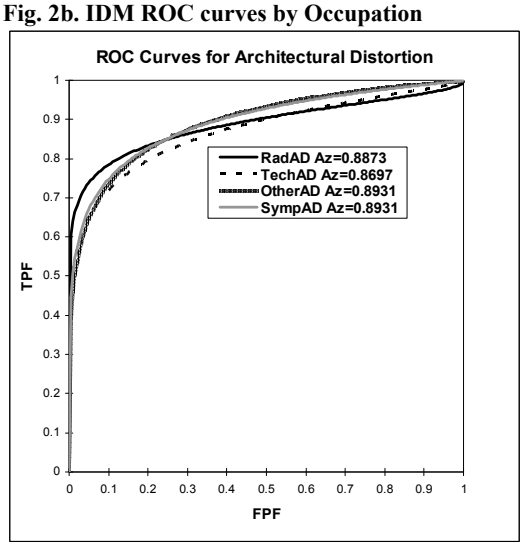

Fig. 2d. Architectural Distortion ROC curves by Occupation

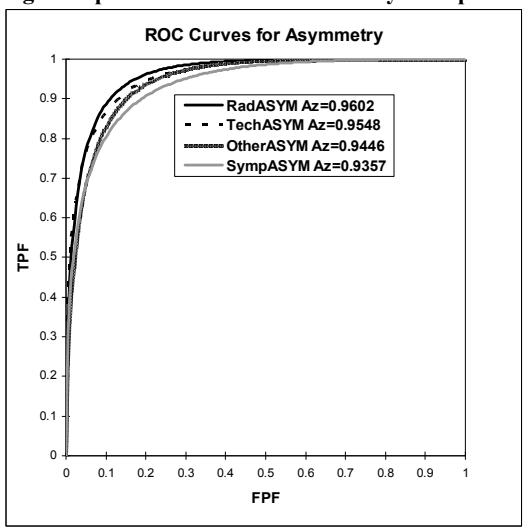

Fig. 2e. Asymmetry ROC curves by Occupation

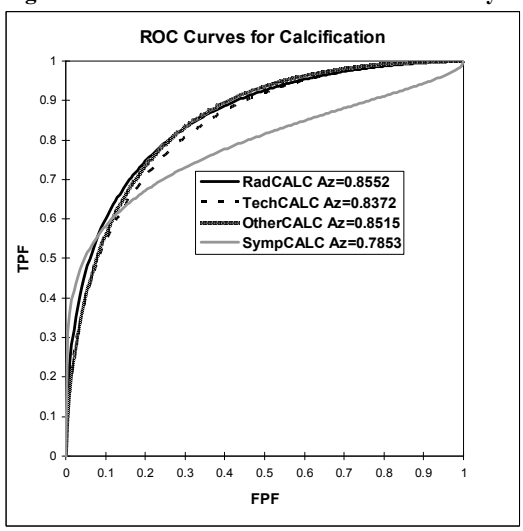

Fig. 2f. Calcification ROC curves by Occupation

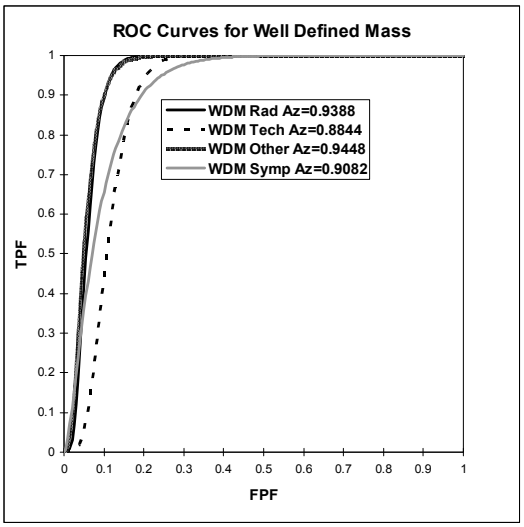

Fig. 3a. WDM ROC curves by Occupation

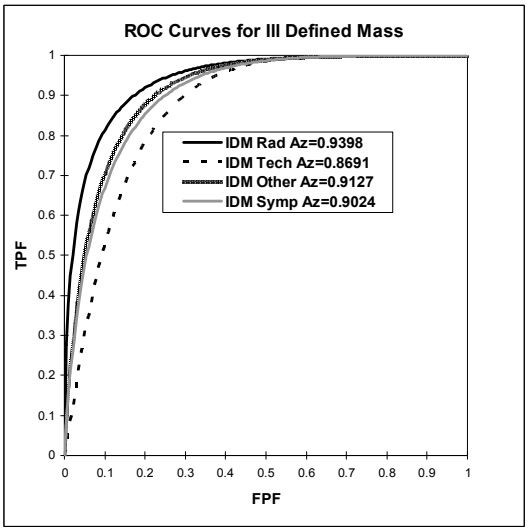

Fig. 3b. IDM ROC curves by Occupation

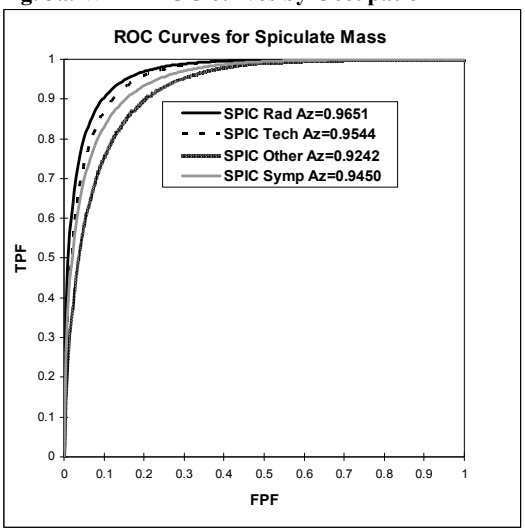

Fig. 3c. Spiculate Mass ROC curves by Occupation

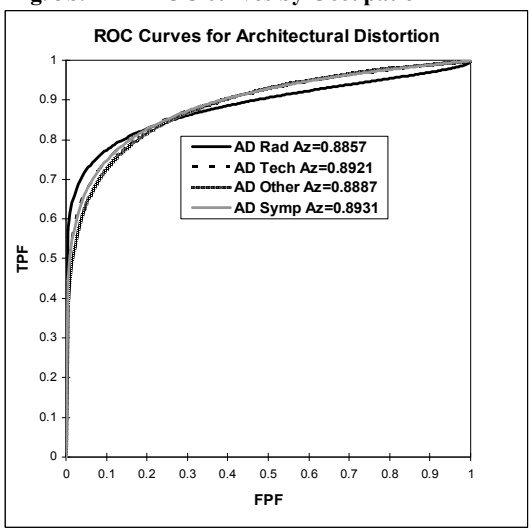

Fig. 3d. Architectural Distortion ROC curves by Occupation

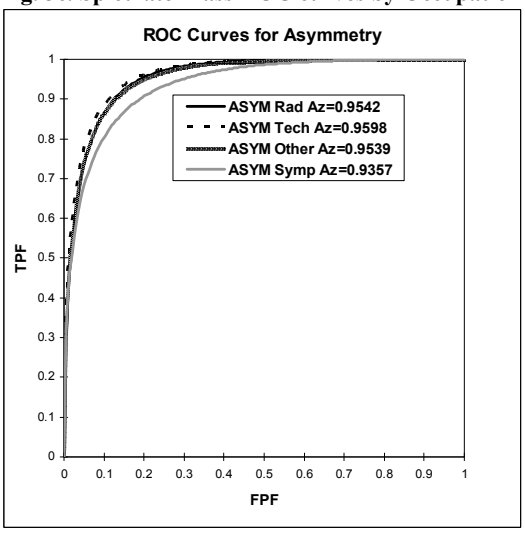

Fig. 3e. Asymmetry ROC curves by Occupation

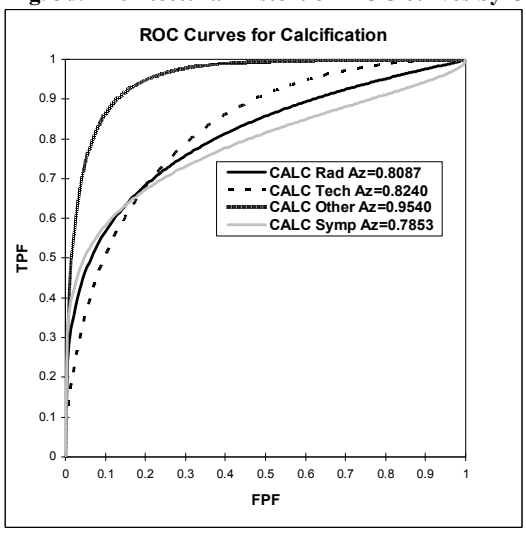

Fig. 3f. Calcification ROC curves by Occupation

4. DISCUSSION

Results from the ROC analysis reveal feature detection to be relatively robust across occupational groups. This demonstrates what we have previously observed, that all medical personnel, Technologists and Symptomatic Radiologists show excellent overall performance.

For the overall analysis, using data from the entire population, occupational groups showed similar performance, there were significant differences in performance for feature types, as for most PERFORMS sets certain exemplars are more difficult than others and this can fluctuate across schemes (Scott & Gale, 2006; Scott, Gale & Hill,2008).

Occupational group differences across feature type (although non-significant) showed some notable trends. Radiologists' performance overall was generally the highest AUC. Symptomatic Readers' showed lower AUC notably (but not statistically) for several feature types including well defined masses, spiculate masses and calcification. These results may be due to the differences in clinical practice, a lower case volume or perhaps reading in a way not akin to normal reading practice (where some knowledge of lesion location is often available). This may also be due to recent change in symptomatic practice whereby a growing population of women are being referred with benign features (Michell, 2006). This current trend may affect Symptomatic Reader sensitivity levels, as even a slight tendency to over-read (which we have observed in previous studies, Scott and Gale; 2006) has both human consequences (perhaps unnecessary follow-up) as well as the consequences of increased workload. Symptomatic Readers read, in general, a lower case volume which has been shown to be a predictor of case accuracy (Scott and Gale, 2007). Radiologists and Technologists (who read high volumes of cases in real life) did not show the same descriptive dip in AUC for any of the mammographic feature types.

As the identification of feature types, where some occupational groups were descriptively more sensitive than others ,could simply be due to the real-life difference in occupational group size, groups were matched on sample size. Significant differences were found for feature types but not for occupational group. Interestingly, post hoc tests revealed different feature types were performed better and other, least well- compared to the overall groups analysis. Although observer variations in the classification of feature types has been well reported (Simpson, Neilson & Kelly, 1995) what constitutes a difficult case in terms of mammographic descriptors is variable across PERFORMS sets, with feature types showing the highest FN alternating from year to year. These two analyses reveal that even changing reader numbers to create sub-samples will highlight the fluidity and perhaps difficulty of identifying which abnormalities are the more challenging ones and for whom.

Bi-variate chi-square paired comparisons did reveal significant differences between AUC for several feature types (although there were no significant overall differences so these results must be treated descriptively), firstly for spiculate masses, Radiologists' comparisons with Other Medical Specialists showed significant differences (and approaching significant differences for Technologists) whereby Other Medical Specialists were shown to have less sensitivity for this feature type than these groups. Contrastingly, for suspicious calcification, Other Medical Specialists showed significantly different results compared to all other groups, their Az scores were larger. They were significantly better at the detection of suspicious calcifications than all other occupational groups. This may be an indication of the range of skill sets apparent in a multi-disciplinary team. The Other Medical Specialist group includes Breast Surgeons and it has been argued by Viya and Dixon (2001) that combining Radiologist/Surgeon mammographic reader reports resulted in an increase in real life sensitivity (when Breast Surgeons' decisions are combined with that of Consultant Radiologists).

There were some methodological issues with the data set, which, in an attempt to keep an even number of normal vs abnormal cases (coupled with the exclusion of those cases with double-features) significantly reduced the 'cases' available for ROC analysis. The method of reporting PERFORMS necessarily must follow real life procedures and as such lesion rating was not recorded explicitly, ongoing analysis however aims to control for this and investigate these data using a JAFROC paradigm. Future analysis with a larger sample size –perhaps over several PERFORMS sets is warranted.

Results support the current NHSBSP practice of utilising multi-disciplinary teams (although they are costly in terms of time) as well as variations in double-reading. Results suggest that different occupational groups (when matched by

sample size) show fluctuations in sensitivity for feature types and that a mix of radiological reading styles may be advantageous to reporting accuracy.

5. CONCLUSIONS

It was concluded that although all occupational groups in this sample showed excellent overall performance there were some differences apparent in the detection of specific feature types mediated perhaps by real life practice.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Scott H.J., Gale A.G., & Wooding D.S.: Breast Screening Technologists: does real-life case volume affect performance? In: Image Perception, Observer Performance, and Technology Assessment, D.P. Chakraborty & M.P. Eckstein (eds.) Proceedings of SPIE Vol. 5372, 2004.
[2] Scott H.J., Gale A.G, Griffiths, C.E.: Breast screening radiographers and Radiologists: performance and confidence levels on the PERFORMS film sets. Symposium Mammographicum 2004 Breast Cancer Res 2004, 6(Suppl 1):P10
[3] 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.
[4] 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". in SPIE Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment, Berkman Sahiner; David J. Manning, Editors, 72631E
[5] Scott H.J., Gale A.G: Does mammographic practice affect film reading style – Breast Screening vs. Symptomatic Radiologists? In: Image Perception, Observer Performance, and Technology Assessment, D.P. Chakraborty & M.P. Eckstein (eds.) Proceedings of SPIE Vol. 5374 2006.
[6] NHS Cancer Plan, Department of Health, 2000.
[7] House of Commons Committee of Public Accounts, The NHS Cancer Plan: a progress report, January 2006.
[8] 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.
[9] 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.
[10]Scott H.J., Gale A.G.,:Breast screening: PERFORMS identifies advanced mammographic training needs. British Journal of Radiology, 79 (2006), S127-S133.
[11] Scott H.J. & Gale A.G.: How much is enough: factors affecting the optimal interpretation of breast screening mammograms, In Image Perception, Observer Performance, and Technology Assessment. Y Jiang and B Sahiner (Eds.) Proceedings of SPIE 2007.
[12] 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.
[13]Scott H.J., Gale A.G. & Hill S.: How are false negative cases perceived by mammographers? Which abnormalities are misinterpreted and which go undetected? Image Perception, Observer Performance, and Technology Assessment. D. Manning and B Sahiner (Eds.) Proceedings of SPIE 2008
[14] Michell, M.J.: Symptomatic breast clinics: the radiologist's perspective, Breast Cancer Research 2006, 8(Suppl 1):P6
[15] Vidya, R.& Dixon, J. M.: Should surgeons as well as Radiologists report mammograms in symptomatic patients? The Breast, Vol 10. Issue 2, 2001.