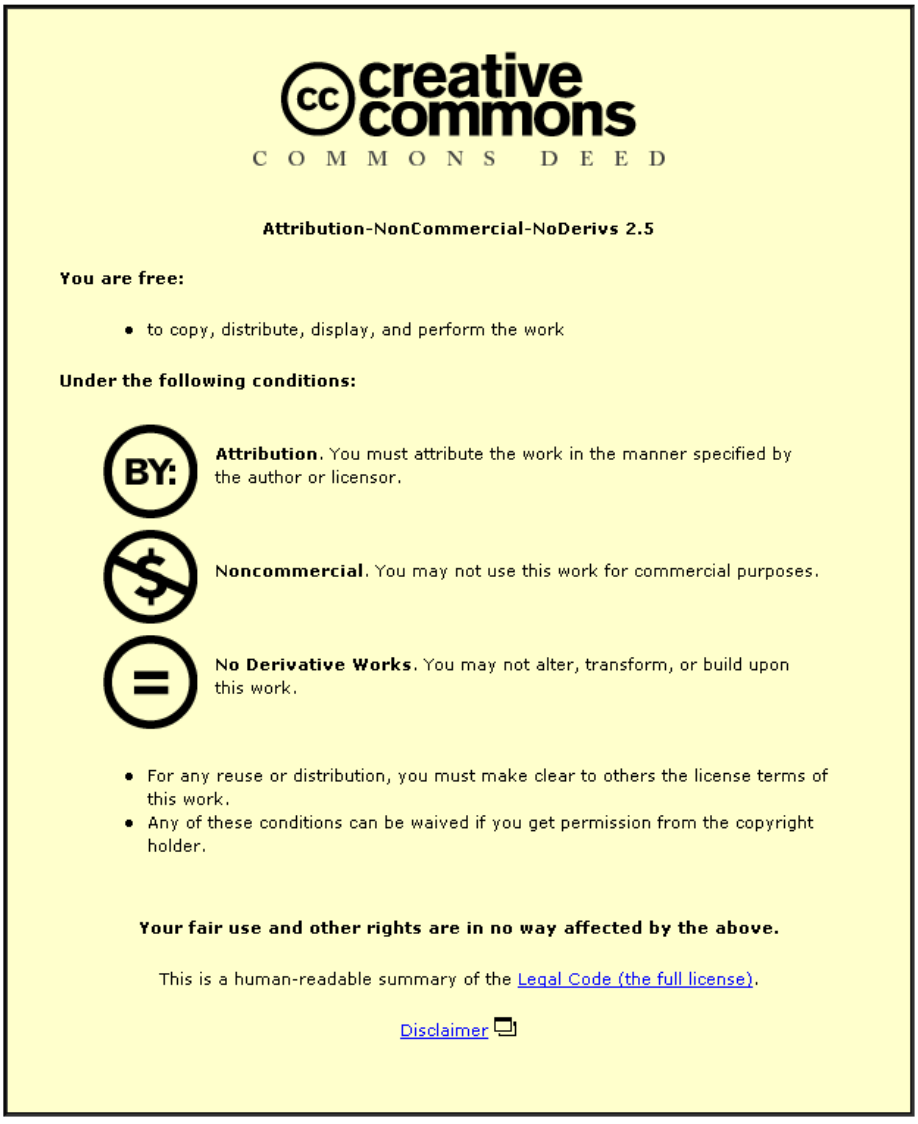

creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

How are False Negative Cases perceived by Mammographers? Which Abnormalities are misinterpreted and which go undetected?

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

ABSTRACT

A radiographic 'false negative' or a case which has been 'missed' can be categorised in terms of errors of search (where gaze does not fall upon the abnormality); detection (a perceptual error where the abnormality may be physically 'seen' but remains undetected) and misinterpretation (a perceptual error whereby an abnormality, although detected, is not deemed worthy of further assessment). This study aims to investigate perceptual errors in mammographic film-reading and will focus on the later of the two error types, namely errors of misinterpretation and errors of non-detection. Previous research has shown, on a self-assessment scheme of recent and difficult breast-screening cases, that certain feature types are susceptible to errors of misinterpretation and others to errors of non-detection. This self assessment scheme, 'PERFORMS' (Personal Performance in Mammographic Screening), is undertaken by the majority (at present over 90%) of breast-screening mammographers in the UK Breast Screening Programme. The scheme is completed bi-annually and confidentially and participants receive immediate and detailed feedback on their performance. Feedback from the scheme includes information detailing their false negative decisions including case classifications (benign or malignant), feature type (masses, calcification, asymmetries, architectural distortions and others) and case perception error (percentage of misinterpretation and percentage of non-detection). Results from a recent round of PERFORMS (n=506), revealed that certain feature types had significantly higher percentages of error overall (including architectural distortion and asymmetries), and that these feature types also showed significant differences for error type. Implications for real-life screening practice were explored using real-life self-reported data on years of screening experience.

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

1. INTRODUCTION

PERFORMS (Personal Performance in Mammographic Screening) a self-assessment scheme for mammographers is undertaken as an educational tool by film-readers involved in reading breast-screening films in the UK. The scheme has been running as a bi-annual exercise since its inception in 1991(c.f. Gale and Walker, 1991)[1]. It is both a free and anonymous exercise consisting of difficult screening cases and provides immediate and confidential feedback to all film readers on their respective performance based on a radiological "gold standard". Feedback from the scheme takes the form of information relating to true positives or mammographic 'sensitivity' as well as information on true negative decisions or 'specificity'. In addition detailed information on 'false negatives' is provided, fractionated by radiological feature type as well as error type (misinterpretations versus non-detections).

There are two kinds of error that are possible on this self-assessed scheme, a 'false positive' error, which can be described as a 'normal' or non-suspicious case which has been incorrectly recalled (in real life this would result in unnecessary and expensive follow-up). Conversely, a 'false negative' decision occurs when a participant records a case as not suspicious enough for recall - when that case is, in fact, a malignant or suspicious case (in real life this has the consequence of a cancer that has been missed). In this study we will be concerned only with those missed 'false negative' cases as being the more serious of the two errors.

PERFORMS cases that are suspicious enough for recall contain a range of mammographic features namely; well defined and ill defined masses, spiculate masses, calcification, asymmetries and architectural distortions. Previous work

* h.scott@lboro.ac.uk; phone, 0044 1509635733; fax, 0044 1509635736; www.appliedvision.org

Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment, edited by Berkman Sahiner, David J. Manning, Proc. of SPIE Vol. 6917, 691713, (2008)
1605-7422/08/$18 · doi: 10.1117/12.771137

(Savage, Gale, Pawley and Wilson, 1994)[2] has shown, from the UK National Interval Cancer Database, that several feature types are more commonly missed than others. In this example, asymmetry had the highest percentage of error, followed by ill defined masses, architectural distortions and calcifications. In justification of these figures, the study goes onto describe how certain features or radiological abnormalities (through Positive Predictive Value scores) are more likely to have malignant pathologies. Lone asymmetry and well defined masses are defined as having a low Positive Predictive Value (PPV) for malignancy and therefore even if the abnormality is perceived, it may be misinterpreted as non-suspicious. Architectural distortions and ill defined masses are described as having a medium PPV and therefore, if these features are missed, the error could be one of misinterpretation or of non detection. Those features with high PPV then are more likely to be non-detected, when missed, as in real life those kinds of abnormalities are likely to be malignant.

In a later paper (Cowley, Gale and Wilson, 1996)[3], written at the beginning of PERFORMS inception, the authors detail those features that (in the first two rounds of PERFORMS) were shown to be the most problematic. In those instances the features which were most commonly missed were architectural distortions and asymmetric densities. In addition, they found that specific feature types were sensitive to misinterpretation and others to non-detection. Architectural distortions and asymmetries were overwhelmingly undetected but spiculate masses and calcifications were largely misinterpreted. When comparing PERFORMS data with real-life interval cancer database, a strikingly similar pattern emerges – those features most commonly missed on both were ill defined masses, architectural distortions and asymmetric densities.

In a matched study (Scott and Gale, 2005)[4], looked at a smaller cohort (n=90) of PERFORMS participants (Radiologists and Technologists matched on real-life case volume and years of screening experience) over a more recent time span (2001 to 2003) and found that the pattern of feature types which showed the most radiological error differed compared to the earlier studies. In this later study, the features which proved to be the most difficult (in recallable cases) were asymmetries and ill defined masses, with architectural distortions as one of the *least most* problematic abnormalities. This difference in feature difficulty could possibly be attributable (even though the study was matched) to an overall increase in the years of reading experience of the participant group. In general, in 2005, there were a large number of film readers who had been reading far longer than the three-five years of the original studies. The mean years of reading experience in this study was over four years, with more than half the participant group reading for far longer.

It was aimed to update this body of research, with a larger cohort of participants, in order to investigate which of these trends are still apparent in a body of screening film-readers (who presently have an overall level of reading experience far greater than previously). In order to address the question of reading experience over time- as the initial two studies were carried out when breast screening in the UK had only been running for 3-5 years (it has now been running for over 18 years)- the effect of years of screening experience was also investigated.

In addition to completing the PERFORMS cases, a large percentage of film-readers also provide information, via questionnaire, on their regular mammographic reading practice, including their years of reading experience. In light of observations from previous research, this information was compared with 'false negative' performance on the most recent PERFORMS set.

Work is presented with a view to understanding perceptual errors on specific mammographic feature types (for the interpretation of a test set of breast screening mammograms) and how this may relate to normal screening practice.

2. METHODOLOGY

Results, from over 500 UK breast-screening Radiologists, Technologists (specially trained in mammographic film-reading) and other health professionals, for the most recent PERFORMS sets (SA07) were analysed by feature type (in terms of percentage of errors and error type). Specifically this study focused on six main error types, calcification, masses (well defined, ill defined and spiculate), architectural distortions and asymmetries and examined their percentage of error for instances of misinterpreted and non-detected feature type.

Although at present a voluntary activity, the majority of film-readers on the UK NHSBSP, completes PERFORMS. Breast Screening Units throughout the UK provided cases for PERFORMS, a potion of which were included in the final PERFORMS set (following peer and technical review).

An initial radiological standard was gleaned from the cases review of an experienced panel of 5 radiologists as well as, where relevant, with reference to case pathology. Latterly, a fairer 'gold standard' is implemented from the majority decisions of all participating film readers (over 500 in this instance) and a 'national radiological opinion' about each case is utilised (also in accordance with case pathology).

In completing PERFORMS, individuals entered their decisions about each feature and case classification, into a tablet PC, whereupon they receive detailed feedback (via the tablet PC) on all aspects of their performance (case by case) compared with the radiological standard.
Following the completion of the scheme when all participants have read all 120 cases, performance, was measured against the 'national opinion'. Individual reports were disseminated back to these participants (comparing their performance to the anonymous data of their peers). Specifically, individuals received detailed feedback on their number of 'Correct Recall' decisions (a measure of sensitivity), 'Correct Return to Screen' decisions (a measure of specificity), percentage of correct malignancy's detected as well as ROC measures such as d' and d' for pathology. In addition they received information on features missed in a detailed analysis of their false negative data.
For this recent set, a majority of individuals invited to complete a computerised self-report detailing their most common reading practice chose to do so.

2.1 Overall Levels of Film Reading Experience

This body of work aimed to expand on that of previous studies in this area with a cohort which is approximately twice as large and therefore with a greater range in experience level. Consequently, this participant group largely included the same participants (as in previous studies) with many more years of experience. Information on years of screening experience was gleaned from questionnaire data reported on the PERFORMS computer at the beginning of the (SA07) set. Not all participants chose to complete the questionnaire (n=410), see Figure 1. A wide range of experience levels was apparent; the largest groups were those reading for less than six years and those reading for more than 16.

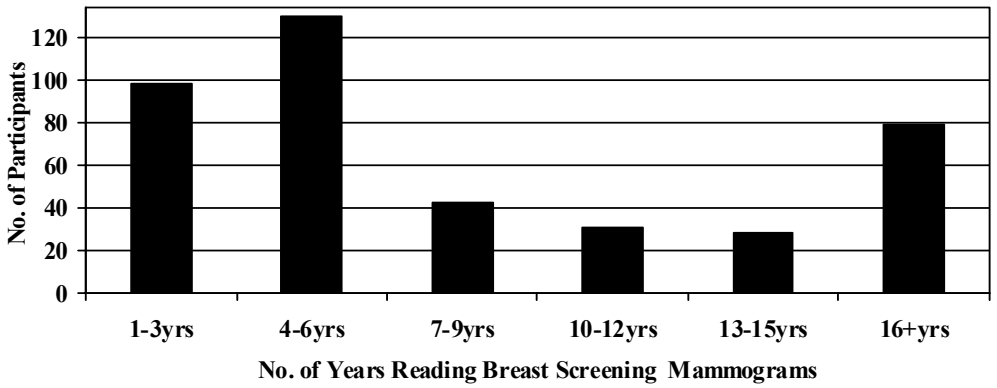

Figure 1: Years of Screening Experience

3. RESULTS

Inclusion criteria for this study were those participants (n=506) that completed the most recent round of PERFORMS (120 difficult cases). For this analysis, all false negative errors were calculated from the 'National Radiological Opinion' (the 'gold standard' gleaned from the opinion of all participating film readers) as well as from case pathology.

3.1 False Negative Errors

This study concentrated on those participant errors which were categorised as 'false negatives'. All cases that were deemed 'false negatives' were those cases where **both** Pathology and National Radiological Opinion agreed that the case was suspicious and should have been recalled. If the participant failed to do so, they were recorded as having made a 'false negative' classification for that instance (feature).
There were some instances where a case was 'missed' (or where a false negative error was made) as it being either an undetected (where no feature was detected and the case was misread as a normal case) or a misinterpreted case (in these instances a feature was detected but misinterpreted as not suspicious enough for recall). Percentage of missed error type was calculated for each of the case's four views (as a mammographic feature may appear on both the oblique and the CC views for example). Figure 2. indicates the mean overall 'false negative' error percentage. Also shown is the 'false negative' error by instances where the false negative was either misinterpreted or was undetected.

All false negative errors were pooled as a percentage of actual instances of suspicious features for the entire PERFORMS set (a percentage out of 120 cases). The mean number of false negatives on the PERFORMS set was relatively low and for all 120 cases this averaged at mean=10.65%. However, there was a significant difference of error type. A paired t-test for related samples showed that there were more false negatives errors for instances of non detection (mean=5.72) compared to misinterpretation (mean=4.93); $t(505)= -3.338$, $p<.01$.

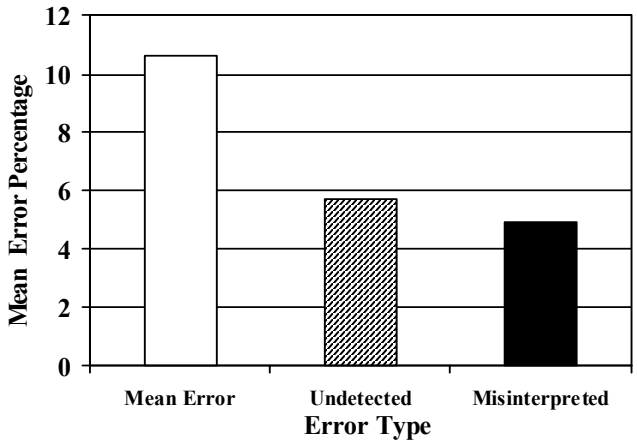

Figure 2: False Negatives by Error Type.

3.2 False Negatives by Case Classification

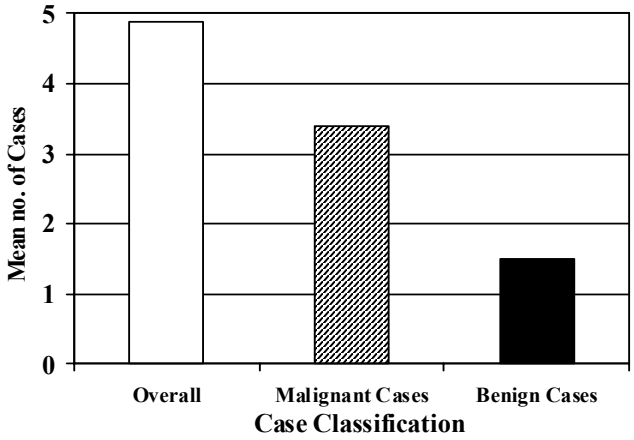

Figure 3: False Negatives by Case Classification.

False negatives were analysed as a function of case classification. For this analysis the number of cases was observed rather than mean percentage of feature instance. False negative cases can be of either benign or malignant pathology (they cannot be from normal cases as an error in a normal case would be classified as a 'false positive' error). There were significantly more errors in malignant cases owing to the higher percentage of malignant cases in the set compared to benign cases that required recall. This difference of case classification was significant - $t(505)=-21.96$ $p<.05$. Therefore, for the following analysis, the false negatives error type were, for the most part, those cases with a malignant pathology.

3.3 False Negatives by Feature Type

False negatives were analysed by mammographic feature type namely; well defined Masses (WDM), ill defined masses (IDM), spiculate masses (Spic), architectural distortion (AD), Asymmetry (Asym), and Calcification (Calc). A repeated measures, within-subjects, Analysis of Variants (ANOVA) with two IV (feature type and error type) and one DV (error percentage) revealed a main effect of feature type [$F(5,2525) = 70.58$, $p<.01$]. A priori Bonferroni Pairwise comparisons revealed significant differences ($p<.05$) between all feature types with the exception of well defined masses which did not differ significantly from either ill defined or spiculate masses. The most prominent feature error was architectural distortions and asymmetry - see Figure 4.

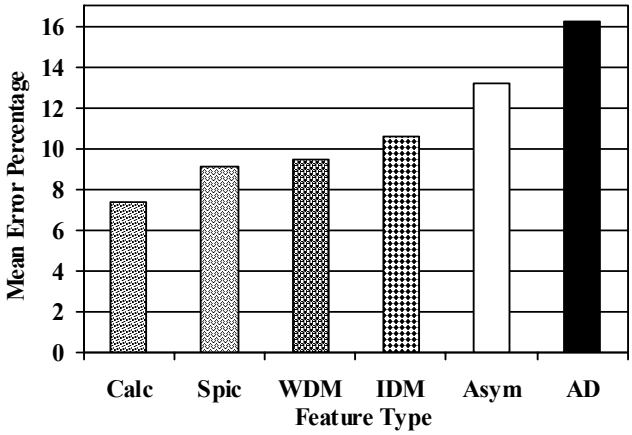

Figure 4: Error by Feature Type

3.4 False Negatives by Feature and Error Type

Performance on each feature type was fractionated by error type (instances of misinterpretation or non detection). The repeated measures, within-subjects, ANOVA with two IV (feature type and error type) and one DV (error percentage) also revealed a main effect of error type [$F(1,2525) = 10.49$, $p<.01$]. More cases were undetected than misinterpreted overall - see Figure 5. There was also a significant feature x error type interaction [$F(5,2525) = 85.59$, $p<.01$]. A priori Bonferroni Pairwise comparisons revealed that there were significant differences between AD undetected and all other feature types, AD undetected was significantly less well performed than all other feature types. Calcification showed the least error and was significantly better detected than all other features (with the exception of WDM). For misinterpreted features, asymmetry was significantly worse than all other feature types except IDM and WDM (where the difference did not reach significance). For misinterpreted features, spiculate masses were the best performed showing significantly less error than all other features ($p<.001$).

In order to establish which features are more prone to errors of misinterpretation or errors of non detection post hoc t-tests were employed. All results are shown in Table 1. There were significant differences for WDM, IDM, spiculate masses, architectural distortions and calcifications. WDM, IDM and calcifications showed significantly more errors of misinterpretation, however spiculate masses and architectural distortions show significantly more errors of non-detection. There were no significant differences between error type for asymmetries, however a one-sample t-test

revealed that for asymmetry both errors of non-detection ($t(505)= 2.384$, $p<.05$) and misinterpretation ($t(505)= 4.384$, $p<.01$) were significantly higher than the average error score for either factor.

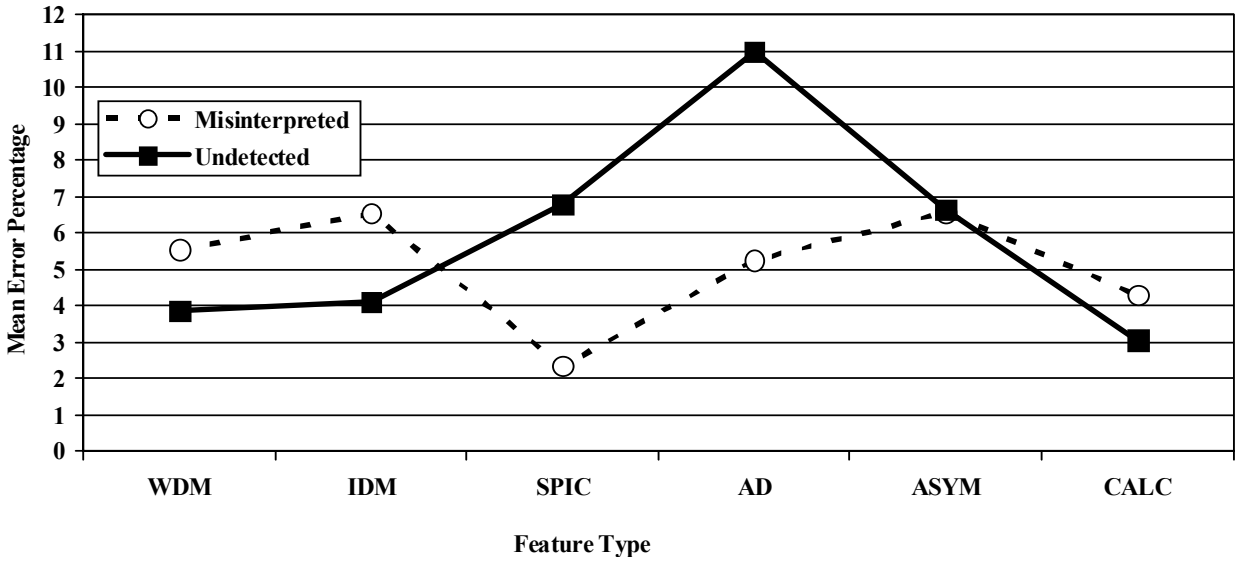

Figure 5: Feature by Error Type

KEY		
SPIC: Spiculated Mass	**AD**: Architectural Distortion	**Calc**: Calcification
IDM: Ill Defined Mass	**WDM**: Well Defined Mass	**ASYM**: Asymmetric Density

Table 1: Pairwise Comparisons for Feature Type by Error Type

	Error Type			
Feature Type	**Undetected**		**Misinterpreted**	
	Mean	*SD*	**Mean**	*SD*
WDM	3.87**	*6.34*	**5.55****	*9.02*
IDM	4.08**	*6.10*	**6.55****	*7.49*
SPIC	**6.78****	*8.21*	2.34**	*4.20*
AD	**10.97***	*10.36*	5.25*	*6.53*
ASYM	3.05	*4.85*	4.30	*5.77*
CALC	3.05	*4.85*	**4.30****	*5.77*

* Sign. differences at the $p<.05$ level.
**Sign. differences at the $p<.01$ level.

Results showed significant differences for all feature types (with the exception of well defined masses) as well as for perception type. Well defined and ill defined masses showed higher percentage of errors of misinterpretation, as did calcifications whereas error rates for both spiculate masses and architectural distortions were highest for errors of non-detection.

3.5 False Negative Errors by Experience

False negative errors were also analysed by mammographic experience. However, for this analysis we only included those who had also completed the PERFORMS self-report, this was a cohort of the original group of participants who

completed only the PERFORMS set (n=409). Years of screening experience were sub-divided into six categories for ease of analysis.

A two-factor, mixed ANOVA with one within subjects factor (error type) and one between subjects factor (experience group) revealed a main effect of group [$F(5,403) = 13.397$, $p<.001$] and error type was approaching significance [$F(1,403) = 3.668$, $p=.056$]. The group x error interaction was non significant (p=n.s). Post hoc SNK (Student Newman Keuls) tests revealed that those groups with lower experience (1-3yrs, 4-6yrs and 7-9yrs) showed significantly more errors ($p<.05$) than those who had longer reading experience (10 years or more) – see Figure 6.

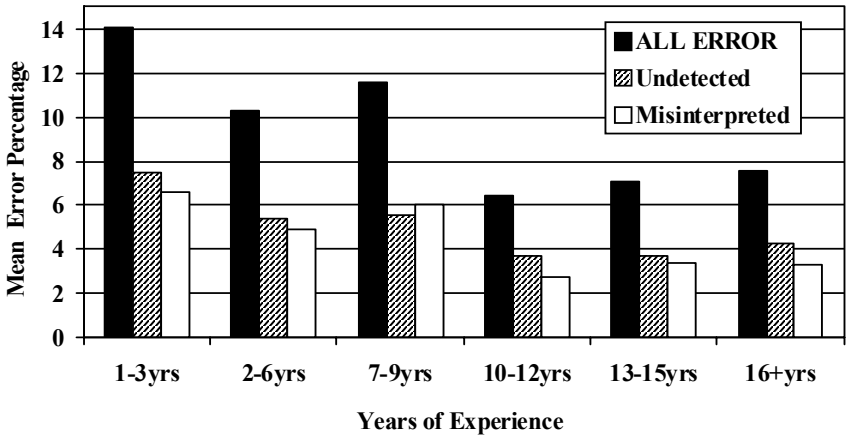

Figure 6: Error Type by Perception and Experience Group

3.6 False Negatives by Feature Type and Experience

False negatives were analysed by feature type and experience. A three-factor mixed ANOVA, with two within groups measure (feature and error type) and one between groups measure (experience group) revealed a significant main effect of both feature type [$F(5,2015)=32.214$, $p<.01$] and experience group [$F(5,403)=12.805$, $p<.01$]. In addition, there was a significant feature x experience group interaction [$F(25,2015)=2.582$, $p<.01$]. For all groups, AD was the feature type displaying the most errors. AD overall was significantly less well performed than all other feature types, followed by asymmetries which, although better performed than AD showed poorer performance compared to other types of feature - see Figure 7. Descriptive statistics suggest here that, although error rate for all feature decreases with experience type – error rate for the two most difficult features, in this sample AD and asymmetry, appears to equate for the most experienced group.

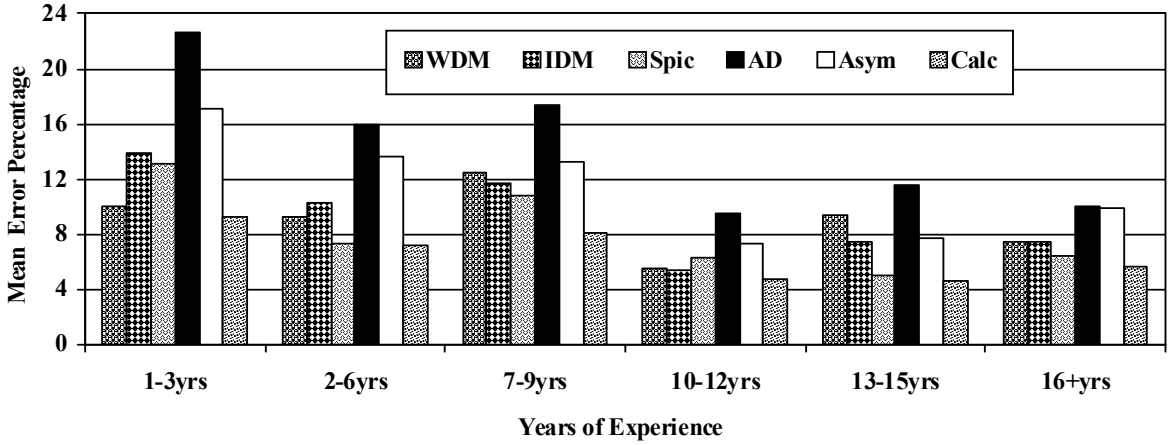

Figure 7: Error by Feature Type

3.7 False Negatives, AD and Asym Feature Type by Experience

In the analysis of false negatives by feature, error type and experience groups a three-way mixed ANOVA revealed a main effect of feature and group but no main effect of error type (p=n.s). There were significant feature x experience group [F(25,2015)=90.923, p<.01] and error type x feature type interactions [F(5,2015)=36.847, p<.01]. However, these were subsumed by a significant three-way error x feature x experience interaction [F(25,2015)=1.9, p<.01].

In the previous section, those features which were most problematic for all experience groups were AD and asymmetries, therefore we looked at these features separately in order to map any possible trends in performance over years of experience - see Figure 8 and Table 2. Post hoc t-tests revealed that undetected AD showed significantly higher percentage error for all age groups (p<.05) compared with undetected asymmetry. There were no differences (p=n.s) for any of the experience groups for AD misinterpreted and Asym misinterpreted, the pattern on misinterpretation for these features appears very similar, see Figure 8.

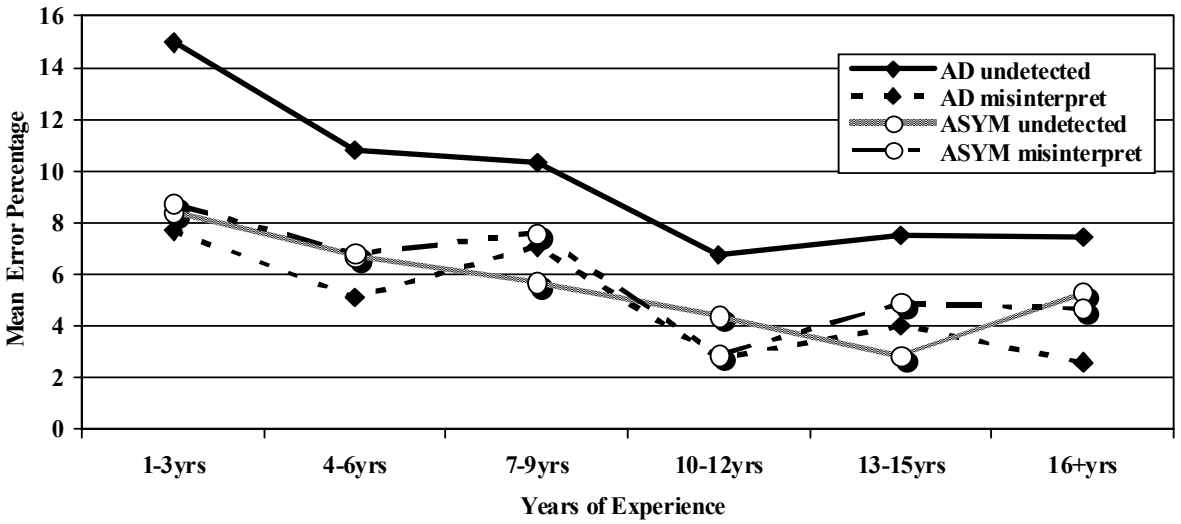

Figure 8: Feature by Error Type by Experience

Table 2 details all post-hoc pairwise comparisons for feature by error by experience. For architectural distortions, the lower experience groups and the most experienced group showed significant differences between undetected and misinterpreted percentages. For these groups, significantly more architectural distortions were undetected compared with those who were misinterpreted. For asymmetries, the opposite pattern was observed for groups 1-3yrs and 13-15 years. Although not all results reached significance, a general descriptive trend can be observed whereby architectural distortions tend to remain undetected, whereas asymmetries show greater errors for misinterpretation.

Table 2: Pairwise comparisons for feature type by error type

	Error Type			
Feature Type	**Undetected**		**Misinterpreted**	
AD	**Mean**	*SD*	**Mean**	*SD*
1-3yrs	13.90**	*15.7*	6.65**	*10.17*
4-6yrs	9.32**	*11.0*	4.70**	*7.90*
7-9yrs	8.48	*11.06*	5.88	*8.21*
10-12yrs	3.41	*5.42*	1.32	*3.29*
13-15yrs	7.77	*9.34*	4.41	*7.44*
16yrs+	4.91**	*7.73*	1.71**	*4.01*

ASYM	Mean	*SD*	Mean	*SD*
1-3yrs	5.48*	*9.45*	8.36*	*12.41*
4-6yrs	4.06	*6.71*	5.43	*9.23*
7-9yrs	3.82	*6.10*	5.31	*7.96*
10-12yrs	0.46	*2.56*	2.53	*6.53*
13-15yrs	1.27*	*3.91*	4.33*	*7.61*
16yrs+	1.62	*4.27*	2.17	*4.76*

* Sign. differences at the p<.05 level.
**Sign. differences at the p<.01 level

4. DISCUSSION

False negative cases show a distinct profile, in this sample, of specific features more likely to be misinterpreted and others which are more likely to remain undetected.

For false negatives, overall error was low (under 11%) and was characterised by non-detection of features, rather than of misinterpretation (Figure 1). With the exception of well defined masses, each feature type was significantly different in terms of apparent severity of error. Significantly, architectural distortions and asymmetries were the most problematic features, with calcifications showing the least error percentage. These results are in some way representative of those from the initial two studies[2,3] where the highest percentage of false negative cases were asymmetry and ill defined masses (taken from the UK National Interval Cancer Database) and asymmetry and architectural distortion (from the first two rounds of PERFORMS). These results, in part, support our previous study[4] whereby asymmetries were the most problematic feature (although they do not explain why in this previous study[4] architectural distortions were one of the least problematic features).

Pinpointing which features were not detected as opposed to misinterpreted was particularly clear. Overall, undetected architectural distortions showed significantly higher error percentage than any other feature, followed by spiculate masses and asymmetries, with calcifications showing as the least problematic feature. For misinterpretation, asymmetries showed the highest error percentage with spiculate masses showing the lowest error score for any feature. Pairwise comparisons revealed that certain features such as architectural distortions and spiculate masses were significantly more prone to errors of non detection whereas other features namely, well defined masses, ill defined masses and calcifications were significantly more likely to be misinterpreted. However, asymmetries showed no significant difference between undetected and misinterpreted errors however. These data are somewhat supported by the argument that PPV for malignancy should be able to predict whether a feature is non-detected or misinterpreted. By this argument (Savage et al, 1994)[2] spiculate masses, ill defined masses and architectural distortions (as the features with the higher PPV for malignancy) should not be misinterpreted and if missed would be due to errors of non-detection of the abnormality. Conversely, well defined masses, asymmetry and calcification show a low-to-very-low PPV for malignancy so are more likely to be misinterpreted as benign non-suspicious features. In the present study, errors on ill defined masses were significantly higher for misinterpretation and asymmetries were evenly spread between both error types.

It has been argued that the differences in PERFORMS false negative results throughout the years (notably the differences in the error rates for architectural distortions) could be due to the different periods in which the participants were sampled. This, it has been suggested, poses a possible developmental influence on feature detection by years of screening experience. In order to examine this probability, false negative errors were profiled by years of screening experience in the current sample. There were no significant differences for error type when analysed by experience group, although descriptively the non detected errors were higher than the misinterpreted. However, those who had been reading for less time (1-3yrs, 4-6yrs and 7-9yrs) showed significantly higher error percentage than those who has been reading for 10 years or more. All of which strongly suggests that years of reading experience is tantamount to 'expert' performance. This supports our previous research into optimal characteristics for film reading (Scott and Gale 2007)[5] which stated that, years of screening experience, over all other possible factors, affected performance.

When looking at feature type error by years of screening experience, it was noted that although error rates decreased inversely to years of experience, all groups showed significantly more errors for architectural distortions and

asymmetries. However, these data suggest (see Figure 7) that the difference in performance for architectural distortions and asymmetries appears to parallel-out in the most experienced group. This suggests that those reading films for a long time may have a 'visual vocabulary' for mammographic abnormalities so sufficiently large as not to be challenged by difficult features.

When examining these difficult feature types for this sample (asymmetries and architectural distortions) by error type, architectural distortions showed significantly higher error percentage (over asymmetries) for all experience groups, but there were no such differences between the features when looking at misinterpreted error type. Where misinterpretation is the larger part of the asymmetric error type across experience groups (although not all groups reached significance). For architectural distortions, all individual groups showed a higher percentage for non-detection than for misinterpretation (which again did not reach significance across all groups). These group results largely support the overall trend in terms of error type and problematic feature type, and, although error percentage is reduced over years of experience, the characteristic or profile of the false negative errors remains relatively intact.

An obvious difficulty in comparing different years of PEFORMS sets (across several studies) is due to the nature in which cases on a yearly basis are acquired - it is not possible to control for case difficulty to any great degree. Performance could be different on separate sets due to the inclusion of particularly difficult examples of any one feature type. Any future study which aims to map developmental differences over several PERFORMS sets should control for this as far as possible. Individuals are often described as having a particular reading style (some participants are better at certain features) and it may be prudent to control for individual difference in future studies. In addition, a meta-analysis of data from all studies including data from real life interval cancer databases (extracted from the same time frame) would validate these results still further.

Implications for real life screening from these data is to elucidate which features which should be 'red flagged' for initial training and also which feature types prove consistently challenging regardless of years of screening experience. Architectural distortions are very often not detected, for example. Although there may exist particular developmental differences in false negative profiles, film-reading experience best characterises peak radiological performance.

5. CONCLUSIONS

It was concluded false negative responses could be clearly characterised in terms of not only mammographic feature but of error type. If certain feature types, as in this self-assessment pool of cases, could be universally identified as susceptible to specific types of errors, this may have implications for the monitoring of accuracy in radiological performance in real life. In order to provide an accurate profile of feature type and perceptual difficulties it is suggested that a further developmental study relating PERFORMS to real life 'false negative' data was warranted.

ACKNOWLEDGEMENTS

This work is supported by the UK 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 E. Lovesay (Ed.) Ergonomics - design for performance 1991, Taylor & Francis, London.
2. Savage, C.J., Gale, A.G., Pawley, E.F. & Wilson, A.R.M. To err is human, to compute divine? In A.G. Gale, S.M. Astley, D.R. Dance & A.Y. Cairns (Ed) Digital Mammography .Proceedings of the 2nd International Workshop on Digital Mammography, 1994, Elsevier, Amsterdam.
3. Cowley, H., Gale A. & Wilson, R. Mammographic training sets for improving breast cancer detection. In: Medical Imaging 1996: Image perception and performance. Miguel P. Eckstein & D.P. Chakraborty (Ed.). Proceedings of SPIE Vol. 2712, 102-112.
4. Scott, H.J., Gale, A.G.: Breast-Screening Technologists: when is a difficult case truly difficult and for whom?, In: Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment. Eckstein MP and Y Jiang (Eds.) Proceedings of SPIE Vol.5749, 557-65.

5. 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. Y Jiang and B Sahiner (Eds.) Proceedings of SPIE Vol.6515, 65150F.