Breast screening technologists: when is a difficult case truly difficult and for whom?

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- This is a refereed conference paper.

Metadata Record: https://dspace.lboro.ac.uk/2134/2645

Publisher: © SPIE

Please cite the published version.
Breast Screening Technologists: when is a difficult case truly difficult and for whom?
Hazel J. Scott and Alastair G. Gale*
Applied Vision Research Institute, University of Derby, Kingsway House, Kingsway, Derby, DE22 3HL, UK

ABSTRACT

Each year almost all film readers in the UK Breast Screening Programme voluntarily read a set of difficult mammographic cases as a means of self-assessing their film reading skills. We set out to investigate what case characteristics, if any, actually constituted a ‘difficult’ or ‘easy’ case in the opinion of radiological experts. We also examined how UK Breast Screening personnel performed on those cases which the experts deemed were difficult, in order to build up a profile of the types of cases that provide film readers with the most problems. We examined two main elements of case diagnosis, case classification and case features and investigated if there were any group differences in terms of case difficulty and the percentage of incorrectly reported cases. Data from over 15 radiological experts and approximately 400 film readers were compared on 180 cases. Significant differences were found between the expert and screening populations (p < .05) in terms of these case characteristics. These data contribute to the understanding of just what constitutes a difficult case as considered by experts and other film-readers, with a view to elucidating the type of cases most appropriate for advanced mammographic training.

Keywords: PERFORMS, Breast Screening, Performance, Case Difficulty, Expertise.

1. INTRODUCTION

It is recommended that a film reader in the UK Breast Screening Programme examines a minimum of 5,000 screening cases per annum. Given the low incidence of breast cancer, circa 6.8 per 1,000 women, (1) and a typical case volume of well over 10,000 per annum a film-reader can expect to see few malignant cases per week. In the UK each film reader must decide whether to recall a case for further assessment or return that case to normal screen (where the woman will come back for future screening in 3 years time). Feedback on those cases that have been recalled is relatively rapid, pending further assessment or biopsy report, but feedback on whether a case is a false negative is not (as that woman will not return for a future screen for several years).

As part of a response to this difficulty of performance feedback the PERFORMS (PERsonal PerFORmance in Mammographic Screening) scheme was established 15 years ago (2). This functions as a free and anonymous self-assessment scheme whereby film readers have the opportunity to evaluate 120 difficult cases per annum in two sets of 60 cases. This educational exercise allows immediate and confidential feedback to each film reader. Each year some 94% of these film-readers in the UK participate. The film sets are amassed nationally and renewed annually. Cases are initially assessed by a panel of five radiological experts who identify (among other things) the type of features apparent in the case as well as the case difficulty. They also provide the ‘gold standard’ for the scheme by classifying each case on a measurement akin to the BI-RADS system.

We set out here to investigate what case characteristics, if any, constituted a ‘difficult’ and ‘easy’ case in the opinion of the radiological experts over several years of different PERFORMS film sets. We also examined how participant film-readers performed on those cases which were deemed difficult in order to build up a profile of the types of cases

* a.g.gale@derby.ac.uk; phone 0044 1332593131; fax 0044 1332593176; http://ibs.derby.ac.uk
that provide film readers with the most problems and also to determine whether radiological experts can pinpoint challenging cases accurately.

2. METHODOLOGY

Each year we amass a number of exemplar mammographic screening cases which are submitted by UK screening centres as potentially diagnostically interesting cases to use in our self-assessment scheme. A panel of five expert radiologists first independently examines these cases and rates them on several scales, including their perceived degree of difficulty in interpretation (in terms of radiological appearance), as well as identifying the potential presence of a number of mammographic features. From these separate expert opinions we determine a single overall ‘radiological expert opinion’ for each case and from the collected large number of cases we then select a range of difficult examples to use in our self-assessment scheme.

We deliberately select those cases which represent a range of normal, benign and malignant appearances and which are classed as ‘difficult’ by the panel so as to stretch the skills of the participants and to make the scheme both interesting and challenging. Participants subsequently read the set of cases when their opinions on case classification and mammographic features are compared to this initial expert opinion, as well as to their peers. Further details of the scheme are given elsewhere (3,4,5).

Each year a new set of cases is derived. Additionally, the experts serving as panel members can change, depending upon their everyday workload and their availability. Here, data from over 15 radiological experts and approximately 400 other film readers on 180 difficult cases were examined over three PERFORMS film sets, from the years 2001 to 2003. Although the experts were instructed explicitly to determine which cases were difficult and which were not, the participants in the scheme were not requested to do so as part of their task.

The expert radiological opinion was collected, by questionnaire, on each case so as to determine which cases the experts rated as difficult. Case difficulty ranged from ‘very easy’ (rated 1) to ‘very difficult’ (rated 4). In order to build up a profile of difficult cases we examined two main case descriptors, case classification (normal, benign and malignant) and case features (‘none’, well-defined mass, ill-defined mass, architectural distortion, spiculate mass, asymmetry and calcification). As each case can potentially contain a mixture of several features - for these analyses we identified those features which were the most prominent.

Initially the data were analysed in terms of identifying a benchmark for case difficulty - as established by an explicit difficulty rating by a group of radiological experts. Data concerning 180 difficult recent screening cases were examined over the past three audited years (2001-2003) of the scheme. The data, for ease of analysis, from each year half (60 cases) were pooled and analysed in terms of overall performance for the three years. In each year individuals (including Radiologists and Technologists) participated on two occasions, reading 60 cases each time. For the purposes of these analyses we selected the second half of each annual film set across the three years.

Secondly, we also looked at how difficult the population of film readers subsequently found the cases to be. In order to investigate which cases posed the most problems for participants over the three years of the PERFORMS scheme we examined participants’ performance levels in terms of mean percentage of cases incorrectly recalled or returned to screen (i.e. where the participants failed to identify whether the case was suspicious or normal according to expert radiological opinion and case pathology).

3. RESULTS

3.1 Experts’ Opinion

3.1.1. Case Classification

For case classification (Figure 1) radiological experts rated cases with abnormal features as the most difficult (on a scale of very easy (rated 1) to very difficult (rated 4)). A one-way Analysis of Variance (ANOVA) was performed on these data, with mean difficulty rating as the dependent variable (DV) and case classification as the independent variable (IV). Results showed that there was a significant difference for type of case classification \(\text{F}(2,70724) = 4419, \ p<.001\) and post hoc Student Newman Keuls (SNK) tests revealed that normal cases were rated significantly less difficult than benign and malignant cases (Normal cases, mean = 2.17; Benign cases, mean = 2.54; Malignant cases,
mean = 2.54%). There were no significant differences between benign and malignant cases which were rated equally demanding.

![Mean difficulty rating by case classification](image1)

**Figure 1: Expert difficulty rating by case classification**

3.1.2. Case features
Radiological experts rated case features (both benign and malignant) which were split into seven main types (‘none’, well defined mass (WDM), ill defined mass (IDM), speculated mass (Spic), architectural distortion (AD), asymmetry (Asym) and calcification (Calc)). Note that ‘none’ indicates the absence of key mammographic features in a case and in the analyses presented here this absence is treated as a ‘feature type’.

![Mean difficulty rating by case feature](image2)

**Figure 2: Expert difficulty rating by case feature**

A univariate between-subjects ANOVA was carried out with one DV variable (case difficulty) and one between groups factor (feature type). The results indicated that there was a significant effect of feature type \( F(6,70724) = 2774.81, p< .001 \). Post hoc tests (SNK) showed that there were significant differences between all the feature types \( p< .05 \) with those cases which had no features rated the easiest and architectural distortions rated as the most difficult (mean scores: None = 2.11; mean Calc = 2.31; mean WDM = 2.34; mean IDM = 2.5; mean Asym = 2.58; mean Spic = 2.64, and mean AD = 2.75). The pattern of difficulty is illustrated in Figure 2 (above).
3.1.3. Case difficulty by case classification and feature type

Some mammographic cases are highly irregular and are not included in PERFORMS sets, such as a malignant case that has no apparent key mammographic features or a normal case which potentially has, what may appear to be, say a spiculated mass. However, as some features are more dominant within particular case classifications feature type was also analysed as a function of case classification (see Figure 3 below).

![Figure 3: Expert difficulty rating -case feature by classification](image)

A univariate ANOVA with one DV (case difficulty) and two IVs (case classification and feature type) indicated that there was a significant effect of case classification \[F(2,7078) = 18.4, p < .001\] and feature type \[F(6,7078) = 508.54, p < .001\]. There was also a significant interaction between case classification and feature type \[F(8,7078) = 179.9, p < .001\]. Post hoc analysis (SNK) revealed that the pattern of difficulty for different features was unique to case classification. There were significant differences between all the feature types within case classification with two exceptions, for the benign cases there were no significant differences between asymmetries and ill defined masses and for malignant cases there were no significant differences between spiculated masses and asymmetries (however, these groupings were significantly different from their adjacent groupings). The mean difficulty patterns are detailed overleaf (Table 1) and show that for normal cases those features rated easiest for the PERFORMS sets were cases without features and those most difficult contained asymmetries. For benign cases the easiest cases were rated to be those with calcification and the most difficult were those with spiculated masses. The pattern was again different for malignant cases where well defined masses were rated as the easiest and those cases with architectural distortions were rated as the most difficult (over the three year period).
Table 1: Mean difficulty scores and significance levels for case classification by feature

<table>
<thead>
<tr>
<th>Feature Difficulty</th>
<th>Normal</th>
<th>Benign</th>
<th>Malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Easiest</td>
<td>None*</td>
<td>2.08</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Calc*</td>
<td>2.25</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>AD*</td>
<td>2.37</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>WDM*</td>
<td>2.43</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Asym*</td>
<td>2.64</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Spic*</td>
<td>3.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Most Difficult</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Sign. differences between features at the p<.05 level.
**Sign. differences between adjacent groups p<.05 but not from each other.

3.2 Participant performance

3.2.1 Case classification

A one-way Analysis of Variance with one DV (cases incorrectly reported) and one IV (case classification) showed that there were significant differences for case classification \[F(2,73498) = 1137.07, p<.001\].

Student Newman Keuls post hoc tests revealed that there were significant differences between all case classifications (see Figure 4) whereby malignant cases showed the lowest percentage of incorrect cases reported compared to benign and normal cases (normal cases showed significantly higher percentage of incorrect cases reported than benigns). Malignant cases (percentage cases incorrect = 13%) were reported significantly better than benign cases (26% incorrect) and normal cases (28% incorrect). Normal cases were reported less accurately than benign and malignant cases. Therefore, those classifications that radiological experts reported as being the most difficult (benign and malignant) were not those which the participants found the most difficult in these cases.

3.2.2 Case features

Figure 5 shows the mean percentage of cases incorrectly reported cases reported by feature type for approximately 400 participants over three years of PERFORMS sets.
A one-way ANOVA indicated that there was a significant effect of feature type \([F_{6,73498}=276.00, \ p<.001]\). Post hoc tests (SNK) showed that there were significant differences between all the features apart from ill defined masses and none (which although they did not differ from each other, differed from those features whose scores were adjacent to them such as well defined masses and asymmetries). The pattern of features that participants found difficult to report correctly overall, shown here to be asymmetries and none, were different from those that radiological experts reported as difficult (which were spiculate masses and architectural distortions).

### 3.2.3. Case classification and feature type

Participants’ performance levels were analysed by case classification and feature type in order to establish a profile of those cases that they found the most challenging (Figure 6).

A univariate ANOVA with one DV (incorrectly reported cases) and two IV (case classification and feature type) revealed that there was a significant effect of both case classification and feature type (already reported) as well as a
significant classification by feature interaction $[F(8,73482)=216.32, p<.001]$. Post hoc tests showed that there were significant differences between groups of features (see Table 2).

### Table 2: Mean incorrectly reported scores and significance levels for case classification by feature

<table>
<thead>
<tr>
<th>Case Classification</th>
<th>Normal</th>
<th>Benign</th>
<th>Malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correct = 0</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Mean</td>
<td>SD</td>
<td>Type</td>
</tr>
<tr>
<td>None**</td>
<td>0.26</td>
<td>0.44</td>
<td>AD**</td>
</tr>
<tr>
<td>Asym**</td>
<td>0.27</td>
<td>0.44</td>
<td>IDM**</td>
</tr>
<tr>
<td>AD**</td>
<td>0.36</td>
<td>0.00</td>
<td>Calc**</td>
</tr>
<tr>
<td>WDM**</td>
<td>0.38</td>
<td>0.48</td>
<td>Spic**</td>
</tr>
<tr>
<td>Calc*</td>
<td>0.52</td>
<td>0.50</td>
<td>WDM**</td>
</tr>
<tr>
<td>Asym*</td>
<td>0.65</td>
<td>0.47</td>
<td>IDM*</td>
</tr>
</tbody>
</table>

* Sign. differences between features at the p<.05 level.
** Sign. differences between adjacent groups p<.05 but not from each other.

Here again the pattern of results differed from those set by radiological experts. For normal cases the experts rated ‘none’ as the easiest feature type and asymmetry as the most difficult. For performance measures both ‘none’ and asymmetry tied as the least problematic feature types for normal cases. Calcification was significantly different from all other feature types for normal cases and was the one which participants found the most challenging (as compared to asymmetries which were rated as the most difficult feature by experts). For benign cases experts rated calcification as the easiest feature with spiculate masses as the most difficult, in reality participants found those cases with architectural distortions and ill defined masses as the least challenging to assess with asymmetries posing the most problems. As for malignant cases participants found architectural distortions, calcifications and spiculate masses the easiest to report with ill defined masses as the most difficult. Compare this to the experts’ decisions of well defined masses and architectural distortions (respectively).

### 4. DISCUSSION

Our initial question was concerned with what constituted a difficult case in the opinion of radiological experts and how this related to the performance of other breast screening film readers in interpreting these cases. We assume here that a case judged to be difficult should then be more likely to give rise to more errors of interpretation than a case judged to be easy.

The results indicated that, unsurprisingly, the expert radiologists considered the normal cases as significantly less difficult than either benign or malignant examples even though all the cases used in these film sets were considered to be of a difficult and challenging nature. Although all three types of classifications received a fairly similar overall mean difficulty score, normal cases were rated as those exemplars which would be easiest for the breast screening population of film readers. Somewhat in contrast, participants subsequently performed far better on the malignant cases than on both benign and normal cases and unlike experts’ predictions (on which cases would be the most challenging) there were significant differences between how well they performed on the malignant and benign cases.

In the second part of the study we looked at how well breast screening personnel performed in terms of key radiological feature types and here again there was an inverse finding of which features were problematic. Whereby experts judged those cases with no features to be the easiest and those with architectural distortion and spiculate masses to be the most difficult, in practice, the film-readers found ‘no features present’ and asymmetries to be problematic to identify and made less errors with architectural distortions and spiculate masses.

For the final analysis we looked at case classification by feature type (as certain features are more dominant within certain classification type). For normal cases the experts predicted that the absence of features would be the easiest and this coincided with the least number of incorrectly recalled cases. In practice the normal cases with calcification gave rise to the most errors whereas those with asymmetry did not create the problems that the experts had predicted.
Over all of the various types of cases, benign cases with asymmetries were the most demanding of the set with the percentage of case errors on over half of the cases (65%) which was 13% higher than the next most challenging feature type (for normal classification: calcification mean = 52%) and were markedly different from other benign feature types, although the experts had judged asymmetries to be one of the easiest of the benign appearances. In practice, fewer errors were found for malignant cases (as such there was a lower overall percentage of feature type error) in contrast to the high difficulty rating of the experts. The experts had rated architectural distortion as the most difficult but these proved to be the easiest.

In summary, the results seem to indicate that there is a strong disassociation between case difficulty and participants’ performance on cases. This unexpected finding seems to raise far more questions than it answers, as follows.

Why should experts’ ratings of difficult cases and film readers predicted difficulty (in terms of error rates) differ so much? One explanation is that experts really do find different things difficult than the normal reader. Which raises the question - do differing levels of knowledge in reading mammographic cases determine which cases are the most demanding? Do certain feature types (such as asymmetries) become less problematic as radiological skill is amassed?

In order to answer these questions further study of the data is planned whereby individual groups will be compared on expertise as measured by years of experience and volume of cases read.

Another explanation as to the dissociation of results lies in the methodology itself. In this study the experts were asked to rate difficulty of a case (an explicit measure) whereas the population of film readers were not, therefore in order to attain a measure of difficulty we measured their incorrect responses (implicit). This could indicate a number of things, firstly that experts cannot explicitly measure case difficulty. One must measure how they actually performed on these cases and examine how well their results tally with what they measured as difficult in order to ascertain if their perceptions of what is difficult are actually what they find difficult.

Couples with this is the question of who is an expert? Experts can be defined as individuals who are clearly internationally respected as expert breast screening radiologists based on their extensive knowledge, experience and their demonstrated skill in correctly identifying cases. At a national level, similar criteria also apply and in our study we utilise as experts those individuals with long standing experience in breast screening in the UK. The radiologists used by us as experts also subsequently participate in the scheme themselves, at a much later date than when they originally view the cases in order to provide an initial opinion, and so it is possible for us to examine their original opinions concerning difficulty and their own performance in the scheme - a topic for future consideration.

Secondly, although one definition of difficulty would assume that when a task is arduous, and demanding of much effort and skill, that there would be a greater number of instances where an individual fails at this task than when it is an easy one. If experts’ results on the set do not tally with their difficulty ratings this raises an interesting question of whether perceived difficulty is not what is actually difficult. Perhaps this does not go far enough, perhaps no-one can explicitly measure case difficulty - just because a case was correctly reported does not mean it was not perceived as difficult. If we were to ask people about cases they incorrectly classified for example –they might not perceive them to have been difficult.

This raises the further question of whether a measure of case difficulty is a predictor of how well people perform (which was its intention). We include this question in the radiological opinion in order to pick out those discerning cases for the set. Perhaps there must be two measures here, one of actual difficulty as measured by correct responses (and perhaps also by observation time on each case - a more difficult case may take longer to classify) and one of perceived difficulty (those cases which were correctly identified but marked as difficult).

Another possible methodological issue is the difficulty rating itself, cases are chosen for the PERFORMS scheme if these have a majority decision on classification and feature type. Experts may tend to agree more on the cases with more obvious features. It is clearly difficult to get experienced radiologists to agree on mammographic features (c.f. 5).

5. CONCLUSIONS

In summary, the findings are not as clear-cut as we expected and possible reasons for this are offered. Although these data go some way to elucidating which characteristics of judged case difficulty would allow tailoring cases into specialised training sets for specific film-reading populations, further work is needed to clarify many of the issues raised by the non-conformity of results reported here.
ACKNOWLEDGEMENTS

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

REFERENCES


