Breast screening technologists: does real-life case volume affect performance?

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/2134/2643

Publisher: © SPIE

Please cite the published version.
Breast Screening Technologists: does real-life case volume affect performance?
Hazel J. Scott*, Alastair G. Gale, and David S. Wooding
Institute of Behavioural Sciences, University of Derby, Kingsway House, Kingsway, Derby, DE22 3HL, UK

ABSTRACT
In the UK fewer radiologists are now specialising in breast cancer screening. Consequently, a number of technologists have been specially trained to read mammograms so as to double-read with existing radiologists. Each year the majority of these film-readers examine a set of difficult cases as a means of self-assessing their skills. We investigated whether the technologists performed as well as breast-screening radiologists on this difficult test set. We also investigated technologists’ performance over a number of years to compare the performance of those technologists who have read a greater number of breast screening films and those who have had less experience. Finally, we investigated real-life experience and performance on the scheme by comparing; volume of cases read, experience, and technologists’ performance over time versus radiologists’ performance. Data for approximately 250 breast screening Radiologists and 80 specially trained technologists over three years for six sets of 60 difficult recent screening cases were examined. Overall, those technologists who have not read the same volume of cases as radiologists did not perform as well on this particular task. Although when the group was fractionated by volume of cases read in real-life and the number of years reading cases, then the technologists performed at a level similar to the radiologists.

Keywords: PERFORMS, Breast Screening, Performance, Case Volume, Double-reading, Radiologist, Technologist.

1. INTRODUCTION
Widespread breast screening began in the UK some 15 years ago, initially with the introduction of single view mammography for all women between the ages of 50 and 64 every three years. Subsequently this was then extended to double view mammography and currently the upper age range limit is being raised, up to, and including, 70 years of age.

The interpretation of mammographic screening images is primarily the domain of the specialised consultant radiologist. However, in the UK, reading screening cases is performed mainly by three groups of specialists: consultant radiologists, breast clinicians, and specially trained technologists. Breast clinicians are medically qualified personnel specialised in breast screening and the last group are technologists with experience in breast screening and who have also undertaken a specialised training programme. As well as radiologists, approximately 33% of Breast Screening Units employ breast clinicians and 65% employ such trained technologists.

In the UK fewer radiologists are now entering this speciality. Within the last seven years two surveys have been carried out which examined various issues concerning breast screening, in 1997 (by the Breast Screening Programme across all of the UK) and in 2003 by the Royal College of Radiologists in England1. Of those Breast Screening Units that responded, some 59% in 2003 had vacancies for radiologists (in 1997 this figure was 41%). Whilst some of the tasks for these specialists could be performed by other personnel it is clear that there exists a shortfall in experienced radiologists in breast screening. Consequently the specially trained technologists should help alleviate the growing shortfall in the availability of experienced radiologists. These personnel typically double-read cases, usually with a radiologist. Anecdotal reports indicate that they perform well. This is further supported by a recent study at one UK Breast Screening Unit which showed that such technologists performed as well as radiologists2. Whilst these data are limited to the evaluation of only two technologists and only one screening centre the findings are very supportive of technologists reading screening cases.

* h.scott@derby.ac.uk; phone 0044 1332593139; fax 0044 1332593176; http://ibs.derby.ac.uk
The incidence of breast cancer within the screening age range in the UK is fairly low, in 2001/2 circa 6.8% cancers were detected per 1,000 women screened. Consequently, any individual reading mammographic cases has to examine a large number per year in order to have both current and progressing experience of variations in normal and abnormal mammographic appearances. In the UK all breast screening personnel must read a minimum number of 5,000 screening cases per annum. We have previously argued that it is not only important for an individual to read a considerable number of screening cases per year but also that the volume of cases read relates to the skill with which a radiologist has in interpreting the cases examined.

As part of the UK Breast Screening Programme all individuals who read mammographic cases are invited annually to participate in a voluntary self-assessment scheme, ‘PERFORMS’ (PERsonal PerFORmance in Mammographic Screening), where they are presented with a series of recent and difficult cases. The majority of film readers invited agree to take part in this scheme. In this present study we examined the data from three years of this scheme and investigated whether technologists performed as well as breast-screening radiologists on these difficult test sets. Furthermore, we also determined the technologists’ performance over these years in order to compare data from those technologists who have read a greater number of breast screening films and those who have had less experience. Finally we investigated real-life experience and performance on the PERFORMS scheme by using a matched study paradigm in order to compare the performance of technologists and radiologists.

2. METHODOLOGY

In the UK all practising breast screening Radiologists, and other suitably qualified individuals involved in interpreting screening cases, are annually offered the opportunity to participate in a confidential self-assessment film-reading scheme. In this scheme the individual reads, on two occasions each year, a series (60 cases) of difficult recent screening cases by entering data into a computer system. Participants are immediately given detailed feedback on how well they have judged each case as compared to a standard based on the radiological opinions of a group of experienced breast screening radiologists and also as against known pathology, where appropriate. All 120 cases are amassed nationally from the UK breast-screening programme and each year a new set of cases are employed. Although PERFORMS is a voluntary scheme, the majority of film-readers involved in the Breast Screening programme choose to take part. After all individuals have participated then a ‘national radiological opinion’ about each case is determined, namely that based solely on the radiological appearance should that case be recalled or not, and each individual’s decisions concerning the radiological appearance of every case are then calculated against this national opinion.

Data are subsequently fed back to individuals on how well they have performed as compared to the anonymous data of their peers. These data include information on performance measures such as the number of correct recall decisions (equivalent to sensitivity), correct return to screen decisions (equivalent to specificity), cancers correctly detected and ROC measures such as d’. ROC d’ values are a measure of overall performance (from correct recall and correct return to screen results) as calculated from the national opinion (d’ for national opinion) as well as for pathology alone (d’ for Pathology). In addition all participants in the scheme are required to complete a computerised questionnaire covering details of their individual real-life screening practices such as weekly volume of cases read and years of screening experience.

Here we examine the radiologists’ and technologists’ data from three recent rounds of the PERFORMS scheme; namely the SA2000, SA2001 and SA2002 mammographic film sets which were circulated around the UK in the corresponding years 2000-2002. This yields information concerning how 360 difficult screening cases had been reported by 250 breast screening radiologists and 80 specially trained technologists.

3. RESULTS

3.1 Overall performance

Data concerning 360 difficult recent screening cases were examined over the past three audited years (2000-2002) of the scheme. In each year individuals (radiologists and technologists) had participated on two occasions, reading 60 cases each time. Initially the results were analysed in terms of overall performance for each group over this three-year period. The results, for ease of analysis, from each year half (60 cases) were pooled and analysed in terms of annual
performance (120 cases). Several performance measures were examined, namely; number of malignancies detected, correct recall (CR) and correct return to screen (CS) decisions, and ROC measures of d’ for the national radiological opinion and d’ for pathology.

For malignancies detected (see Figure 1) the mean performance of the radiologists was higher than the technologists consistently over a three-year period and annual performance, for both groups, was higher in recent years.

![Graph showing mean malignancies detected by occupation](attachment)

**Figure 1: Overall performance – malignancies detected by occupation**

A two-way mixed Analysis of Variance (ANOVA) was performed on these data, with one between groups factor (occupation) and one within groups factor (PERFORMS year). Results showed that there was a significant within groups difference of PERFORMS year \( F(2,1976) = 113.28, p<.001 \) and post hoc Student Newman Keuls (SNK) tests revealed that the percentage of malignancies detected significantly increased for each consecutive PERFORMS year (SA2000 mean = 86.1%, SA2001 mean = 89.27%, SA2002 mean = 94.53%). Therefore, although both groups’ performance increased for each PERFORMS case set, the technologists’ performance was significantly below that of the radiologists.

**Table 1: Mean Scores and significance levels for overall performance measures**

<table>
<thead>
<tr>
<th>Performance measures</th>
<th>Correct Recall (CR)</th>
<th>Correct Return to Screen (CS)</th>
<th>d’ values for National Opinion (d’NO)</th>
<th>d’ values for Pathology (d’Path)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Radiologists</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA2000 (n=514)</td>
<td>Mean 81.62**</td>
<td>Mean 87.71</td>
<td>Mean 0.001**</td>
<td>Mean 0.98**</td>
</tr>
<tr>
<td></td>
<td>SD 10.42</td>
<td>SD 7.65</td>
<td>SD 0.98</td>
<td>SD 0.040**</td>
</tr>
<tr>
<td>SA2001 (n=518)</td>
<td>Mean 88.53**</td>
<td>Mean 88.71</td>
<td>Mean 0.783**</td>
<td>Mean 1.04**</td>
</tr>
<tr>
<td></td>
<td>SD 7.81</td>
<td>SD 7.87</td>
<td>SD 0.089**</td>
<td>SD 1.11</td>
</tr>
<tr>
<td>SA2002 (n=500)</td>
<td>Mean 91.91**</td>
<td>Mean 90.85</td>
<td>Mean 0.150**</td>
<td>Mean 1.01**</td>
</tr>
<tr>
<td></td>
<td>SD 6.73</td>
<td>SD 6.24</td>
<td>SD 0.156**</td>
<td>SD 1.11</td>
</tr>
<tr>
<td><strong>Technologists</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA2000 (n=86)</td>
<td>Mean 81.98**</td>
<td>Mean 87.37</td>
<td>Mean -0.008**</td>
<td>Mean -0.239**</td>
</tr>
<tr>
<td></td>
<td>SD 10.40</td>
<td>SD 9.08</td>
<td>SD 1.19</td>
<td>SD 1.39</td>
</tr>
<tr>
<td>SA2001 (n=160)</td>
<td>Mean 85.21**</td>
<td>Mean 89.19</td>
<td>Mean -0.253**</td>
<td>Mean -0.289**</td>
</tr>
<tr>
<td></td>
<td>SD 9.45</td>
<td>SD 7.76</td>
<td>SD 1.09</td>
<td>SD 1.20</td>
</tr>
<tr>
<td>SA2002 (n=204)</td>
<td>Mean 88.46**</td>
<td>Mean 90.54</td>
<td>Mean -0.328**</td>
<td>Mean -0.355**</td>
</tr>
<tr>
<td></td>
<td>SD 9.06</td>
<td>SD 6.07</td>
<td>SD 1.10</td>
<td>SD 1.25</td>
</tr>
</tbody>
</table>

**Sign. differences between occupational groups at the p<.001 level.**

Similar results were found for the correct recall performance measure, where the radiologists’ mean scores for all years was above that of the technologists (see Table 1). A two-way mixed ANOVA was carried out with one between groups factor (occupation) and one within groups factor (PERFORMS year). The results indicated that there was a significant effect of both year \( F(2,1976) = 90.84, p<.001 \) and occupational group \( F(1,1976) = 18.89, p<.001 \). Post hoc tests (SNK) showed that performance increased with PERFORMS year. There was also a significant interaction between
Performs year and occupational group. Post hoc analysis showed that there were significant differences between the groups (p< .001) for the SA2001 and SA2002 film sets. However, the technologists’ performance overall was significantly below that of the radiologists, whereby overall radiologists’ sensitivity level for the test sets exceeded that of the technologists.

A two-way mixed ANOVA with one between groups factor (occupation) and one within groups factor (Performs year) was also performed on the correct return to screen percentages. Results indicated that there were no differences between the occupational groups (p = n.s.), although there was a significant effect of year [F(2,1976) = 19.54, p< .001], with the percentage correct increasing throughout the three years. There was no interaction between any of the factors. Therefore, the technologists’ specificity was on a par with that of the radiologists.

For the overall ROC measures for the national radiological opinion (d’NO) and for pathology (d’Path) the results were similar. Two-way mixed analysis of variance revealed that for both d’NO [F(2,1976) = 21.77, p< .001] and d’Path [F(2,1976)=35.46, p< .001] there were significant group differences as radiologists outperformed the technologists. There were no main effects of year for either measure, although there was a significant interaction between year and occupational group for d’NO. Post hoc analysis revealed that the group differences were not apparent for the SA2001 and SA2002 film sets.

3.2 Technologists – within group analysis
In order to investigate the effect of volume of cases read and years of experience in reading mammograms on the performance of the technologists, whereby those who have read more cases may perform better than those with less experience, a smaller sample of matched individuals (n = 10) for each year experience group, and (n = 30) for each PERFORMS year group was selected.

![Figure 2: Technologists' malignancies detected by case volume and years of experience](image)

Figure 2 shows the percentage of malignancies correctly detected fractionated by volume of cases read and years of experience. Years of experience were divided into three groups, ranging from reading from one year or less to reading for over five years. The volume of cases read was divided into two groups, those who had read less than 200 and those who had read over 200 cases per week.

A two-way between subjects ANOVA indicated that there was no effect of years reading (p = n.s) but that there were significant differences in the volume of cases read per week and the percentage of malignancies detected [F (2,84) = 6.64, p<.05]. The interaction between volume of cases read per annum and years of experience was non-significant. Mean differences in the volume of cases read (see Figure 2) reveal that those who have read a greater number of cases did detect significantly more percentages of malignant cases overall.
Table 2: Mean scores and significance levels for technologists’ performance measures

<table>
<thead>
<tr>
<th>Volume of cases</th>
<th>Correct Recall (CR)</th>
<th>Correct Return to Screen (CS)</th>
<th>d’ values for National Opinion (d’NO)</th>
<th>d’ values for Pathology (d’Path)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=5,000-10,000</td>
<td>Mean SD</td>
<td>Mean SD</td>
<td>Mean SD</td>
<td>Mean SD</td>
</tr>
<tr>
<td>1 yr or less (n=21)</td>
<td>82.80* 10.72</td>
<td>89.69 6.42</td>
<td>-0.463* 1.00</td>
<td>-0.609* 1.09</td>
</tr>
<tr>
<td>2-3 yrs (n=14)</td>
<td>86.07* 9.02</td>
<td>88.94 9.02</td>
<td>-0.044* 1.04</td>
<td>-0.128* 1.08</td>
</tr>
<tr>
<td>5+ (n=21)</td>
<td>80.53* 12.77</td>
<td>87.96 7.54</td>
<td>-0.822* 0.96</td>
<td>-0.930* 1.18</td>
</tr>
<tr>
<td>&lt;=10,000-36,000+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 yr or less (n=21)</td>
<td>87.24* 7.54</td>
<td>85.63 6.42</td>
<td>-0.204* 0.93</td>
<td>-0.473* 0.99</td>
</tr>
<tr>
<td>2-3 yrs (n=14)</td>
<td>88.90* 6.75</td>
<td>91.46 4.34</td>
<td>0.488* 0.69</td>
<td>0.725* 0.87</td>
</tr>
<tr>
<td>5+ (n=21)</td>
<td>88.76* 8.13</td>
<td>88.43 6.00</td>
<td>-0.035* 0.79</td>
<td>-0.085* 0.71</td>
</tr>
<tr>
<td>Overall Years</td>
<td>Mean SD</td>
<td>Mean SD</td>
<td>Mean SD</td>
<td>Mean SD</td>
</tr>
<tr>
<td>1 yr or less (n=30)</td>
<td>84.13 9.96</td>
<td>88.47 6.58</td>
<td>-0.385 0.97</td>
<td>-0.568 1.05</td>
</tr>
<tr>
<td>2-3 yrs (n=30)</td>
<td>87.58 7.88</td>
<td>90.28 6.92</td>
<td>0.239 0.93</td>
<td>0.327 1.05</td>
</tr>
<tr>
<td>5+ (n=30)</td>
<td>83.00 12.0</td>
<td>88.10 7.20</td>
<td>-0.586 0.97</td>
<td>-0.677 1.12</td>
</tr>
</tbody>
</table>

Similarly, significant differences in correct recall performance were found when fractionated by volume \([F(2,84) = 5.39, p<.05]\) but not by years of experience \((p = n.s.)\). Mean scores (see Table 2) indicate that the higher the volume of cases read then the higher the performance levels for correct recall of cases. There were no differences for either years of experience or volume for correct return to screen percentages. There were, however, significant effects of both years of experience and volume of cases read for both d’NO (volume \([F(2,84) = 6.29, p<.05]\), years \([F(2,84) = 3.95, p<.05]\)) and d’Path values (Volume \([F(2,84) = 6.90, p<.01]\),Years \([F(2,84) = 5.88, p<.005]\)). All year by volume interactions were non-significant. Mean scores (see Table 2) suggest that, although overall, those reading mammograms for longer have higher d’ values, when fractionated by volume those who have read more cases have a higher overall score than those who have read less for all year groups.

3.3 A matched study of experienced technologists and radiologists

From the foregoing, the more experienced technologists (particularly those who consistently read a high volume of cases) performed significantly better than less experienced technologists. However, the question remains as to whether those experienced technologists perform as well as radiologists (who overall consistently score higher than technologists).

![Figure 3: Occupational group match by case volume and years of experience](https://example.com/figure3.png)
In order to explore this issue, in the light of previous research as well our previous findings on technologists’ performance, both occupational groups were matched on both volume of cases read and number of years reading cases. For this analysis each individual was closely matched on both volume and years of experience. Unlike the first study reported here in section 3.1, no differences between years of experience and volume of cases read (p = n.s.) were found, the mean scores for these measures for each PERFORMS year were identical (see Figure 3).

A two-way mixed ANOVA revealed that when the participants were matched on volume and years of experience there were no significant differences between the groups for malignancies detected (p = n.s.). Figure 4 shows the mean score for each over three years, both group means were similar for all three year groups. This pattern of results was repeated for each of the performance measures (see Table 4). There were no significant differences between the groups for any of the performance measures on this matched study (see Table 3).

Therefore, when volume and years experience are controlled, occupational group differences for sensitivity and overall performance measures disappear. In order to further support these findings, Analysis of Co-variance (ANCOVA) was carried out for each of the performance measures. Even when the variance for volume and number of years experience was co-varied from the analysis there were still no significant group differences for any of the performance measures. This suggests that performance is mediated by factors other than simply occupational group.
Although there are no overall group differences a further question remains as to the specific effect of level of experience (volume and years of reading) on performance. A two-way ANOVA with two between groups factors, occupation group and years of experience (with three levels of experience, ‘1 or less’, ‘2-3 years’ and ‘5+’) revealed that there were no group differences for malignancies detected. However, there were significant differences for years of experience \( [F(2,174) = 3.08, p<.05] \) and there was a significant interaction between occupational group and years of experience \([F(2,174) =3.00,P<.05 ]\). Mean values (see Figure 5) were higher for technologists in the 2-3 yrs category and higher for radiologists in the 5+ year category. Post-hoc analysis revealed that group differences in the 5+ years category were significant \( (t = 2.39, p<.05) \).

![Figure 5: Mean % of malignancies detected by year group](image)

This result was repeated for overall performance measures such as d’NO and d’Path where a two-way Analysis of Variance revealed significant differences for years of experience \( ([F(2,174)=7.164, p<.001],[F(2,174)=8.373, p<.001]) \) and significant interactions between year of experience and occupational group \( ([F(2,174)=5.3,p<.05],[F(2,174)=5.22,p<.05]) \). Post hoc tests also indicated that there were significant group differences for the 5+ year category only \( (d’NO \text{ radiologists’ mean}=0.228, sd=1.01, \text{ technologists’ mean}=0.586, sd=0.97), \text{ d’pathology [radiologists’ mean }= 0.2727, \text{ SD }= 1.01, \text{ technologists’ mean }= -0.677, \text{ sd }= 1.12]) \). These results suggest that although overall technologists perform as well as radiologists, those expert radiologists who have been reading for over five years outperform the technologists. For CR and CS performance measures however there were no significant differences for years of experience nor were there any significant interactions.

4. DISCUSSION

Performing meaningful real life screening comparisons between radiologists and technologists without them all reading identical cases can prove difficult, not least because of the potential high variability in case appearances. Therefore the purpose of this study was to examine data from this self-assessment scheme to see how well those technologists who have participated compared to the radiologist participants when reading the identical cases with the purpose of contributing to the debate over whether technologists should be reading screening cases.

For overall analysis the technologists, in terms of sensitivity and d’ measures did perform well, although not as well as the radiologists. Their specificity levels were however similar to that of the radiologists. The pattern of results indicated systematic under-recalling on the part of the technologists. Technologists, who in general have not read the same volume of cases as radiologists, did not perform as well on these particular test-sets.

Analysing the technologists’ data further indicated that although years of experience is an important factor (especially for overall performance measures) case volume is consistently a defining factor in high performance levels for sensitivity. There were no significant differences in performance (as defined by volume and years) of experience for specificity measures. Highest mean scores on all performance measures were achieved by those technologists who had
over two years experience and typically read a large volume of cases. This suggests that technologists who consistently read a high volume of cases for at least two or more years achieved higher sensitivity and overall performance levels than those who had read for one or less years, or who were reading a lower volume of cases.

However when radiologists and technologists were matched on volume of cases read and years of experience the differences between the groups disappeared and technologists performed (on all measures) at the same level of radiologists.

It is important to note the limitations of this work. Each year different cases are utilised and so it is inherently difficult to match case difficulty on a year by year basis which can yield variable annual difficulty. This could contribute to some of the differences between the various PERFORMS case sets. Data for the volume of cases read in real life, at the time of participating, are based upon participants’ self-reports using a response scale of 100s of cases per week.

How comparable are data from the PERFORMS scheme to real life screening? Cowley and Gale previously showed some similarities for radiologists between their performance in the scheme and real life. However, importantly, it must be noted that the manner in which a participant undertakes such self-assessment schemes can vary, depending upon their motivation at the time and other work pressures.

5. CONCLUSIONS

The results indicate that the technologists on this self-assessment scheme performed well. This finding adds further support to the current UK practice of technologists double-reading with radiologists. Additionally it suggests that experienced technologists may be skilled sufficiently to function as single-readers, able to double read with other technologists. Clearly this conclusion must be subject to a caveat, namely that these data are from what is essentially an experimental situation and not everyday screening practice.

ACKNOWLEDGEMENTS

This work is supported by the UK National Health Service Breast Screening Programme. We acknowledge the contribution of Helen Cowley to these data.

REFERENCES