Does routine breast screening practice over-ride display quality in reporting enriched test sets?

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


Additional Information:

- Copyright 2014 Society of Photo-Optical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

Metadata Record: [https://dspace.lboro.ac.uk/2134/20075](https://dspace.lboro.ac.uk/2134/20075)

Version: Published

Publisher: © SPIE

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: [https://creativecommons.org/licenses/by-nc-nd/4.0/](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Please cite the published version.
Does routine breast screening practice over-ride display quality in reporting enriched test sets?

Yan Chen* a, Alastair G. Gale a, Michael Evanoff b

a Applied Vision Research Centre, Loughborough University, Loughborough, UK;
b American Board of Radiology, Tucson, USA

ABSTRACT

The performance of a group of 16 American (US) breast screening radiologists in interpreting a number of cases from a recent PERFORMS self-assessment case set which had been carefully selected to exclude small calcifications, using sub-mammographic resolution displays, as compared to a British (UK) group of radiologists using mammographic displays has previously been reported. It was found that the UK group performed better, detecting more cancers with the US participants correctly recalling less. These results were interpreted as due to differences in the displays employed by each group as well as to routine screening differences between the two countries. This current study extended that work with 11 of these experienced US breast screening radiologists further interpreting 20 new PERFORMS mammographic cases using a suitable mammographic clinical workstation. The PERFORMS cases were selected so as to show a range of normal, benign and abnormal appearances. Data from these radiologists were compared to their earlier performance on different PERFORMS cases and sub-clinical displays. Their data were also compared to recent data of 11 UK radiologists reading the same cases, again on clinical workstations as well as to all UK screeners. Despite using equivalent clinical monitors, data indicate differences between the UK and US groups in recall decisions which is not just a function of the countries’ screening approaches. Lower detection of abnormal cases by the US group was found here and reasons for this are explored.

Keywords: breast screening, FFDM, performance

1. INTRODUCTION

In both the UK and the US approximately one in eight women are affected by breast cancer during their lives 1,2. Consequently, routine breast screening using two-view mammography is commonly implemented to identify early signs of this disease. In the UK over two million women aged between 45 and 74 are screened every three years by the National Health Service Breast Screening Programme (NHSBSP) 3 with all screening personnel (i.e. breast screening radiologists and advanced practitioners - technologists trained to read and interpret screening cases) having to read a minimum of 5,000 cases a year 4. A key performance measure, in addition to cancer detection performance, is the number of women recalled after screening for follow up examinations which is low ~ circa 4.2% nationally 5. This low figure means that overall few UK women are needlessly recalled, thus minimizing psychological trauma to them of any false positive screening outcome. Despite this, there has been widespread recent debate concerning ‘over diagnosis’. This is where a malignant tumor has been correctly identified but where it would probably have done no harm yet women in such a situation have often opted for a surgical solution, possibly needlessly. This has led to a recent review of the breast cancer screening program which demonstrated that 1,300 lives were saved every year through screening whilst also leading to almost 4,000 women having unnecessary treatment. This in turn has led to women who attend for screening being better informed in future of their various treatment choices.

In the USA breast screening radiologists typically need to read a much lower annual volume of cases; circa 960 cases over two years’ with generally more women being recalled for further investigation – one comparative study between the US and the UK indicating almost twice as many women being recalled (13.3% as compared to 7.2%) 6. This reflects differences between the two countries in the approach to screening practice where in the USA there is more emphasis on ensuring that a potential abnormality is not overlooked and with the UK trying to minimise unnecessary recalls.

*y.chen@lboro.ac.uk

* © 2013 SPIE · CCC code: 1605-7422/13/$18 · doi: 10.1117/12.2007846


Proc. of SPIE Vol. 8673 86730V-1

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 01/06/2016 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx
In the UK all screeners undertake the annual PERFORMS scheme\textsuperscript{9,10}. This is an educational exercise where enriched test sets of recent difficult cases containing a range of mammographic features are examined. Previously we have compared how individuals from the UK and US have read a PERFORMS scheme test set of mammographic films using multi-viewers\textsuperscript{11}. Recently, a related study\textsuperscript{12} has compared UK and USA radiologists reading the same full field digital mammographic cases selected from a recent PERFORMS test scheme. However, whereas the UK radiologists used clinical mammographic workstations, the American group unfortunately only had access to lower resolution displays (as it had not been feasible to acquire clinical mammographic workstations for this study) necessitating the cases being selected so as not to disadvantage this group by not including small calcifications. Other previous work of ours has demonstrated that using such sub-clinical workstations is feasible to report mammographic cases in this type of task as long as suitable image manipulation software is used\textsuperscript{13}.

This present study investigated what differences in performances occur when experienced breast screening radiologists from both countries examined the same FFDM case set using appropriate mammographic workstations. Data were examined both between countries and between this current test and the previous one. In undertaking the test, participants examined the image case set and for each case recorded various decisions using bespoke software. This software not only records decisions related to identifying abnormal appearances but also decisions about whether they would recommend that case for recall for further investigations or not. Thus abnormality identification and approaches to patient recall can be examined separately. Differences in recall would be expected on the basis of differences in routine screening approaches. Data were analysed to elucidate what effect, if any, the different clinical mammographic displays made to the American group. In addition differences between the US and UK radiologists were also investigated.

It was hypothesised that the US and UK radiologists would identify malignant abnormalities and truly normal cases with similar skill but that the US group would recall more cases (specifically benign cases) than the UK group.

\section{2. METHOD}

\subsection{2.1 Previous study}

In the UK, the PERFORMS self-assessment scheme is undertaken by every screener who participates in the UK Breast Screening Programme twice per year. Each time, 60 difficult exemplar cases are examined on clinical workstations in the 109 screening centres. The cases are reported by each person and appropriate feedback is provided. In the previous comparative study\textsuperscript{12} a set of PERFORMS cases were also examined by each US participant at the American Board of Radiology meeting in Louisville in 2011. With the unsuccessful attempt to acquire clinical mammographic workstations on that occasion, the participants interpreted the cases using PCs running dual 20" DICOM calibrated monitors. Forty difficult PERFORMS cases were reported by 16 experienced US breast screening radiologists using tablet PCs running in-house bespoke PERFORMS software. To allow for the limitation of low screen resolution, any micro-calcification cases were carefully excluded from the test set. The room lighting was reduced to reflect typical screening room light levels for digital mammography reporting and measured.

\subsection{2.2 Present comparative study}

A number of experienced US breast screening radiologists were invited to take part at the American Board of Radiology meeting in Louisville in 2012. A Barco clinical mammographic workstation was acquired for carrying out the study. Considering the high difficulty of the test set and limited time slot that each participant had, then a set of 20 challenging PERFORMS cases were selected and loaded onto the workstation. Each participant first viewed two practice cases to familiarize them with the particular DICOM viewer being used as well as the new PERFORMS reporting software. The case set was a mixture of normal, benign and malignant cases and contained various mammographic features, such as spiculate masses, micro-calcifications, etc. A dedicated laptop with the PERFORMS software installed was used for recording the responses and giving each participant appropriate feedback after they had taken part (figure 1). Each participant examined the set in a randomized order in a quiet experimental room with suitable controlled lighting level.
For each case, they had to decide if any key mammographic features were present and if so to locate these on case mimic images. Also, they had to rate whether they would recall each breast or not by using a five point rating scale equivalent to the BIRADS scale\textsuperscript{14,15} and finally rate the overall case density. In the event 11 radiologists took part and there was no time limit for them individually to read all 20 cases.

For comparison purposes the anonymous data were used of 11 experienced UK breast radiologists randomly selected from out of approximately seven hundred participants who had also read the same cases as part of the PERFORMS scheme using their usual clinical mammographic workstations.

In addition, the US data were compared to the anonymous data of all UK participants ($n=685$) who have read these cases in the PERFORMS scheme. This included data for both breast screening radiologists and advanced practitioners.

Figure 1. Experimental set up in Louisville showing the PERFORMS reporting software running on a laptop.

3. RESULTS

Data were examined in terms of cancer detection (CD) where the number of cancers correctly identified as compared to the known case pathology was scored. ROC analysis was performed to calculate $A_z$ values using the pathology information. Additionally the measures of correct recall (CR), correct return to screen (CS), positive predictive value (PPV) and negative predictive value (NPV) were recorded. These are calculated as judged against the radiological decisions of a panel of expert UK breast screening radiologists as to whether a case should be recalled or not. These four measures refer to the proportion of suspicious cases correctly identified and recalled (CR: this could include both benign and malignant cases); the proportion of non-suspicious cases correctly identified and not returned to screen (CS. i.e. in real life that woman would have been told she was disease free and would then attend at the next screening round); PPV and NPV refer to the proportion of correct recall compared to all recall decisions and the proportion of ‘correct return to screen’ decisions compared to all such judgments respectively. The data of the US group on this occasion were first examined with consideration of their performance on the previous occasion where they read a different case set on sub-clinical monitors. Then the US group was compared to 11 UK breast radiologists and finally they were compared to all the UK screening participants who had read these cases.
3.1. US group now and in previous study

Quantitative examination of measures between the US individuals on these two occasions are not presented here as it is difficult to enumerate performance data when two different sets of cases using different monitor resolutions were used. However it is possible to consider their relative performances as compared to the UK radiologists who read the same cases on the two occasions. In the previous study, using sub-clinical monitors, the US group, not surprisingly, were significantly worse than the UK group of radiologists in terms of number of cancers detected, CS, and PPV decisions. In the present study, using suitable monitors then the only difference in performance between the US and UK groups was in CR decisions (figure 1).

3.2. Comparison of US and UK performances

The data of the 11 American radiologists were compared on the screening metrics to that of randomly chosen 11 UK radiologists who had read these cases as soft copy images. There was no significant difference (figure 2) between these two groups in correct return to screening (CS) decisions (UK: $M = 85.3\%$, $SE = 0.02$; USA: $M = 80.3\%$, $SE = 0.02$, $p = n.s.$). Also, there was no significant difference between these two groups in Positive Predictive Value (PPV) decisions (UK: $M = 75.0\%$, $SE = 0.04$; USA: $M = 65.9\%$, $SE = 0.05$ $p = n.s.$) or in Negative Predictive Value (NPV) decisions (UK: $M = 92.8\%$, $SE = 0.03$; USA: $M = 89.0\%$, $SE = 0.01$, $p = n.s.$).

Furthermore, there was no significant difference found (figure 3) between these two groups in $A_x$ score (UK: $M = 0.93SE = 0.01$; USA: $M = 0.87$, $SE = 0.03$, $p = n.s.$) nor cancer detection (CD) decisions (UK: $M = 84.5\%$, $SE = 0.03$; USA: $M = 70.9\%$, $SE = 0.07$, $p = n.s.$). However, there were significant differences in correct recall (CR) decisions (UK: $M = 89.1\%$ $SE = 0.03$; USA: $M = 79.5\%$, $SE = 0.02$, $p < 0.05$) between the two groups (figure 2).
3.2 Further UK – US comparisons

The data of the 11 US radiologists were then compared on the screening metrics to that of all UK screeners \((n=685)\) who had read these cases as soft copy images (figure 4). The K-S test was carried out and showed that data were all significantly non-normal, \(p<.05\). Therefore, the Mann-Whitney test was used which found that there was no significant difference (figure 4) between these two groups in correct recall (CR) decisions (UK: \(Mdn = 82.3\%\); US: \(Mdn = 80.0\%\), \(z = -1.011, p=n.s., r = -0.038\)). Also, there was no significant difference between these two groups in CS decisions (UK: \(Mdn = 84.6\%\); US: \(Mdn = 79.3\%\), \(z = -1.824, p=n.s., r = -0.069\)). No significant difference was also found between these two groups in PPV percentages (UK: \(Mdn = 75.0\%\); US: \(Mdn = 62.5\%\), \(z = -1.834, p=n.s., r = -0.069\)). Furthermore, there was no significant difference found between these two groups in NPV percentages (UK: \(Mdn = 91.6\%\); US: \(Mdn = 87.5\%\), \(z = -0.186, p=n.s., r = -0.007\)). However, there were significant differences in cancer detection (CD) decisions (UK: \(Mdn = 90.0\%\); US: \(Mdn = 70.0\%\), \(z = -1.99, p<.05, r = -0.07\)) and in \(A_z\) score (based on known pathology) as shown in figure 5 (UK: \(Mdn = 0.93\); US: \(Mdn = 0.90, z = -1.90, p<.05, r = -0.07\)).

4. DISCUSSION

In the previous study the use of lower resolution displays (approximately half that of a mammographic workstation) by the US group was found to be offset by their experience. Not surprisingly, the UK group overall performed better on these cases as they were using high resolution mammographic workstations. However, the US group performed well but still recalled more, which was interpreted as reflecting their real life screening practice. The set of cases used had been deliberately selected not to contain significant small calcifications so that the US group using the lower resolution displays would not be too disadvantaged and also we have previously shown that acceptable performance can be obtained using sub-clinical mammographic displays. There are various possible explanations for the differences found between the two groups – the most likely one being due to the differences in monitor resolutions. Hence in this follow up study equivalent clinical monitors were employed. It was now found that the US and UK groups performed well in detecting cancers, with the UK detecting more but not significantly so, but differed in CR decisions (as judged against an expert UK panel in whether a case should be recalled) – again reflecting screening differences between the two countries. The previous study also explored possible differences between the two groups in terms of the volume of cases reported as read annually by each participant. Volume of cases read per year and years of reading screening cases are important predictors of screening performance\(^{13}\). This aspect was not further investigated here.

Figure 3. \(A_z\) values comparing 11 US radiologists and 11 UK radiologists

![Az Values](image-url)
Figure 4. Performance values comparing 11 US radiologists and UK screeners
Figure 5. Az values comparing 11 US radiologists and UK screeners.

Figure 6. US and UK reporting of malignant, benign and normal cases.
When the 11 US and 11 UK radiologists were compared then the UK group scored higher on the various metrics (figure 2 and figure 3) with a significant difference in correct recall (CR) decisions (as judged against an overall UK decision to recall). There was no difference in terms of the other metrics of cancers detected, CS, PPV, NPV and $A_z$. When compared to data from across the UK then significant differences were found in the number of cancers detected and in $A_z$. The performance metrics of CR, CS, NPV, and PPV are based on scoring decisions on each case as true positive or true negative decisions based upon the overall decisions of the UK expert panel on that case on whether it is abnormal and should be recalled or not. Thus, it would be expected that the US group would appear to over-recall on these measures as compared to UK radiologists. To understand this more, figures 3 and 5 show the $A_z$ values for each group using $A_z$ measures calculated only on the known normal and known malignant (based on pathology) cases. Whilst both groups scored very highly the UK group of 11 radiologists scored higher with the overall UK data UK being significantly higher. Overall these data indicate that in this study the US group did not identify as many true abnormal cases as their UK counterparts. Figure 6 elaborates more on this with the 11 US participants compared to all the UK data on these cases. The US group better identified normal cases and the UK better identified malignant and benign cases. Why should this be? It may be pedantic to make too much of these differences on such a small case set but nevertheless there are several interesting possible explanations. Differences in decisions to recall or not were to be expected however differences in identifying abnormality per se were not. There are differences in how the US and UK undertook the study. In the UK the participants read the cases in their own clinics using their own DICOM viewing software etc. whereas here the US group used a DICOM viewer which was not familiar to them although the room was at a typical radiological room lighting level. In reporting the cases the PERFORMS software, as implemented here, used a five point scale similar to BIRADS and the US group had little problems in quickly using this. However, the US group performed the test in amongst their ABR examining schedule and the UK radiologists carried out the test in amongst their clinical screening. It may well be that the US radiologist at a disadvantage.

5. CONCLUSIONS

Better performance was found here when suitable clinical monitors were employed which was to be expected. However comparative data between the two groups still indicate differences which may well relate to differences in screening practices.

It was predicted that using comparable clinical monitors as compared to the previous study where the US radiologists had access to lower resolution displays then the US and UK groups would score similarly in cancer identification but differ in recall decisions so reflecting routine clinical practices in the two countries. The data indicate lower cancer identification here by the US group and potential reasons for this are explored. Further studies using larger datasets to yield appropriate statistical power are required to explore this in more detail.

The PERFORMS scheme was specifically designed for individual screeners in the UK to enable them to gain insights into their performance in identifying early signs of cancer as well as how well they perform as compared to colleagues. This study demonstrates that breast screeners from outside the UK can use the scheme equally well and reported finding it to be a very useful educational aid.

ACKNOWLEDGEMENTS

This work is supported in the UK by the UK National Health Service Breast Screening Programme. The study in Louisville was kindly facilitated by the ABR and for which Barco kindly loaned a mammography workstation. PERFORMS is a registered trade mark.

REFERENCES


