
This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Breast screening: visual search as an aid for digital mammographic interpretation training

PLEASE CITE THE PUBLISHED VERSION

PUBLISHER

© 2010 Society of Photo-Optical Instrumentation Engineers

VERSION

VoR (Version of Record)

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Chen, Yan, Ann Turnbull, Jonathan James, Alastair G. Gale, and Hazel J. Scott. 2019. "Breast Screening: Visual Search as an Aid for Digital Mammographic Interpretation Training". figshare.
<https://hdl.handle.net/2134/6286>.

This item was submitted to Loughborough's Institutional Repository (<https://dspace.lboro.ac.uk/>) by the author and is made available under the following Creative Commons Licence conditions.



For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

Breast Screening: visual search as an aid for digital mammographic interpretation training

Yan Chen^{*a}, Ann Turnbull^b, Jonathan James^c, Alastair Gale^a, Hazel Scott^a

^a Applied Vision Research Centre, Loughborough University, Loughborough, UK;

^b Breast Unit, Derby Royal Hospital, Derby, UK;

^c Nottingham Breast Institute, Nottingham City Hospital, Nottingham, UK

ABSTRACT

Digital mammography is gradually being introduced across all breast screening centres in the UK during 2010. This provides increased training opportunities using lower resolution, lower cost and more widely available devices, in addition to the clinical digital mammography workstations. This study examined how experienced breast screening personnel performed when they examined sets of difficult DICOM two-view screening cases in three conditions: on GE digital mammography workstations, on a standard LCD monitor (using a DICOM viewer) and an iPhone (running Osirix software). In each condition they either viewed the full images unaided or were permitted to use the post-processing manipulations of pan, zoom and window level/width adjustments. For each case they had to report the feature type, rate their confidence on the presence of abnormality, classify the case and specify case density. Their visual search behaviour was recorded throughout using a head mounted eye tracker. Additionally aspects of their real life screening performance and performance on a national self assessment scheme were examined. Data indicate that screening experience plays a major role in doing well on the self assessment scheme. Task performance was best on the clinical workstation. However, the data also suggest that a DICOM viewer that runs on a PC or laptop with a standard LCD display allows viewing digital images in full resolution support impressive cancer detection performance. The iPhone is not ideal for examining full images due to the amount of scrolling and zooming required. Overall, the results indicate that low cost devices could be used to provide additional tailored training as long as device resolution and HCI aspects are carefully considered.

Keywords: Mammographic interpretation training, iPhone, low resolution devices, eye movements, HCI

1. INTRODUCTION

Breast screening for women aged 50-64 has been undertaken across the UK for over 20 years using two view mammographic film as the imaging medium¹. Recent developments will see the screening age range being increased to encompass women aged 47- 73 years². Partly to cope with this additional workload, as well as enabling the improved investigation of dense breasts (which typically occur in younger women), Full Field Digital Mammographic (FFDM) imaging is being rolled out, following successful trials in selected screening centres, so that in 2010 all UK screening centres will have some digital imaging ability, alongside the existing analogue film-based screening, with full digital imaging subsequently ensuing.

The change to digital requires that current screening personnel be further trained in examining these images as, whilst early signs of cancer appearance are essentially the same on either digital or analogue images, the appearance of digital images are somewhat different to mammographic film. Additionally the extended screening age range means an increased workload, potentially necessitating further screening staff being trained. There is also the usual need for ongoing CME (Continuing Medical Education) training for existing staff. Most training comprises utilising well known textbooks³ or some form of interactive education where selected key images are presented to be examined, followed by critical and reflective feedback. Such interactive training ideally should be undertaken on the digital clinical workstations themselves but this is not always possible due to time and cost constraints as the workstations are primarily

*y.chen3@lboro.ac.uk

used for the clinical practice of screening. In the UK a national annual self assessment scheme can help identify the training needs of particular groups or individuals⁴ which can then lead to the development of particular training sets of cases⁵.

Mammographic interpretation training

Interpreting mammographic images can be considered to comprise a range of perceptual and cognitive skills which include the recognition of certain mammographic features⁶. Low resolution displays could potentially be employed usefully in training as long as the training regimes were planned around the known resolution of the displays. For instance, parts of mammograms which contain an abnormality could well be displayed on such devices, in a similar way to many CAD research approaches, as a means of familiarising individuals with specific aspects of abnormality appearances. In order to appreciate the location of an abnormality or to ensure that an individual has fully examined the whole image on a low resolution display would potentially require considerable image manipulation panning to move around the image and zooming in to specific areas for further detailed analysis. Additionally window level manipulation may also be required to elucidate certain appearances. Such tools are a standard part of many DICOM viewing software which run on PC or MAC computers. Osirix DICOM viewing software runs on the iPhone and provides a similar facility on this very portable device using a fairly intuitive interface. The iPhone, and similar PDA devices, are gaining in popularity in being used, with constraints, for examining medical images⁷. Previous work, with a view to training purposes only, has shown that radiologists can examine mammographic images on PDA devices with some success^{8,9}. Examining large mammographic images on non-clinical displays inevitably has some limitations. These include the speed of display (important for image manipulation as well as changing amongst the different mammographic views), resolution, and displayable grey scale levels.

This study is part of a series of investigations which examine the use of a range of display devices for training purposes in screening¹⁰. Specifically the investigation examined how well screeners performed when examining cases on a clinical workstation as compared to smaller and lower resolution displays and whether their ability to perform on such latter displays was related to their screening experience and whether they interacted with the images manipulation (i.e. HCI) or simply viewed the images with no manipulation. Whilst examining images their visual search behaviour was monitored to investigate in detail how they visually examined the same images on the different displays. Furthermore, the performance of participants was also related to their data from actual breast screening and also from a national self assessment scheme (PERFORMS) where all UK screeners examine sets of difficult exemplar cases.

2. METHOD

2.1 Materials

Firstly, an expert breast radiologist selected two sets of 20 challenging two view recent digital screening cases. Each set demonstrated difficult examples of normal, benign and malignant appearances. Mammographic features present included: masses, calcifications and architectural distortions. The two sets were closely matched according to case difficulty and feature type. Each case included both the medio-lateral oblique [MLO] and the cranio-caudal [CC] screening views. All images were stored as DICOM files.

2.2 Participants

There were fourteen participants; nine consultant radiologists and five advanced practitioners (i.e. technologists who have been specifically trained to read screening cases) from two major UK breast screening centres who volunteered to undertake the experiments. In the UK all breast screening centres have used analogue film for screening since screening was established and in recent years various centres have begun to introduce FFDM digital mammography. These two centres had each had digital mammography for at least four years and all participants were familiar with the appearance of such images.

The participants were then divided into two groups according to the different screening centres where they primarily worked. The study was approved in advance by the university's ethics committee; appropriate NHS hospital ethical

clearance was also sought and it was determined by the hospital that such clearance was not required for the study as it was judged to represent clinical audit.

2.3 Procedure

Over a period of eight months each group undertook three rounds of trials. For each round, every participant examined these DICOM cases on one of three different display devices. These were: GE digital mammography workstations (with 5 megapixel dual monitors; resolution 2048 x 2560 pixels each); a standard LCD monitor (images were shown using a DICOM viewer running on a laptop, screen size: 21.5", resolution: 1050 x 1680) and an iPhone (images were shown using Osirix DICOM viewing software, screen size: 3.5", resolution: 480 x 320) respectively. The image files shown on each modality were identical.

All participants from one centre first examined the images on one of the screening centre's workstations then at least two months later participants from one centre examined images on the iPhone followed at least two months later by the standard monitor and the other centre did this in reverse order. There was a gap of at least two months between each individual's trials at each centre and for some individuals this was three months. Between trials then each participant would have examined approximately over 1,000 routine screening digital cases. No feedback was given to participants on their performance on each case. The experiment was carried out in darkened reporting rooms and other windowless rooms with controlled ambient lighting levels which were recorded. For the monitor and iPhone conditions an offset desk lamp was used to provide some additional low level ambient illumination¹¹.

For each case set, there were two conditions. Individuals were either only allowed to view the full two view cases unaided as displayed on each device (i.e. view either both the MLO or the CC views and also be able to switch between them) and then for the other case set they were also able to use post-processing image manipulations (here termed the HCI condition) – namely zoom, pan and window level/width adjustment. Each case was first presented as two MLO views; on the workstation and standard monitor these views fully filled the displays whereas on the iPhone these were initially shown as small joint images and the participant had to tap the relevant image for it to be displayed larger (see figure 1).

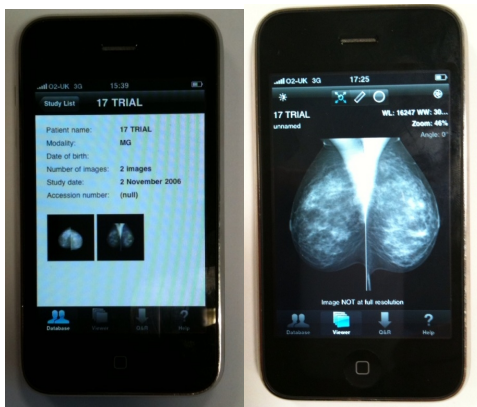


Figure 1. iPhone running Osirix DICOM viewing software

Participants were videotaped using a fixed camera to monitor their behaviour in interacting with the displays. Additionally, when examining the images the participants wore a head mounted eye tracker (ASL 504) to record their visual search behaviour throughout (figure 2). A head mounted system was used as remote eye tracking devices, which would sit beneath the displays unconnected to the observer, do not have the overall spatial recording range to encompass accurately the subtended visual angle at the observer's eye of the two workstation monitors. In figure 2 the top row illustrates the experimental setting at one centre for the workstation task together with an extract from the recorded eye movement record of one person; the central row illustrates the standard LCD monitor task and an associated eye movement record; the bottom row shows the iPhone task and a related eye movement record.

When using the iPhone, the device was fixed on an angled board in front of the observer, both to facilitate user interaction with the displayed images as well as to enable appropriate recording of their visual search behaviour and their

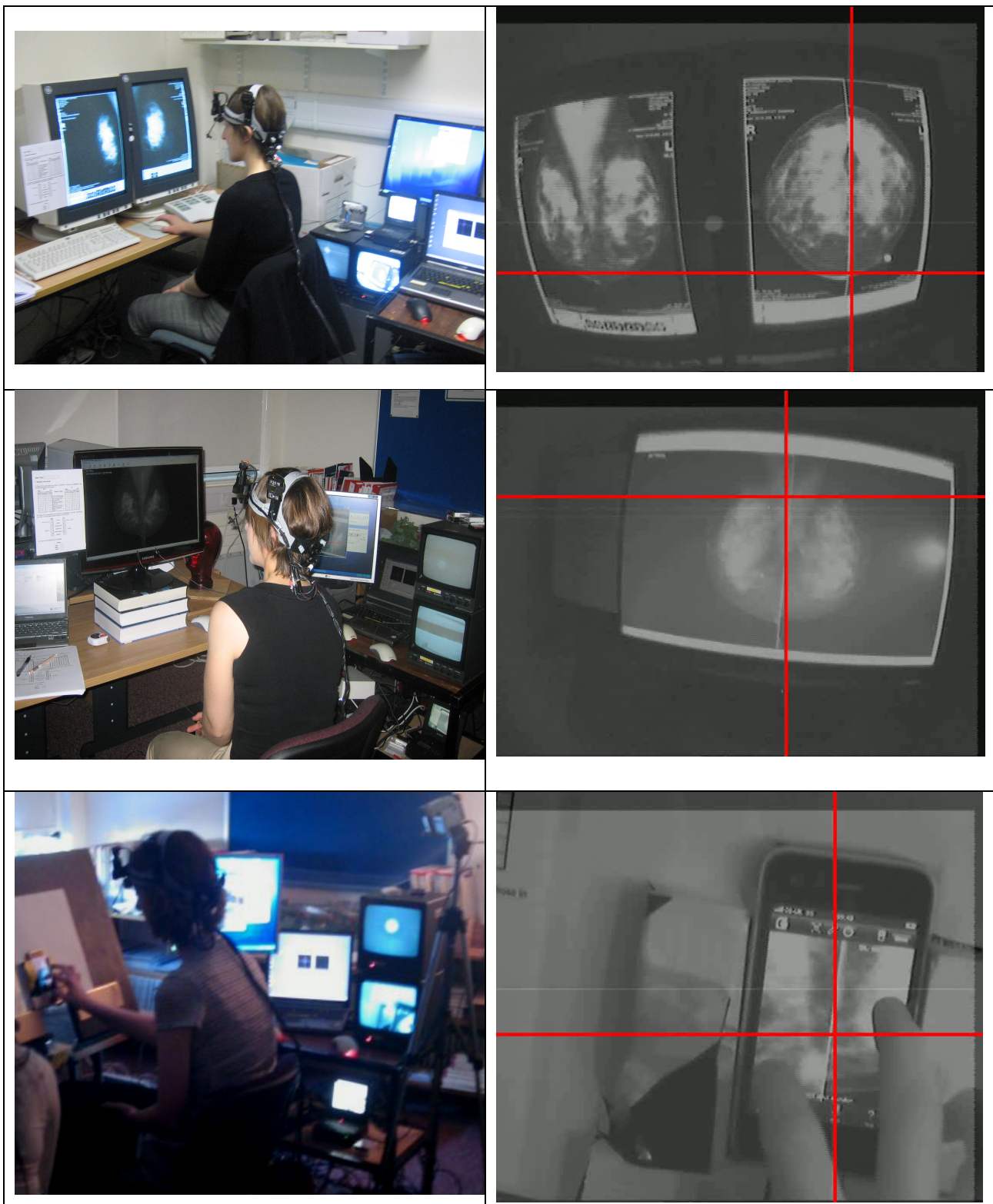


Figure 2. Examples of participants examining images on the three different displays. The ambient lighting levels were altered for photographic purposes.

interaction. Such interaction with the iPhone involved tapping the screen to select images to view; two finger movements to zoom and a single finger movement to make window/level adjustments. The height of the iPhone on the board was adjusted appropriately to suit each participant. Somewhat similarly the height of the monitor was adjusted for participants to facilitate their inspection of the images, interaction with the DICOM viewing software was by mouse. Interaction with the workstations was by means of the standard GE workstation interaction keyboard.

For each case, the participant was invited to report verbally if it was normal or abnormal, specify mammographic features, rate their confidence of abnormality presence, classify the case (six classes from Normal to Malignant) and report its density (either dense, mixed, or fatty). In the standard monitor and iPhone tasks participants first practised using the relevant DICOM viewing software. Each trial took approximately 45-75 minutes depending on each individual. For visual search comparison purposes the expert who originally selected the case set also examined the cases under the same experimental conditions.

The performance of each individual was treated anonymously and then related to their known recent performance in the UK PERFORMS self assessment scheme (where each UK screener reports on a set of difficult exemplar screening images) as well as their known real life performance data from everyday clinical screening.

3. RESULTS

Participants' performance was compared on both levels of reading experience (high or low), reading modality type (i.e. workstation, standard LCD monitor and iPhone) as well as image manipulation (i.e. with and without HCI). A repeated measures ANOVA with one between groups measure (experience level) and two within groups measures (modality type and with/without image manipulation) revealed a significant main effect of modality [$F(2, 24)=19.880, p<.001$] and a significant main effect of image manipulation [$F(2, 12)=5.803, p<.05$] but no significant effect of experience ($p=n.s$). Pairwise post-hoc statistics (Bonferroni) showed no significant differences between workstation and standard monitor modalities ($p=n.s$) but found significant differences for the iPhone and both workstation and standard monitor comparisons (see Figures 3-6 below).

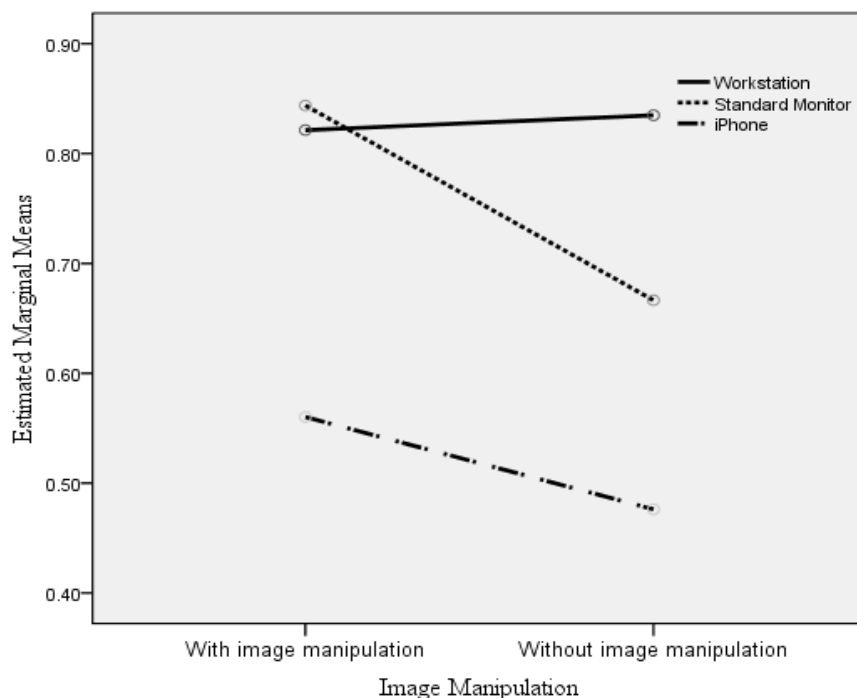


Figure 3. Performance on the three modalities with and without HCI

3.1 Modalities

Figure 4 shows the plots for each modality. JAFROC¹² analysis showed that the mean figure-of-merit (FOM) averaged over all readers was 0.9073, 0.7654, and 0.5928, corresponding to performance on the digital mammography workstation, standard LCD monitor, and iPhone respectively. A repeated measures ANOVA revealed a significant main effect of modality [$F(2, 20)=27.489$, $p<.001$]. Also, pairwise post-hoc statistics (Bonferroni) showed significant differences between all modality types ($p<.05$) whereby the workstation FOM was significantly higher than the standard LCD monitor FOM and both were significantly higher than the iPhone FOM – see Figure 3 below. To perform the JAFROC analysis two participants' data were dropped due to their lack of false positive responses. The FROC plots below (figures 4 and 5) use all 14 participants' data.

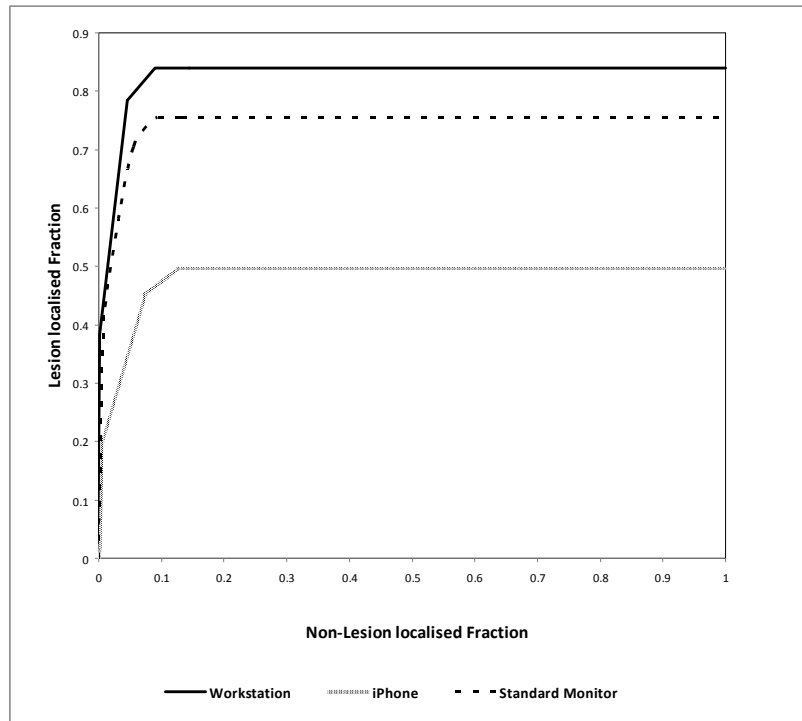


Figure 4. FROC curve of performance in digital workstation, standard LCS monitor, and an iPhone

3.2 Image Manipulation

Whilst a significant difference was found overall between performance with/without the support of image manipulation ($p<.05$), when the data were analysed by considering modality/image manipulation, further post-hoc analysis (t-tests) elicited that there was little difference in performance whether or not HCI was used when the cases were examined on the workstation although surprisingly not using HCI here was found to be slightly better but not statistically significant. No significant differences were found between the workstation with HCI and the standard monitor with HCI, i.e. with the standard monitor using HCI increased performance significantly to mirror the workstation levels. In contrast there were significant differences between workstation and standard monitor without HCI ($p<.05$). All other modality/image manipulation comparisons were significant ($p<.05$). HCI on the iPhone again increased performance although this was always a lot lower than on the standard monitor. For details see figure 5.

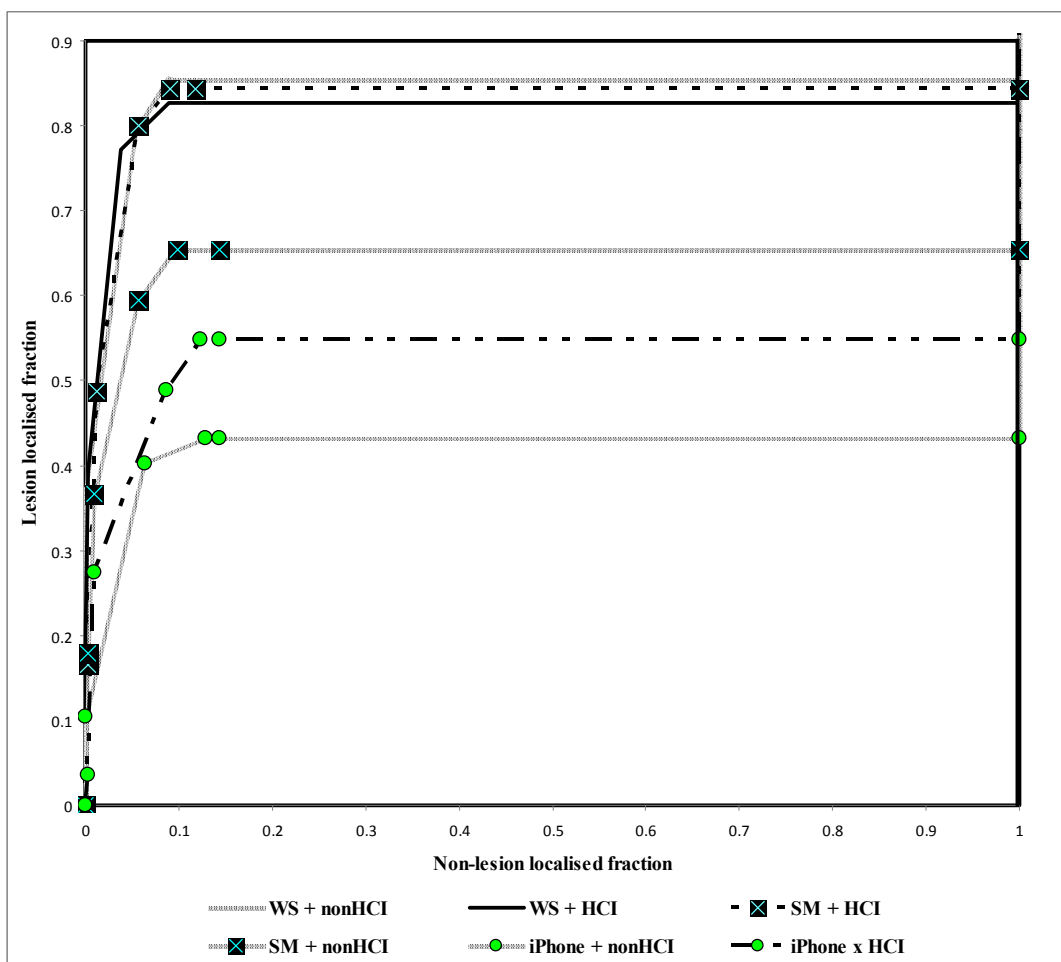


Figure 5. FROC curves of performance on workstation (W/S), monitor (LCD) and iPhone with/without HCI (nonHCI).

3.3 Experience

Participants were separated into two groups: those with over 10 years experience and those with less than ten years experience in reading screening cases. Examining the performance data split into these two experience groups then FROC analysis demonstrated little difference overall across the three displays. A repeated measures ANOVA with one between groups measure (experience level) and one within groups measure (modality type) showed no significant effect of experience level and no experience level/modality interaction. When the effect of experience was examined with each modality (figure 6) then the performance of the less experienced group using the workstation was somewhat similar (n.s) to the more experienced group using the standard LCD display. The performance of both groups with the iPhone was comparable and poor. For each modality the more experienced group performed better. Other data from the two centres show that participants' cancer detection results from the PERFORMS scheme (figure 7) are related to their real-life years of screening experience. A one-way ANOVA with one IV (group: less or more screening experience) and one DV (scores on cancer detection for self-assessment) revealed a significant group difference [$F(1, 23)=5.4, p<.05$] whereby those in the more experienced group scored significantly higher than those in the less experienced group.

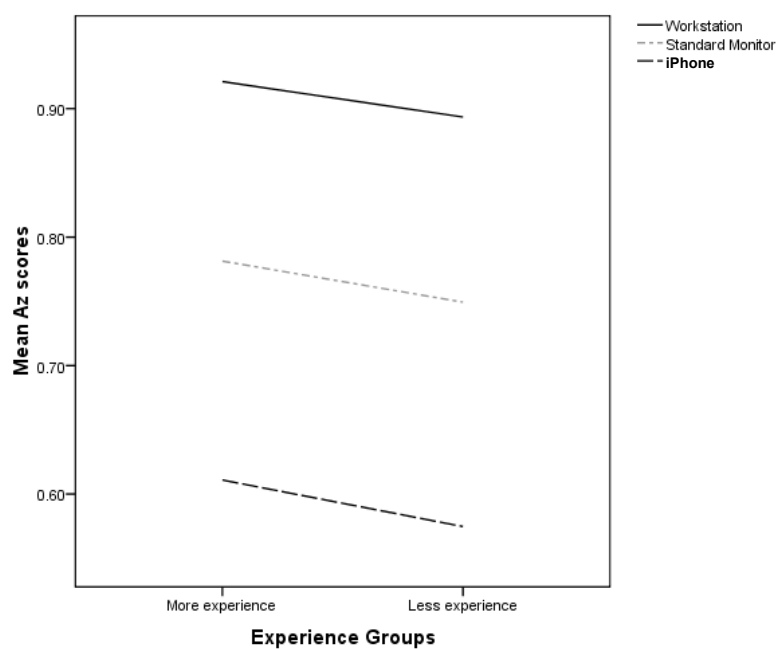


Figure 6. Experience Groups by Modality Type

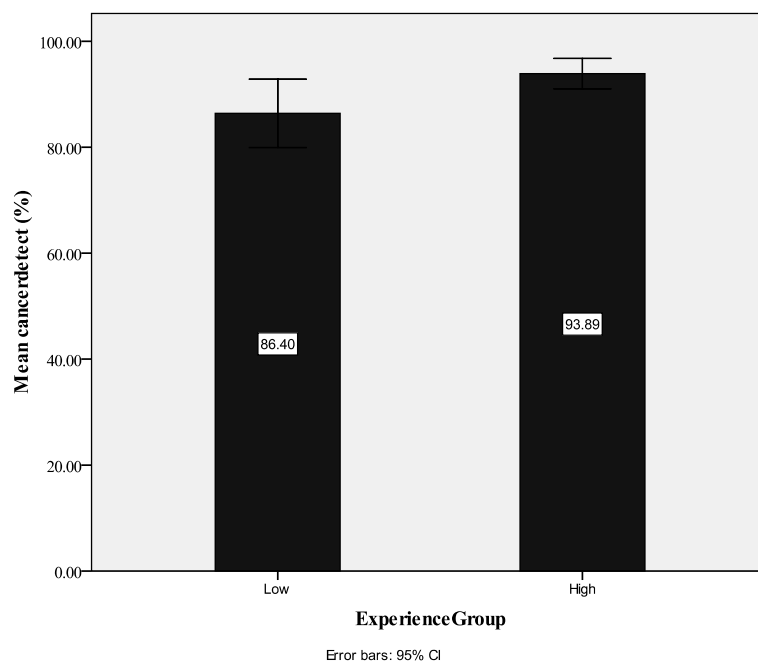
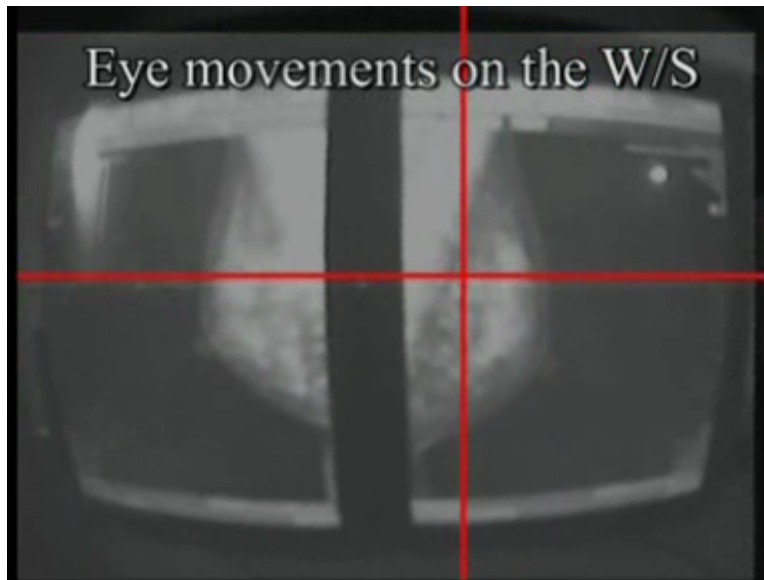


Figure 7. Mean cancer detection for the low and high experience groups on PERFORMS

3.4 Visual search behaviour

A key part of the interaction with the images was how individuals visually examined and interacted with the cases when HCI was used or not used. Detailed analyses of such visual search and interaction behaviour will be given elsewhere. However, the main finding was that the more experienced participants made fewer and longer fixations in key mammographic areas as compared to the less experienced participants. Video 1 illustrates the type of eye movements made on each modality. In each case the participant magnifies and scrolls around the image.



Video 1. Eye movement video clip of participant examining images on LCD monitor: <http://dx.doi.org/10.1117/12.843820.1>

4. DISCUSSION

This study examined how radiologists and advanced practitioner radiographers performed when examining sets of difficult recent screening cases on different modalities. The research interest is in whether a variety of display devices, which are less high resolution than clinical screening mammography workstations, can be used for training purposes in breast screening. Although specially selected recent screening cases were used here as test images we are not proposing that monitors with less resolution, or physical size, than workstations should be used for clinical screening.

A key question then is whether mammographic features can actually be displayed appropriately on such modalities so that an individual can perceive them? If when viewing a test set of images on different modalities it is possible to actually perceive key mammographic features then such modalities could be used for training purposes. Assuming this possibility, the question would then be how individuals actually interact with such modalities and whether they can navigate such displays effectively and appropriately to easily bring areas of interest into view for inspection. Another issue is how the workstation level of performance of an individual is affected when the same images are viewed on such other modalities? Although the study required the same image set to be viewed on three separate occasions no participant indicated that they remembered any case from having been presented with it previously. Additionally, no feedback was given on whether any decisions concerning features present in a case or case classification were correct or not.

Best performance in the study was, not surprisingly, attained on the clinical workstation. Whether the images were simply viewed or manipulated made a statistically significant difference for these test cases. Overall mean performance was very high on the workstations and participants were essentially reporting as they would do in routine screening. When the cases were examined on the standard LCD monitor then using HCI served to improve their performance. Using HCI with the monitor, whether participants were experienced or not, they performed well, almost as good as their

performance on the clinical workstation. This implies that using such a monitor with HCI would be useful for training purposes. Here, we used a readily available standard DICOM viewer. Performance on the iPhone was poor with or without using HCI. The iPhone performance, with or without HCI, was significantly lower than either workstation or standard monitor performance.

The iPhone is representative of a growing number of PDAs and similar devices which are increasingly being used in radiology for various purposes. Here performance on the iPhone using the Osirix software was uniformly poor. This is far from unexpected. With full DICOM mammograms being viewed on the iPhone, even with the device's excellent interaction capabilities it is hard for an individual to cognitively remember whereabouts they are when zooming in and panning around the breast images. Of particular interest was the performance of one person who reported and located correctly all the small calcifications on the images on the iPhone. This then demonstrates that the iPhone is fully capable of displaying such small features. The poor performance may well then relate to participants not being able to navigate appropriately to that part of the image and therefore not being able to potentially see the features.

In terms of experience, whilst other data indicates that experience improves performance here no difference was found between the two experience groups. This may be due to the low number of participants or cases and this will be explored in a future study.

5. CONCLUSION

Whilst superior performance was attained using the clinical workstations, participants were able to identify abnormal features on both the standard LCD monitor and iPhone. In general, using image manipulation improved performance across the modalities. On the standard monitor it actually increased performance to workstation levels indicating that using such displays with suitable manipulation software is realistic adjunct to workstations for training purposes. Results for the iPhone were disappointing possibly reflecting the difficult task of displaying very large images on this device. Further analysis of the eye movement data will yield insight into such difficulties. It is argued that lower resolution displays are useful for training purposes only. Improved mammographic interpretation training software for the iPhone would render it more useful.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Patnick J. (ed.) Celebrating 20 years of screening, NHS Cancer Screening Programmes
2. NHS Cancer Plan, Department of Health, 2000.
3. Tabár L., Tot T., and Dean P.B. Breast cancer: the art and science of early detection with mammography, Thieme Medical
4. Scott H.J., Gale A.G.; Breast screening: PERFORMS identifies advanced mammographic training needs. British Journal of Radiology, 79 (2006), S127-S133.
5. Cowley, H., Gale A. & Wilson, R. "Mammographic training sets for improving breast cancer detection." in Medical Imaging 1996: Image Perception and Performance edited by Miguel P. Eckstein & D.P. Chakraborty, Proceedings of SPIE Vol. 2712, pp. 102-112.
6. Gale A.G.: Human response to visual stimuli. In Hendee W. & Wells P. (Eds.) Perception of Visual Information - second edition, (New York) Springer Verlag, 1997.
7. Boonn WW, Flanders AE. Survey of Personal Digital Assistant Use in Radiology1. Radiographics 2005;25(2):537.
8. Chen Y. & Gale A.G.: Mammographic interpretation training: how useful is handheld technology? In Image Perception, Observer Performance, and Technology Assessment. D. Manning and B Sahiner (Eds.) Proceedings of SPIE 2008

9. Chen Y., Gale A.G. & Scott H.: Anytime, Anywhere Mammographic Interpretation Training. In P.D. Bust (Ed.) Contemporary Ergonomics 2008. Taylor & Francis, London, 2008.
10. Chen Y, Gale A, Scott H, Evans A, James J. Computer-Based Learning to Improve Breast Cancer Detection Skills. In Jacko J.A. (Ed.). Proceedings of the 13th International Conference on Human-Computer Interaction. Part IV: Interacting in Various Application Domains; Springer; 2009, pp 49-57
11. RJ Toomey, JT Ryan, MF McEntee, J McNulty, et al. The impact of faceplate surface characteristics on detection of pulmonary nodules. In Image Perception, Observer Performance, and Technology Assessment. D. Manning and B Sahiner (Eds.) Proceedings of SPIE 2009
12. Chakraborty, D.P., Jackknife Free-Response Receiver Operating Characteristic Analysis Software, [computer software], Available at: www.devchakraborty.com [accessed on 15th January 2010].