Intelligent Computing Applications based on Eye Gaze: their Role in Mammographic Interpretation Training

by

YAN CHEN

A Doctoral Thesis

Submitted in partial fulfilment of the requirements

for the award of

DOCTOR OF PHILOSOPHY of Loughborough University

February 2011

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge my supervisor, Prof. Alastair Gale, for introducing me to medical imaging research, his guidance (although he changes his mind a lot), tremendous amount of patience (towards my endless questions), kindest encouragement (towards my millions of 'exciting' ideas) and his clever supervising strategy (he often shows me his own HUGE thesis, in order to stop me complaining and get back to work). His passion, dedication, imagination and great knowledge motivate me and have provided me with a path to follow.

I would also like to express gratitude to all my colleagues in AVRC for their continuous help and support. Especially, I would also like to thank Kevin, from whom I have learned the necessity of being pessimist; Dr. Darker is whom I have learned a lot from and also argued a lot with (only over statistics and KitKats); and Sarah and Ruth, who make me realize how little common sense I have.

Most certainly Dr. Andrew Evans, Dr. Anne Turnbull, Dr. Jonathan James and Prof. Dev Chakraborty deserve to be mentioned here. I have been very fortunate to receive their invaluable input and their kindest support. Most importantly, I am inspired by their enthusiasm and commitment within the research area.

.

Special thanks go to Mrs Ann Flynn and Baobao who have always been there for me throughout my whole time of doctorate studies.

My last and greatest thanks go to my parents for their incomparable love, endless patience and generous support for my dream.... they kept me going till the end.

雨季奉献给大地
岁月奉献给季节
我拿什么奉献给你, 我的爹娘

My PhD journey has been most enjoyable because of all of you:

ABSTRACT

Early breast cancer in women is best identified through high quality mammographic screening. This is achieved by well trained health professionals and appropriate imaging. Traditionally this has used X-ray film but is rapidly changing to utilise digital imaging with the resultant mammograms visually examined on high resolution clinical workstations. These digital images can also be viewed on a range of display devices, such as standard computer monitors or PDAs. In this thesis the potential of using such non-clinical workstation display devices for training purposes in breast screening has been investigated. The research introduces and reviews breast screening both in the UK and internationally where it concentrates upon China which is beginning screening. Various imaging technologies used to examine the breast are described, concentrating upon the move from using X-ray film to digital mammograms. Training in screening in the UK is detailed and it is argued that there is a need to extend this.

Initially, a national survey of all UK mammography screeners within the National Health Breast Screening Programme (NHSBSP) was undertaken. This highlighted the current main difficulties of mammographic (film) interpretation training being tied to the device for inspecting these images. The screeners perceived the need for future digital imaging training that could be outside the breast screening centre; namely '3W' training (Whatever training required, Whenever and Wherever). This is largely because the clinical workstations would logistically not be available for training purposes due to the daily screening demand. Whilst these workstations must be used for screening and diagnostic purposes to allow visualisation of very small detail in the images, it is argued here that training to identify such features can be undertaken on other devices where there is not the time constraints that exist during breast screening.

A series of small pilot studies were then undertaken, trialling experienced radiologists with potential displays (PDAs and laptops) for mammographic image examination. These studies demonstrated that even on a PDA small mammographic features could be identified, albeit with difficulty, even with a very limited HCI manipulation tool. For training purposes the laptop, studied here with no HCI tool, was supported.

Such promising results of display acceptability led to an investigation of mammographic inspection on displays of various sizes and resolutions. This study employed radiography students, potentially eventual screeners, who were eye tracked as they examined images on various sized displays. This showed that it could be possible to use a small PDA to deliver training.

A detailed study then investigated whether aspects of an expert radiologist's visual inspection behaviour could be used to develop various training approaches. Four approaches were developed and examined using naïve observers who were eye tracked as they were trained and tested. The approaches were found to be all feasible to implement but of variable usefulness for delivering mammographic

interpretation training; this was confirmed by opinions from a focus group of screeners.

On the basis of the previous studies, over a period of eight months, a large scale study involving 15 film readers from major breast screening centres was conducted where they examined series of digital mammograms on a clinical workstation, monitor and an iPhone. Overall results on individuals' performance, image manipulation behaviour and visual search data indicated that a standard monitor could be employed successfully as an alternative for the digital workstation to deliver on-demand mammographic interpretation training using the full mammographic case images. The small iPhone, elicited poor performance, and was therefore judged not suitable for delivering training with the software employed here. However, future software developments may well overcome its shortcomings.

The potential to implement training in China was examined by studying the current skill level of some practicing radiologists and an examination of how they responded to the developed training approaches. Results suggest that such an approach would be also applicable in other countries with different levels of screening skills.

On-going further work is also discussed: the improvement of performance evaluation in mammography; new visual research on other breast imaging modalities and using visual search with computer aided detection to assist mammographic interpretation training.

Key Words: mammography, training, visual search, eye tracking, Human-Computer Interaction, PERFORMS, performance evaluation, breast screening, NHSBSP, Breast screening in China

TABLE OF CONTENTS

ABBREVIATIONS

3D	Three-dimensional
3W	Whatever required, Whenever, Wherever
ACR	American College of Radiology
AD	Architecture Distortion
AGR	Average Glandular Dose
AOIs/ROIs	Area of Interests/Region of Interests
ASYM	Asymmetry
AUC (Az)	Area-Under-the-Curve
BI-RADS	Breast Imaging Reporting and Data System
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CAL	Micro-calcification
CBT	Computer-based Training
CC	Cranio Caudal
CPD	Continuing Professional Development
CR	Correct Recall
CS	Correct return to Screen
CT	Computed Tomography
CME	Continuing Medical Education
CR	Computed Radiology
DM	Digital Mammography
DICOM	Digital Imaging and Communications in Medicine
EOG	Electro-OculoGraphy
FFDM	Full Field Digital Mammography
FOM	Figure-of-Merit
FROC	Free-response ROC
HCI	Human-Computer Interaction
HRT	Hormone Replacement Therapy
IDM	Ill-defined Mass
JAFROC	Jackknife Alternative Free ROC

LROC	Localization ROC
MDM	Multi-disciplinary Team Meetings
MLO	Media Lateral Oblique
MRI	Magnetic Resonance Imaging
MRMC	Multiple-Reader-Multiple-Case methods
NBCCEDP	National Breast and Cervical Cancer Early Detection Programme
NBSS	National Breast Screening Service
NHSBSP	National Health Service Breast Screening Programme
NICE	National Institute for Health and Clinical Excellence
NPV	Negative Predictive Value
PACS	Picture Archiving and Communication System
PDAs	Personal Digital Assistants
PERFORMS	PERsonal perFORmance in Mammographic Screening
PPV	Positive Predictive Value
RCR	Royal College of Radiology
RIS	Radiology Information System
ROC	Receiving Operating Characteristic
ROI	Region of Interest
SDT	Signal Detection Theory
SFM	Screen-Film Mammography
SPIC	Spiculate Mass
UFOV	Useful Field of View
WHO	World Health Organization

SYMBOLS

PCR	Pupil and Corneal Reflection
θ	The eye line of gaze angle with respect to the incident light source and recording camera
K	The distance between the pupil centre and the centre of corneal curvature
S	A length on the scene
D	Distance from the subject's eye to the scene being viewed
d	Cohen's *d*, i.e. the difference between two means divided by the standard deviation of the data

CHAPTER 1

Breast Cancer: Screening and Training

1.1 Breast Cancer and Breast Screening

1.1.1 Breast Cancer

According to the World Health Organization (2006), cancer is responsible for 13% of all deaths globally and is the leading cause of death. Breast cancer, the most common type of cancer amongst women, causes around 502,000 deaths each year. It was estimated that worldwide in 2008 (Ferlay, *et al.*, 2008) some 1,380,000 million women were diagnosed with breast cancer, which is 10.9% (approximately a tenth) of all new cancers and 23% (nearly a quarter) of all female cancer cases. Figure 1-1 shows the estimation in 2008 of female breast cancer mortality rates in some selected countries around the world.

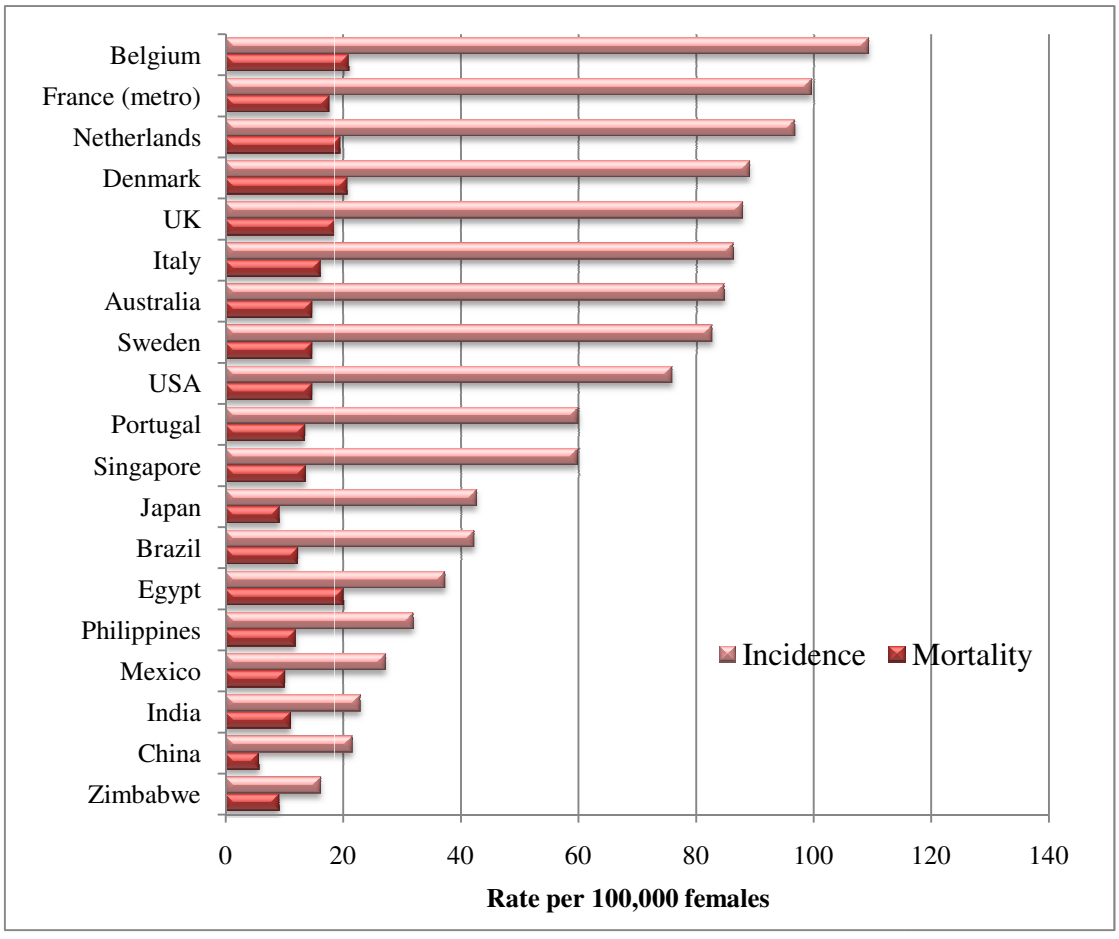


Figure 1-1. Breast Cancer, age-standardised (world) incidence and mortality rates, selected countries, 2008 estimates (Ferlay, *et al.*, 2008)

One in eight women in the USA will develop invasive breast cancer over the course of their lifetime with a 1 in 33 chance of breast cancer causing their death (American Cancer Society, 2006). The situation is similar in the UK; for instance approximately 37,000 new cases were diagnosed in England in 2004, with over 10,000 deaths in 2005 (Department of Health, 2007). The mortality rate amongst women in the UK is 28 per 100,000 women (National Statistics Online, 2007). The European Union, one of the highest breast cancer incidence areas, is estimated to have had 332,000 new cases of breast cancer in 2008 (Ferlay*, et al*., 2008); with another high incidence country, the USA, having an estimated 182,460 cases occurring each year (American Cancer Society, 2008). With the advent of screening for breast cancer and better ways of treatment death rates from the disease have been declining in recent years in the USA (Espey*, et al*., 2007) and the UK (Society for Women's Health Research, 2005), however breast cancer still remains the most feared disease.

Somewhat in contrast, the incidence of breast cancer is increasing in developing countries. To take China as an example (addressed in the research in this thesis) with a fifth of the world's female population, the incidence of breast cancer among its 1.3 billion population is rising steadily, see figure 1-2, partly as techniques to detect the disease are implemented (e.g. some breast cancer screening is now being undertaken - but not nationally) and partly as Asia increasingly adopts more Western lifestyles and eating habits. Li*, et al*. (2010) recently reported that in the past 20 years the incidence of breast cancer in young women in China has increased by 80%. In a somewhat related fashion earlier research (Yang*, et al.,* 2005) pointed to the incidence of breast cancer in Chinese women having increased from 19.9 patients in 100,000 women in 2002 to 24.5 patients in 100,000 in 2005. In 2005 the World Health Organisation predicted that 170,000 new cases would be diagnosed, resulting in a 155% rise in detected breast cancers in China between 2000 and 2005 (Diagnostic Imaging, 15th August, 2005). Zhang*, et al*. (2010) confirmed that the diagnosis and treatment of breast cancer in China is less than ideal and in researching the coping styles of Chinese breast

cancer patients point to a positive relationship between coping styles, their hope and financial income which allows them access to better healthcare.

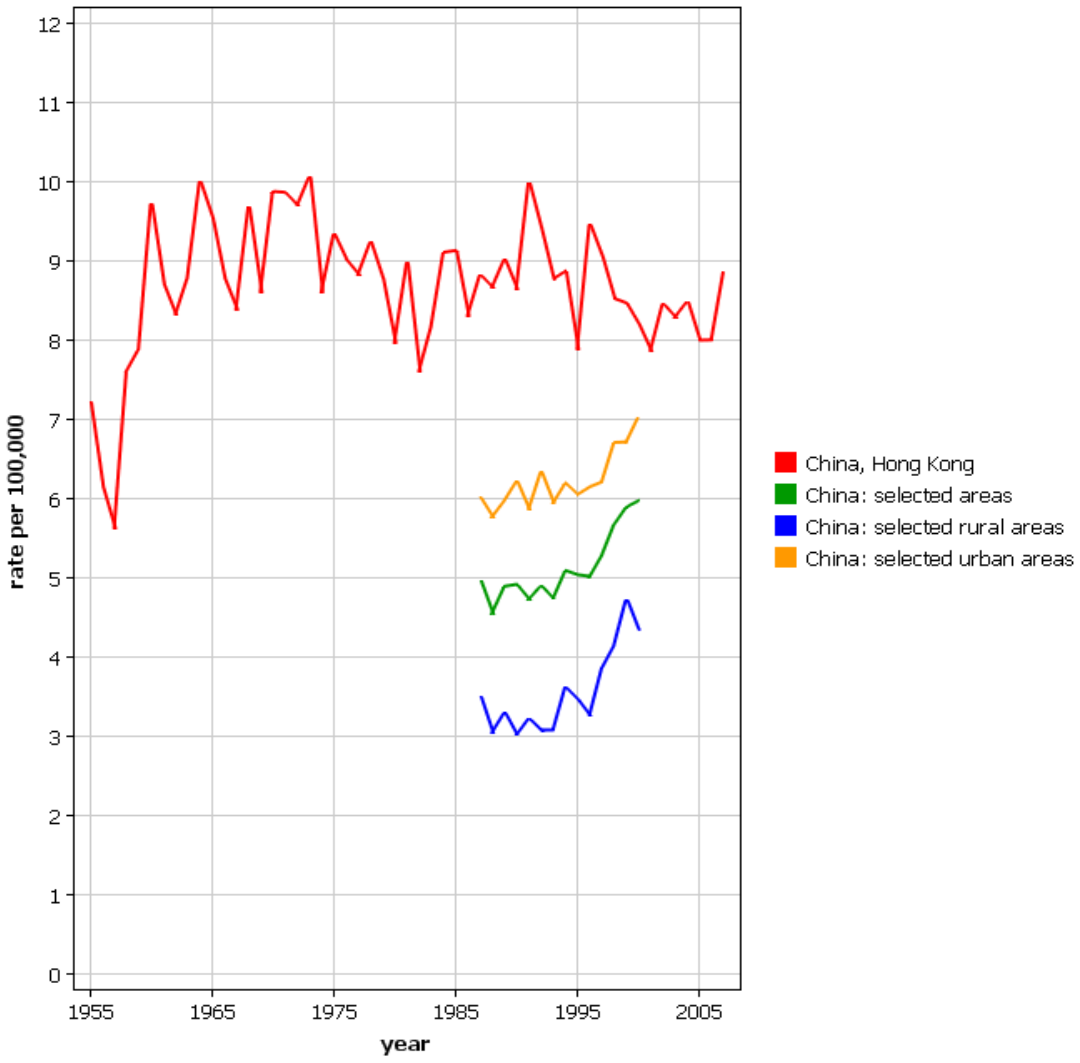


Figure 1-2. Breast Cancer, age-standardised (World) incidence and mortality rates at all ages, in different areas of China, 2008 estimates (Ferlay, *et al*., 2008)

Zheng, *et al.* (2005) detailed the epidemiological Shanghai Women's Health Study which recruited some 74,942 adult Chinese women from selected urban communities from 1996 to 2000. Data from this cohort were then used by Linos (2008) to predict breast cancer incidence in China in future years. This estimated that the incidence rate of the disease would increase significantly from the then current estimated rate of 10 – 60 cases per 100,000 women to more than 100 cases per 100,000 women aged 55 – 69 years by 2021 (with 2.5 million cases of breast cancer by 2021 among Chinese women who were

aged 35 – 49 years old in 2001) This dramatic increase of incidence rate was linked to changes in reproductive behaviour (after the one child policy took place in China in 1979 the average birth rate fell from 5.9 births per woman in 1970 to 2.9 in 1979 and 1.7 in 2004 (Hesketh*, et al.*, 2005) and lifestyle risk factors such as weight gain, alcohol consumption and the use of hormone replacement therapy (HRT).

Variations exist in estimated figures for the incidence of breast cancer in China as it is difficult to adduce full detailed figures for the whole country due to its geographic size and large population. Some data have been produced for Hong Kong and parts of China by Ferlay*, et al.* (2008) as illustrated in figure 1-2.

Wide ranging suggestions, e.g. encouraging a healthy diet, physical activity and control of alcohol intake, overweight and obesity have been proposed by the WHO to have a positive impact in reducing the incidence of breast cancer in the long term. However, in low-income and middle-income countries, such as China, simple strategies which aim to eliminate the risk of having breast cancer simply will not reduce the majority of breast cancers. Breast cancer control, in all countries, will fundamentally rely on early detection in order to improve breast cancer outcome and survival (Anderson*, et al.*, 2008).

1.1.2 Breast Cancer Screening

The best way to detect breast cancer at an early stage is by breast screening. This is the process whereby healthy women are regularly invited for routine breast screening, usually by using mammography. The flow chart in figure 1-3 illustrates the screening process in the UK.

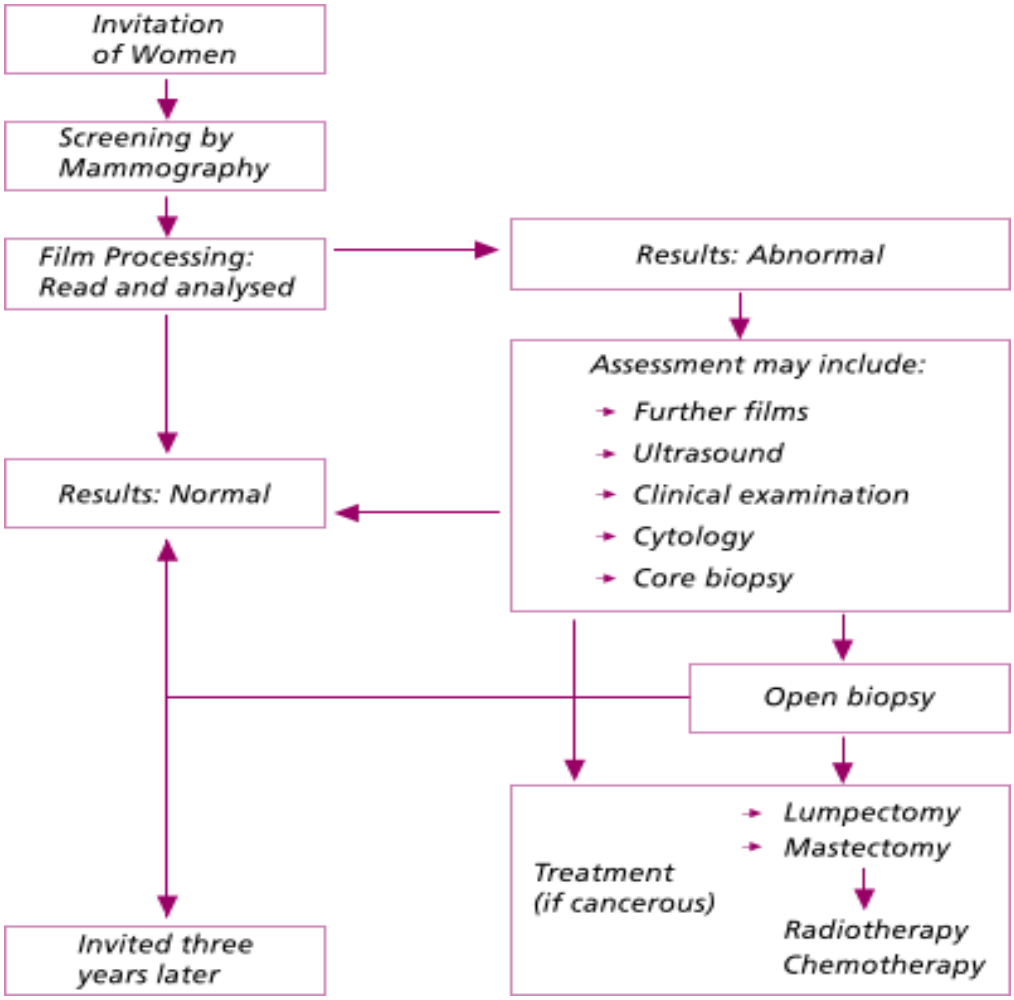
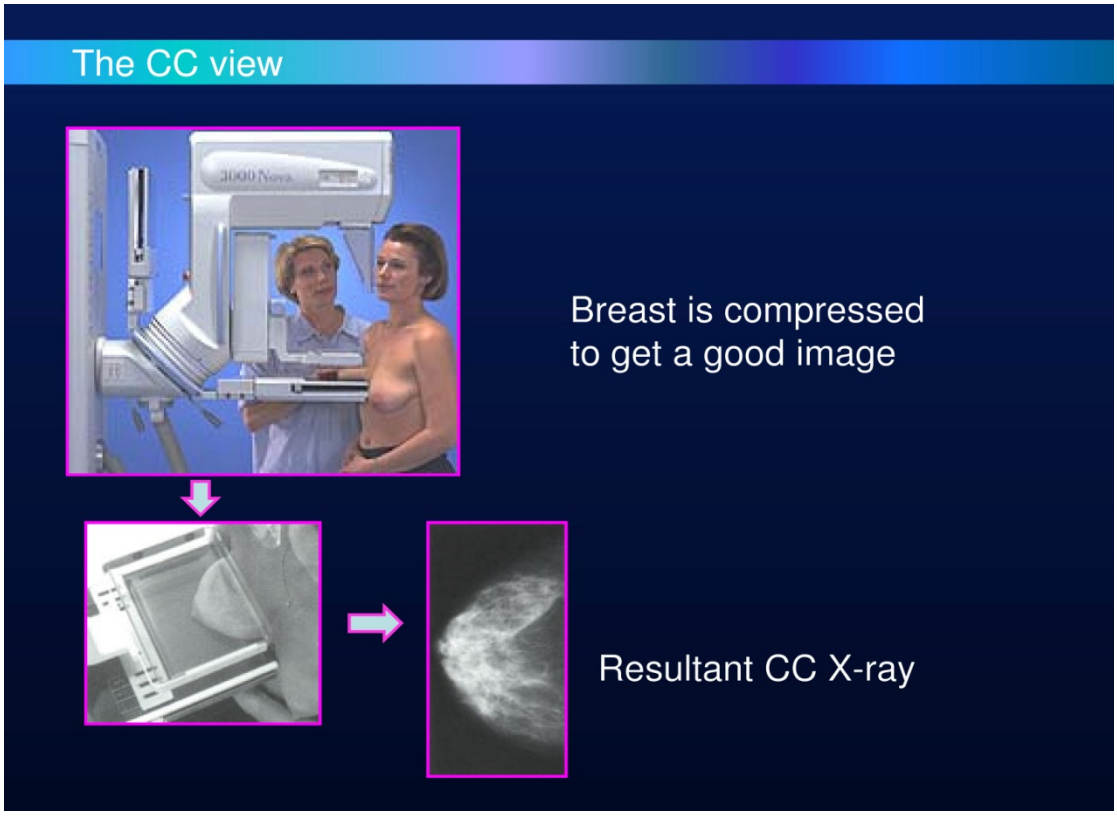


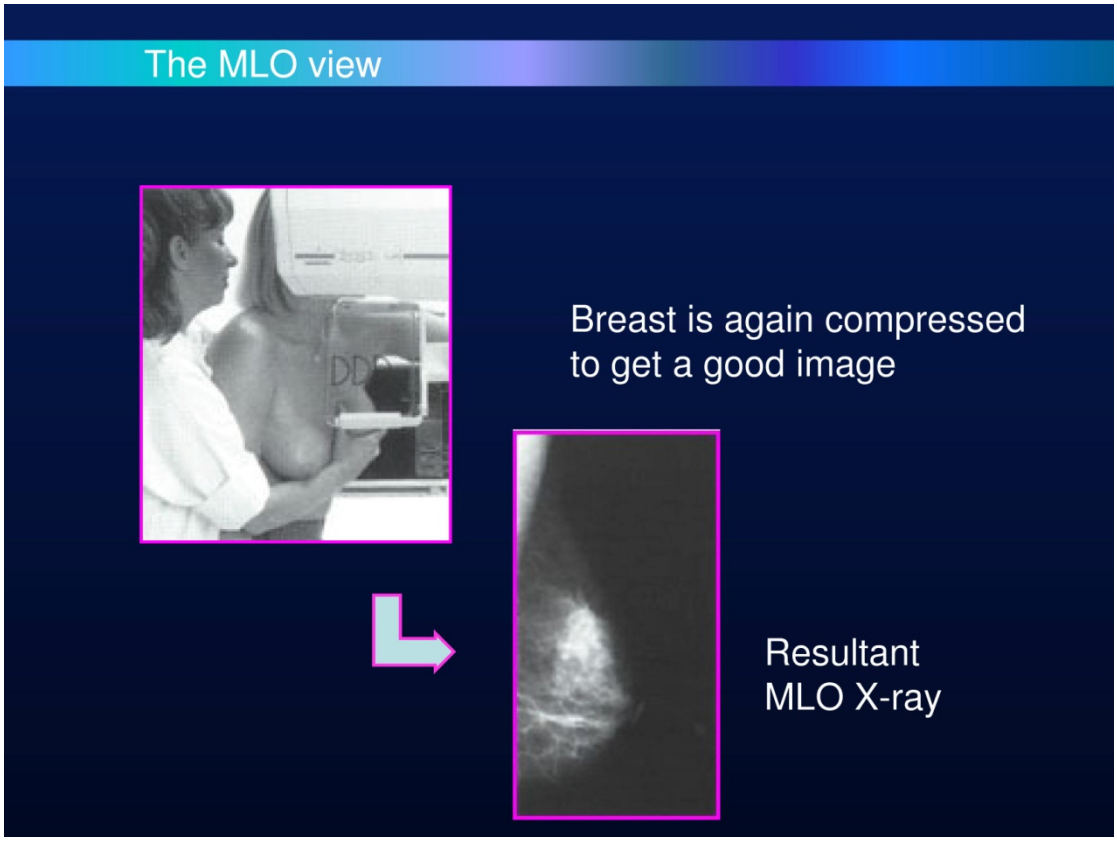
Figure 1-3. The breast screening process (NHSBSP, 2010).

A woman is invited and attends for screening by mammography. The first key step is taking mammographic images of both breasts. Currently, two X-ray images of each breast are taken which provide different radiological views through the breast:

- The Cranio Caudal (CC) view - this is a vertical view through the breast (figure 1-4a);
- The Media Lateral Oblique (MLO) view - this is an angular view at 45^{o} – it allows imaging of the glands under the arm (figure 1-4b). For each view, the breast is extended and compressed gently in the breast imaging unit to achieve somewhat uniform imaged breast tissue thickness so that a good image of the whole breast is obtained



a)



b)

Figure 1-4. How both mammographic views are taken for each woman

Films are processed and then read (i.e. interpreted) by experienced personnel. If the case is considered normal then the woman is simply invited for screening again three years later. If abnormal then assessment is carried out which may result in biopsy, surgery or therapy appropriately.

The efficient and accurate interpretation of these images for the presence of abnormalities that are indicative of cancer is the focus of the research presented in this thesis. Figure 1-5 illustrates how mammograms are examined by a film reader on a mammogram multi-viewer. Each woman is represented by four images - the two views of each breast.

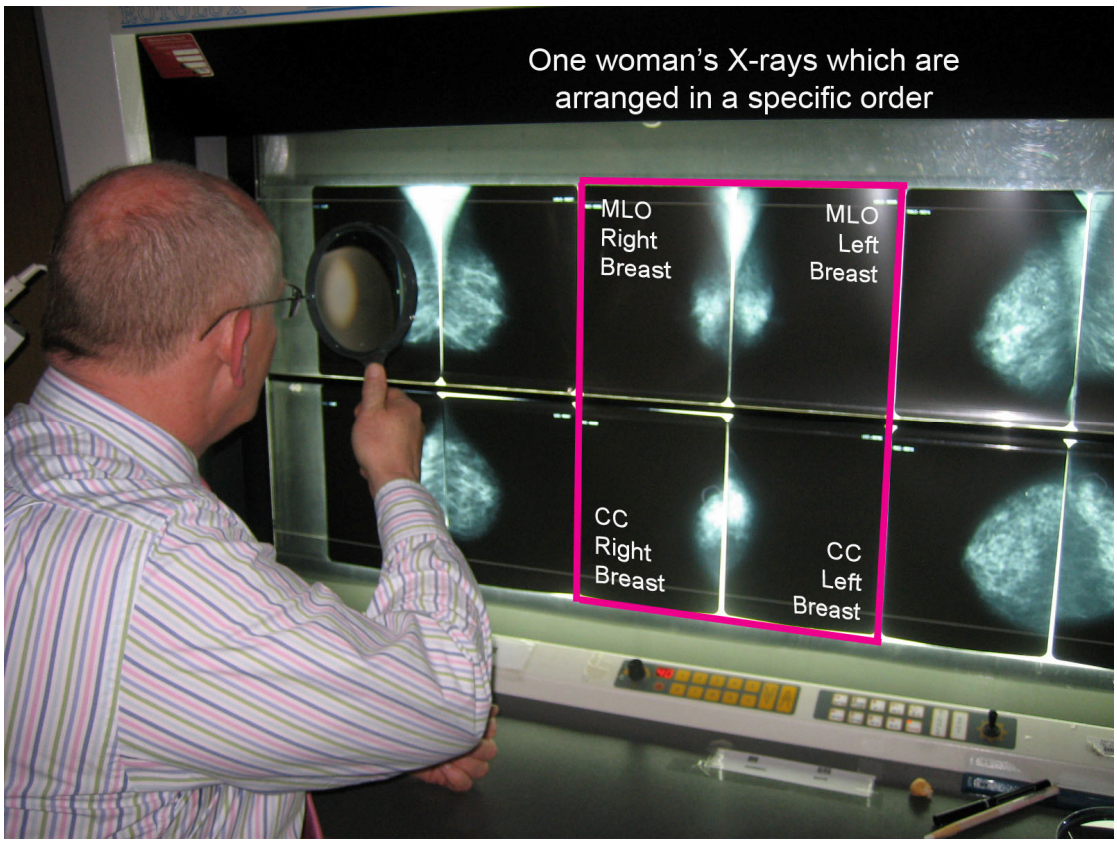


Figure 1-5. A film reader examining films on a mammogram multi-viewer

The mammographic interpretation task is carried out by trained mammographic film-readers (in the UK these are primarily specially trained radiographers or consultant radiologists, although increasingly other clinical specialists are also involved). Detecting early signs of breast cancer is an especially difficult task due to the rarity of the disease within the screening population. Despite a woman having a one in nine chance of contracting

breast cancer at some time in their lives (which inherently implies widespread incidence), an abnormality will be present in only approximately seven cases per 1,000 screened women (Patnick, 2005). Furthermore, the numerous subtle ways in which an abnormality can present increases the difficulty of correct identification of malignancies.

Breast Cancer Screening in the UK

The timescale of the development of breast cancer screening in the UK is shown in figure 1-6. In 1986, an expert committee chaired by Professor Sir Patrick Forrest examined the available evidence on breast screening and then presented a report to the Minister of Health. In the report, it was concluded that:

'Screening by mammography can lead to prolongation of life for women aged 50 and over. There is a convincing case on clinical grounds for a change in UK policy on the provision of mammographic facilities and the screening of symptom-less women.' (Department of Health, 1986)

This then led directly to the initiation in the UK of the National Health Service Breast Screening Programme (NHSBSP) by the Department of Health in 1988. This was one of the first nationwide schemes of its kind in the world. With the purpose of facilitating the early detection of breast cancer and improving treatment, the scheme was established to provide free breast screening every three years for all women in the UK aged 50 to 64 years initially using a single mammographic view of each breast (the MLO view). A subsequent report by the Pritchard Committee (1988) set up appropriate quality assurance guidelines to ensure high quality screening. One of these requires that all individuals participating in the UK screening programme read at least 5,000 screening cases a year. In 1992/3 the second screening round took place since screening was initiated and this successfully screened 1.2 million women detecting 6,597 cancers. In 2008 twenty years of screening was celebrated (Patnick, 2008) with data from 2007/8 showing that two million women were screened with 16,500 cancers detected of which 6,878 were small (<15mm) invasive cancers. The overall figures show that (in 2009) 18

million screening cases were examined and 100,000 cancers detected which approximates circa 100 cancers being identified nationally per week since 1988.

In 1995, two view screening, i.e. two view images (MLO and CC) of each breast, was introduced instead of the single MLO view, for the incident round of screening (i.e. the first time a woman attends for screening) after a randomized controlled trial indicated a 24% increase in cancer detection rates for two-view mammography as compared to single (MLO) view mammography (Wald*, et al.*, 1995). Subsequently, two views screening at each screening round was introduced in England in 2000 with the publication of the NHS Cancer Plan (2000); this also recommended extending the screening age range to 50-70 years of age by 2014.

By 1992/3 around one-and-a-half million women were screened in the UK each year. By 2007/8, some five million women had been screened since the NHSBSP began in 1988 (Patnick, 2008). With the introduction of the scheme, the death rates from breast cancer began to fall as a result both of early cancer detection along with improved treatment. Based on the predictions of pre-screening rates in various age groups, by 1998 the death rate was about 20% lower than it would have been without the screening scheme (National Statistics Online, 2007). Digital mammography (detailed further in Section 1.2.2) was first introduced in the UK in 2007 following publication of the Cancer Reform Strategy (2007) which also further recommended extending the screening age range to 47-73 years of age. Women younger than 50 years tend to have dense breasts which are hard to interpret using Screen-Film Mammography (SFM) but digital mammography allows such images to be interpreted easier and so this technology allows the starting screening age to be lowered. By 2010 it was planned for all breast screening centres to have some digital mammography with all UK screening to be fully digital by 2012.

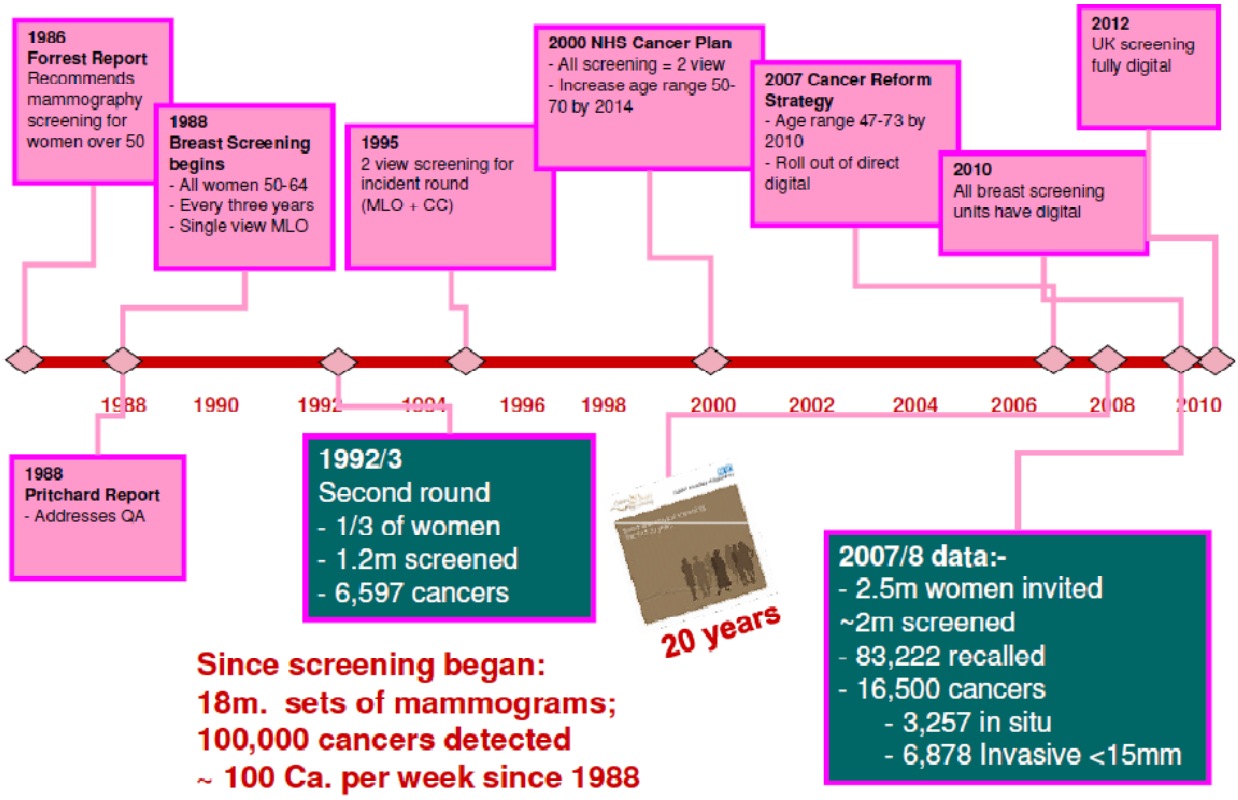


Figure 1-6. The process of breast screening in the UK

Breast Screening Workforce in the UK

Screening nationally takes place via some 110 screening centres which are manned primarily by radiographers and radiologists. Before October 2000, the organization of the breast screening centres ran under a two tier structure. Radiographers were responsible for taking the mammograms and assessing image technical quality; with radiologists, or sometimes breast clinicians, in charge of examining the mammograms and deciding which women should be recalled for further assessment (Department of Health, 2007).

As the NHSBSP expanded and more women were screened then to cope with the increased workload changes were necessary for the organisation of the screening workforce within each centre which was expanded into four tiers:

- A lead practitioner - registered practitioner (for example, radiologist, breast clinician, radiographer) who leads the clinical team;

- Advanced practitioner - registered practitioner (e.g. radiologist, breast clinician, radiographer) with advanced training to undertake film reading, breast ultrasound and breast investigative procedures;
- Practitioner - state registered (e.g. radiographer) undertaking all practical aspects associated with mammography imaging with an additional supervisory role for assistant practitioners, including mentoring and training support;
- Assistant practitioners - someone trained to carry out mammograms under supervision of a practitioner.

Every screening centre now has several radiologists and advanced practitioners who can read and interpret the screening cases, together with a range of other staff who take the mammograms and carry out other tasks. Screening centres implement some version of 'double reading' of every screening case. This is where at least two individuals read and report on a case. If a disagreement exists between their opinions then a third individual arbitrates. Double reading should be carried out where each person is blind to the opinions of the other but this varies between centres on what they do in practice. In the 'New ways of working in the NHS Breast Screening Programme' report (Nickerson & Cush, 2004) good progression since 2002 was found, both in terms of the implementation of new workforce changes in the breast screening programme and in the expansion of the programme.

Breast screening has been undertaken across the UK for over 20 years now using mammographic film as the imaging medium. Recent developments are seeing the age range being increased. As mentioned above the upper age limit of women invited to take up screening in the UK has been extended from 65 to 70 years to encompass women aged 47-73. This increased age range increases the number of cases annually examined in the UK's breast screening centres. Thus, the combination of increased demand and limited capacity places pressure on the NHSBSP. It was estimated in 2002 by the Royal College of Radiologists Breast Group that there was a 40% shortfall in radiology staff (Castledine, 2002). This shortfall has largely been addressed by the increased roles undertaken by radiographers as Advanced

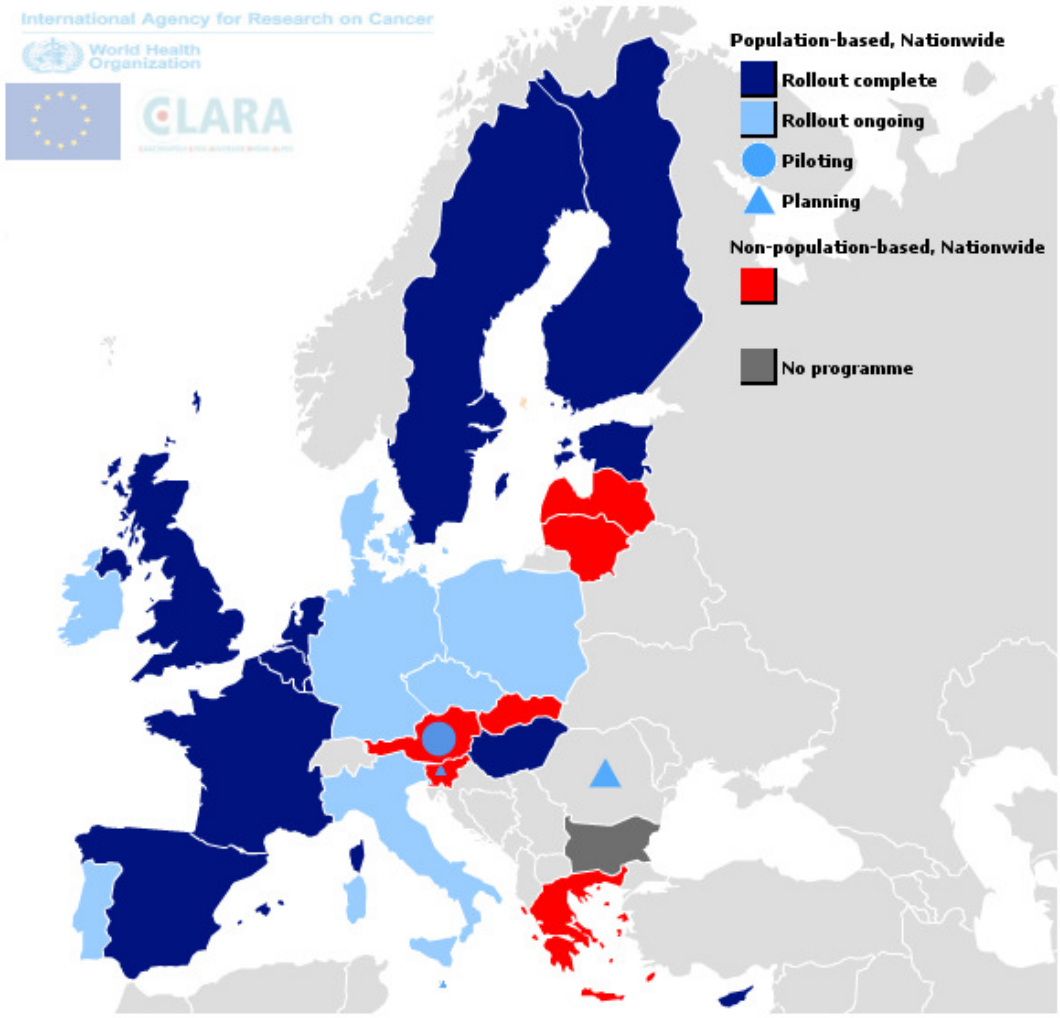
Practitioners. However there is a constant ongoing need for training and the digital mammography introduction also requires training implementation.

International Breast Cancer Screening

In a similar fashion to the NHSBSP, there have been breast cancer screening programmes initiated in other countries. In the European Union (27 member countries) in 2007, some 22 population-based programmes were running or being established and in two countries there was nationwide non-population based screening. See figure 1-7 for details.

According to the European Guidelines on Quality Assurance in Mammography (Perry*, et al*., 2008), all women aged 50-69 are recommended to have mammography screening. Tables 1-1 and 1-2 give various details of breast cancer screening programmes in 22 different countries (including 19 European countries, Japan, USA, Canada and Uruguay) that have population-based breast cancer screening programmes. From table 1 it is clear that in 1995 broadly similar age ranges of women were being screened; generally over 50 years old but with some countries screening from the age of 40 years even though using SFM to screen such young women would make identifying small cancers difficult. The introduction of FFDM some 10 years ago (Bick & Balleyguier, 2010) facilitates screening these younger women.

In general, most countries screen more regularly (every two years or even every year) than the three year period used in the UK. A recent review of evidence for screening has proposed that the UK should move to a two year screening interval as this would enable detecting more cancers at an earlier treatable stage (Evans, 2010). Table 1-2 shows further details of how individual screening programmes are organised in the International Breast Cancer Screening Network.



Breast screening programmes in the European Union in 2007, by programme type (population-based; non-population-based; no programme) and country implementation status (population-based: nationwide or regional, rollout complete or ongoing, piloting and/or planning; non-population-based: nationwide or regional). Programmes shown use screening test (mammography) recommended by the Council of the European Union in 2003

Figure 1-7. Overview of Breast cancer screening programmes in the EU Member States in 2007 (IARC, 2007)

Table 1-1 Summary of guidelines most commonly used or recommended to use in population-based breast cancer screening programmes in 22 countries surveyed in 1995 (Shapiro, *et al.*, 1998).

Country	Age groups covered[a]		Screening interval-years[b]		Detection methods[c]
	Lower limit	Upper limit	Age 40–49[d]	Age 50+	
IBSN programme countries					
Australia[e]	40	69	2	2	MM
Canada[f]	50	69	1[f]	2	MM,CBE,BSE[f]
Finland	50	59	NA	2	MM
Hungary	50	64	NA	1	MM,CBE
Iceland	40	69	2	2	MM,CBE[g]
Israel	50	74	1 High risk	2	MM
Italy	50	69	NA	2	MM
Japan	30	None	1	1	CBE, BSE
The Netherlands	50	69	NA	2	MM
Sweden	40–50[h]	64–74[h]	1.5	2	MM
United Kingdom[i]	50	64	NA	3	MM
United States[j]	40–50	None	1	1–2	MM,CBE,BSE
Uruguay[k]	45	None	1	2	MM, CBE,BSE
European Network pilot projects					
Belgium	50	69	NA	2	MM
Denmark	50	69	NA	2	MM
France	50	65–69	NA	2–3	MM
Germany[l]	50	None	1 High risk	2	MM
Greece[m]	40	64/none	2	2	MM,CBE,BSE
Ireland	50	65	NA	2	MM
Luxembourg	50	65	NA	2	MM,CBE
Portugal	40	None	No Rec	2	MM
Spain	45	64	2	2	MM

[a] Age groups covered for screening mammography except for Japan which does not use mammography for screening.

[b] Refers to screening interval for mammography if mammography is used.

[c] Mammography (MM), Clinical Breast Examination (CBE), Breast Self-examination (BSE).

[d] NA = Not applicable as age group was not included in screening guidelines. No Rec = No recommendation regarding screening interval was included in national guidelines.

[e] Women under or over the stated age limits may be screened if they so request. State policies regarding screening interval for age group 40–49 vary.

[f] Although national guidelines recommend screening begin at age 50, provinces establish province-specific policies. In British Columbia, policy was to start screening at age 40 on an annual basis. CBE is added in three provinces; BSE is added in seven provinces (passively with use of a video in four and actively by an individual in three).

[g] See footnote[g] in Table 2 regarding use of CBE in screening programmes.

[h] Age groups that are covered are decided by each county.

[i] Women over age 65 may be screened if they so request.

[j] In the US guidelines differ and no single source is used for national policy. The American Cancer Society (ACS) and several professional organizations specify age 40 for the initiation of screening. The US Preventive Services Task Force recommendation is 1–2 years for women aged 50 and over. ACS guidelines note annual screening for women aged 50 and over; 1–2 years for women aged 40–49 (modified in March 1997 to 1 year for women aged 40–49).

[k] In Uruguay recommended screening interval is 1 year for age group 45–55 and 2 years for women over age 55. In Uruguay there are two programmes. One is based only on CBE and BSE; the other uses MM,CBE, and BSE.

[l] Describes plans in Germany as of 1995 which were not further implemented.

[m] Both screening programmes in Greece recommend initiating screening at age 40 and advise 2 years as the screening interval for all ages. In Greece I, MM is the only detection method in the screening programme, the upper age limit recommended in guidelines is 64 years for women initially screened, and presence of risk factors does not alter either the age or interval recommended for screening. Once a woman enters the screening programme she is included in rescreening irrespective of age. In Greece II, MM, CBE and BSE are included in the programme, there is no upper age limit for inclusion in the screening programme, and presence of risk factors reduces the interval recommended for screening to 1 year.

Table 1-2 Organization of screening mammography programmes represented in the International Breast Cancer Screening Network, 1998 (Klabunde, *et al.*, 2002)

	Organized Screening	Year implemented	Target population[1] (%)	Programs or registries (#)	Facilities[2] (#)	X-ray units (#)	Organization of quality assurance[3]	Linkage to cancer registry[4]
National programs								
Australia	Yes	1991	54	**1**	35	35 +	National	Yes[5]
Finland	Yes	1986	100	**1**	11	97	National[6]	Yes[5]
Iceland	Yes	1987	100	1	3	5	National	Yes[5]
Israel	Yes	1997	100	1	42	44	National	Yes[5]
Luxembourg	Yes	1992	98	**1**	10	10	National	Yes[7]
The Netherlands	Yes	1989	100	1	55	62	National[6]	Yes[5]
United Kingdom	Yes	1988	100	1	94	315	National, state	Yes[5,7]
State, provincial, local programs								
Established program								
Belgium[8]	No			1	35	40	Facilities	No
Canada[9]	Yes	1988	30	10	163	179	National, state	Yes[5,7]
Denmark	Yes	1991	18	**1**	1	1	State	Yes[5]
France[9]	Yes	1989	25	26	900	1000	National[6]	Yes[5,7]
Italy[6]	Yes	1990	10	22	26	43	State	Yes[7]
Norway	Yes	1995	40	1	7	15	National	Yes[5]
Portugal	Yes	1990	20	**1**	7	7	State	Yes[7]
Spain	Yes	1990	60	1	2	2	State[6]	Yes[5,7]
Sweden[9]	Yes	1986	100	26	26	60	National[6]	Yes[5]
Pilot program								
Germany[10]	Yes	1999	2	3	8	8	National, state	Yes[5]
Greece	Yes	1989	25	3	7	7	State	No
Hungary	Yes	1991		1	10	10	State[6]	No
Japan	Yes	1989	30	**1**	2	3	State	Yes[7]
Uruguay	Yes	1996	20	1	1	2	Facilities	Yes[7]
Registry program								
United States[8]	No			8[11]	128	166	National, state[6]	Yes[5]

[1] Percentage target population: the proportion of the national population of women deemed eligible to attend for screening and covered by organized screening programs within the country.
[2] Mammography facility: a location at which women obtain screening mammography. Contains one or more X-ray units plus the staff required to perform the procedure.
[3] Organization of Quality Assurance: state — state, provincial, local
[4] Data for the screening program or mammography registry system are linked to cancer registry data.
[5] Linkage of screening and cancer registry data is computerized.
[6] Countries in which screening mammography quality assurance is required by law.
[7] Linkage of screening and cancer registry data is manual.
[8] Year implemented and percentage of target population covered by organized screening are blank because at the time of the survey, Belgium and the United States had not implemented organized, population-based screening programs.
[9] Countries with a decentralized (subnational) organization of screening mammography, but in which the totality of organized screening programs attains national coverage.
[10] Program in implementation phase.
[11] The number of geographical regions participating in the mammography registry program.

Table 1-3. Cancers Detected per 1000 Screening Mammograms by Age, Setting, and Screening Cycle (Smith-Bindman, *et al*., 2003)

	Rate per 1000 (95% Confidence Interval)					
	First Screening Mammogram			Subsequent Screening Mammogram		
Age, y	BCSC	NBCCEDP	NHSBSP	BCSC	NBCCEDP	NHSBSP
			Total			
50-54	5.8 (4.5-7.3)	5.9 (5.0-6.8)	6.3 (6.1-6.5)	2.6 (2.4-2.9)	2.8 (2.4-3.1)	3.8 (3.5-4.0)
55-59	7.4 (5.4-9.8)	8.1 (6.8-9.3)	9.2 (8.2-10.2)	3.6 (3.3-4.0)	3.5 (3.1-3.9)	4.9 (4.7-5.1)
60-64	10.1 (7.5-13.2)	11.9 (10.2-13.6)	11.9 (10.6-13.4)	3.9 (3.5-4.4)	3.7 (3.3-4.2)	5.9 (5.6-6.1)
≥65	14.4 (12.3-16.8)	8.8 (6.3-11.3)	16.6 (12.2-22.0)	5.2 (4.9-5.6)	4.4 (3.5-5.3)	8.7 (7.7-9.8)
All*	8.6 (7.9-9.4)	8.3 (7.7-8.7)	10.1 (9.4-10.7)	3.6 (3.5-3.7)	3.4 (3.3-3.6)	5.4 (5.2-5.5)
			Invasive			
50-54	4.5 (3.4-5.9)	4.6 (3.8-5.4)	4.9 (4.7-5.1)	1.9 (1.6-2.1)	1.7 (1.4-2.0)	3.0 (2.8-3.2)
55-59	6.8 (4.9-9.1)	6.0 (4.9-7.1)	7.7 (6.8-8.6)	2.9 (2.6-3.3)	2.4 (2.1-2.8)	3.9 (3.8-4.1)
60-64	7.7 (5.5-10.5)	8.9 (7.5-10.4)	9.5 (8.3-10.8)	3.0 (2.6-3.4)	2.6 (2.2-2.9)	4.9 (4.7-5.1)
≥65	12.4 (10.4-14.6)	7.1 (4.9-9.4)	14.9 (10.7-20.0)	4.2 (3.9-4.5)	2.9 (2.2-3.6)	6.9 (6.1-7.9)
All*	7.2 (6.5-7.8)	6.3 (5.9-6.7)	8.4 (7.8-9.0)	2.8 (2.7-2.9)	2.3 (2.2-2.4)	4.3 (4.2-4.5)
			In Situ			
50-54	1.3 (0.7-2.1)	1.3 (0.90-1.7)	1.4 (1.3-1.5)	0.77 (0.6-0.9)	1.1 (0.86-1.3)	0.70 (0.60-0.80)
55-59	0.63 (0.20-1.6)	2.1 (1.4-2.7)	1.4 (1.0-1.8)	0.73 (0.6-0.9)	1.1 (0.83-1.3)	0.90 (0.90-1.0)
60-64	2.4 (1.2-4.1)	3.0 (2.1-3.8)	2.2 (1.6-2.9)	0.96 (0.8-1.2)	1.2 (0.93-1.5)	1.0 (0.90-1.0)
≥65	2.0 (1.3-3.1)	1.7 (0.6-2.8)	1.8 (0.60-4.1)	1.0 (0.9-1.2)	1.6 (1.0-2.1)	1.7 (1.3-2.2)
All*	1.5 (1.2-1.8)	1.9 (1.7-2.2)	1.6 (1.4-1.9)	0.83 (0.77-0.90)	1.2 (1.1-1.3)	0.99 (0.92-1.1)

Abbreviations: BCSC, Breast Cancer Surveillance Consortium (US); NBCCEDP, National Breast and Cervical Cancer Early Detection Program (US); NHSBSP, National Health Service Breast Screening Program (UK).
*Adjusted to a standard age distribution.

Various studies have been carried out examining performance variations between different countries' approaches to screening women for cancer. One of these is shown in table 1-3. This shows data for 5.5 million women from 1996-1999 for the UK and USA: the Breast Cancer Surveillance Consortium (BCSC, *n* = 978,591) and the National Breast and Cervical Cancer Early Detection Program (NBCCEDP, *n* = 613,388), both in the United States; and the UK (NHSBSP, *n* = 3.94 million). Some 27,612 women were diagnosed with breast cancer from these three groups. Whilst screen detected cancers were similar between the two countries, the number of women recalled or who had a negative biopsy was twice as high within the USA. The authors concluded that screening in the USA should concentrate on lowering the recall rate whilst maintaining the detection rate.

In China, the Chinese Ministry of Health, together with the Chinese Anti-Cancer Association (CACA), has examined the viability of breast screening. The result was the 'One Million Women' project which was undertaken to determine the impact of screening and started in 2005 running until 2010. The breast cancer screening program is part of the country's ambitious health care

reform plan which expects to cost 850 billion Yuan (£85 billion). The aim was to select 100 high-quality hospitals nationwide as designated institutions and offer one million women aged 35 to 70 standard mammography screening over a six year period (People's Daily Online, 2005; Li, 2009; Mao, 2010). This is further discussed in Chapter 8.

1.2 Information Technology in Breast Screening

Over the last twenty years, radiology has undergone a major development in information technology such that nowadays virtually all radiological investigations utilize digital imaging with the resultant images then examined on high resolution digital workstations.

Information technology has enabled the transition from X-ray film based radiology departments to a new digital organization, extending beyond the confines of a hospital, and opens up completely new opportunities for clinical radiology. Efficiency in radiology has been improved (Smith, 2006) through the use of:

1. Information management (e.g. Radiology RIS and PACS deployment);
2. Computer-Aided Diagnosis (CAD);
3. Remote access, e.g. teleradiology.

These three areas are described next. Mammography has been the last area to develop digital imaging as a routine radiological tool because of its special requirement of high contrast and high spatial resolution images.

1.2.1 RIS and PACS

Two main computer systems exist in the digital radiology department, the Radiology Information System (RIS) which is responsible for most text-based computing functions and the Picture Archiving and Communication System (PACS) which deals with image related computing functions (Smith, 2006). Figure 1-7 illustrates the general overview of the basic functions of PACS and

its relationship with RIS. Although RIS and PACS are two separate systems, these work together as one package.

The RIS is responsible for scheduling patient orders, capturing the clinical reporting information, preparing prior patient' exams (if needed), and providing the PACS with the information. The basic functions of PACS (Picture Archiving and Communication System) include image acquisition, image presentation for interpretation, image storage, and local image distribution which can extend outside a hospital into the NHS enterprise.

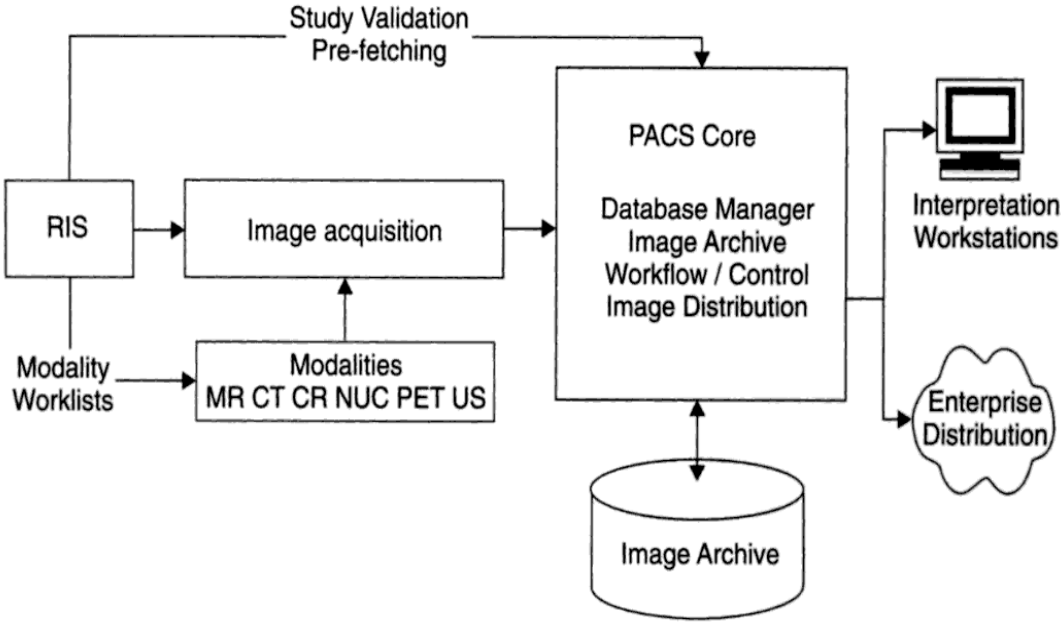


Figure 1-7. RIS - PACS architecture (Smith, 2006)

The Digital Imaging and Communications in Medicine (DICOM) standard specifies the general format for PACS image files storage and transfer. It has two main purposes. Firstly, it includes the actual pixel image data information, as well as details such as pixel size, image slice distance (e.g. for computed tomography (CT), Magnetic Resonance Imaging (MRI)) along with patient demographic information (this being contained within the image 'header' information); the second purpose of the DICOM standard is to specify the information that is used for image retrieval and transmission.

In digital mammography, the DICOM working group, DICOM 3.3-2003, has standardized the headers and other information which is stored. These digital mammographic images are stored as two types: *raw data images* for processing, which is an unprocessed image as recorded when the woman is imaged; and *presentation images*, which are images with the inclusion of additional display information which is necessary for the images to be viewed appropriately on a PACS workstation. Some of this information is proprietary depending upon the manufacturer of the PACS system – this raises problems when an image is generated on one manufacturer's system and is then viewed on another manufacturer's PACS monitors. A raw data image is produced when the mammogram is taken and the radiographer checks this image immediately for image quality (e.g. ensuring the woman has not moved during the X-ray exposure, ensuring that the whole breast is imaged appropriately). This information is then transformed into a presentation image which is stored in the PACS system and is the image which is then clinically reported. The raw data image is also essential for computer based analysis of breast density or for CAD analysis. The typical size of a DICOM standard digital case for a woman is about 200Mbyte per four view case (i.e. circa 50M byte per image/view).

1.2.2 Digital Mammography and CAD

Using X-ray film, which is then examined on an illuminated multi-viewer, is being replaced by digital imaging of the breast with examination of the resultant images on very high resolution digital workstations. In order to examine fine detail in mammograms, radiologists used to use a magnifying glass with the X-ray film mammograms, whereas now they can utilize numerous software interaction tools. In the USA about 60% of centres which conduct breast screening use FFDM. Many other countries are also using FFDM and in countries where breast cancer screening has recently been introduced then they have implemented FFDM without employing SFM.

There is a second form of digital mammography known as Computed Radiology (CR) which does not have the same sensitivity as FFDM does and also is reported as generating a higher radiation dose to the screened woman

compared to FFDM. However, to use CR in an existing screening centre which uses SFM simply means that the X-ray film cassette which is inserted under the breast in the mammographic unit is replaced by a digital detector plate. This means that an existing screening unit which is using X-ray film can then produce digital CR images immediately and at the fraction of the cost of FFDM which requires that the SFM unit be totally replaced by a FFDM screening unit.

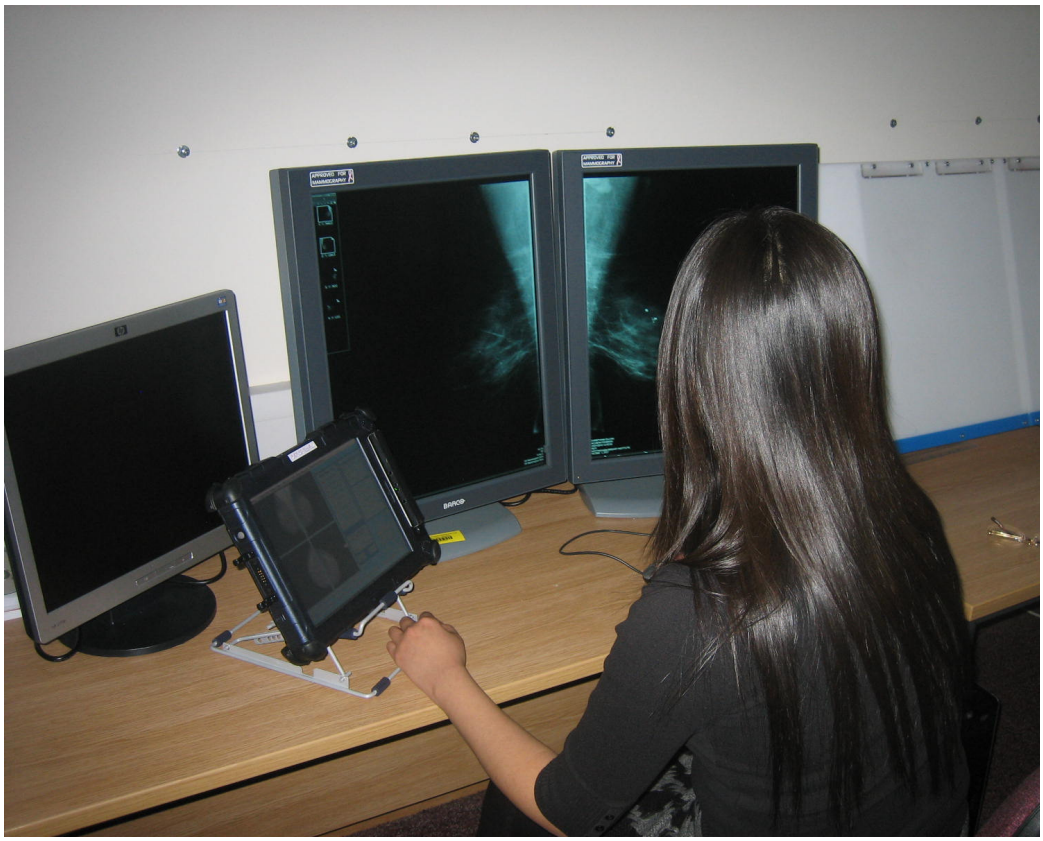


Figure 1-8. The digital mammogram workstation with dual monitors. On the left is the monitor used to display the RIS information. The small tablet computer is part of the PERFORMS scheme.

Typically these FFDM workstations (figure 1-8) have a dual monitor set up where each has a resolution of 5 Mega pixels (2,048 x 2,560) and is capable of displaying 10-bit greyscale images with a high contrast ratio. Each monitor can display various combinations of mammographic views and is used primarily to display a single mammogram which can then be zoomed and panned into and manipulated (e.g. contrast window levels adjusted).

Some large-scale clinical trials have compared the performance on FFDM and SFM. The first major trial, the Colorado-Massachusetts screening trial, involved 6,768 paired examinations on 4,521 women over a 30 month period at two institutions. The results showed a significant decrease of recall rate on the digital and a non-significant cancer detection rate on the film. The trial however used a digital workstation with dual 21" monitor with 1,000 x 1,500 pixels resolution (Lewin, 2002) – this is considerably lower than is now judged to be clinically acceptable.

Built on this trial, the Digital Mammography Imaging Screening Trail (DMIST) involved 49,528 women at 35 sites in the USA over 25.5 months. The project enrolled four different digital imaging systems from multiple manufacturers (GE, Fischer, Lord/Trex, and Fuji CR). A significant advantage of FFDM mammography was found, both in cancer detection and overall ROC analysed performance in detecting cancer in younger women with dense breasts (Pisano, *et al*., 2005; Pisano, *et al*., 2008). Some other trials based on a European population, such as the Oslo I study (Skaane*, et al*., 2003) and the Central East London Breast Screening Service Study (Vinnicombe, 2009) have also revealed similar results.

Apart from supporting better diagnostic performance as shown by such clinical trials, digital mammography has the potential to offer several advantages in screening. One of the major ones is allowing a substantial reduction in radiation dose to the breast as compared to X-ray film screening. It is reported that in practice the radiation dose can be reduced between 25%-30% for thin (30–40 mm - compressed breast thickness) and thick (>70 mm) as compared with film screening without compromising the image diagnostic accuracy (Samei*, et al*., 2007; Svahn*, et al*., 2007; Yaffe, 2010). Figure 1-9 shows one experimental result which compared the radiation dose between SFM and FFDM on different compressed breast thicknesses (Gennaro & Di Maggio, 2006); clearly FFDM produces a lower dose for all breast thicknesses.

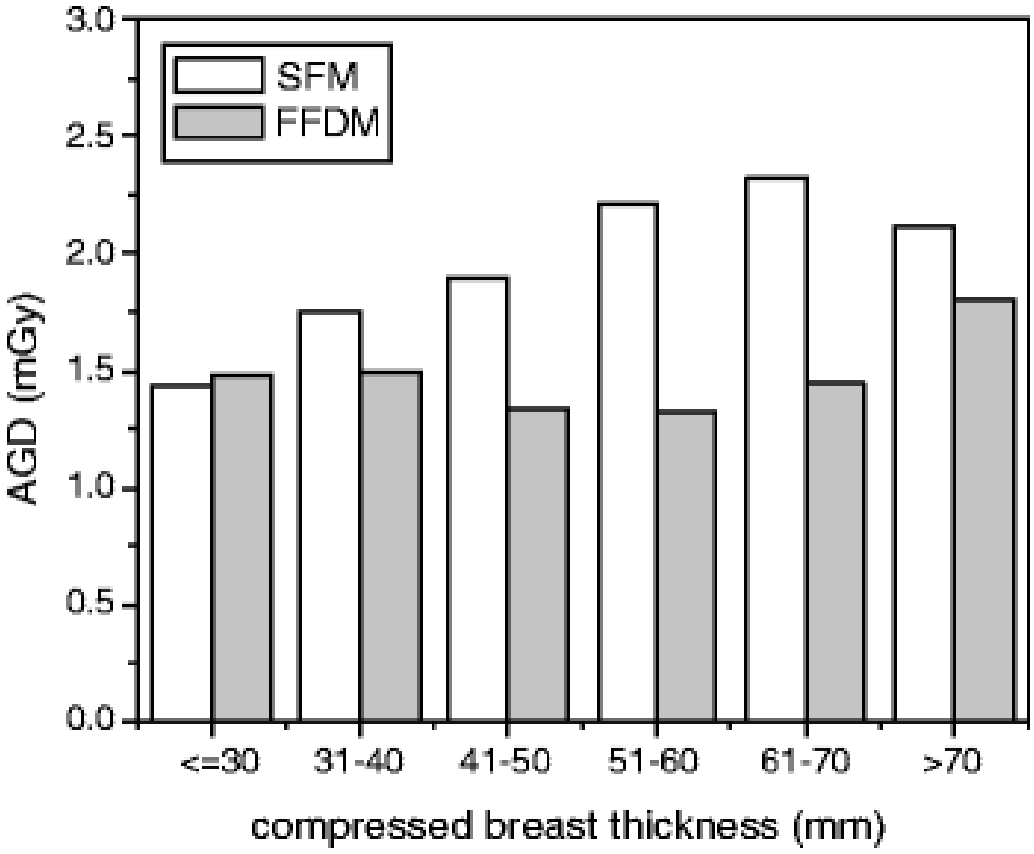


Figure 1-9. Radiation dose comparison for different intervals of compressed breast thickness. Comparison between SFM and FFDM (Gennaro & Di Maggio, 2006)

Several other benefits of FFDM over SFM have been summarized as follows (Skaane*, et al.,* 2003; Skaane, 2010): elimination of technical recall as the image quality can be checked immediately; simpler image storage, image retrieval and transmission; improved diagnostic quality, especially for younger women with dense breasts; the possibility of implementation of advanced technologies, e.g. computer-aided detection.

Computer-Aided Detection & Computer-Aided Diagnosis

Medical image interpretation is heavily dependent on intelligent computing approaches which typically pre-process images prior to these being examined by experienced radiologists. Key relevant approaches to potential radiological abnormality identification and error reduction are exemplified by computer aided detection (CADe) and computer aided diagnosis (CADx).

Such methods have been well researched and aid radiological imaging interpretation by suitably applying image processing algorithms in order to identify known key image features. The result of such approaches is that the original image can be viewed by the radiologist with the option to visualize computer generated overlays of key feature identifiers (i.e. 'prompts' - see figure 1-9) which help the reader in identifying abnormality presence (CADe) and classify if it is benign or malignant (CADx). Both commercial CAD systems are now implemented in several radiological domains, including breast cancer detection (Astley & Gilbert, 2004) which is concentrated upon here. Typically in the past such mammographic CADe systems have produced many false positive detections per woman, which have detracted from the usefulness of the approach, but recent advances have improved markedly upon this. Mammographic CADe systems are widely used routinely in America where they improve the performance of individual radiologists. Such CADe systems have also been trialled in some UK centres as an intelligent aid to the screening radiologist. A recent large study compared performance when two screeners examined cases as compared to a single screener with a CADe system and found similarity between the two approaches in terms of performance (James*, et al.,* 2010). However the CADe approach produced a small but significant increase in recall rates (i.e. false positive decisions).

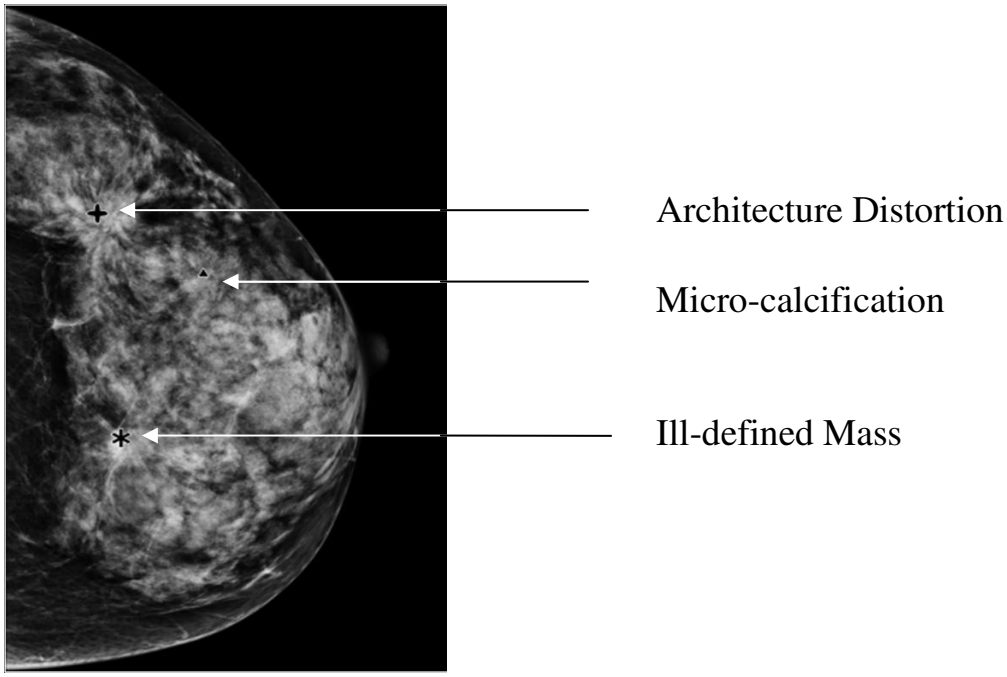


Figure 1-9. Example of digital mammogram (LCC) with CADe prompts. The potential abnormalities that were identified by CADe system (Hologic R2) are highlighted.

1.2.3 Teleradiology

The move to digital radiology in general has enabled electronic transmission of radiological patient images, such as X-rays, CTs, and MRIs, from one location to another (e.g. another hospital or even a radiologist's home) for the purposes of interpretation and/or consultation (Ruggiero, 1998). This is known as teleradiology. Tele-mammography is the transmission of mammographic images between different sites. In the USA some companies are setting up 'expert centres' where a group of experienced radiologists are based. Screening images are taken of women elsewhere and then these images are electronically transmitted to the experts for reporting, who then return their report electronically to the screening centre. Whilst ostensibly a simple process the image sizes cause some technical challenges in transmission.

Teleradiology enables a range of computing devices to display clinical images apart from the high quality clinical workstations within a radiology department. In particular there has been considerable interest in whether

small hand held PDAs could be used in teleradiology. Up to 70% of medical trainees currently use handhelds, mainly as electronic textbooks, or for medication reference databases, medical calculators, and patient-tracking. For hospital residents, PDAs have been found to help patient treatment by providing real time image viewing. However, information on the prevalence and usage of these devices among radiologists is limited. It has been found that slightly less than half of the radiologists in North America used PDAs on a daily basis. Possible reasons include:

1) The limitation of hardware, for example memory capability, low image resolution, etc. This limits the usefulness of PDAs in radiology. The use of a PDA to access imaging data is not as persuasive because it does not replace or supplement current technologies (i.e. the higher resolution and speed of a clinical PACS workstation is important to handle tasks such as image review and interpretation, management of radiology work flow (via the RIS), speech recognition, and image processing (Wiggins, 2003) which are today common place in many radiology departments.

2) Shortage of software: there is a relative lack of PDA software designed for radiology as compared with software designed for other medical specialties. Only a few PDA users had radiology-specific applications installed on their devices, which could due to poor application design.

According to Boonn's (2005) survey on PDAs use in radiology in North America, radiologists expected future PDAs usage to be limited to the review and sign off of reports, access to e-mail or the internet, access to radiology references with images and access to teaching materials. However, with handheld technology and networking performance evolving, the future direction for radiology and portable computing could be focused on portable offices for a 'mobile' radiologist to increase the efficiency of work-flow monitoring and maintenance, increase productivity, support group work for radiologists, and improve communication with referring physicians and patients (Raman, 2004; Wiggins, 2003).

Some recent research projects have been carried out to develop mobile teleradiology systems using a PDA. These projects are for better managing Digital Imaging and Communications in Medicine (DICOM) image data and to support group work. Teleradiology using such devices allows efficient management of lossless DICOM image data and is useful for collaborative work by radiologists in education, conferences, and research (Schweitzer*, et al*., 2002; Nakata*, et al*., 2005). PDAs have also been shown successfully to support the interpretation of CT images, whose small physical size and resolution is adequately handled by the PDA screen's resolution and size (Toomey*, et al*., 2007).

Tele-mammography offers the potential for more freedom in mammography interpretation training which could use such hand held devices - providing suitable resolution images can be transmitted and displayed appropriately without loss of required resolution. Some projects have already aimed to employ advanced technologies to support mammographic interpretation training. For example, GIMI (Generic Infrastructure for Medical Informatics) was a collaborative project in the UK with the purpose of developing a prototype training tool for screening mammography which could offer radiologists a tailored educational experience based around the intelligent selection of training activities (Yap & Gale, 2009; Simpson*, et al*., 2009; Scott*, et al*., 2008). GIMI is based on using advanced grid technology to deliver training to individuals however this was based on using clinical digital workstations. A related training system has also been further developed by Taylor*, et al*. (2010) which uses sections of mammographic images. Furthermore, a computer-based training (CBT) system has been investigated to support both the improvement of the skills of experienced film-readers and the training of inexperienced ones using advances in high-quality computer displays at the mammogram viewing workstation and high-speed networking (Soutter*, et al*., 2003).

These projects all aim to implement training at the digital mammography workstation itself. However, technological advancements have enabled such complex medical images to be viewed on a laptop, or even on a PDA – either

at the original pixel resolution or using a reduced resolution; both employing image pan and zoom in order to view the whole image. This raises the possibility of having some aspects of mammographic interpretation training delivered both whenever, and wherever, it suits the individual.

1.3.3 Other Relevant Imaging Techniques

Apart from mammography several other radiological imaging techniques are used in breast screening.

Tomosynthesis

Breast tomosynthesis is a new tool that has been recently introduced to help breast cancer detection. It is a modification of digital mammography which enables the acquisition of a three-dimensional (3D) volume of thin section data in a similar fashion to a CT scan, thus it reduces or eliminates the perceptual ambiguities caused by imaged tissue overlap. The basic principles are illustrated in figure 1-10 (Park, *et al*., 2007; Smith, 2005; Niklason*, et al*., 1997).

Clinical trials have been conducted which indicate that breast tomosynthesis generates a lower recall rate and a higher positive predictive value for a biopsy recommendation than mammography (Park, *et al*., 2007; Poplack, *et al*., 2007; Fornvik, *et al*., 2010). This indicates that the use of breast tomosynthesis may make earlier detection of breast cancer possible (Svahn, *et al*., 2007). However imaging takes longer than mammography and the overall 3D image volumes generated (circa 2 Gigabytes per woman) is very considerable for hospital PACS systems to handle easily.

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) of the human body was first accomplished in 1980 (Smith, *et al*., 1981). The potential of this technology implemented in detecting breast cancer was demonstrated in the early

eighties (El Yousef, *et a*l., 1984; Kaiser & Zeitler, 1989; Heywang, *et al.,* 1989).

Today, MRI is commonly accepted to be the most sensitive imaging technique in breast cancer detection, especially for measuring lesion size (Kuhl, *et al*., 2005; Boetes, *et al*., 1995; Mann, *et al*., 2008; Van Goethem*, et al.,* 2006). It has been shown to assist in detecting multifocal disease (such as in dense breasts, invasive lobular cancer, and discordant findings), recurrent disease, discordant imaging and monitoring chemotherapy.

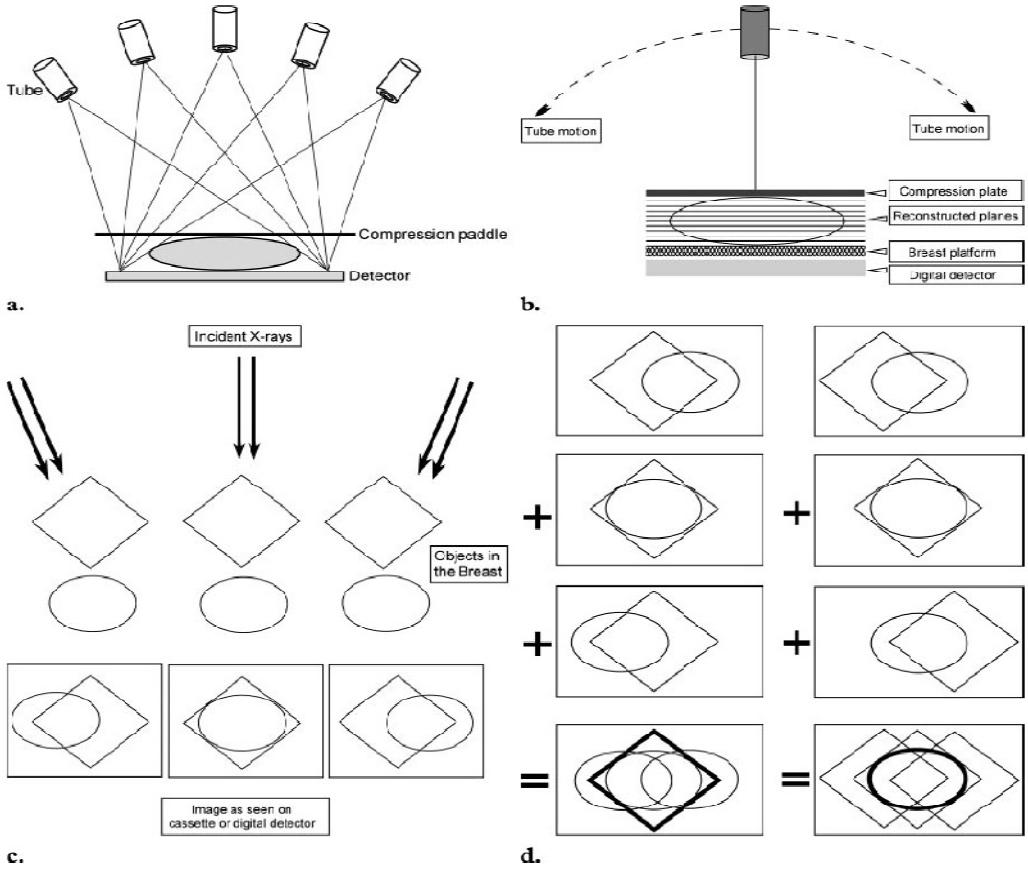


Figure 1-10. Basic technologic principles of breast tomosynthesis. (a, b) Schemas show how image data are acquired from various angles as the x-ray tube moves in an arc. Either the step-and-shoot method (a) or the continuous exposure method (b) may be used, and the detector may be moving or stationary during image acquisition.
The 3D image data are subsequently reconstructed as conventional mammographic projections (craniocaudal, mediolateral oblique, and mediolateral views). (c, d) Diagrams show how different 3D image data acquired from different angles (c) are reconstructed to provide separate depiction of two overlapping structures located in different planes (d) (Park*, et al*., 2007).

Also, it has very high sensitivity and the advantage of having no radiation exposure during the scan (as it uses non-ionizing radio frequency signals) enabling it to be an ideal tool for screening women under 50 years of age or who are at high risk of developing breast cancer (Gilbert, 2010). Imaging a woman takes about 20 minutes and so the technique is not ideally suitable for mass screening. The guidelines of how this technique needs to be used in breast cancer detection, along with its limitations, have been provided by: the National Institute for Health and Clinical Excellence (NICE, 2009); the American Cancer Society Breast Cancer Advisory Group (Saslow*, et al.*, 2007); and the European Society of Breast Imaging (Mann*, et al.*, 2008).

Ultrasound

Ultrasound has been used for screening women but it is not very sensitive and typically is used mainly as a follow up investigation after routine mammography screening. It is slower to perform than mammography for screening and not good at visualising calcifications or small lesions. However, it is very helpful with identifying whether a mass is abnormal or benign depending upon the appearance of the echoic shadow (Stavros, 2004)

1.3 Mammographic Interpretation

The main purpose of mammography is detecting cancer at an early stage before it can grow large enough to be palpable and thus cause the woman to present symptomatically to a General Practitioner (GP). However, to identify cancer at the earliest possible stage, when cancers are very small and subtle is very difficult and requires great radiological skill. For example see figure 1-10 which shows a small portion of a single mammogram containing some early indicators, calcifications, of a small cancer.

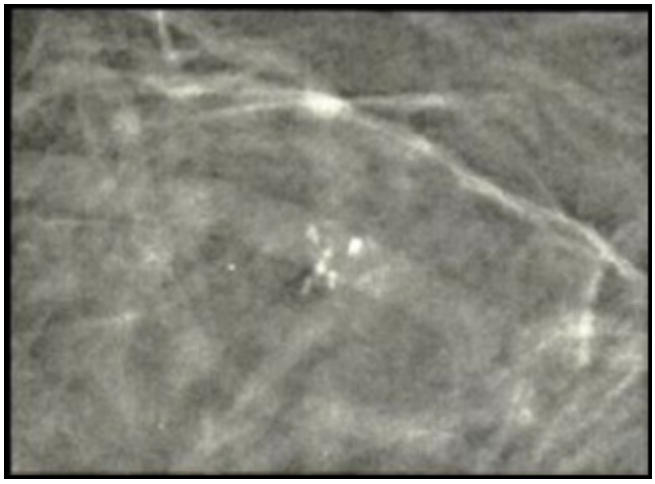


Figure 1-10. Seven clustered calcifications, one of which is a thin linear (rod) shape. On biopsy this was a 2mm intraductal carcinoma (Sickles, 1984).

It is therefore not surprising that mammographic interpretation is a very difficult task, which is partly due to: the diverse anatomical characteristics of the woman being screened, e.g. different breast density; the size and shape of any potential lesions, and partly due to the low incidence of breast cancer. As an example of this, the incidence of invasive breast cancer in the USA is only 4.9 in every 1,000 women who are over 40 years old (Jemal*, et al.,* 2007); even for women with all ages included in the UK, age-standardised breast cancer estimated incidence rate is 8.6 in every 1,000 women (Ferlay*, et al.,* 2008).

It is also known that performance on the task varies between individual film readers. Variability in interpretive performance in screening mammography exists not only between general radiologists and specialised mammography film readers (Sickles*, et al.,* 2002) but between mammography film readers (Elmore*, et al.,* 1994; Elmore*, et al.,* 2009; Skaane*, et al.,* 2008). This is not surprising as individual variations in imaging interpretation performances is well known and have been reported in every medical imaging domain for many years. For example, Robinson (1997) reported individual performance variation while examining skeletal, chest and abdominal images; individual variation was also described while examining chest images (Manning*, et al.*, 2004). Interestingly, Scott*, et al.* (2004) examined the performance differences between consultant radiologists and advanced practitioners in reading a test

set of screening cases and whilst performance differences between individuals existed, overall they found many similarities in performance between these two groups of film readers.

One aspect of performance in real life is where a cancer which is actually present on a screening film is missed. This may then not be detected until the woman presents for screening again, some three years later, when the cancer will have grown. If the malignancy presents in between screening rounds (for instance if the woman or her GP feels a lump which is then investigated) then this is called an interval cancer. Interval cancer figures can be used to gauge the efficiency of screening. An interval cancer can also be due to a rapid tumour growth but examination of the previous screening mammograms can clarify if in fact there was a failure in detecting it on that round.

Previous research on the retrospective evaluation of interval cancer cases, has suggested that 38% (van Dijck*, et al.,* 1993) to 67% (Warren Burhenne*, et al*., 2000) of such cases show evidence of an abnormality on the prior screening films. Of course in such re-evaluation the cancer detection task is easier as one knows precisely where on the previous mammogram to look for early signs of malignancy. Notwithstanding that, these figures demonstrate that mammographic interpretation is a challenging task. In the USA, it was estimated that, 10–30% of breast cancers were unreported during screening mammography (Bird*, et al.,* 1992). Furthermore, a recent report from the Norrbotten Mammography Screening Programme showed that overall interval cancer rate was 11/10,000, constant by age and the overall interval cancer rate ratio was 38% (Bordas*, et al.,* 2009).

1.3.1 Mammographic Features & Mammographic classification

Identifying malignancy depends upon accurately identify mammographic features indicative of disease. Various authors have proposed lists of key features. Gale*, et al.* (1986) developed one of the first computer aids to mammography diagnosis systems ('MAMCAD') based on radiologists accurately identifying key mammographic features.

Table 1-4. List of mammographic features (Gale, *et al*., 1986)

(a) LESION PATTERN	(b) REACTION PATTERN	(c) CALCIFICATION PATTERN
L1* Lesion detected.	R1* Reaction	C1 Calcification
L2* Dysplasia only	R2 Oedema	C2* High density
L3 Homogeneous texture	R3 Diffuse density	(C3* Size - Predominantly small
(L4* Density greater than gland tissue	R4 "P" Reaction	(C4 Size - Predominantly medium
(L5 Density equal to gland tissue		(C5 Size - Predominantly large
(L6 Density less than gland tissue	R5* Subcutaneous reaction	C6 Size - All equal
(L7* Margins well defined	R6 Local skin thickening	C7 Shape - Round
(L8 Margins segment ill defined	R7* Periareolar skin thickening	C8 Shape - Rods
(L9 Margins ill defined	R8 Distant skin thickening	C9 Shape - Irregular
L10* Spicules present	R9 Hypervascularity	C10 Shape - Mixed
L11 Comet tail present	R10 Parenchymal deformity - pulled	C11 Position - Clusters
(L12 Size less than 1 cm	R11* Parenchymal deformity - pushed	C12* Position - Rows
(L13 Size more than 1 cm		C13 Position - Scattered
		(C14 Number - 5 or less
		(C15 Number - more than 5

A list of 43 features were identified in 500 biopsied cases related to the lesion, a reaction in the breast to the lesion presence, and calcifications discriminate function analysis demonstrated that 12 features were important in predicting malignancy presence. Simply identifying whether or not these 12 key features (marked with an * in table1-4) were present had a sensitivity of 79% and specificity of 88%. This compared to an expert radiologist's sensitivity for the same cases as 87% and specificity of 49%. Had MAMCAD been used as a diagnostic aid when these women had had mammography then 268 fewer biopsies would have been performed.

Since then the terminology in mammography has changed and various other authors have proposed different classifications of features. For instance Sickles (1984; 1986) identified and summarized those non-palpable features into three groups - see table 1-5.

In 1995 the American College of Radiology standardised the reporting of mammographic cases and mammographic features by introducing the Breast Imaging Reporting and Data System (BI-RADS). This is implemented in the

Table 1-5. Mammographic features prompting biopsy in non-palpable breast cancers (adapted from Sickles, 1986)

All calcifications	*All Masses*	*All 'Indirect' sign*
Linear/branching	Spiculate/knobby	Architectural Distortion
Other irregular	Irregular/ poorly defined	Developing density
Indeterminate shape	Relatively well defined	Asymmetry
		Single dilated duct

United States and subsequently in many other countries (American College of Radiology, 1998). BIRADS describes the following mammographic features (Balleyguier, 2007):

- Densities and masses.
- Micro/ Macro calcifications.
- Architectural distortions.
- Special cases including: ductal ectasia, intramammary lymph node, or focal asymmetric density.
- Associated findings: skin or nipple retraction, skin thickening, cutaneous lesions, axillary lymph nodes.

The great strength of BI-RADS is in providing a single standard image description which facilitates communication between health professionals as well as enabling improved health care provision should a patient move from US state to state or even between various countries. For each mammography case, BI-RADS detail the reporting of key mammographic features as well as the location of the malignancy. BI-RADS also provides an internationally widely used classification system, as shown in table 1-6. It presents a standardized classification for each mammography case and accordingly relates image descriptors to the likelihood of breast malignancy (Eberl*, et al.*, 2006).

Table 1-6. Table for Clinical Management Recommendations for Mammograms by the BI-RADS Category

BI-RADS Category	***Assessment***	***Clinical Management Recommendation(s)***
0	Assessment incomplete	Need to review prior studies and/or complete additional imaging
1	Negative	Continue routine screening
2	Benign finding	Continue routine screening
3	Probably benign finding	Short-term follow-up mammogram at 6 months, then every 6 to 12 months for 1 to 2 years
4	Suspicious abnormality	Perform biopsy, preferably needle biopsy
5	Highly suspicious of malignancy; appropriate action should be taken.	Biopsy and treatment, as necessary.
6	Known biopsy-proven malignancy, treatment pending	Assure that treatment is completed

However, in the UK, the BIRADS classification is not implemented (Maxwell*, et al.*, 2009). Instead, each mammography case is classified using a five-point rating scale for the probability of cancer:

1. Normal or Definitely Benign;
2. Probably Benign;
3. Indeterminate finding;
4. Probably Malignant;
5. Malignant.

This rating scale is in many ways similar to the BI-RADS approach.

1.3.2 Mammographic Interpretation Performance Assessment

Performance in interpreting mammographic images can be described via a hierarchical model of efficacy (Fryback & Thornbury, 1991) on the basis of

various previous related works. This is a six-tiered model which is a conceptual continuum of assessing the contribution of diagnostic imaging to the patient management process. The six levels are:

1. Technical quality;
2. Diagnostic accuracy efficacy: agreement between diagnoses and 'truth';
3. Diagnostic-thinking efficacy: impact of diagnostic imaging information on clinician's thinking about each patient;
4. Therapeutic efficacy: impact of diagnostic imaging information on patient management;
5. Patient-outcome efficacy: impact of diagnostic imaging information on patient health;
6. Social efficacy: impact of Diagnostic imaging information on society as a whole.

In the present research the interest is in diagnostic accuracy efficacy which is measured by comparing an individual's decisions on the case images being inspected against some standard ('truth') – usually taken as either known abnormality presence (based on the pathology of the case) or against an expert radiological decision.

The simplest measure of diagnostic decision is the percentage of cases for which the film reader gives the correct answer, i.e.

Accuracy (%) = [Number of correct decisions] / [Number of cases] x 100

However, it has limited usefulness as the disease prevalence has a strong effect on the data. For instance, for a rare disease like breast cancer, a film reader can be considered very accurate by simply calling all the cases negative (Metz, 1978). Therefore, it is important to separate data for positive and negative cases and always consider these together. The simplest situation is the binary decision about a case (table 1-7):

Table 1-7. See text for explanation

		Truth (e.g. abnormality)	
		Present	***Absent***
Decision	***Present***	Yes (true positive – TP)	No (false positive – FP)
	Absent	No (false negative – FN)	Yes (true negative – TN)

An individual's decisions range from agreeing with the truth on an abnormality being present or absent, true positive (TP) and true negative (TN) decisions respectively; and errors in deciding that an abnormality was present when it was not, false positive (FP), or that a case was normal when it was not, false negative (FN). These four categories of decisions can then be used to calculate measures of sensitivity (how often an individual correctly determines that a case is positive compared to all the positive cases) and specificity (how often an individual correctly determines that a case is normal, i.e. does not contain an abnormality compared to all normal cases)

Sensitivity (%) = TP / (TP + FN) x 100
Specificity (%) = TN / (TN + FP) x 100

Both measures have to be used together to describe performances someone can obtain 100% TP simply by reporting every case as positive (and thus making many FP errors) and similarly can obtain 100% TN (and making many FP errors) simply by calling every case normal.

A related measure is known as the "proportional incidence" method (Day, 1985), which gives a different estimation of sensitivity. Additional measures are: Positive Predictive Value (PPV – the percentage of women who are disease positive and who are referred for further assessment, i.e. how often an individual correctly determines that a case contains an abnormality as compared to all the times s/he reports abnormal) and the Negative Predictive

Value (NPV - the percentage of women who are disease negative and returned to normal screening):

PPV (%) = TP / (TP+FP) x 100
NPV (%) = TN / (TN+FN) x 100

Taken together these four measures fairly well describe the performance of an individual in examining images on a particular display and have been used for many years. They are easy to understand by a clinician and practically useful. However, a key problem with sensitivity and specificity measurement is that these do not give any information about how an individual's decisions about cases may vary. Even a single radiologist will use different decision threshold values for the same case; depending on the different clinical situation (e.g. a radiologist's sensitivity may be lower when a case was presented as a normal breast screening case comparing to as a symptomatic case). Furthermore, intra-observer variation on the placement of a decision threshold value may vary as well. Therefore, it is important to compare diagnostic accuracy by means that are independent of the chosen threshold. This is where Receiver Operating Characteristic (ROC) analysis is important.

In ROC analysis it is hypothesised that over a series of cases there is a distribution of signal (abnormality present) and noise (normal) cases and that these can be represented as two Gaussian distributions lying on some axis. The individual in making a decision about a particular case is using a criterion (e.g. A or B) in figure 1-11. For each criterion value then different percentages of each signal/non signal distribution are selected as being TP, FP, TN and FN. Using criterion A will correctly identify more of the signal cases than criterion B will but at the cost of more FP errors. Criterion B will correctly identify more of the normal (non signal) cases than criterion A (with fewer FP decisions) but at the cost of more FN decisions. Thus, the ROC approach shows how making a decision is a trade off between the four decision categories. In contrast, a sensitivity and specificity score simply gives rise to one point in ROC space.

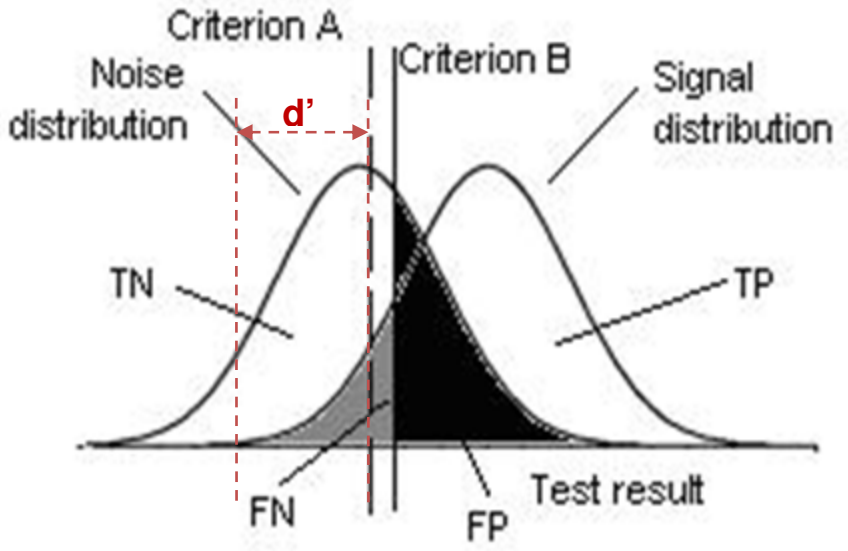


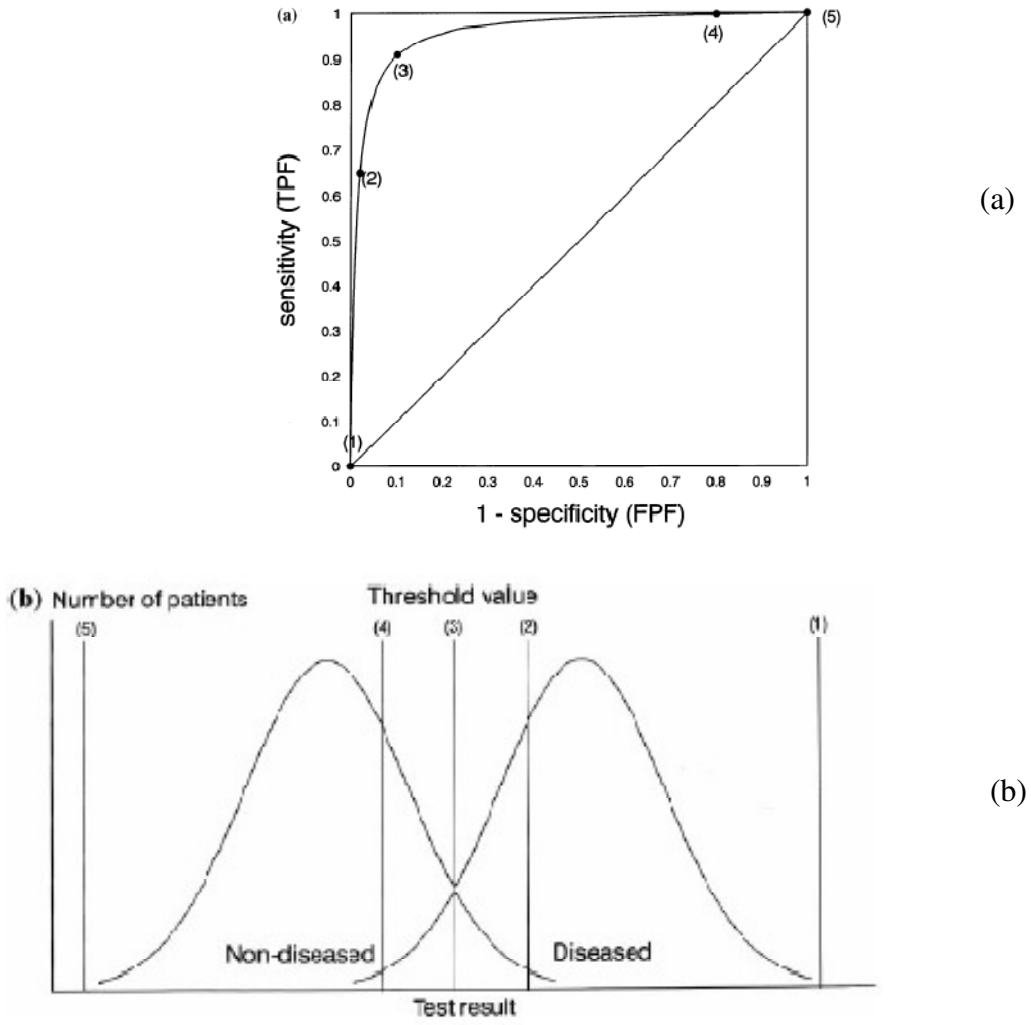
Figure 1-11. Plots of signal and noise distributions

In carrying out a ROC experiment the observer is usually asked to make a rating judgement about each radiological image. Although any scale (from a few points up to a continuous scale) could be used to make a rating judgement, a five- or a six- point scale is often used in radiology studies (Metz, 2008). The five-point ratings could be used to describe the observer's confidence about each image, such as:

1. Definitely normal
2. Probably normal
3. Don't know
4. Probably abnormal
5. Definitely abnormal

These decisions would correspond with the following criterions on the plots of Gaussian distributions (figure 1-12b). Each criterion decision gives rise to different proportions of TP, TN, FP and FN. A response of 5 correctly identifies all signal cases but at the expense of many FP responses and poor normal decisions, whereas a response of 1 correctly identifies all normal cases but yields many FN errors in identifying signal.

The ROC plot is a mapping of [Sensitivity / True Positive Fraction] (on the Y axis) plotted against [1-Specificity / False Positive Fraction] (on the X axis). This gives rise to a space where the diagonal represents chance decisions. The various 5 points (including: a strict threshold (case called positive only if judged almost definitely positive); of a moderate threshold; or of a lax threshold (case called positive if any suspicion of disease)) in figure 1-12a. Each operating point on the ROC curve represents a pair of sensitivity and specificity combination at a certain threshold value (Van Erkel & Pattynama, 1998).



(a)

(b)

Figure 1-12. A typical conventional ROC curve, showing five possible operating points (Metz, 1978)

Typically, in an ROC experiment different individuals will generate different ROC curves representing differences in their performances. Excellent performance is characterised by a curve which starts at point (0, 0) and rises vertically up the Y axis to point (1, 0) and then traverses at this height to the point (1,1). This means that the individual correctly detects abnormalities whilst making no erroneous decisions. This rarely happens in practice and several different curves can be found as in figure 1-13.

As well as giving a pictorial representation of different individuals' performances, various measures can be derived from the ROC approach. Firstly, ***d'*** is a measure of the ability to discriminate abnormal from normal (i.e. the distance between the means of the two Gaussian distributions in figure 1-11). Secondly Area-Under-the-Curve (AUC), the area under the ROC curve, is used as a measure of the accuracy of an individual's decisions. AUC has a value that ranges from 0.5 (chance performance) to 1 (perfect performance).

Ideally, the performance data collected in an experimental study needs to be normally distributed. However, raw data that can be analysed by ROC needs to not be zero at both extreme ends of the scale, e.g. the no noise response in category 1 and no signal response in category 5. So if there is an empty category, then some categories need to be combined. However, d' become very unreliable if categories are collapsed; the value of AUC will also change if categories are collapsed, however it is tolerant for bad data as compared to ***d'*** because curve fitting does not enter into AUC determination (McNicol, 1972).

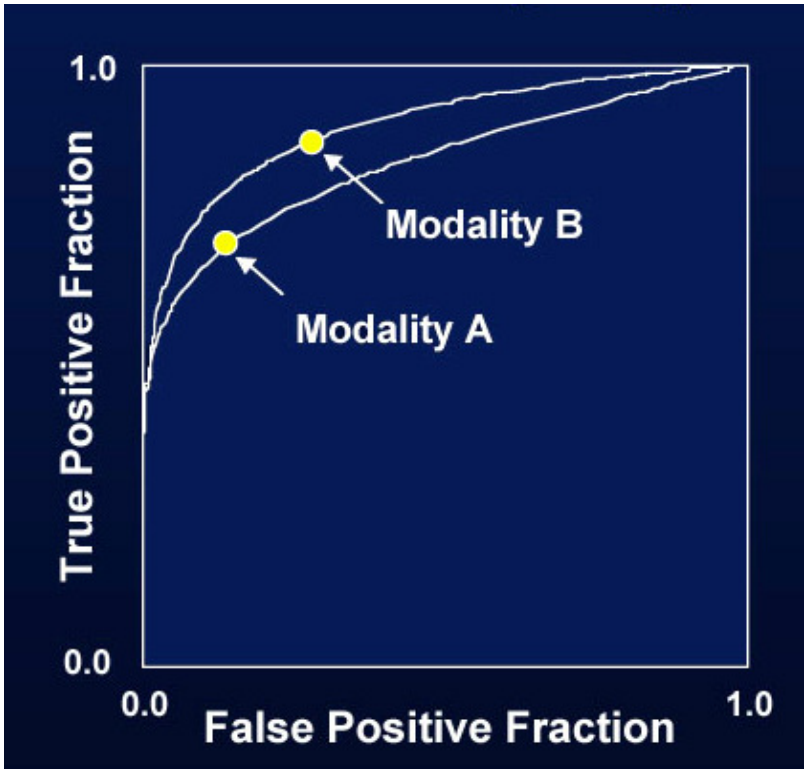
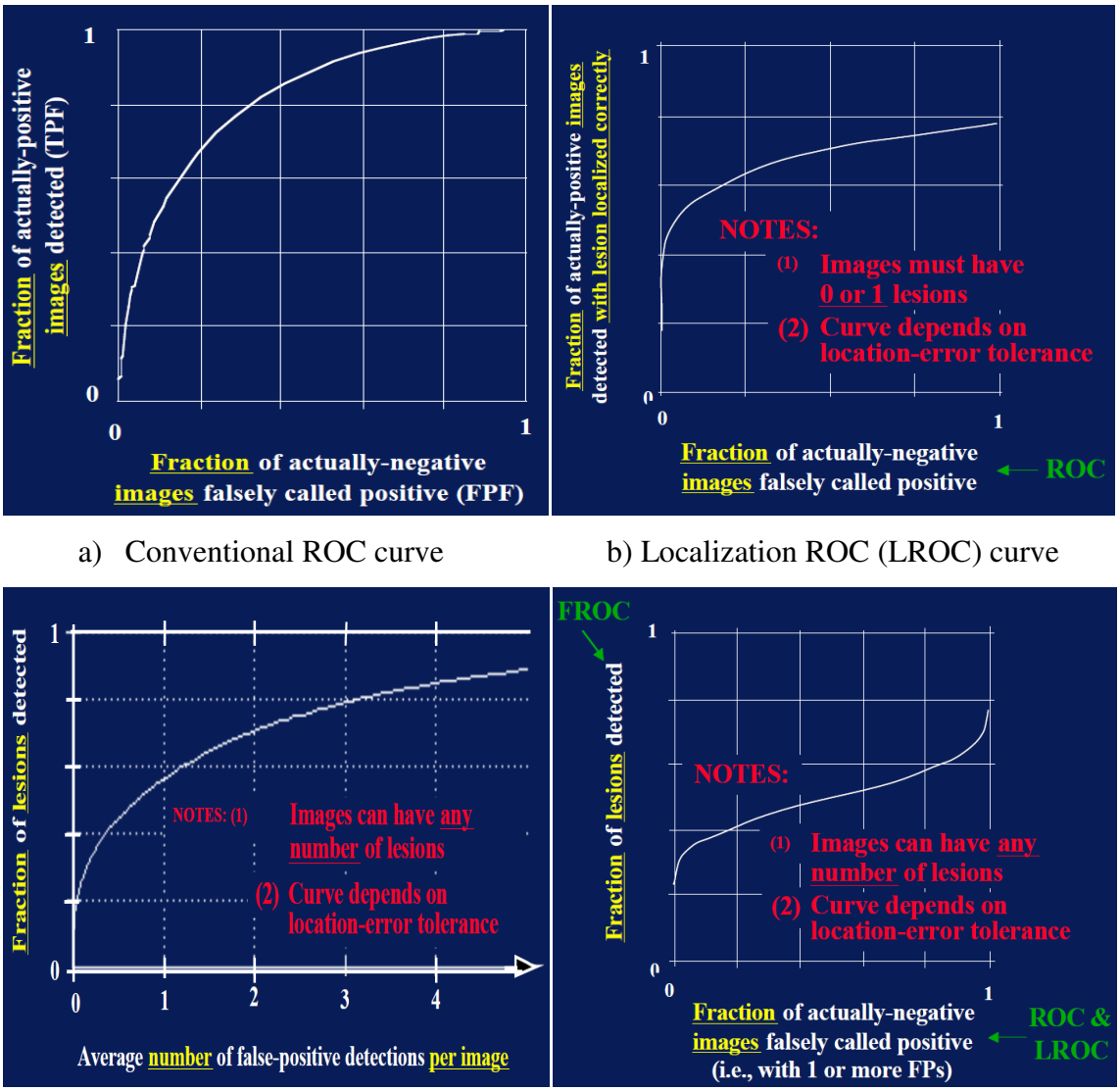


Figure 1-13. ROC curves for modality A&B. It shows Modality B is better, because it can achieve: higher TPF at same FPF, or lower FPF at same TPF (Metz, 2003)

Whilst ROC analysis has advantages over simply using sensitivity and specificity measures it does have limitations. First of all, it suffers from not taking any account of the location of the abnormality. Thus an individual may inspect an image and report it correctly as abnormal but in doing so may well incorrectly be making this decision on the basis of wrongly identifying a normal area of the image as being abnormal. This is known as a 'correct decision/incorrect location' error.

To account for such errors ROC analysis was extended to LROC (i.e Localization ROC) analysis where the location of the abnormality is also used when data are collected. It was first generated to perform a detection-and-location task. For many years researchers in medical image perception have solely concentrated upon images which may contain a single abnormality. LROC exhibits greater statistical power than the conventional ROC. However, each image may contain a maximum of one lesion, which limits its usefulness

in practical life (Swensson, 1996). An example of LROC is shown in figure 1-14b.



a) Conventional ROC curve

b) Localization ROC (LROC) curve

c) Free-response ROC (FROC) curve

d) Alternative FROC (AFROC) curve

Figure 1-14. Examples of conventional and generalized ROC curves

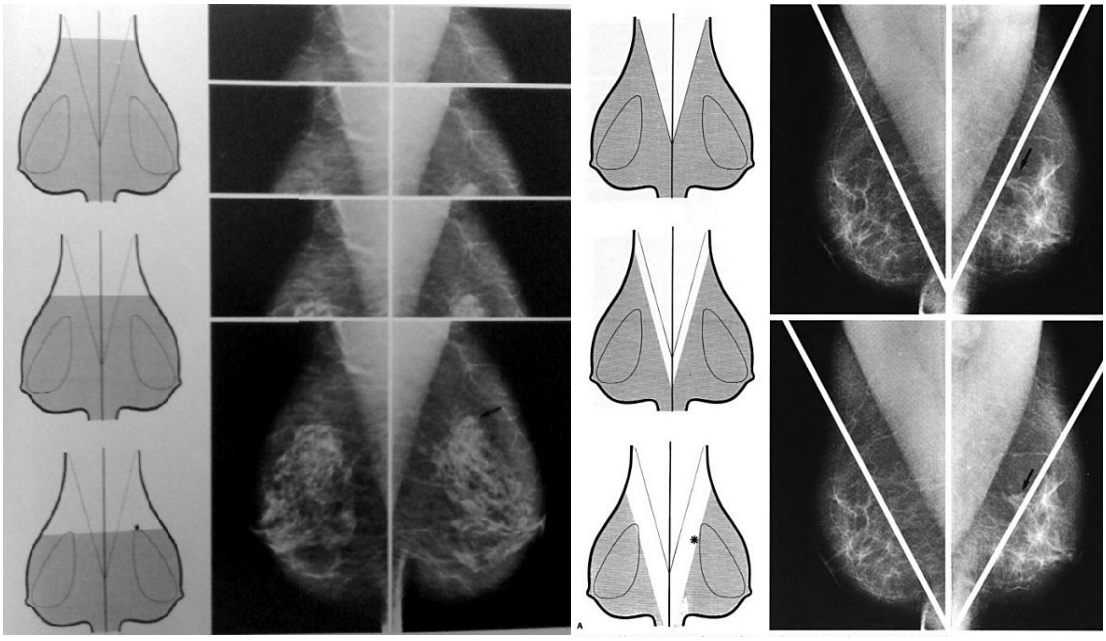
More useful ROC techniques have evolved which allow for the scoring of multiple potential targets/lesions in each image. These approaches are Free-response ROC (FROC) analysis. An example of a FROC curve is shown in figure 1-14c. FROC was designed to perform a detection-and-localization with *any* number of signals (Bunch, *et al.*, 1977; Anastasio, *et al.,* 1998; Chakraborty, 1989; Chakraborty & Winter, 1990; Edwards, *et al.,* 2002). This method allows the observer to make an unlimited number of reports, which

closely mimic radiological reporting in the clinical environment. Most importantly, it increased the statistical power over conventional ROC analysis (Xin & Frey, 2009). Xin & Frey (2009) also presented the challenge of FROC data analysis. Some alternative FROC (AFROC) analysis (example see figure 1-14d), such as jackknife alternative free-response operating characteristic (JAFROC) and initial detection and candidate analysis (IDCA) have been compared on statistic power. A variant of JAFROC (i.e. JAFROC-1) was found to have the greatest statistical power and therefore recommended for use in human observer performance studies (Chakraborty, 2008).

1.3.3 Image Inspection and Mammography

Methods for systematically viewing mammograms have been recommended by mammography experts to enhance the detection of abnormal lesions in routine clinical practice (Tabár & Dean, 2001). They proposed the use of a hand-held viewer (figure 1-15) to help visual masking out areas of the mammographic image so allowing the individual to concentrate on the area viewed, this included: 1) horizontal masking (figure 1-16a); 2) oblique masking (figure 1-16b) in order to: 1) have a detailed comparison of the left and right breast; 2) sequentially view restricted areas of the mammograms.

Figure 1-15. Perception of small and/or low contrast lesions on the mammogram is enhanced by the use of a hand-held viewer, which effectively eliminates extraneous light (Tabár & Dean, 2001).



a) b)

Figure 1-16. Two ways of masking

Evidence from elsewhere in radiology argues that whilst individuals can follow prescribed patterns of examining images this is detrimental to their performance (Gale & Worthington, 1983). In radiology, film readers are required to detect small features in various types of images. The region that is capable of resolving such small features is located in the retina, the area around the foveal fixation point where detailed information is processed – termed the "useful field of view" (UFOV) (Mackworth, 1976). For example, it was reported that fine details such as nodules in chest radiographs can be detected by peripheral vision as far away as 17^{o} from the visual axis, however, most are detected within 5^{o} (Carmody, Nodine & Kundel, 1980) and this figure is typically used in research in medical imaging.

Historically, visual research in radiology has mainly used chest X-ray films. This was because these images are very important clinically and also are large – making it easier to track how a film reader is examining such images and therefore carry out related research (Manning*, et al.*, 2004; Manning, *et al.*, 2005). Recently most work has investigated mammographic image interpretation which involves the examination of screening mammograms for

very early signs (often only a few millimetres in size) of breast cancer (Mello-Thoms*, et al.,* 2001; Mello-Thoms*, et al.,* 2002).

It is well established in image inspection tasks in radiology that experienced observers exhibit measurable differences in visual search strategy as compared to domain inexperienced individuals. A body of research has investigated the underlying reasons for this and how such differences are developed. That expert performance related to specific visual search behaviour implies that, potentially, the visual search behaviour of experts can be utilised in some way to improve the performance of the less experienced, either in general training or in accelerating existing training programmes.

Much of this research has been performed in the domain of medical image inspection and builds upon theoretical foundations which posit the importance of visual search as a key part of the complex development of skilled cognitive performance. Examination of radiological images inevitably produces errors; in particular here false negative errors are of interest. By monitoring the eye movements of individuals as they examine radiological images the types of errors made can be classified into three types; namely errors due to visual scanning (the area of interest is not projected on to the useful field of view), pattern recognition error (the area of interest is projected onto the useful field of view however the potential abnormal features are not separated from the surrounding normal structure) and decision making error (the area of interest is recognized from the surrounding normal structure, however the actual abnormality cannot be separated from normal tissues) (Kundel, Nodine & Carmody, 1978).

1.3.4 Other Factors Affecting Accuracy in Mammography

Film readers' mammographic interpretation accuracy will also be influenced by various factors:

First of all, the importance of ambient light has been greatly emphasised by some recent studies (Brennan*, et al.*, 2007; Kimme-Smith*, et al.*, 1997; Uffmann*, et al.*, 2005). Excessively high levels of ambient lighting will have a

negative effect on the image interpretation accuracy. Accordingly, guidelines have been issued. For hard-copy (i.e. film) reporting, recommendations were issued by World Health Organization (WHO, 1982). Mammography specific guidelines have been made by the American College of Radiology (ACR, 1999) for reporting SFM images, who also issued further guidelines proposing a maximum allowable ambient light level for soft-copy (FFDM) reading as follows:

For film reporting:

"3,000 candelas per square meter (cd/m2) minimum is the standard for screen film." ACR (1999)

For digital reporting:

"Monitor luminance, L is characterized by minimum (L_{min}) and maximum (L_{max}) values. In the presence of reflected ambient luminance (L_{amb}), the monitor luminance is designated as L'. The ratio of the maximum luminance (L'_{max}) to the minimum luminance (L'_{min}) of a mammographic display device should range between 250 and 650 over a 30 degree viewing cone whose principal axis is perpendicular to the image. Ideally, the maximum luminance should be 450 cd/m^2 or higher in order to avoid too low a value for minimum luminance (susceptible to ambient lighting) to maintain a desired luminance ratio." (ACR, 2007a)

For the reading room lighting level:

"*viewing conditions should be optimized by controlling reading room lighting to eliminate reflections on the monitor and lowering the ambient lighting level as much as is feasible. Ambient lights should not be turned off completely nor turned up completely. About 20 lux is generally sufficient to avoid most reflections and still provide sufficient light for the human visual system to adapt to the surrounding environment and the displays.*"(ACR, 2007b)

The potential effects of ambient lighting level are very important and need to be considered in any experimental design involving viewing medical images.

Therefore, in the experimental investigations carried out here the ambient light level has been measured and strictly controlled.

Factors as diverse as room temperature, noise, posture, fatigue, and poor ergonomics may also have significant effects not only on radiologist comfort but also on the quality, accuracy, and consistency of image interpretation performance (Horii*, et al.*, 1989; Siddiqui*, et al.,* 2006).

Time of day is another factor that can potentially influence the radiologist's performance (Gale*, et al.*, 1984). In-depth research reported by Gale & Scott (2010) examining test sets of screening mammograms found a small drop in sensitivity after 6pm, however, not significantly. Details see figure 1-17.

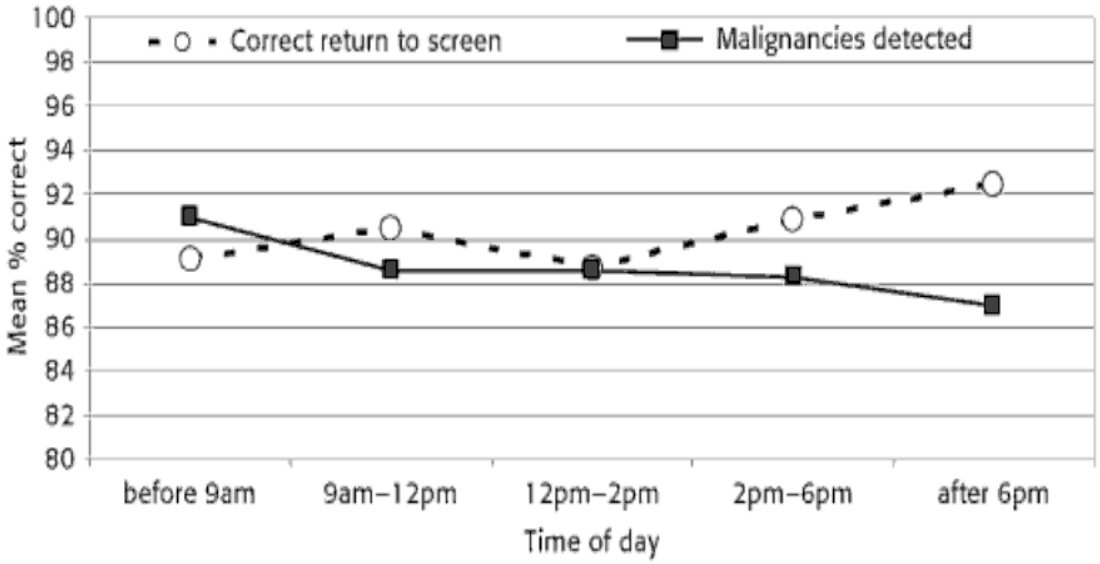


Figure 1-17. Time of day and performance (Gale & Scott, 2010).

Volume of screening cases read per year is also important. Thirty-seven film readers within the East Midlands Screening Programme in the UK had three years of their recent screen reading results compared between readers who have different yearly reading volume (Cornford*, et al.,* 2009). It was reported that the cancer detection rate at first read was significantly lower in the higher volume readers (readers who read ≥25,000 cases per year). Related data from the Florence screening programme (Ciatto*, et al.,* 2005) suggested that

larger cancers missed were due to fatigue or loss of attention. PERFORMS research data (Gale & Scott, 2010) revealed a significant effect ($p<.05$) of fatigue effects on film reader's performance as time on the task approaches three hours.

1.4 Mammographic Interpretation Training

As emphasized above, mammographic interpretation is a very difficult task, which emphasises the importance of efficient mammographic training both to improve, and maintain, individuals' everyday performance. Furthermore, changes within the NHSBSP have increased the need for the availability of a range of training approaches.

Breast screening has been undertaken across the UK for 20 years using mammographic film as the imaging medium. As described earlier, the age range of screened women is being increased which will increase the number of cases annually examined in the UK's breast screening centres - necessitating further screening staff being trained. The gradual implementation of FFDM imaging will help address this but the change to digital requires current screening personnel to be further trained in examining these images as their appearances are somewhat different to mammographic film. Additionally, the Department of Health's (2007) report on breast cancer services over the next five years emphasised:

"Investment in training, both funding and time, is essential to improve cancer services and take account of new evidence to improve the outcomes and experience of cancer patients." (Department of Health, 2007)

1.4.1 Current Training and Recent Changes

Interpreting mammographic images can be considered to comprise a range of perceptual and cognitive skills which include the recognition of certain mammographic features (Gale, 1997). Errors occur in every image interpretation domain and in breast screening error occurrence, particularly false negative decisions of there being no cancer present, when in fact

cancer is present, are particularly serious. For instance, recently (September, 2009) in the UK one radiologist was found to have missed cancer in 14 women, with another 85 needing to be re-imaged and a further 355 women's mammograms requiring re-examination by an expert. In August 2010 another radiologist missed cancer in eight women causing a screening centre to close. Minimising the potential for error by improving performance through better training is therefore vitally important.

Different methods of training are needed to help film readers to improve their expertise in identifying mammographic abnormalities at an early stage. An important practice in aiding film readers to develop the necessary skills could be to expose them to the appearances of a wide range of mammographic abnormalities through having them read a high volume of cases. Also, gaining appropriate experience of attending to specific abnormal features is another key factor which has been shown to help improve an individual's ability to recognise abnormalities (Gale, 2003). However, appropriate mammographic interpretation training opportunities is, of necessity, presently somewhat limited.

The European guidelines for quality assurance in breast cancer screening and diagnosis (EUROPA DONNA, 2007; Perry*, et al.,* 2008) listed guidance on training in mammography. It summarized the requirements for each individual medical staff in a breast cancer screening programme, 1) to undertake specific training in the academic and clinical aspects of screening at an approved training centre before entering any programme; 2) to be offered training in both uni-disciplinary and multidisciplinary settings where they also learn the importance of communicating with their colleagues in other disciplines; 3) to take continuing education, etc.

In the UK, film readers undertake various types of training that supports the development of skills required to visually inspect mammographic images and detect abnormalities. This includes attending mammography training courses, reviewing interval cancer cases, undertaking a national self-assessment scheme (i.e., PERsonal perFORmance in Mammographic Screening

(PERFORMS)) bi-annually, and participating in weekly multidisciplinary meetings.

Mammography Training Course

With various changes to breast screening in the UK, there is a need to train an increasing number of radiographers and to quickly bring them to advanced practitioner level (where they are qualified to examine and report breast screening cases within the national screening programme) in order that they can contribute fully to the NHSBSP. Various training courses exist which are linked to universities offering qualifications such as an MSc or Qualifications for Advanced Practitioners' in breast imaging/mammography. These involve both academic aspects and a clinical work-based learning programme.

There are five major training centres in England which are responsible for training specialized staff working in breast screening. These training centres are: South East London National Breast Screening Training Centre; Jarvis National Breast Screening Training Centre; St George's Hospital NHSBSP National Training Centre; Nottingham International Breast Education Centre, and the Manchester Breast Screening Training Unit. To meet the breast screening programme training standards, these centres are required to work closely with the breast screening quality assurance centres and other breast screening units.

Multidisciplinary Team Meetings

The UK Department of Health defines a Multidisciplinary Team as:

"A group of people of different health-care disciplines, which meets together at a given time (whether physically in one place, or by video or tele-conferencing) to discuss a given patient and who are each able to contribute independently to the diagnostic and treatment decisions about the patient." (Department of Health, 2004)

The Multidisciplinary Teams have regular Multi-disciplinary Team Meetings (MDT Meeting). It "ensures effective coordination between key professionals with all the important knowledge and therefore ensures the optimum treatment planning and good continuity of patient care" (Fleissig*, et al.,* 2006). In the UK such meetings are key to the success of the NHSBSP. Specific interesting screening cases each week are discussed by all team members and they can see how a woman is first identified by screening and then followed up with biopsy, other radiological investigations (ultrasound, MRI etc.) or surgery. From the sharing of knowledge all team members get feedback on their various decisions concerning each case as well as learning for future cases. In the survey of 87 screening centres carried out by Nickerson and Cush (2003) answers to questions about multidisciplinary meetings show the importance of these meetings and that the most common attendees were: breast care nurses, radiologists, radiographers, surgeons and advanced practitioners (see table 1-8 for details).

Table 1-8. Number of Core Assessment team Members in Attendance at MDT Meetings (Nickerson & Cush, 2004)

Profession	***Advanced Practitioners***	***Breast Care Nurses***	***Radiologists***	***Radiographers***	***Prof/Tech Helpers***
Number at MDCM	60	70	69	69	66

* NB figures are out of a total of 70 units

PERFORMS (PERsonal perFORmance in Mammographic Screening)

In the UK there is a national self-assessment scheme which provides an opportunity for film readers to examine a wide range of specially selected, difficult exemplars of previously categorised mammographic screening cases within a short period of time. As part of the quality assurance programme for the NHSBSP, the PERFORMS scheme (PERsonal perFORmance in Mammographic Screening) was established in 1991 in response to the fact that feedback to film-readers on their screening performance on live cases at

that time was of necessity very slow (Gale, 2003) (i.e. feedback on whether a case was a true negative or a false negative report was not obtained until that woman was screened again in the next screening round). Therefore, individuals undertake the scheme which is a free and anonymous self-assessment exercise for all screening film readers in the UK. This is undertaken bi-annually and requires film readers to interpret recent difficult, known screening cases. Figure 1-18 illustrates one individual participating in PERFORMS where they read and interpret the cases, reporting their decisions into a small tablet computer and receive immediate feedback on their decisions. Current developments have produced a digital version of PERFORMS for nationwide roll out in late 2010.

Figure 1-18. Mammogram film reader participates in PERFORMS on the mammogram multi-viewer

The scheme also serves as a training tool. The number of difficult abnormal cases in each set is approximately equivalent to the number of abnormal cases they would see in several years of actual screening. It also provides very informative feedback. This includes: 1) immediate feedback after taking part where the individual can access experts' radiological opinion and full

pathology report for each case; 2) an annual report: each participant receives a report comparing each individual's performance with their regional peers and also each individual's performance with their peers nationally. All data are anonymous and only each individual has knowledge of their own performance. It can identify the training needs of particular groups or individuals which can then lead to the deployment to these individuals of particular training sets of cases.

Both assessment and training functions of the scheme makes it an attractive tool for breast screening personnel. Table 1-9 shows the growth of the number of film readers who participate in the PERFORMS scheme since 2000 which reflects the growth of personnel in breast screening.

Table 1-9. Growth in film reader participates in PERFORMS over the years

	Year	***Participate in both rounds***	***Radiologists***	***Advanced practitioners***	***Other professions***
2010	9	669	377	239	53
2007-2009	8	523	367	231	59
2004	7	506	288	179	39
2003	6	425	263	137	25
2002	5	341	261	117	43
2001	4	348	252	73	23
2000	3	310	246	42	22

1.4.2 Other Ways of Training

Apart from attending training courses and PERFORMS, there are other ways of training. There is also the usual need for ongoing CME (Continuing Medical Education) training for staff. Most training comprises utilising well known textbooks (e.g. Tabár, 2008) or some form of interactive computer based education where selected key images are presented to be examined, followed by critical and reflective feedback. Abnormality appearances in these images are shown, highlighted, demarcated and described; with interactive training producing feedback to the participant. Typically either the abnormal area is shown within the whole breast image or the abnormality is shown magnified. Such approaches train observers by familiarising them with a range of abnormal and normal appearances, demonstrating what to look for as well as indicating potential high probability areas within the mammographic images of where to look for abnormalities. Additionally, in the UK Breast Screening Programme every screener has to examine images of at least 5,000 women a year to help develop and maintain their appreciation of such appearances.

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 used for the clinical practice of FFDM screening.

1.4.3 Limits of Current Training

With SFM, typically examination of mammographic films is undertaken on a mammographic multi-viewer where several hundred cases can be loaded for inspection. This equipment, statically sited at a breast screening centre (see figure 1-3) comprises a back-illuminated surface on which numerous mammographic cases can be presented simultaneously for examination. Alternatively a single light box can be used where only one single mammographic case can be presented at a time. Consequently, this limits the time and places where any detailed training can take place.

The increasing use of digital mammography and its forthcoming widespread adoption in the UK is opening up new opportunities to provide a wider range of training without such restrictions (Department of Health, 2007). For instance, as well as being able to view digital breast images on high resolution monitors at the Breast Screening Centre, it could be possible also to view training images on a range of computer displays - for instance desktop, laptop PCs or even handheld devices (PDA, iPhone, iPad). These could be used to offer mammographic interpretation, anytime, anywhere to fit the individual's needs, provided that it were possible to: maintain the acceptable image quality on the device; devise acceptable interaction methods, as well as performing such viewing in appropriate viewing conditions (Chen & Gale, 2008).

Consequently, this would offer the opportunity for extending the PERFORMS self-assessment scheme to provide increased dedicated and individualised training without any restriction to only doing this in a screening centre's reporting suite. For instance, as well as being able to undertake the bi-annual self-assessment on mammographic interpretation, it would be possible for individuals to view further training images according to the outcome of the self-assessment on a range of computer displays, provided that it were possible to zoom, pan, and otherwise interact with such images appropriately. This raises the possibility of having mammographic interpretation training delivered as required by an individual. Clearly such displays would never be advocated for prime clinical diagnostic purposes but these could be useful for training purposes.

1.5 Summary

Breast screening is important in identifying early treatable breast cancer. In the UK the screening programme is expanding with a parallel expansion of staffing who need training to inspect mammograms. Digital mammography implementation also means that training is required for this new medium. Outside of the UK as other countries start screening or make improvements to their performance in screening then training is also important.

The advent of digital mammography yields high resolution digital images which require clinical examination in breast screening units using appropriate clinical high resolution monitors and associated image enhancement/manipulation software. Such software is linked to the particular mammography vendor. In the UK, and internationally, there are only a handful of such vendors, these include; GE, Hologic, Siemens, and Sectra. For reporting purposes in the UK all digital breast screening units use the same software, termed the NBSS software (National Breast Screening Service). In clinical reporting the NBSS software displays a list of women and the radiologist can quickly select a woman which then automatically brings up her relevant mammographic images. In doing so the system will display the

Figure 1-19. Examples of mammography workstation and dedicated keypads

images according to the hanging protocol that the particular radiologist prefers (this relates to the way in which different views of the breast are shown [typically both MLO views are shown followed by both CC views]). Each vendor has developed sophisticated interaction devices to facilitate the radiologists examining the images. These are variations on the usual computer mouse and a subset of keys (see figure 1-19 for some examples).

Using image manipulation such as: zoom in/out; window/level etc., allows the individual to examine specific image areas in greater detail – this is somewhat equivalent to the radiologist using a magnifying glass when examining mammographic film although additional image interactions are facilitated, such as contrast windowing.

Such digital images lend themselves to being viewed on other non clinical workstations. For instance they can be altered in terms of resolution and size. Such images, or even the original images, could then be used to offer an individual the facility to train outside of the clinic. In considering this there are several potential research avenues that could be explored. For example, one approach would be to consider the physics of such an approach and concentrate upon factors such as pixel spacing, pixel density, resolution etc.

Here the approach taken is to assume that factors related to the device screen's ability to display images will inherently change as technology advances but that the need for the human to interact with the images will not change. For instance, at the start of this research the most complex available mobile screen resolution was 800 x 600 pixels (the highest pixel level on the current general market at that time in the UK). At the time of writing this thesis the most complex PDA screen is that of the 'retina display' of the iPhone 4 (namely 960 x 640 pixels: pixel density of 326 pixels per inch, 3.5" screen size). Consequently the research here is concerned with whether individuals could and would use mobile devices such as PDAs and other display devices (laptops, office monitors) which are less sophisticated than clinical display devices. If so then how would they do so, how useful would such devices be for training and is using such devices useful?

In exploring this topic, the research uses a range of experimental investigations. The approach concentrates upon the need for the use of such devices, the acceptability of such devices to clinicians, explorations of individuals with differing degrees of knowledge and experience (from naive to expert radiologists), and whether image interactivity is a prerequisite for achieving acceptable performance. Additionally, as screening spreads to countries such as China an initial exploration of the feasibility of using such devices for training in China is addressed.

CHAPTER 2

Experimental Methodology

This chapter describes the major methods that were used during this research. It begins with a brief description of the human visual system, eye movements and eye movement recording. The two eye movement recording techniques used are then detailed, together with the reasons for their selection for different aspects of the research work. The identification of eye fixations is described together with how Areas of Interest (AOIs) are derived. The software for analyzing where an individual is looking when free head movements are allowed is covered, as is the software which was used for the analysis of participants' observed behaviour when interacting with images. Performance data were analyzed using Signal Detection Theory derived ROC approaches, which are introduced in the previous chapter. Consequently the relevant background to the ROC methods used here, the varieties of ROC analysis and the two main ROC techniques actually employed in the research are encompassed. Sample size issues for statistical analyses in empirical studies of medical imaging performance are also discussed.

2.1 The Human Visual System

2.1.1 Introduction to the Human Eye

The human eye is a very complex organ that has a very complicated structure (see figure 2-1). The main function of the eye is image formation. When image-forming light passes through the pupil it constricts or expands appropriately so as to adapt to the amount of incident illumination. The lens works with the pupil and accommodates suitably so that the object being viewed is focused as an image on the retina.

The retina contains up to 115 million rods (which sense contrast, brightness, and motion) and 6.5 million cones (which are responsible for spatial resolution and colour vision). The pigments in the rods and cones convert light energy into electrical energy signals which are transmitted via various nerve cells to the visual cortex of the brain, visual perception takes place in a number of brain regions (Curcio, *et al.*, 1990; Hogan, *et al.*, 1971; Goodale & Milner, 1992).

Cones, which are responsible for capturing fine detail, are located highest in the fovea which only covers about 1.5° of the visual field. This itself extends to about 60° nasally (towards the nose) in each eye, to 100 ° temporally (away from the nose), and approximately 60° above and 75° below the horizontal. Therefore, only a very small portion of the visual field in front of an individual is actually focused on the fovea at any one time. For the human eye to appreciate very fine detail, for example for mammogram readers to see micro-calcifications, then the eye must move around appropriately so that such fine details fall upon the fovea (Snell & Lemp, 1998; Atchison & Smith, 2000).

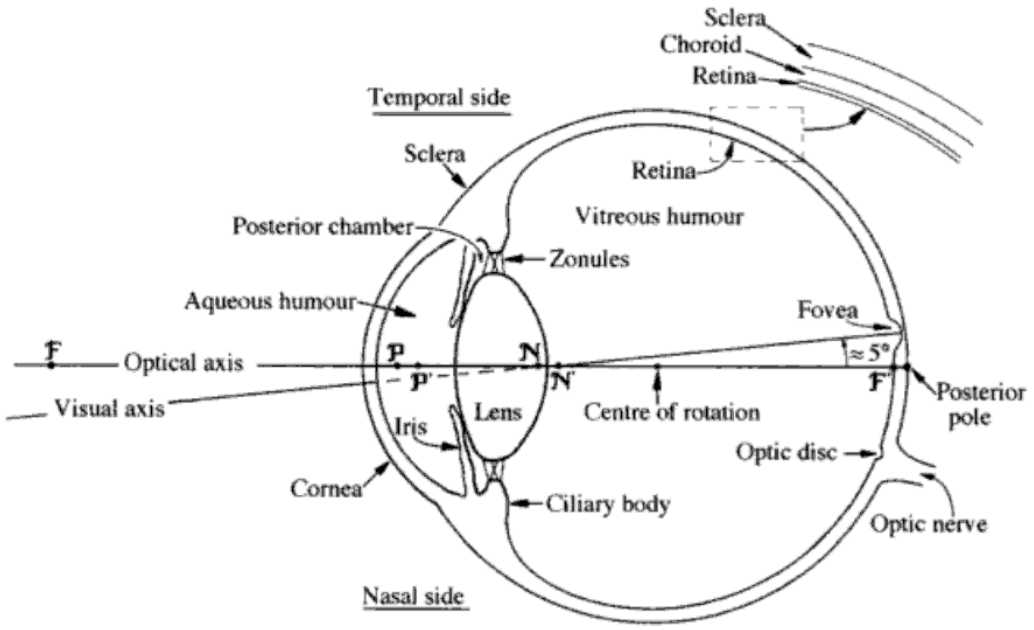


Figure 2-1. The horizontal section of the right eye. The pupil is the opening in the iris (Atchison & Smith, 2000)

2.1.2 Eye Movements & Visual Attention

The human eye makes several different kinds of eye movements and it is never still. The main movements of interest here are saccades which are very rapid ballistic (i.e. pre-planned) movements. These intersperse with eye fixations when the eye is relatively stationary. Visual input is markedly decreased during a saccade. Vision primarily occurs then during fixations and the alternating sequence of fixations and saccadic eye movements are considered to make up the voluntary and involuntary perception of the visual

world. Saccades are generally made under subconscious control although they can also be consciously made.

Generally visual attention, the part of the visual world to which someone is attending, is determined by the location of where they are fixating. Saccadic eye movements largely serve to move the location of visual attention around. Where someone attends or fixates is the basis of many investigations which have studied a whole range of factors.

2.1.3 Eye Movement Recording Techniques

Various techniques have been developed to record eye movements (Mowrer, *et al.*, 1936; Findlay, 1974; Duchowski, 2007). Early approaches used a range of innovative methods ranging from putting a blob of mercury on the eye and recording the reflection of light from it, to affixing mirrors to contact lenses. Mackworth & Mackworth (1958) filmed the reflection of an object in the eye and took the filmed centre of the pupil as indicating where the observer was looking. Such approaches were fairly intrusive, requiring considerable co-operation from the observer. A popular technique, electro-oculography (EOG) which is still used, requires skin mounted electrodes to be affixed to the observer. Such approaches have various levels of accuracy in determining where the observer is looking – for instance EOG is not very accurate (> 1^{o} visual angle). A very accurate technique (< 0.5^{o} visual angle) is to use a search coil mounted within a contact lens and the observer then sits within a large set of coils. Whilst this method is accurate it means that the observer has small wires leading from the eye to the recording device. Another very accurate technique requires that a wax mould be made of the observer's teeth and they then sit biting this within an instrument so as to have no head movements (Liddell, 1919).

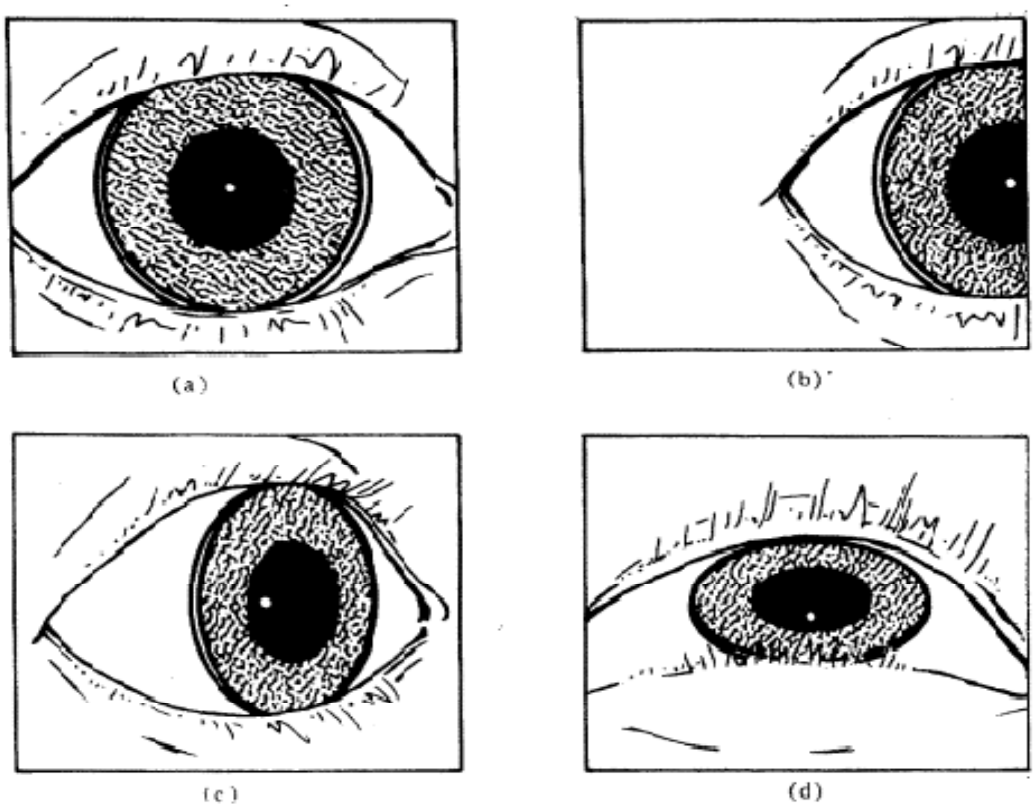


Figure 2-2. Oculometer eye-sensing technique. (a) Eye looking straight ahead. Note corneal reflection is at centre of the pupil. (b) Eye looking straight ahead, but laterally displaced. Note corneal reflection is still at the centre of the pupil. (c) Eye looking to the side. Corneal reflection displaced horizontally from the pupil centre. (d) Eye looking up. Corneal reflection displaced vertically from the pupil centre. (Marchant & Morisette, 1974)

Such techniques have either high accuracy but require the observer to have their head fixed in some way or else have low accuracy and are inconvenient, such as requiring skin-mounted electrodes, or the use of contact lenses. Ideally, a method of eye tracking is required which has no need for attachments to the eye, allows the observer to move their head, and also is capable of providing some acceptable degree of accuracy (< 1^{o} visual angle). Such a technique is based on the use of a pair of eye reflections that move similarly under eye translation but differentially under eye rotation. This is the eye Point-of-Regard measurement, which is based on tracking the centre of a corneal reflection with respect to the pupil centre. It is illustrated by Marchant & Morisette, 1974 (figure 2-2).

Table 2-1. Comparison of some of the major eye movement measuring techniques (Cornsweet & Crane, 1973; Young & Sheena, 1975; Young & Sheena, 1988).

	Measurement Range		Accuracy		Convenient to use
	Vertical	Horizontal	Vertical	Horizontal	
Electro-oculography	±50°	±50°-±80°	2°	1.5°	Medium
Contact lens e.g. using mirror	±10°	±10°	2"	2"	Low
Point-of-Regard (e.g.Honeywell Oculometer)	±30°-10°	±30°	1°	1°	High

2.1.4 Eye Movement Data Recording and Analysis

Two eye movement recording techniques were used in a number of the investigations during this research. These are the Applied Science Laboratories (ASL) Model 501 system and the Tobii X50 eye tracker.

2.1.4.1 ASL Model 501 System

This is a head mounted eye tracker (see figure 2-3) which measures eye position with respect to the head. A centrally mounted miniature camera ('scene camera') records the visual scene in front of the participant. In addition, a small infra-red light source is reflected off the hot mirror and into the left eye of the participant. The light illuminates the retina and is reflected from the retina back via the hot mirror and is picked up by the eye camera. Because the light source and camera are co-aligned the retinal illumination effectively backlights the pupil so that the pupil appears as an illuminated disc. In addition a bright reflection is obtained from the front surface of the eye. From the recorded pupil and corneal reflection of the light source the system automatically calculates both pupil diameter and the observer's line of gaze. The centre of the pupil and corneal reflection are displayed on the 501 system's 'eye monitor' together with cross hairs which indicate the automatically detected pupil centre and the corneal reflection (see figure 2-4).

From these data the system then calculates the observer's point of regard which is displayed as x-hairs overlaid on the scene camera video (see figure 2-5) as well as these data also being recorded in a computer data file.

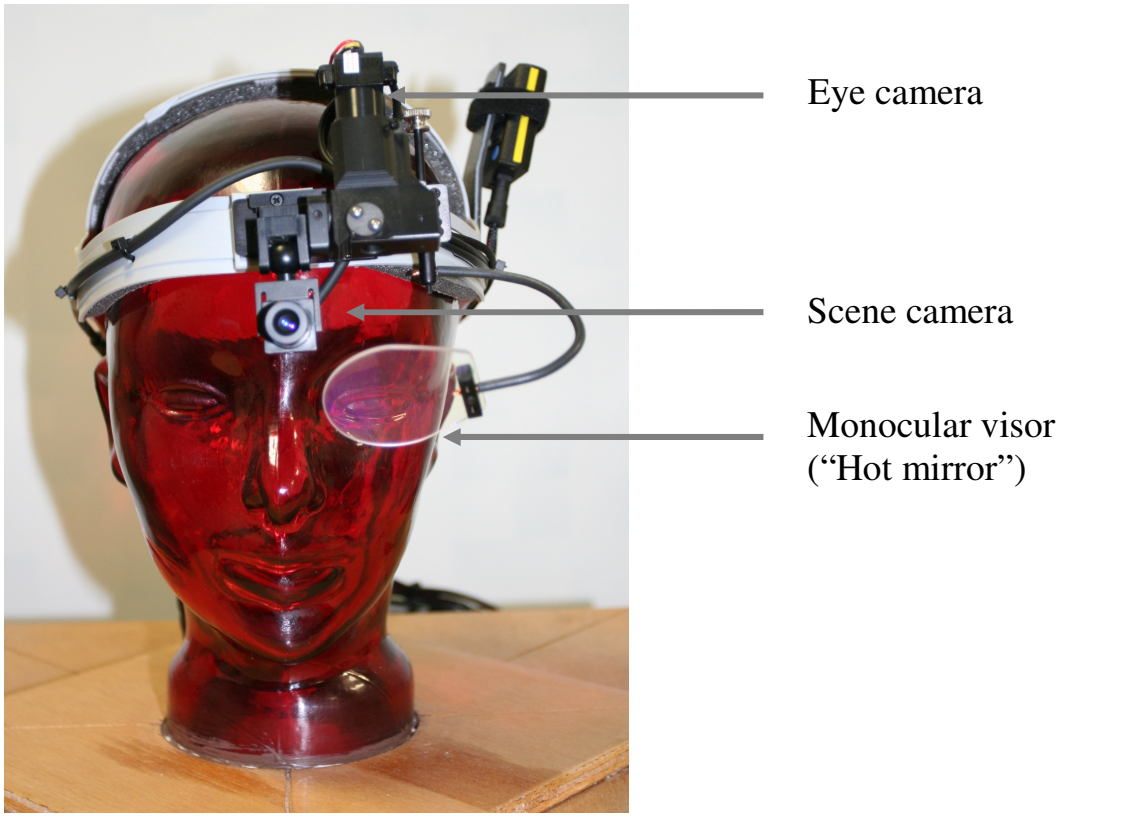


Figure 2-3. The head mounted oculometer

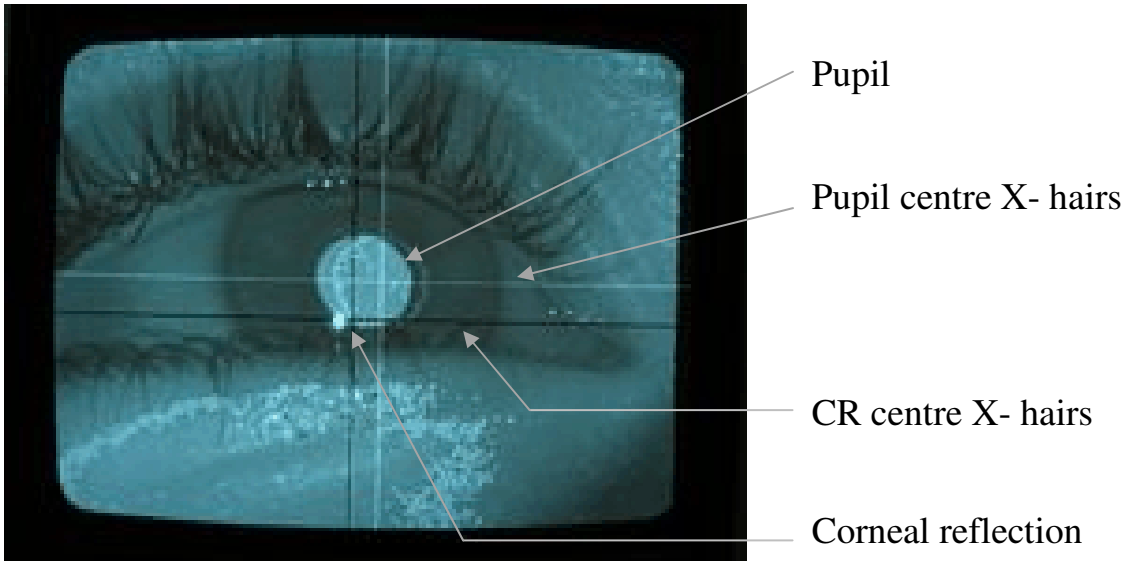


Figure 2-4. Pupil and corneal reflection (Applied Science Laboratories Model 501 Eye Tracker and Gaze Tracker System Setup and Operations Manual)

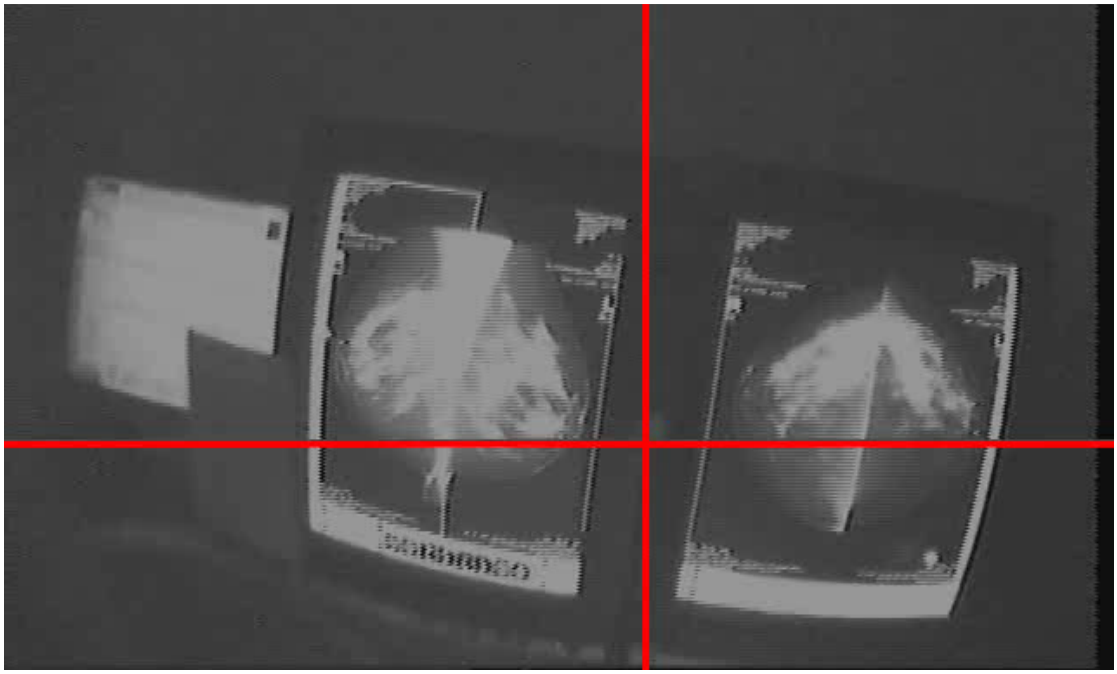


Figure 2-5. Cross hairs overlaid on scene

The calculation of the point of regard or 'line of gaze' is based on the fact (Merchant & Morrissette, 1974) that the measured separation between the centre of the pupil and the corneal reflection changes with eye rotation, however, it does not vary significantly with eye translation (i.e. head movement). The relationship between the line of gaze and the separation between the pupil and corneal reflection (*PCR*) is reduced to (see figure 2-6):

$$PCR = K \sin(\theta) \quad \boldsymbol{\theta} \qquad \textbf{Equation 2.1}$$

Where θ is the eye line of gaze angle with respect to the light source and camera, and K is the distance between the pupil centre and the centre of corneal curvature. The technical description of the ASL model 501 head-mounted eye tracking system is described below (table 2-2):

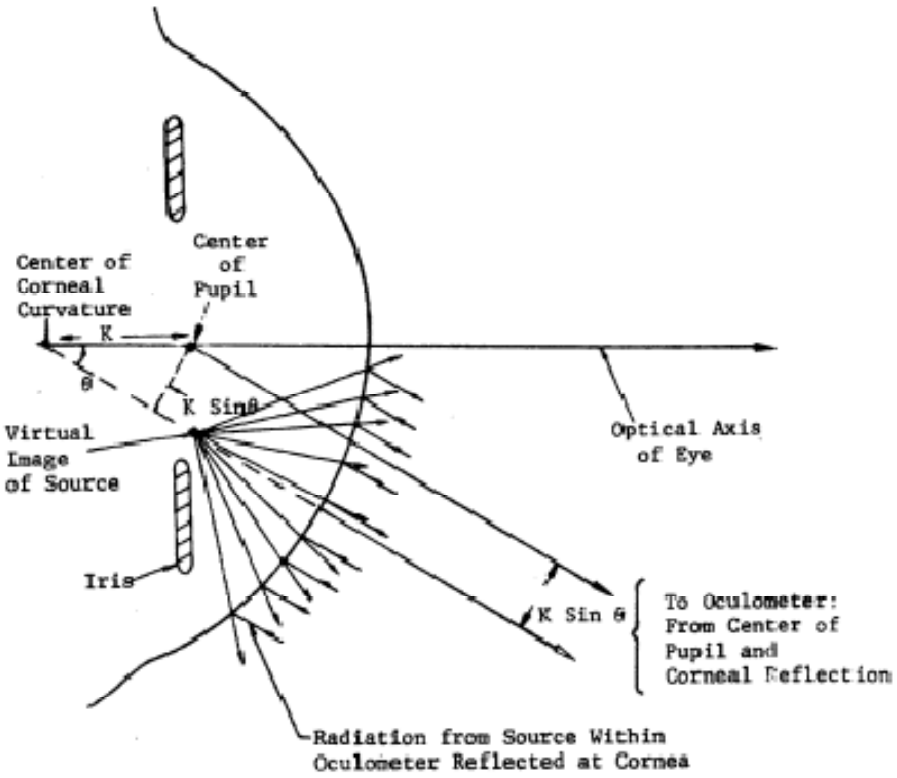


Figure 2-6. Relation between line of gaze and pupil and Corneal Reflection separation (Merchant & Morisette, 1974)

Table 2-2. Technical characteristic of the ASL eye tracker (adapted from ASL Eye Tracking System Instructions Model 501)

Characteristic	***ASL 501***
Data output	Time stamp H-POS: Horizontal Eye Position; V-POS: Vertical Eye Pupil Diam: Horizontal pupil diameter measure in pixels
Accuracy	$< 0.5^{o}$
Spatial resolution	Spatial error between true eye position and computed measurement is less than 1^{o}. Errors may also increase when gaze is beyond the outer boundary of the calibration pattern.
Freedom of head movement	Essentially unlimited due to free head motion
Head-movement compensation error	Errors may increase as the head moves significantly from its initial position.
Frame rate	50Hz in the UK
Ambient illumination	Complete darkness to moderate illumination resulting in pupil diameters greater than 3mm. Brighter environments possible
Maximum gaze angles	Along the horizontal axis, 50^{o} or more. Along the vertical axis, 35^{o} (or more) depending on optics placement and eyelids. (Field will generally be oval in shape.) Optimal performance will be achieved with an eye to camera distance of 20-25” (50.8- 63.5cm).

Fixation Algorithm Criteria Description

Identifying an eye fixation is a fundamental and important part of eye movement data analyses. The ASL system uses proprietary software ('Eyenal') to analyse the raw data generated by the system and turn it into useful measures. Fixations are calculated using spatial and temporal parameters by means of the following algorithm, as descriptively described in figure 2-7. Three key criteria are employed:

Criteria 1: To "start a fixation" the program looks for a specified period (*Min Sample*) during which the eye gaze has a standard deviation of no more than a specified amount. *Min Sample* default setting (of approximately 100msec) is 5 samples (in the UK for 50Hz)

Criteria 2: The horizontal and vertical distance (DX, DY) of the next data sample from the temporary data means (XT, YT)

Criteria 3: For minimum fixation duration Y: 200msec is the recommended value.

Note that figure 2-7 emanates from the second eye movement system used in the present research (Tobii) but the algorithm here is the same as that of the ASL system.

Area-of-Interest (AOI)

The Area of Interest (AOI) is the area around that part of an image of interest (i.e. a 'target') which allows for the particular eye tracking system's accuracy in determining that the person was, or was not, actually looking at the target. Thus, a recorded eye fixation falling within the target would correctly be taken as the person looking directly at it. Additionally, if an eye fixation fell within an area just outside the target would also be taken as the person looking at the target – this area then allowing for the eye tracking system's measurement error. This area is defined according to the accuracy of the eye tracking

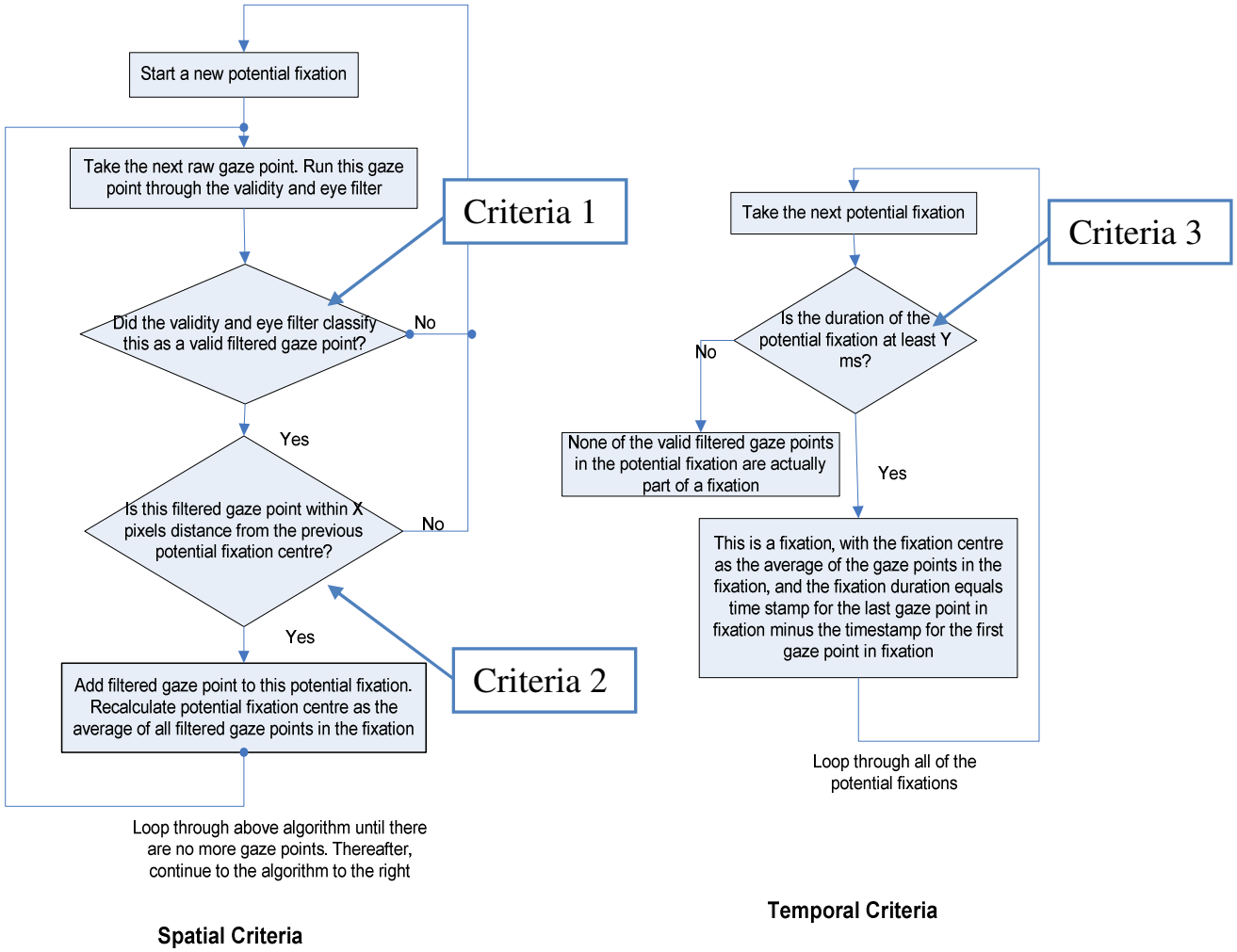


Figure 2-7. Fixation organization method (adapted from the Tobii Analysis Software User Manual)

system. A spatial proximity of within 0.5° visual angle (subtended at the observer's eye) was suggested according to the ASL guidance. Participants viewed the monitor at a viewing distance of 'D'. Consequently, this distance was used to calculate the visual angle subtended by the image. A 'hit' on the target was scored when the measured point of fixation fell on the target or within 0.5° of the edge of the target. Usually in this research the target in the images was an abnormality or a particular mammographic feature.

Figure 2-8 shows the relationship between the distance within the image scene, the subject's eye to displayed scene distance, and the visual angle (for small visual angles in the centre of the scene).

$$S = D * 2 * \text{Tan } 1/2\ \theta \qquad \textbf{Equation 2.2}$$

- S is the distance on the scene,
- D is the subject's eye to scene distance,
- θ is the visual angle, for the ASL system, $\theta = 0.5^{o.}$

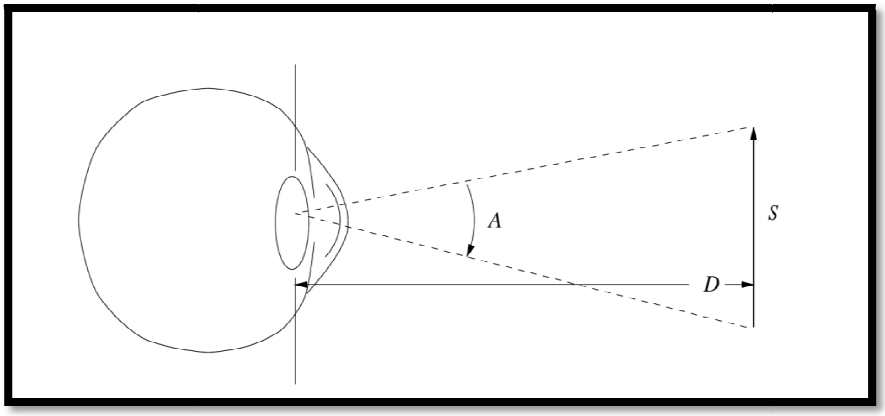


Figure 2-8. Visual angle (Duchowski, 2007)

GazeTracker

GazeTracker™ is further software that is used with ASL's eye tracking systems. It consolidates the extracted eye data with the frame of video that is captured by the ASL 501 scene camera and then generates the eye movement video. It also allows viewing the data easily by creating a database

for the data collected and calculating output statistics based on user-definable regions of interest and fixation data.

GazeTracker was used during the recording of all data in the experiment described in Chapter 8 which was a major study and spanned across 9 months. During the piloting for this study the software worked and during the experimental trials it clearly appeared to be functioning and recording data appropriately during each trial. However, unfortunately at the completion of the experiment it was found that the software had failed to record any useful or meaningful data in the database. This left only the videotaped recording of the scene camera view with the superimposed x-hairs of where the participant was looking moment to moment. A method had then to be devised to extract useful eye fixation data from these videotapes.

To validate this method, a program was written in Visual Basic to record the x and y co-ordinates of the recorded eye fixation positions within every video frame. Then these coordinates were transformed into fixation data using the fixation algorithm criteria described above. After the transformation, one person's eye movement data from the ASL Eyenal software was plotted and compared with the fixation data produced from this developed software. The result (figure 2-10) confirmed that using the developed program produced comparable data to that produced by the ASL Eyenal program. Therefore for this experiment to derive eye movement measures, instead of using the automated GazeTracker software, it was necessary to hand analyse all 140 hours of the recorded data using this new program so as to generate useful eye movement information. The whole process is illustrated in the flowchart in figure 2-9.

The ASL scene camera generates the scene video and the program Eyenal generates raw eye movement (x,y co-ordinates alone with no recording of the scene camera video) and fixation data as well as an output of the scene overlaid with a small white fixation cross indicating the participant's point of gaze. These data are then fed into the GazeTracker software which generates a large red video overlay of the point of gaze superimposed on the scene

camera view. It also inputs information into a database concerning the eye movement data. However, the database failed to save any data. Consequently there was only the GazeTracker generated eye movement video and the Eyenal videos which could be used for any analysis.

A Visual Basic (VB) 'fixation capture' program was written to try to extract from the videos the co-ordinates of the centres of the white and red crosses. It was found that the Eyenal white cross contained too few pixel values to be used reliably. Consequently, the GazeTracker cross co-ordinates were used which gave raw eye data. These data were then converted into fixation data through the use of the ASL algorithms. Then a comparison was made between the fixation data emanating via the VB program with the fixation data generated by Eyenal to confirm comparability (figure 2-10). This demonstrated that the developed approach produced comparable data. The GazeTracker videos for all participants were then fed into a second VB program which allowed the video to be replayed and jump from fixation to fixation whilst simultaneously allowing the experimenter to manually record fixation related information (e.g. fixating within AOI etc.)

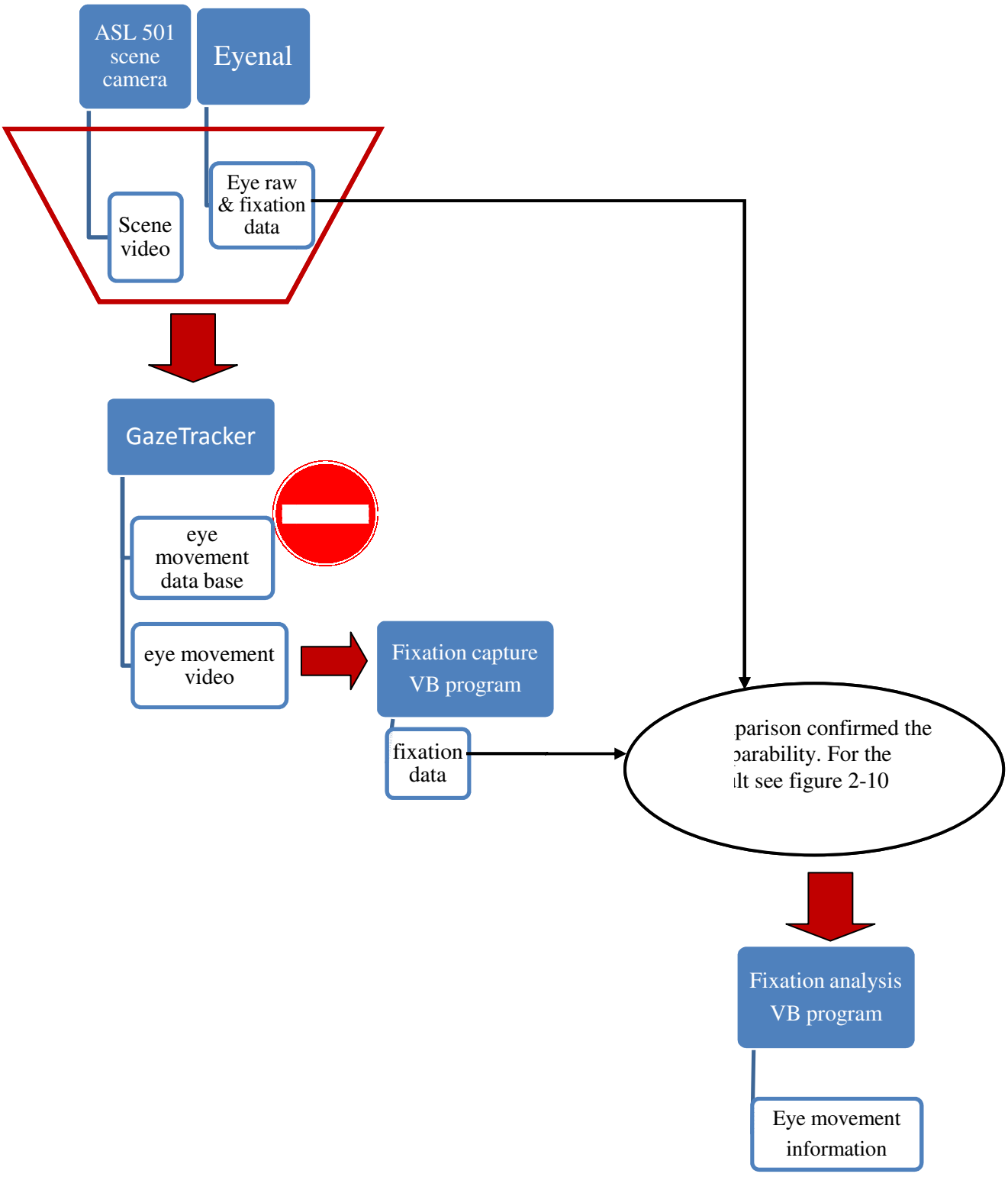


Figure 2-9. The flowchart demonstrates the process of eye movement data analysis

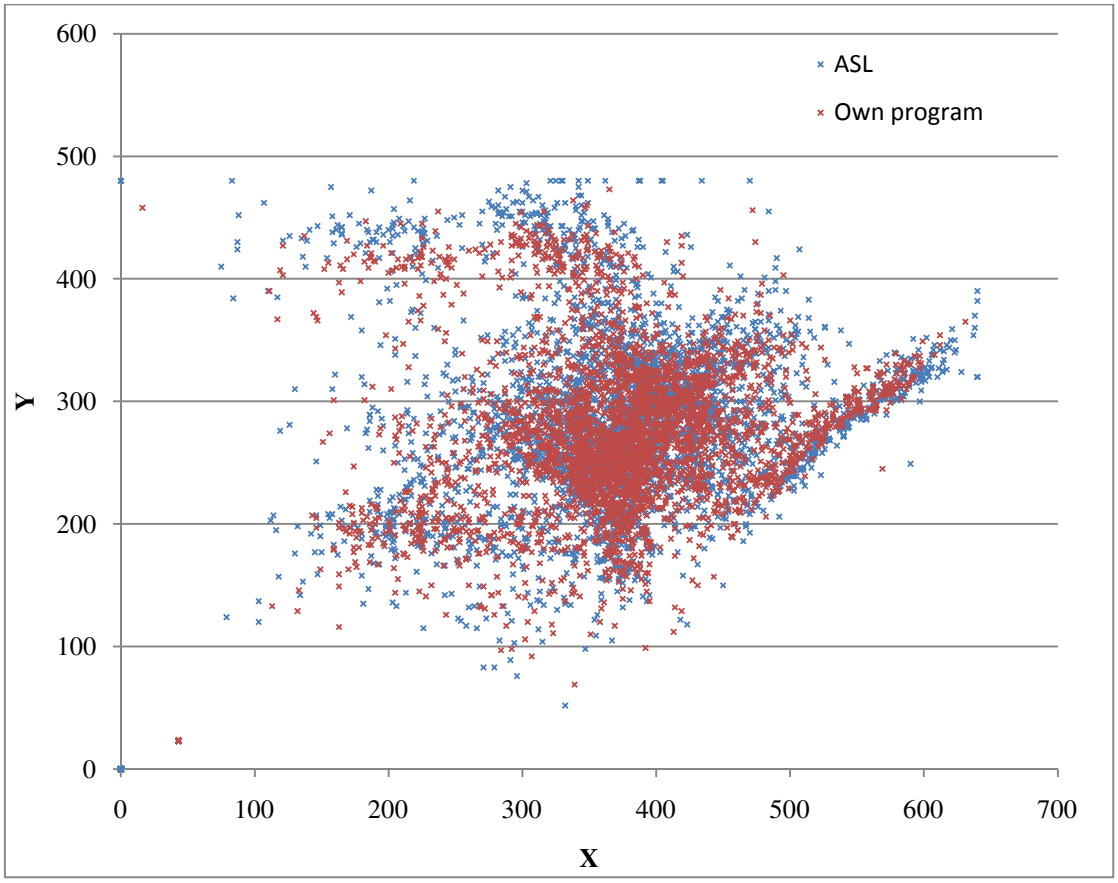


Figure 2-10. Eyenal *vs* VB program

2.1.4.2 Tobii X50 System

The Tobii X50 is a stand-alone eye tracker (see figure 2-11) that can be attached under any monitor or can be used to perform eye-tracking relative to a physical scene. It requires working with the data analysis software - ClearView. Its technical description is detailed in table 2-3.

Fixation Algorithm Criteria Description

Fixation algorithms are velocity-based, dispersion-based or area-based. The Tobii ClearView analysis software uses the algorithm shown in figure 2-7.

For such data analyses it is recommended by ClearView that a fixation radius (X) is used which is equivalent to 30 pixels and with a minimum fixation duration (Y) of 100 ms for viewing mixed content images.

Table 2-3. Technical characteristic of the Tobii X50 eye tracker (adapted from Tobii Analysis Software User Manual)

Characteristic	***Tobii x50***
Data output	Time stamp Gaze position relative to stimuli for each eye (X and Y) Position in camera field of view of each eye (X and Y) Distance from the camera of each eye Pupil size of each eye Validity code of each eye
Accuracy	0.5-0.7^{o}
Spatial resolution	0.35^{o}
Drift	< 1^{o}
Freedom of head movement	30 ×15 ×20 cm from tracker
Camera field of view	20 ×15 ×20 cm from tracker
Binocular tracking	Yes
Head-movement compensation error	<1 degree visual angle compensation error for head translations in three dimensions and rotations across the entire head movement space
Top head motion speed	Approximately 10 cm/s, otherwise smearing effects in the camera image
Frame rate	50Hz
Latency	25-35 ms (the time taken from when the actual eye-position is recorded until data reaches the application)
Maximum gaze angles	± 35^{o}

Figure 2-11. Tobii eye-tracker X50

Area-of-Interest (AOI)

The accuracy of the Tobii eye tracking system is reported to be 0.5-0.7^{o} visual angle (see table 2.3). According to the Tobii ClearView guidance, a spatial proximity of within 1^{o} visual angle (subtended at the observer's eye) is used to define the AOI. Therefore, the range of x, y coordinates which form a circle of 1^{o} of visual angle in any direction from the location of the lesion can also be calculated using Equation 2.2, $\theta = 1^{o.}$.

ClearView

ClearView is software which is used along with the Tobii eye tracker for analyzing eye gaze data thus facilitating the interpretation of participants' behaviour. It is very easy to use and supports in-depth quantitative analysis and also high-level analysis with good data visualizations.

2.1.4.2 Comparison between the Two Eye Tracking Systems

The two eye tracking systems were reviewed and compared here by the system accuracy, the area of visual field, system usability, etc. Details are shown in table 2-4.

Table 2-4. Important characteristic comparison between ASL and Tobii

	Gaze angles	**Accuracy**	**Freedom of head movement**	**Easy to use**	**Restraint on the users**
ASL head-mounted Model 501	Horizontal axis: > 50°; Vertical axis: > 35° depending on optics placement and eyelids. (Field will generally be oval in shape.)	< 0.5 degree	Essentially unlimited due to free head motion	Calibration takes a considerate amount of time	Low, but the eye-tracker is mounted on the user's head
Tobii X50	± 35 degrees	0.5 – 0,7 degree	30 × 15 × 20 cm from tracker	Calibration is quick, automatic and long-lasting.	No restraint

Table 2-4 shows that both systems achieve high eye tracking accuracy, however, the Tobii system supports better usability. It has no restraint on the user and supports quick, automatic calibration. Therefore, it has been used in most of the studies which are described in the following chapters where the participants examined mammographic images displayed on monitors.

Also table 2-4 shows that the ASL head-mounted system allows a wider gaze angle in the horizontal axis (> 50° as compared to 35° with the Tobii system) as well as having unlimited head motion. Therefore, for the study (reported in

Chapter 6) that required participants to examine mammograms on twin large clinical monitors which were too large to be successfully accommodated by the Tobii system and where inevitably the participants would move their heads a lot then the ASL head-mounted system was used to ensure the quality of eye movement tracking.

2.2 Participants Interaction Behaviour

In the experiment reported in Chapter 6 one interest was in how participants examined images on large high resolution clinical twin monitors, an office monitor and an iPhone. It was important to monitor how they interacted with these displays and consequently a fixed video camera was used to film this. In order to analyse this behaviour then Captive L-2100 software was used. This software takes the recorded video and displays it in a window on a computer monitor. The experimenter can decide what activities to record and can then set up buttons on the monitor to identify the start and end points of any such identified activity (figure 2-12). The video is then played and the experimenter can identify the sequence of actions in real time. It is also possible to fast forward or rewind the video. The result is a spreadsheet of actions and their respective timings. An easy way to visualise this is to plot the various actions against time (figure 2-13).

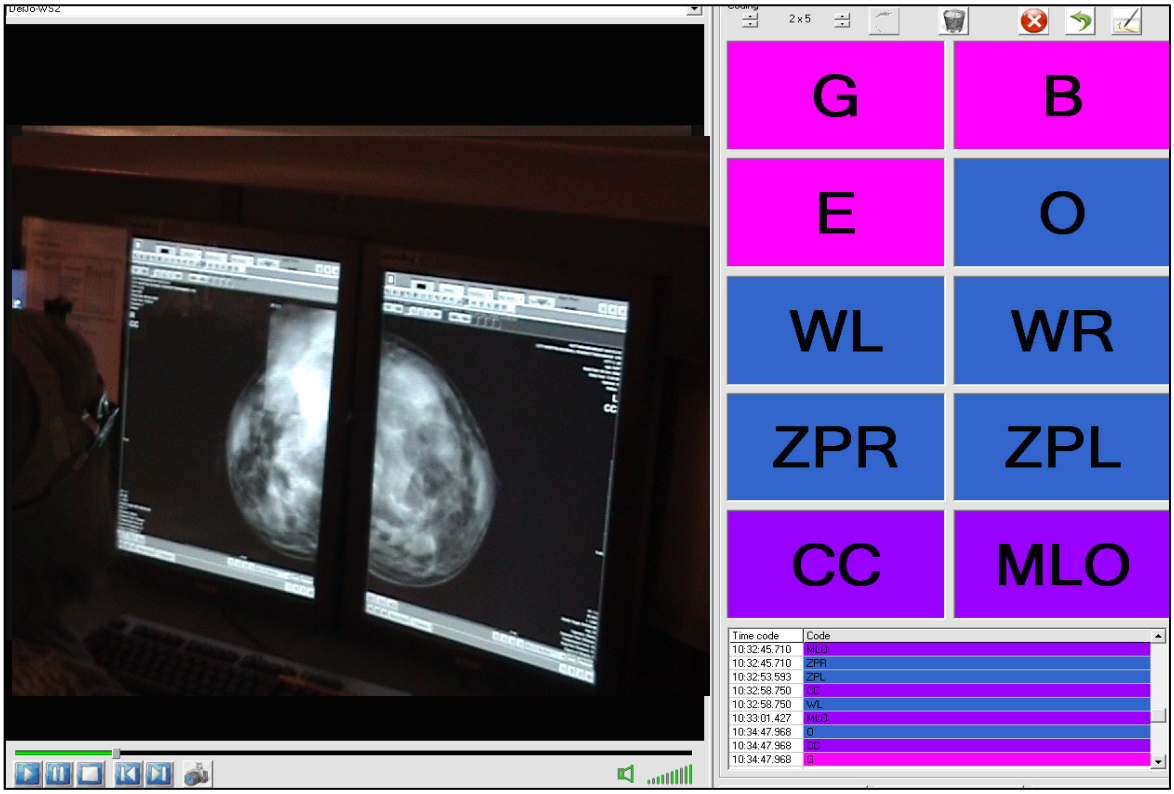


Figure 2-12. Screenshot of behaviour data analysis using Captive L-2100 software

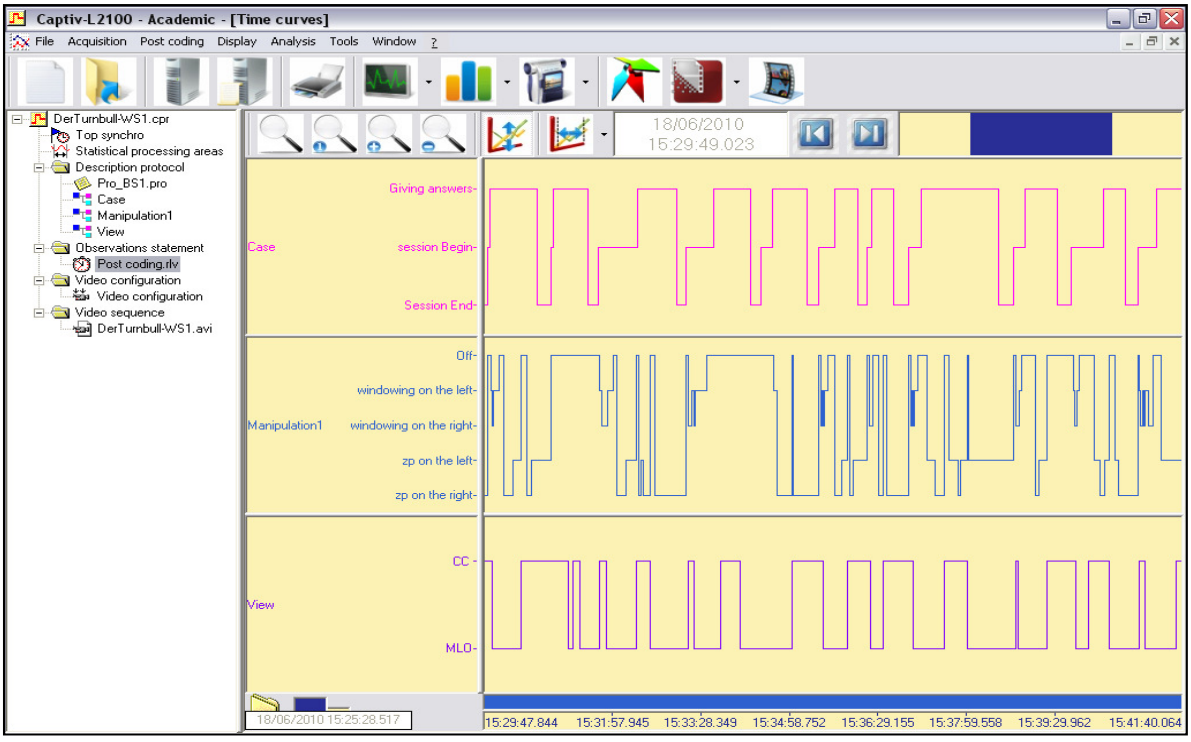


Figure 2-13. Screenshot of interactions sequence of one observer examining several cases

2.3 Receiver Operating Characteristic (ROC) Data Analysis

Many methods have been carried out to analyze ROC performance data. As discussed in the previous chapter (section 1.3.2), the sensitivity measure, d' becomes very unreliable if some rating categories are empty and so need to be collapsed; however, the other measurement, Area Under the Curve (AUC) is tolerant for bad data as compared to d' (McNicol, 1972).

For approximating the area under the ROC curve (AUC), one of the simplest methods is to use the trapezoidal rule. This is used to calculate the area under the ROC curve when each data point is plotted as connected to the next by a straight line. The trapezoidal can be represented as follows (Yeh *et al*., 1991):

$$\mathbf{AUC} = \int_a^b \mathbf{f(x)\,dx} \qquad \textbf{Equation 2.3}$$

The trapezoidal rule is a simple and straightforward way of calculating the AUC, although it is important to realize that the resultant value is normally slightly smaller than if a smooth curve were to be fitted.

Metz *et al.* (1998) proposed a general method for ROC curve fitting and statistical testing. It allows the utilization of unpaired data (for example, some patients' data may be collected in one condition but not in the other condition) and achieves additional statistical power. ROCKIT, is the algorithm based on this method and is an integrated package of all the previous ROC analysis software from the University of Chicago (i.e. ROCFIT, LABROC1, INDROC, CORROC2, and CLABROC).

ROCKIT is used to calculate maximum-likelihood estimates of the parameters of a conventional 'binormal' model, or a 'bivariate binormal' model for the input data; and to calculate the statistical significance of the difference between two ROC curves. The input data can be: unpaired (uncorrelated) test results; fully paired (correlated) test results; and partially-paired test results. It provides

95% confidence intervals for all estimates (for analysis of up to five modalities) (Metz, 1998).

In recent years ROC techniques have expanded to account for multiple readers reading multiple cases. One of the commonly used methods that are designed to allow analyzing data from several observers is the Dorfman-Berbaum-Metz Multiple-Reader-Multiple-Case (DBM MRMC) method (Dorfman *et al*., 1992; Dorfman *et al*., 1998; Roe & Metz, 1997). The method generates pseudo-values of ROC parameters for each reader-case set combination and compares these using a mixed-effect and analysis of variance (ANOVA) (Hillis *et al*., 2005). The null hypothesis of the ANOVA is that the average accuracy of readers is the same for all of the diagnostic tests included. The accuracy measure can be parametric or non-parametric which includes sensitivity, specificity, AUC, partial AUC, sensitivity at a fixed specificity, or specificity at a fixed sensitivity. Software for implementing the DBM method is available to download from the websites of the Medical Imaging Perception Laboratory of the University of Iowa[1]; or from the Kurt Rossmann Laboratories for Radiologic Image Research at the University of Chicago[2].

Although ROC analysis has been successfully implemented to evaluate diagnostic imaging systems, it firmly limits one reader report per case. However, in practice, when interpreting an image for possible breast cancer the reader may identify zero, or more malignant lesions. For instance in the PERFORMS scheme for each image an observer rates whether a number of different mammographic features are present. The experimenter is interested in the performance in correctly identifying each lesion along with its location. The immediate difficulty with this is that whilst it is easy to determine the cases which should be considered for the TP/FP decision it is difficult to determine which cases should be used for the TP/FN decisions (i.e. essentially all cases not containing that particular abnormal feature). In this free response task, a new Jackknife FROC method (JAFROC) (Chakraborty,

[1] http://perception.radiology.uiowa.edu/Software/tabid/109/Default.aspx
[2] http://www-radiology.uchicago.edu/krl/roc_soft6.htm

et al. 2004) has been proposed for the analysis of the ratings and location data. It combines elements of FROC and the Dorfman-Berbaum-Metz (DBM) multi-reader ROC methods and avoids AFROC's justifiable criticism of assuming independence between the multiple ratings obtained on the same case by only using the highest noise responses from normal cases (JAFROC-2).

JAFROC-2 generates greater statistical power than the ROC method (i.e. modified MRMC), especially pronounced in difficult case sets (Chakraborty & Berbaum, 2004). In JAFROC analysis when a case is jackknifed (Dorfman *et al*., 1992) all marked ratings on the case are removed from the analysis and each case yields one pseudo-value. However, the jackknife pseudo-values are only asymptotically independent and normally distributed and therefore do not satisfy ANOVA assumptions. The bootstrap method is therefore used for significance testing in the latest release of JAFROC software for achieving a correct Null Hypothesis without sacrificing statistical power (Chakraborty & Yoon, 2008).

JAFROC uses the non-parametric Mann-Whitney-Wilcoxon U-statistic applied to the lesion ratings and the ROC-equivalent ratings of normal images generating a figure-of-merit θ (equivalent to the trapezoidal area under the ROC curve) that credits the observer for good decisions (true positives and true negatives) and penalizes the observer for bad decisions (false negatives and false positives) (Chakraborty & Berbaum, 2004). JAFROC software is available to download from Dev Chakraborty's FROC web site[3].

On this website, some practical suggestions are also given for designing and conducting a free-response study. The most relevant suggestions for the studies that are described in this thesis are summarized as follows:

1) The level of difficulty of the test set needs to be strictly controlled. If the images are too easily interpreted, then the observers may not generate

[3] http://www.devchakraborty.com/downloads.html.

appreciable numbers of non-lesion localizations (i.e. false positives) and the data set cannot be analyzed. To achieve difficulty, using images from a previous screening round where the lesion may be less visible is proposed;

2) It is very important to ensure the localization accuracy;

3) It is important that observers are familiar with the tasks of the observer study and the user interface. Also, the use of a rating scale that observers are familiar with is suggested, such as the Breast Imaging Reporting and Data System (BI-RADS). However, BI-RADS is not implemented in the UK. Therefore, the UK rating scale is employed in the studies described in the following chapters;

4) Ideally, 50 normal images, and 50 abnormal images with one to two malignant masses per image for each modality should be included in the study.

2.4 Sample Size in Clinical Studies of Medical Imaging

In a comparative research study, the statistical power of the statistical analysis test is greatly affected by the sample size. Therefore, it is important to mention here that a great deal of previous literature published within the research area of medical imaging has obtained valuable results by using a relatively small number of participants, e.g. three or four expert mammogram film readers.

Obuchowski (2004) recommended multiple-observer studies being critical to clinical studies of medical imaging. Three phases were proposed to evaluate the diagnostic performance of a medical imaging system. Phase I is mainly to determine if it is worthwhile to further pursue the study. As a pilot study, it is suggested to include a small sample of patients, often 10–50; and 2-3 observers; phase 2, often has between 5 and 10 observers examining 50-200

challenging cases, with the purpose of comparing the accuracy of the tests and report differences in observers; the final phase is for testing the generality of the sample to a population of observers. Therefore, it requires collecting performance data from more than 10 observers from several different institution/ hospital sites. The required sample sizes for these studies can be estimated by analyzing data collected in the pilot study using the formal Cohen's ***d*** calculation and ***r*** calculated by Pearson's Correlations for the strength of relationship between scores at each level of the provision of information factor (Maxwell *et al.*, 1990).

Cohen's ***d*** is defined as the difference between two means divided by the standard deviation for the data

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}, \qquad \textbf{Equation 2.4}$$

In the research presented here these figures were then used to derive estimates of the required sample sizes based on designing a study with a power of .80 at an alpha level of .05 by consulting the look-up table (Maxwell, 1990. table 13.10).

CHAPTER 3

Mammographic Interpretation Training in the UK

3.1 Introduction

Chapter 1 introduced the difficulty of examining mammograms for early signs of breast cancer and the perceived need for mammographic interpretation training. This chapter expands on this by investigating breast screening personnel's views on current training available in the domain and their perceived need for training in the future. This especially concentrates upon the implementation of digital mammography and the potential use of a range of smaller computing devices as adjuncts to high resolution clinical display monitors. Focus groups and a user requirements study paved the way for a national survey of opinions.

3.1.1 The Mammographic Interpretation Task

As described in Chapter 1 mammographic interpretation is well known to be a difficult task that can only be carried out by specially trained personnel. In the UK, the interpretation of screening cases used to be only performed by consultant radiologists. In recent years some radiographers have been specially trained, as advanced practitioners, to also undertake this screening reporting role. Additionally, developments within radiography have led to the institution of the role of consultant radiographer who takes on aspects of the broader consultant role (e.g. clinical research, education and training role). A third group of health professionals also now undertakes screening reporting and these are mainly physicians who have a special interest in screening and have been trained to read screening cases. The expansion to include other professions has come about due to the growth of screening itself coupled with the limited number of consultant breast radiologists.

In the UK all such screening personnel are invited annually to participate in the PERFORMS self assessment scheme and from the recorded participant details it is then possible to derive an overall view of the various categories of professionals who are actually routinely screening. Data from the scheme has recently been reported (Gale, 2010) indicating that in 1991 some 150 consultant radiologists were screening nationally and took part in the scheme

(Gale & Walker, 1991) with this number increasing to 250 radiologists in 1996 (Gale, *et al*., 1996). Since then, other professions have taken up breast screening and currently well over 700 individuals are actively screening and take part in the PERFORMS scheme. Table 3-1 shows these professions as identified in the 2009/2010 round of the PERFORMS scheme.

Table 3-1. The number of occupations reading mammograms in 2009 (PERFORMS, personal communication, 2010).

Profession	***Numbers***
Advanced Practitioner	239
Consultant Radiologist	378
Associate Specialist	4
Breast Clinician	19
Breast Physician	13
Symptomatic Advanced Practitioner	13
Symptomatic Radiographer	45
Registrar	17
Other	3
TOTAL	**731**

The difficulty of the breast screening task is evidenced by data from the most recent PERFORMS scheme report for 2009/2010 where participants read 120 recent difficult screening cases. It was found that for these particular cases the mean national Correct Recall percentage across all UK participants was 79.6% with a 95% confidence interval of 65.4% to 93.8%; additionally, the mean national Correct Return to Screen percentage across all UK participants was 87.6% with a 95% confidence interval of 75.1% to 100.2%. See figure 3-1 and figure 3-2 respectively (PERFORMS, personal communication, 2010).

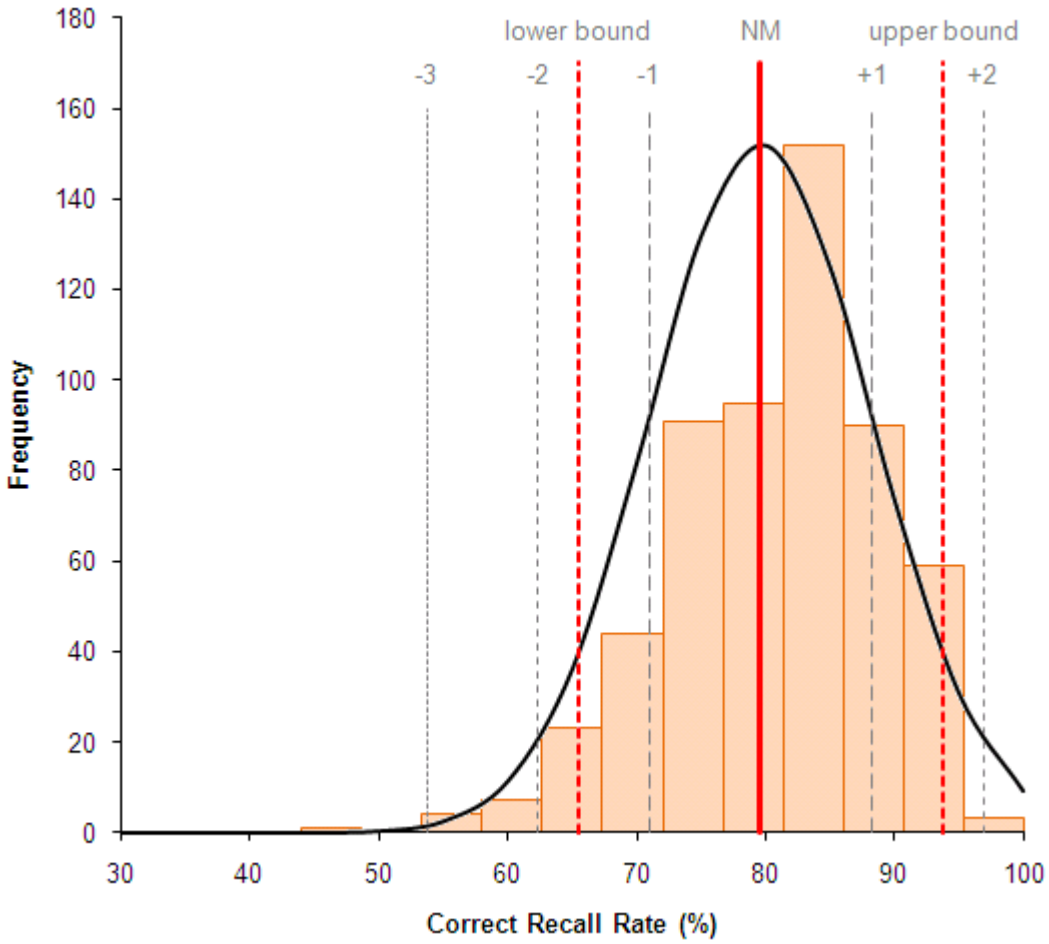


Figure 3-1. National variations in Correct Recall percentage for the PERFORMS scheme 2009-2010

These PERFORMS data evidence that on average over 20% of these difficult cases which contained an abnormality and that had originally been identified successfully in screening were incorrectly reported nationally. Also, in a similar fashion, over 10% of those cases which in screening had been correctly identified as being normal were nationally incorrectly judged to need recalling for further investigation. These findings serve to highlight the difficulty of screening and the wide variability in actual skill levels in screening which emphasises the importance of efficient training both to improve, and maintain, screeners' everyday performance.

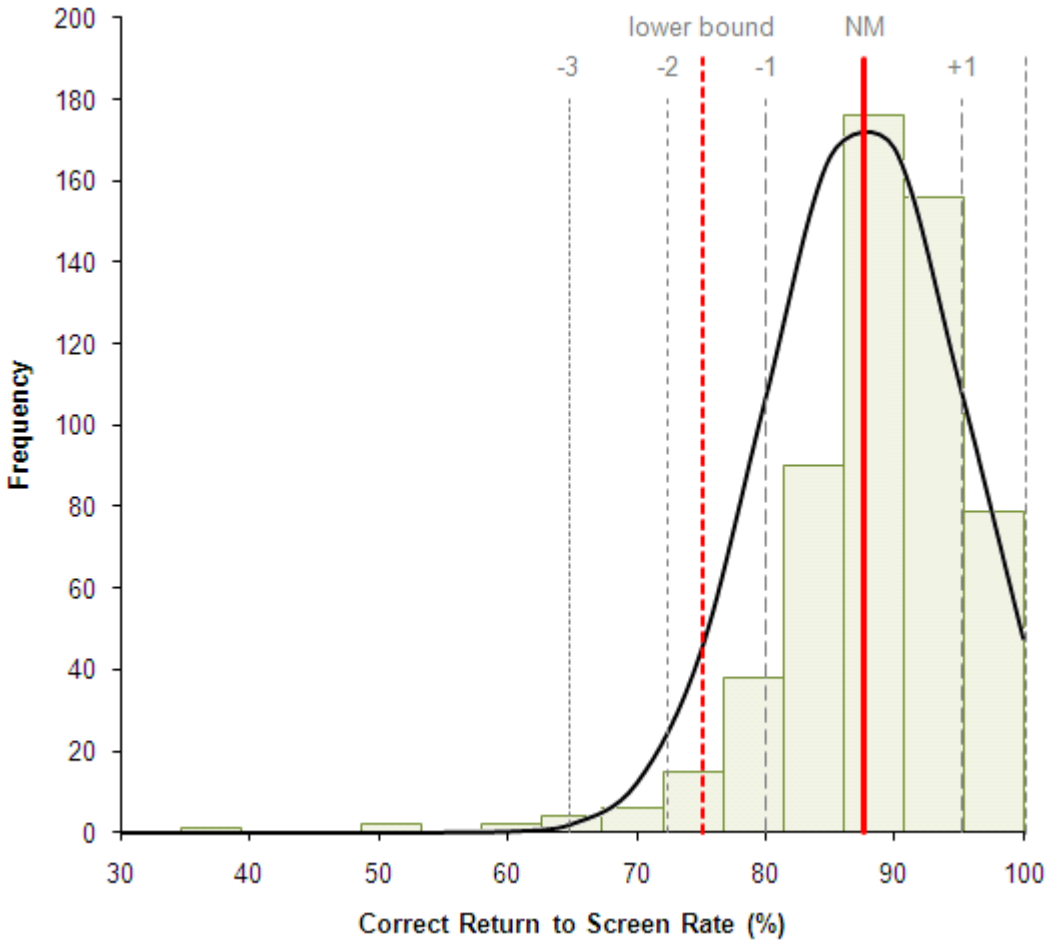


Figure 3-2. National variations in Correct Return to Screen percentage for the PERFORMS scheme 2009-2010

Furthermore, recent changes within the NHS Breast Screening Programme have increased the need for more training. The upper age limit of women invited to take up screening in the UK has been extended from 65 to 70 years, and is now being further increased to encompass 47-73 year olds (Department of Health, 2007). This increased age range will substantially increase the number of cases annually examined in the UK's breast screening centres. In 2007-2008 this number was 2.5 million women invited for screening with 1,994,651 actually screened (Patnick, 2009).

The combination of such increased demand for screening, coupled with the fixed limited capacity of available screening centres and mobile screening units, put pressure on the manning of the national screening programme, particularly with regard to the examination and interpretation of the screened

images. Consequently, there is an additional need to train more screening personnel. As there is a limited number of screening consultant radiologists then an increasing number of radiographers need to be trained to quickly bring them to the advanced practitioner level, where they are qualified to examine and report breast screening cases within the national screening programme, in order that they can contribute fully to the NHSBSP.

At the same time, digital mammography is being introduced nationally (Department of Health, 2007), which means that both new and existing staff need further training in interpreting digital mammograms as breast appearance in these images is somewhat different from that on analogue film mammograms.

3.1.2 Mammographic Interpretation Training

In the UK, most current mammographic interpretation training needs to be undertaken where there is a mammo-alternator or other suitable radiographic light box to facilitate the viewing of film mammograms; consequently limiting the times and places, typically to the breast screening centres themselves, where training can take place. In contrast, the introduction of digital mammography (Department of Health, 2007) opens up new opportunities of providing such training potentially without the restriction of having to use current viewing devices and therefore tying training to the screening centres themselves. Whilst high-resolution specialist mammographic digital monitors in appropriate viewing environments are de rigueur for actual clinical reporting of screening cases it is hypothesized here that additional advantages of the digital image over film are in the flexibility of potential training opportunities afforded. For instance, training could take place whenever and wherever suits the individual using a range of computer display devices. Whilst such devices would not have the clinical resolution and grey scale depth necessary for clinical investigation these devices could present sufficient image quality to enable them to be used in a variety of training situations.

Accordingly, a questionnaire study was undertaken to identify what the current situation was in the UK of mammographic interpretation training and furthermore to determine potential users' opinions concerning the training opportunities that could be enhanced by the introduction of digital mammography.

3.2 Questionnaire Survey

3.2.1 Initial Pilot Study

Due to the shortage of literature concerning the current situation of mammographic interpretation training in the UK, an initial pilot study was undertaken using focus groups to investigate this. Focus groups are a means of collecting qualitative data and have the advantages of providing the opportunity of collecting in-depth information about multifaceted aspects in a relatively short space of time (White & Thomson, 1995). The approach has been an increasingly popular qualitative data collection method in health and nursing research (Krueger, 1995; Dilorio, *et al*., 1994; O'Donnell, *et al*., 2007; Perez, *et al*., 2007).

Some 18 mammographic film readers (including 12 radiologists and six other professions including advanced practitioners) took part from three of the main UK Breast Screening Training Centres, namely: the Jarvis National Breast Screening Training Centre; the Nottingham International Breast Education Centre, and the Manchester Breast Screening Training Unit. Participants were recruited to take part voluntarily. In addition, four film readers (two radiologists and two advanced practitioners) from the Breast Unit of Derby Royal Hospital also voluntarily took part. This Unit is one of the pioneers in the introduction of digital mammography in the UK and these individuals then had considerable experience of interpreting both analogue and digital mammographic images. In particular, their perceived training requirements and preferences given the national introduction and implementation of digital screening were well worth investigation.

Four focus group discussions were carried out to elicit information about film readers' understandings and perceptions of current training in the UK. Two of the four centres (Derby and Manchester) involved in this pilot study had digital mammography experience.

The questions covered three main aspects and are listed as follows:

- *Current mammographic interpretation training*
  1) What is your current mammographic interpretation training?
  2) What kind of training do you have at the moment?
  3) Where do you currently read your mammograms?
  4) Do you carry out any of your mammographic interpretation training in your spare time?
  5) Where do you normally carry out your training?
  6) Apart from attending training courses, what other type of training have you had or does your centre offer?

- *Identify any current difficulties about training*
  1) Do you have any difficulties/problems with the current training? If so then what kind of training problems do you have? What aspects are you not happy about?
  2) You have a full-time job as well as needing to undertake a lot of training. Is this problematic? How do you cope with this?

- *Ideal training*
  1) Ideally, what would you like for your training?
  2) What is the ideal situation for you to examine mammograms or to undertake this kind of training?
  3) Do you want to do training at home?

The film readers who participated were encouraged to raise any training related questions and associated issues. The discussion was tape-recorded, transcribed and analysed according to the three aspects listed above. Data

suggested that mammographic interpretation training for film readers (especially for advanced practitioners) was relatively problematic, although a number of different types of mammographic interpretation training were potentially available. The problems reported mainly focused on limited time for training and limited access to the training display devices (especially for the digital viewing clinical monitors). A lot of comments were made about ideal training which concentrated on what was needed to be involved in training. For example, being able to view good exemplar cases with the guidance of experienced film readers was on the top of the wish list of ideal training. However, very little information was collected about the delivery of such training.

One disadvantage of focus group is the risk of biasing data if equal participation is not completely guaranteed (Straw & Smith, 1995). It is then possible that the information collected from the pilot study was biased towards the actual participants (who were mostly senior film readers). Therefore, it was important to have a follow-up survey to collect information from a much wider population.

3.2.2 User Requirements Study

To build upon the pilot study, a second more detailed investigation of a group of film readers (*n*=4) from the Breast Unit, Derby Royal Hospital was undertaken. They were questioned in semi-structured interviews about user requirements with regards to training issues. This Unit is one of the pioneers in the introduction of digital mammography in the UK and these individuals then had considerable experience of interpreting both analogue and digital mammographic images. In particular their perceived training requirements and preferences given the national introduction and implementation of digital screening were investigated. Content analysis of the data revealed several main categories of training preferences which included: the requirement for individualized and tailored training, the location and timing of training opportunities to suit the individual, and their perspectives on the possible

applicability to training of a range of image display devices including desktop monitors and hand-held devices.

3.2.3 National Questionnaire

On the basis of the previous studies, a questionnaire was then designed to amass full details nationally of breast screeners' current and future training needs. These needs included four main categories on the basis of the foregoing studies:-

- Details of an individual's current screening role (e.g. profession; experience of digital mammography);
- An individual's current usage of mammographic interpretation training (e.g. the types of training available; the number of training opportunities; any difficulties with current training approaches; the advantages and disadvantages of current training);
- Future training outlook (in particular their attitude towards a 'whenever and wherever' training approach; views on personalised tailored training that could be based on outcomes from an individual's performance in the PERFORMS scheme);
- Views towards what aspects of digital mammographic interpretation training could potentially be delivered on different types of computer displays and any further suggestions about mammographic interpretation training using digital images.

A draft questionnaire was first produced and this was then discussed and reviewed by the lead Quality Assurance radiologist of the East Midlands Region. The final version of the questionnaire (see Appendix A) was then sent out to in August, 2008 to all 109 Breast Screening Units in the UK. This meant that in total 601 current screeners (the figure reported from the PERFORMS scheme in 2008 as actively screening) were approached to seek their views on current and future mammographic interpretation training.

The questionnaire achieved a 45% response rate (n = 273) and encompassed the main screening professions, i.e. 152 consultant radiologists, 78 advanced practitioners, as well as other professions whose job involved mammographic interpretation, e.g. radiographers, radiology specialist registrars (SpRs). (See table 3-2 & figure 3-3 for full details). This confirmed that most screeners were radiologists or advanced practitioners and also ensured that the responses represented all types of screeners.

Table 3-2. The number of each profession who took part in the questionnaire

Profession	***Number***
Consultant Radiologist	152
Advanced Practitioner	78
Consultant Radiographer	4
Radiographer	15
Radiology SPR	3
Breast Clinician	15
Others	6
TOTAL	**273**

Data were also collected about the participants' experience of digital mammography as shown in figure 3-4. It was found that, at this time, only 7% of the participants were experienced in digital mammography. The majority had very little experience of digital mammography (i.e. 54% with none and 10% with less than six months experience). Some 8% of participants did not answer this particular question.

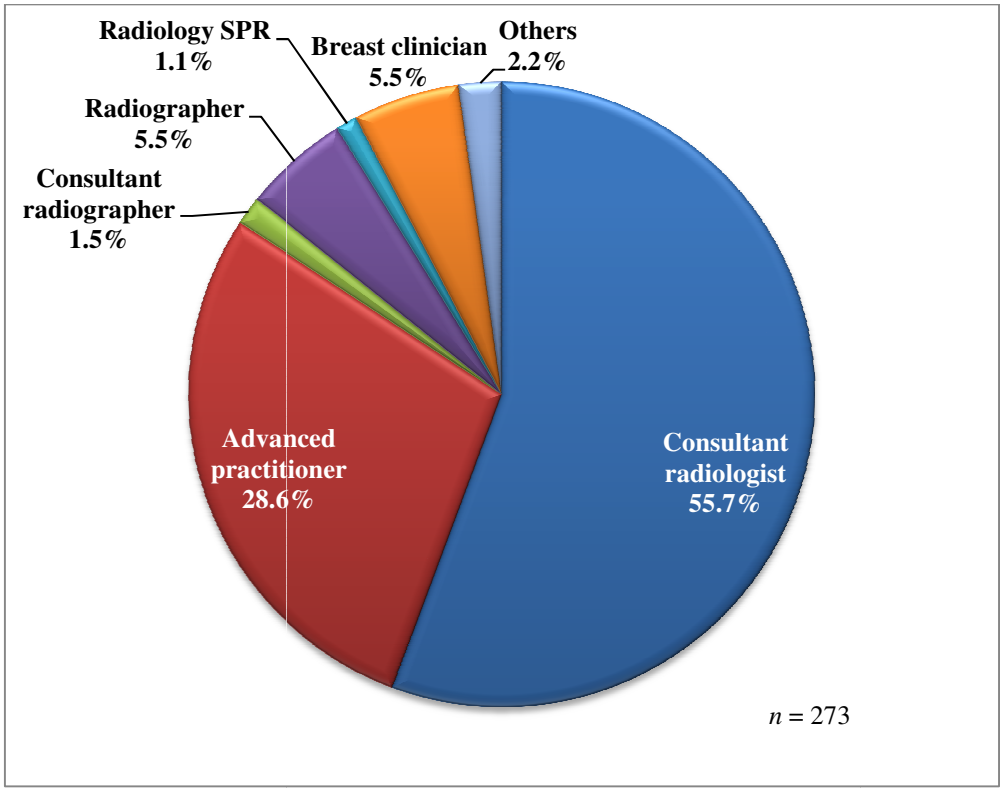


Figure 3-3. The percentage of participants' professions

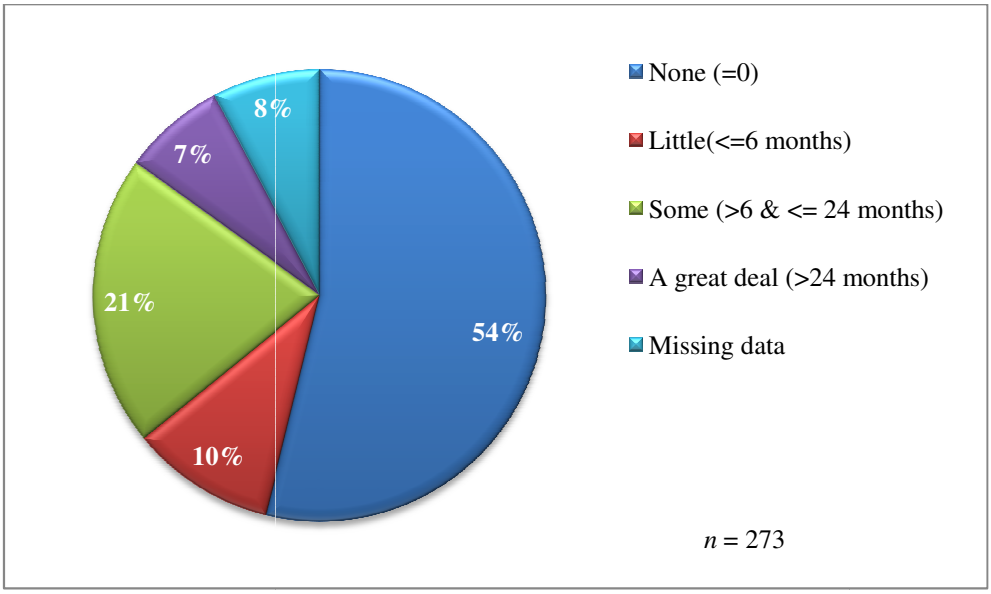


Figure 3-4. The percentage of participants' digital mammography experience

3.2.3.1 Current UK Mammographic Interpretation Training

The participants were then divided into three main groups for further analysis. This comprised 152 consultant radiologists (55.7%), 97 practitioners (including advanced practitioners, consultant radiographers, and radiographers: 35.5%), and 21 others professions (this included radiology specialist registrars [SpRs], breast clinicians, and other film readers: 7.8 %.)

The questionnaire data were analysed firstly as a whole and then examined in terms of these three different main participant groups; '*Radiologists'*, '*Practitioners'* and '*Others'*. Participants were asked to rate the current amount of mammographic interpretation training available to them. It was found that in general, 40% of participants considered that the opportunity for mammographic interpretation training "could be more" or "not enough". A Kolmogorov-Smirnov test was performed that the training opportunity rating deviate significantly from normality. The test revealed that the data was significantly non-normal data ($D(270) = 0.212$, $p<.01$). To explore the difference between these three independent groups on training opportunity rating, the Kruskal-Wallis test was used. This indicated that there were significant differences ($H(2) = 28.47$, $p<.05$)) between the three profession groups on their rating of training opportunity. 58.2% of participants from the 'Practitioners' group considered the opportunity "could be more" or "not enough". See figure 3-5 for the distribution of the opportunity rating for each group. Overall the radiologist's considered the opportunity to be adequate and the 'Others' group saw it as being more than adequate. However, the Practitioners group saw such opportunity as lower than the other groups. This result agrees with the data from the pilot study, which may well reflect the Practitioner group's growing demand for mammography interpretation training. It could be because the group has taken more and more responsibility for mammography film reading with the current workforce changes in screening in the UK.

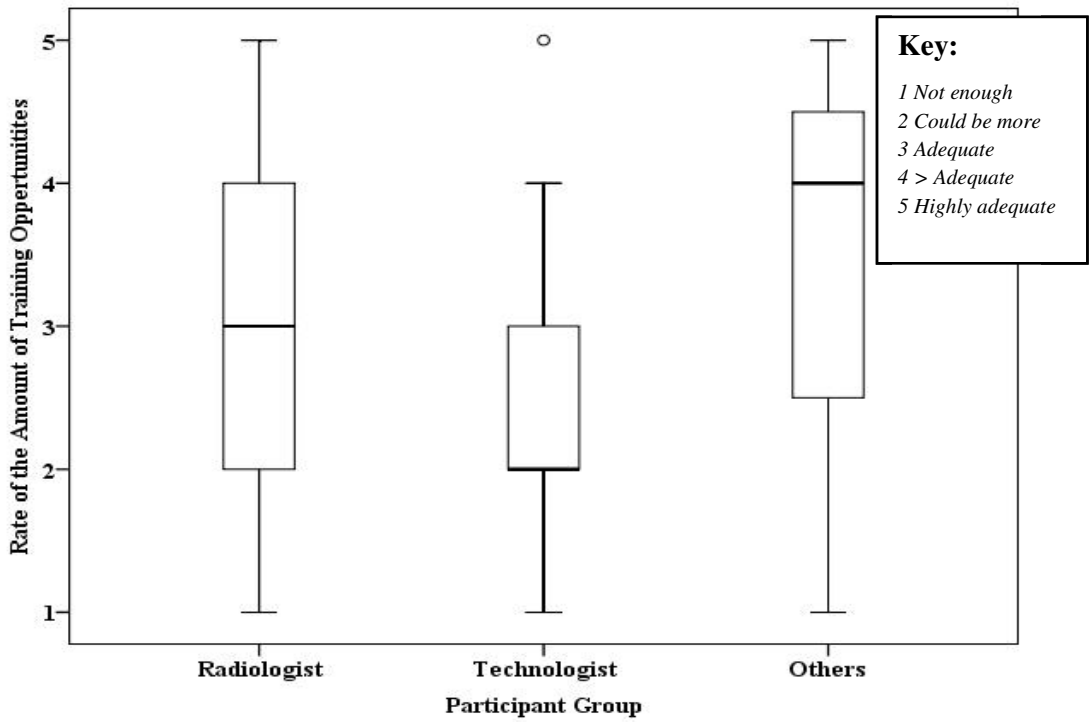


Figure 3-5. Rating distribution for mammographic interpretation training opportunity between professional groups.

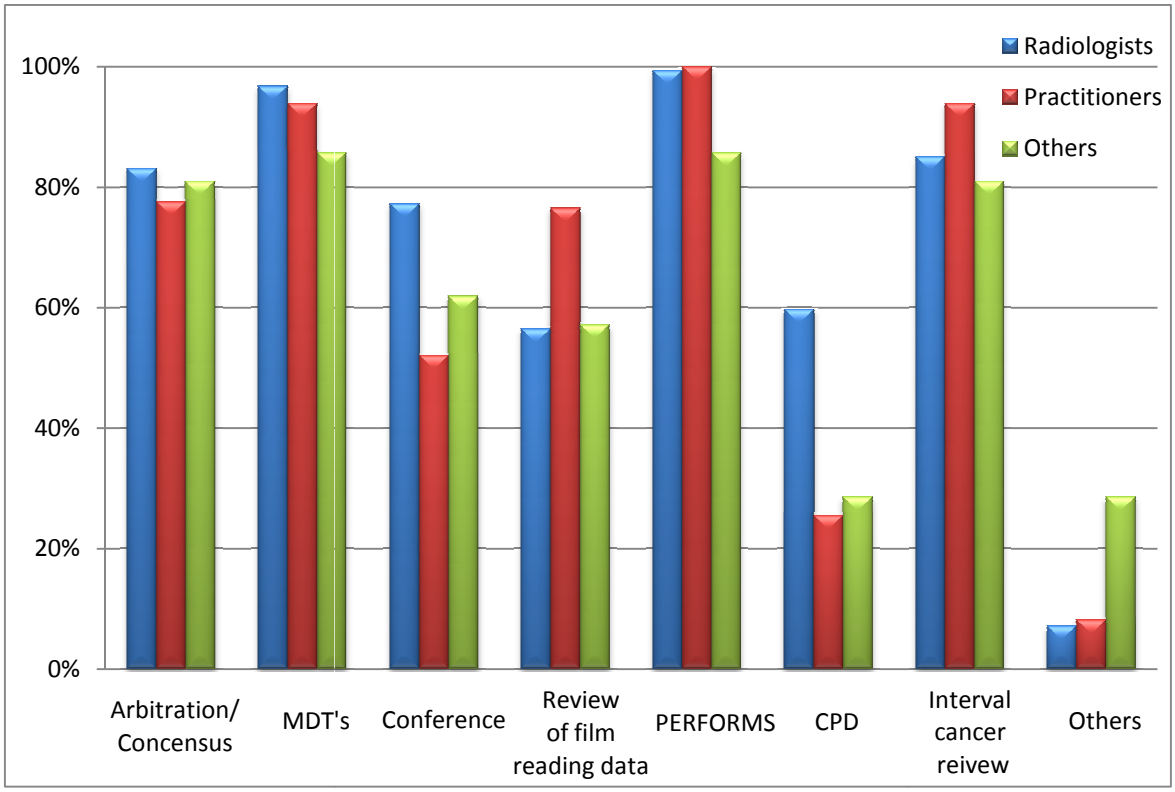


Figure 3-6. The profile for undertaking different forms of mammographic interpretation training

Data collected from these groups were then compared on the different forms of current available training, their ratings of the amount of mammographic interpretation training, and any identified difficulties when undertaking training. Current training included: arbitration/consensus; MDT's (multi-disciplinary team meetings); interval cancer review; conferences; review of individual film reading data; the PERFORMS scheme, CPD (Continuing Professional Development) courses, and any others. These are shown in figure 3-6 for each group. There was no significant difference ($p>.05$) between the groups for different types of mammographic interpretation training each group undertaking.

For current formal training opportunities, then multi-disciplinary team meetings (MDTs), interval cancer reviews (both held within the screening centres) and the PERFORMS scheme were the three most commonly used types of training. The commonly shared characteristic amongst all these three training types is that the readers are able to access a large number of representative difficult mammogram cases and feedback on each case by undertaking such training. It is therefore suggested that good example cases along with appropriate feedback is essential for a successful training method.

Difficulties Identified in Mammographic Interpretation Training

Some 142 participants (52%) reported training-related difficulties (66 radiologists, 67 Practitioners and 9 others). These difficulties were then grouped into four main sections, namely: limited time, limited access to a mammographic film roller-viewer, limited access to a digital workstation (where appropriate), and other difficulties (see figure 3-7). The reported other difficulties included; financial issues (e.g. costs of attending training courses), lack of validated training sets, lack of management support, etc.

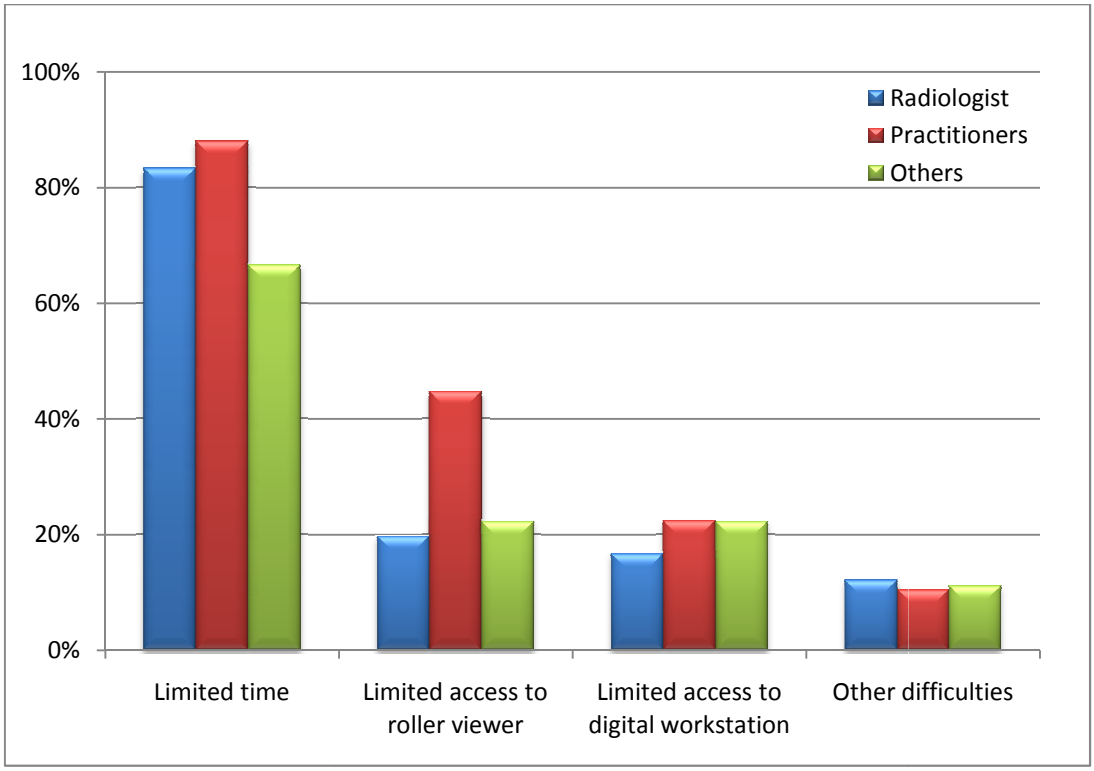


Figure 3-7. Training difficulty types: the percentage of participants who identified training difficulties

From figure 3-7 it is apparent that whilst all three groups had limited time this was primarily due to the limited access of the 'Practitioners' to roller viewers.

Attitudes to Potential Future Digital Training

Apart from details concerning current mammographic interpretation training, information was also sought on the screeners' opinions on possible future digital training; in particular in receiving training whenever, wherever convenient, and whatever suits the individual.

For the possible 'whatever required' training (e.g. tailored training) poll, 222 participants (81.3%) showed their strong interest. For the possible 'whenever, wherever' training poll (see figure 3-8), 213 participants (78%) responded positively. Within these participants, 79 of them specified the details of when they would like to undertake such training. This included 'anytime convenient', 'a specifically allocated time', 'during general working time', 'after-work', and 'others'. Of these, 135 participants have given detailed answers on

ideal places for their training (one or more answers were given), such as, 'in the breast screening centre'; 'anywhere convenient'; 'using the digital workstation', and 'using home PC', etc.

Additionally, 222 participants (81.3%) were very interested in undertaking tailored training based specifically upon data from their latest PERFORMS performance evaluation.

Furthermore, the correlation was examined between participants' experience of digital mammography and their attitude to different possible tools that could deliver digital mammographic interpretation training (i.e. digital workstation, desktop PC, laptop PC, and handheld device). Data showed that the respondents' attitude plotted against their experience of digital mammography (i.e. how many months of digital experience). Overall, not surprisingly, a Spearman's rank correlation test indicated that there was a positive relationship between all three groups' digital mammography experience, and their attitude to using the digital workstation as a tool to deliver mammographic interpretation training (*rho* (242) = 0.146; $p<.05$. $r^2 = 0.02$). It was shown that the screeners' confidence rose when they gained more digital mammography experience.

In terms of using other smaller displays for training purposes the results were more variable. For the Practitioners, a positive attitude to using smaller computer displays as a possible training tool was significantly correlated with their digital mammography experience: *rho* (67) = 0.252; $p<.05$. $r^2 = 0.06$ (desktop PC; figure 3-9a); *rho* (78) = 0.327; $p<.05$. $r^2 = 0.10$ (laptop PC; figure 3-9b); *rho* (54) = 0.278; $p<.05$. $r^2 = 0.07$ (handheld device; figure 3-9c). The data confirmed that increased digital mammography experience helped to improve screeners' confidence and enthusiasm on possible training delivered on different devices other than the traditional viewing device (i.e. digital mammography workstation and the multi-viewer). For the radiologists there was no significant correlation between their digital mammographic experience and attitude to using smaller displays for training purposes.

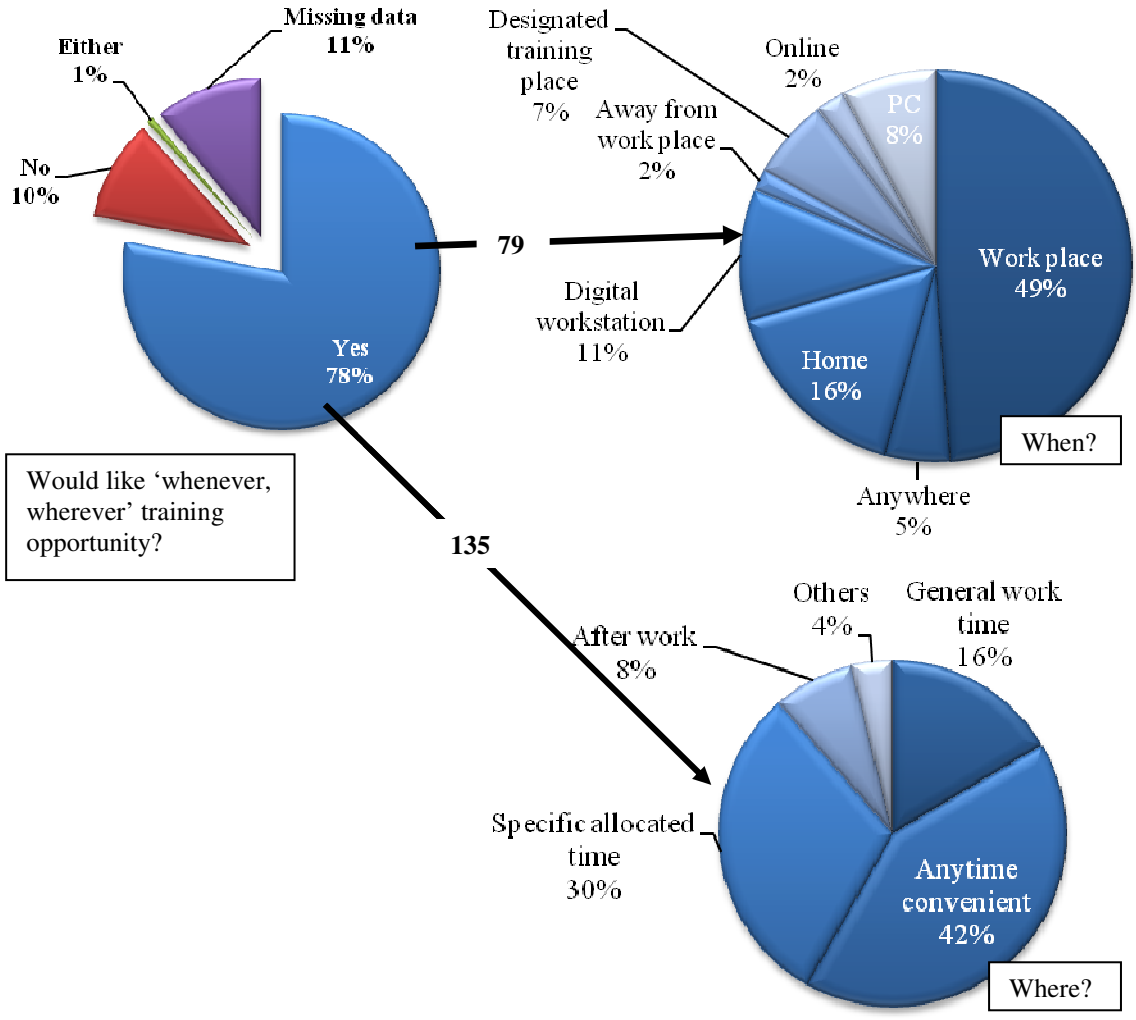
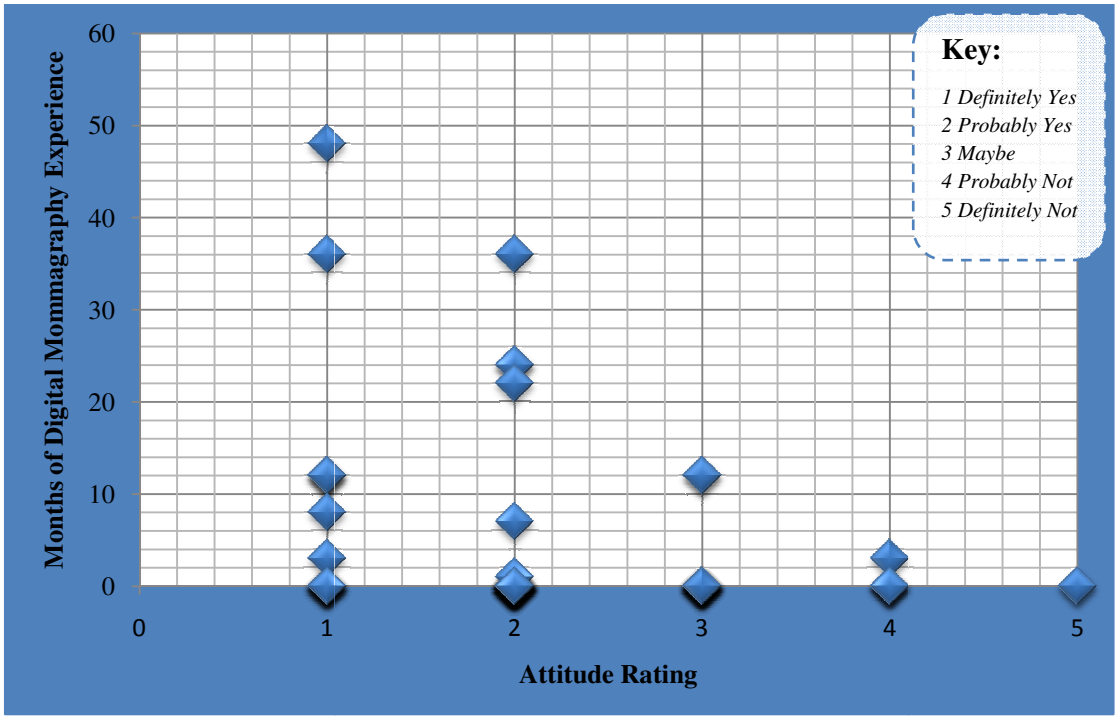
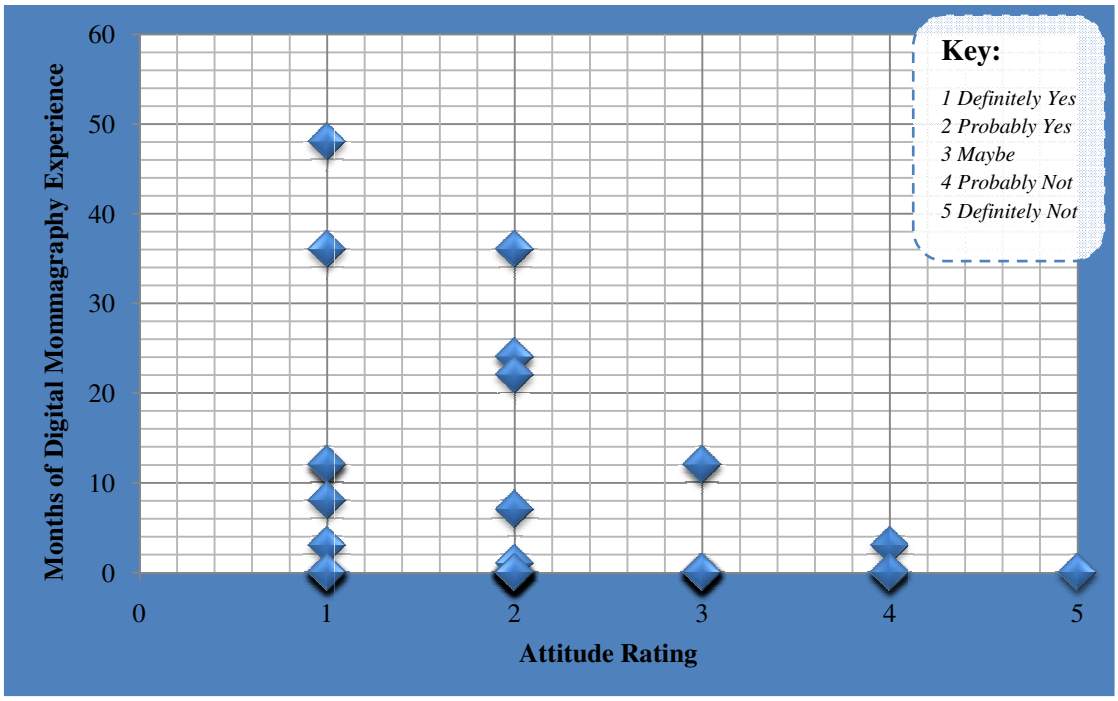


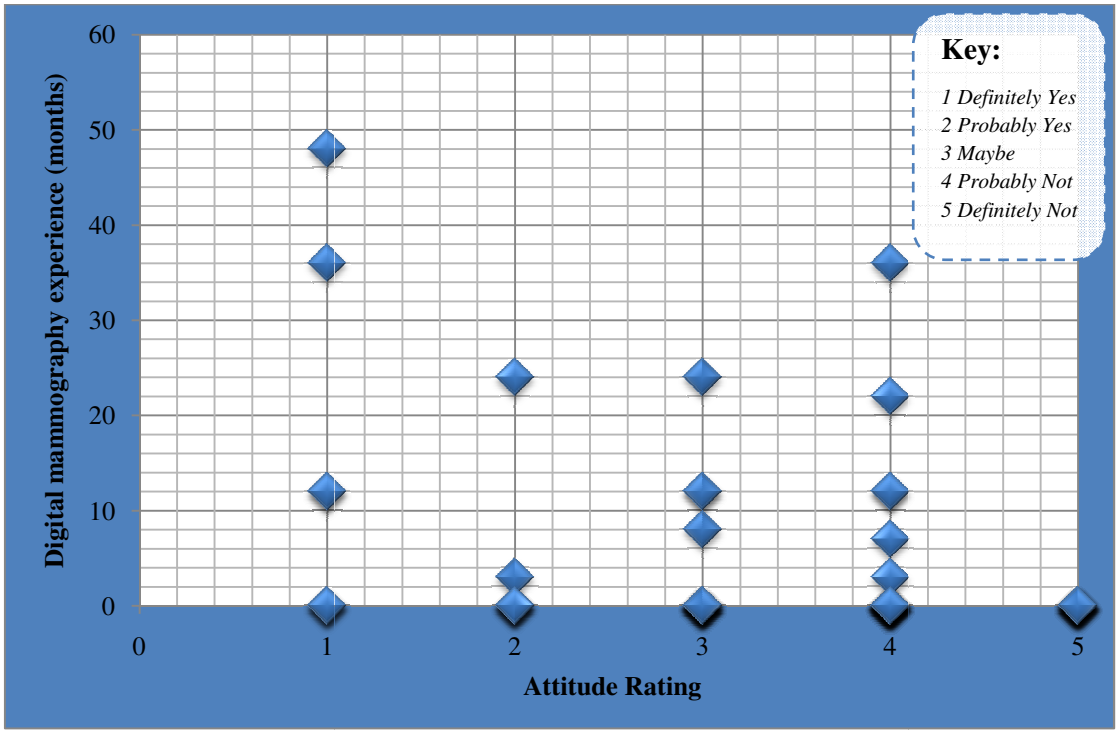
Figure 3-8. Attitudes towards 'whenever, wherever' mammographic interpretation training


Months of Digital Mommagraphy Experience
60
50
40
30
20
10
0
0
1
2
3
4
5
Attitude Rating
Key:
1 Definitely Yes
2 Probably Yes
3 Maybe
4 Probably Not
5 Definitely Not

a)


Months of Digital Mommagraphy Experience
60
50
40
30
20
10
0
0
1
2
3
4
5
Attitude Rating
Key:
1 Definitely Yes
2 Probably Yes
3 Maybe
4 Probably Not
5 Definitely Not

b)



c)

Figure 3-9. The Practitioner's digital mammography experience x Attitude rating
a) Practitioners' digital experience and attitude rate on Desktop PC;
b) Practitioners' digital experience and attitude rating on laptop;
c) Practitioners' digital experience and attitude rating on handheld device.

3.3 Discussion

To investigate the current mammography interpretation training situation within the UK Breast Screening Programme, a pilot study in three main Breast Screening Units was carried out followed by a semi-structured interview with groups of experienced breast screening film readers, along with a detailed and wide ranging national questionnaire. The questionnaire had an unusually high response rate (45.5%) which encompassed all the different professions in breast screening in the UK. This provided insight into the current and potential future availability of mammographic interpretation training in the UK. The majority of respondents were consultant radiologists (57%) with advanced practitioners making up 29%. Overall, 38% had had some experience of digital mammography: 10% having less than six months experience with the majority (21%) having between 6 and 24 months experience. Some 7% had more than 24 months experience. At the time of the questionnaire (August

2008), this broadly reflected the gradual introduction of digital mammography into the UK.

Considering the data in terms of the three main groups of respondents then over 80% took part in arbitration/consensus meetings on specific cases, interval cancer reviews and multi-disciplinary meetings where all professionals involved (including pathologists and surgeons) meet to discuss specific cases on a regular basis. Virtually all screeners took part in the PERFORMS scheme. Review film reading and conference attendance was seen variably by over 50% to be important. Continuing professional development (CPD) was rated as more important by the radiologists than the other two groups. Some 40% of respondents considered the amount of current training as being less than ideal, with the main difficulties classified as 'limited time' and 'limited access to image viewing facilities'. The Practitioners in particular indicated less current opportunities for training than did the other professional groups.

With regard to future digital training, some 81.3% of participants showed their strong interest in receiving tailored training on the basis of their individual recent PERFORMS scheme outcome and additionally 78% of participants responded positively towards having training whenever and wherever convenient. The questionnaire results showed a positive attitude to the use of small computer devices for further training. However, there appeared to be relatively low confidence, especially from the radiologist group, on training delivered on handheld devices. For the Practitioners, the data showed that a positive attitude to the use of smaller computer displays as a possible training tool was significantly correlated with their digital mammography experience, which suggested that such difficulties may be overcome with increasing digital mammography experience. It also suggested that Practitioners require more training opportunities if not an equal amount as the radiologists. However, they have less time and opportunities to gain access to the digital mammography workstation. Therefore, this leads to the Practitioners' higher confidence and requirement on the training that could be delivered on more widely accessible facilities.

3.4 Conclusion

This initial investigation into the current situation of mammographic interpretation training in the UK identified the difficulties of training availability, which could be due to the high UK breast screening workload. A positive attitude towards using some potential digital displays as a technology for delivering 3W ('whatever required, whenever, wherever') training was found. It also showed that Practitioners were very supportive of the potential of using different types of computer displays (e.g. PC, laptop, handheld devices) for delivering future mammographic interpretation training; however, radiologists were more reserved, which could potentially be as a result of: 1) more training opportunities; 2) better access to image viewing equipment (these are also showed by the questionnaire data); they did not divorce training from actually making screening identification decisions.

The gradual introduction of digital mammography opens up new opportunities for delivering mammographic interpretation training. For example, it increases training opportunities using lower resolution, lower cost and more widely available devices, in addition to the clinical digital mammography workstations. However, the potential for mammographic interpretation training on different viewing devices has not yet been comprehensively examined. Therefore, further research is needed to investigate how such training could be delivered as well as what type of training could be so delivered. The following chapter begins to investigate these issues.

CHAPTER 4

On-demand Mammographic Interpretation Training: initial studies

4.1 Introduction

Chapter 3 investigated the current mammographic interpretation training situation in the UK Breast Screening Programme and highlighted the importance of digital mammography training. This chapter describes a series of small pilot studies to investigate the possibility of delivering on-demand mammographic interpretation training using digital mammographic images.

4.1.1 Mammographic Interpretation

As discussed in Chapter 1, breast screening is a method of detecting breast cancer at a very early stage. The key step in the process is the visual examination of mammograms for the presence of abnormalities that are indicative of cancer. Detecting early signs of breast cancer is an especially difficult task due to the rarity of the disease within the screening population: an abnormality will be present in only approximately seven cases per 1,000 women (Patnick, 2005). Furthermore, the various subtle ways in which an abnormality can present increases the difficulty of correct identification. Therefore, this task is carried out only by trained mammographic film-readers. In the UK there are broadly three types of readers: advanced practitioners who are specially trained radiographers; consultant radiologists; also, there are increasing numbers of other professionals that undertake film reading, such as breast clinicians.

4.1.2 Supporting Mammographic Interpretation Training

Currently, in the UK, there are a few projects that have begun to focus on supporting training in mammographic interpretation. First of all, the training development in the PERFORMS scheme. The increased use of digital mammography, and its forthcoming widespread adoption in the UK, offers the opportunity for extending the PERFORMS self-assessment scheme to provide increased dedicated and individualized training. Also, the success of implementing teleradiology in the domain of medical care, which allows the electronic transmission of radiological patient images from one location to another for the purposes of interpretation and/or consultation, offers the

potential for more freedom in mammography interpretation training (Flanders, *et al*., 2003) - providing appropriate resolution images can be transmitted and displayed suitably without loss of required resolution.

Furthermore, there are some other projects which have already aimed to employ advanced technologies to support mammographic interpretation training. For example, GIMI (Generic Infrastructure for Medical Informatics) was a collaborative project in the UK with the purpose of developing a prototype training tool for screening mammography which could offer radiologists a tailored educational experience based around the intelligent selection of training activities (Gale, 2003; Yap & Gale, 2009). This is based on using grid technology to deliver training to individuals based at clinical workstations. Also, a related computer-based training (CBT) system has also been investigated to support both the improvement of the skills of experienced film-readers and the training of inexperienced ones using advances in high-quality computer displays at the mammogram viewing workstation and high-speed networking (Hellinger, *et al*., 2004; Soutter, *et al*., 2003).

A development from the GIMI approach has been a recent mammography computer-based training system, 'Lesion Zoo', which allows trainee radiologists to access a reasonable number of areas of suspicious mammographic lesions (300 digitized mammogram images, lesion being either a mass or micro-calcification) via the web. The system invites the user to classify the lesion; and then provides confidence scoring feedback based on three experts' BI-RADS ranking on each case. In the end, an overall performance summery is given over the set of images (Taylor, *et al*., 2010)

All these projects aim to implement training at the digital mammography workstation itself. However, technological advancements have enabled such images to be viewed on a laptop, a PDA or any computing platform (e.g. iPad). This raises the possibility of having mammographic interpretation training delivered both whenever, and wherever, it suits the individual.

It is convenient for screeners to access images via the web, however, in Lesion Zoo only an area of interest of the lesion is displayed. Previous research has reported that the majority of false negative errors in mammography interpretation are due to a 'search' error. This is where the individual has simply not looked at or near the abnormality. Readers' visual search characteristics while examining mammograms was reported in recent research, especially for inexperienced readers. It was found that only 23% of false negative errors made by experienced mammogram readers are 'decision making' errors (Mello-Thoms, 2010), i.e., a great deal of mammography diagnostic error is because readers missed abnormalities because they failed to look at or near them or look at the lesion but only for a very short time period. Naive readers were reported as having more problems looking for the abnormality, i.e. making more 'search' errors (65%) (Chen & Gale, 2009). This research suggests that the readers' main difficulty is being able to identify the abnormality instead of interpreting the lesion. Therefore, it remains unconvincing that only interpreting the lesion area is effective.

4.1.3 On-demand Mammographic Interpretation Training

According to the findings from the national questionnaire (for details see Chapter 3), mammographic interpretation training would ideally be on-demand; that is whenever and wherever an individual decides to undertake it. To use a portable device for such a purpose would be attractive on many levels. Such devices are very low-cost as compared to expensive digital mammographic workstations, which make them widely available for individuals. They are also portable, which frees the individual out of the clinical location limit. Thus they allow individuals to carry out training at any time that suits him/her.

Previous studies have successfully indicated the potential for viewing certain medical images which have fairly low spatial resolution (e.g. CT, MRI) on PDAs. For example, a few recent research projects have been carried out to develop mobile teleradiology systems on a PDA, which are independent of stationary and cable-bound computers. These projects are for better

managing Digital Imaging and Communications in Medicine (DICOM) image data and to support group work. The wireless system allows efficient management of heavy loads of lossless DICOM image data and could be useful for collaborative work by radiologists for instance in education, conferences, and research (Ikeda, *et al*., 2003; Istepanian, *et al*., 2006). PDAs have been shown successfully to support the interpretation of CT images, whose small physical size and resolution is adequately handled by the PDA screen's resolution and size (Toomey, *et al*., 2007)

However, the potential for mammographic images to be viewed on PDAs has not yet been comprehensively researched, mainly due to an unanswered key question: can PDAs provide sufficient image quality? In particular, the potential for using PDAs and other portable devices to deliver mammographic interpretation training has not been investigated. Therefore, an initial investigation into the possible factors which might affect the feasibility of using some portable devices as a training technology for examining large high resolution mammographic images were examined here.

4.2 Initial Pilot Studies

Overall, the pilot studies employed digitized versions of mammographic cases which have previously been used in the PERFORMS scheme and for which both the clinical outcome, as well as the opinions of virtually all UK (~ 95%) screeners on their mammographic appearance, was known. Therefore the decisions of participants in these studies could be compared against these data, whilst bearing in mind that the PERFORMS scheme data were gained from screeners inspecting mammographic film images of these particular cases.

Each participant's opinion was derived for the identification of the presence of a range of key mammographic features, namely: ill-defined mass (IDM); spiculate mass (Spic); architectural distortion (AD); calcification (Calc); asymmetry (ASYM); or the presence of no key mammographic features (none).

The studies comprised three parts. As a starting point, a pilot study was conducted at a major UK breast screening conference. This was to determine if a small PDA screen can support mammographic interpretation amongst expert radiologists and also to collect feedback and comments from these individuals about the potential use for a PDA in screening training. Subsequently, a series of mammograms were presented on a laptop and examined by a group of screeners at another major breast screening conference. This investigation explored how individuals interpreted larger images (as JPEG files) than were presented on the PDA, but where they were not able to manipulate (e.g. magnify) these images at all. Finally, some breast screeners examined a series of mammograms (as DICOM images) presented on a laptop using a DICOM viewer which did permit mammographic image manipulation.

4.2.1 Pilot Study One

The study aimed to investigate the performance of experienced screening radiologists in making screening judgments using a PDA. This would then provide an indication of the feasibility of displaying mammograms on PDAs and potentially suggest directions for further research.

4.2.1.1 Methods and Materials

Participants

Twenty experienced screeners were recruited from an opportunity sample which presented itself at the Royal College of Radiologists' Breast Group Annual Scientific Meeting (2007).

Materials

Visual Stimuli: Seven pairs of mammographic images were selected from the PERFORMS archive of digitised screening cases which have previously been categorized and reported by large numbers of screeners. Five of the pairs featured a specific abnormality (namely: ill-defined mass (IDM), Spiculate mass (SPIC); Architecture Distortion (AD); Calcification (Calc), and

Asymmetry (ASYM)) and two of the pairs featured no abnormality (i.e. normal cases which had also had a three years' follow-on mammogram that proved to be normal). The cases were selected as suitable exemplars of images showing these abnormalities. Each image pair comprised the Medio-Lateral Oblique (MLO) views of both breasts.

Each pair of case images was combined into one large image with a resolution of 7,200x4,800 pixels and were then transformed using Adobe Photoshop into a 320x240 pixels sized image which satisfied the resolution of the PDA that was used in the study. The Cranio Caudal (CC) view of these images was not included in consideration of the envisaged limited time that participants would have to view the images, additionally radiologists typically report that abnormalities are mostly visible on the MLO view.

Viewing Device: An SPV M700 3G PDA phone was used. At the time of the study this PDA was commonly available in the UK and has a 2.8" screen with a resolution of 240 x 320 pixels, a 65,536 colour display, and a MiniSD-expandable 64MB of memory (see figure 4-1).

Design

Each of the seven mammogram pairs was presented to participants, using Microsoft PowerPoint on the PDA. This allowed the display of the MLO pair of images to fill the PDA screen (figure 4-1) and also permitted the participant to zoom in, using a single fixed zoom level, and pan around the image Participants' decisions concerning each case were recorded by the experimenter on paper. After the study, each participant's feedback on viewing these images on such a viewing device were tape recorded.

Procedure

Each participant viewed each mammogram pair, in sequence, and made a decision as to whether an abnormality was present. Where an abnormality was identified, then its location was indicated and noted by the experimenter. Also, the participant's overall feedback concerning the examination of

mammogram images on the PDA was collected after they had viewed all of the cases. Given the experimental environment, the ambient lighting was not controlled but was normal room lighting.

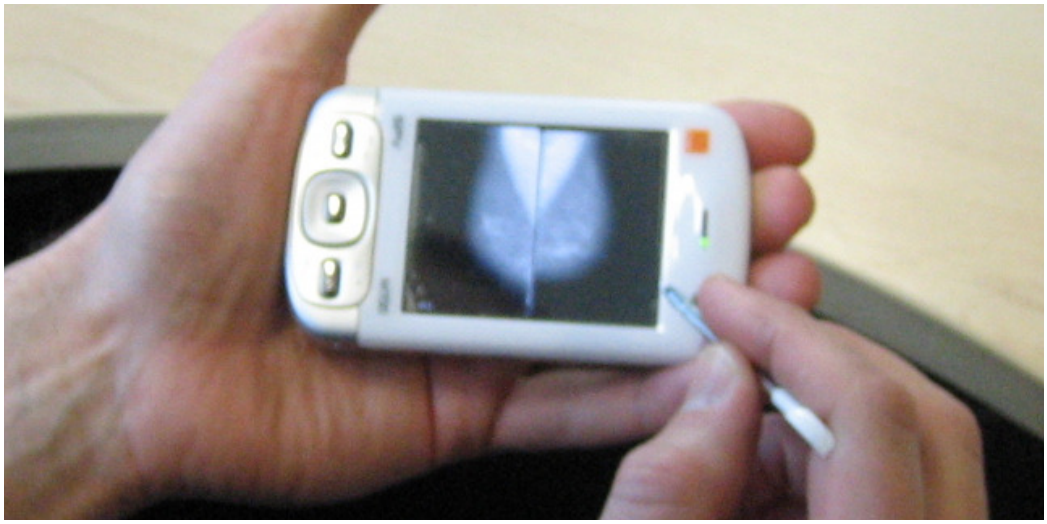


Figure 4-1. The PDA used to display mammograms in part 1 of the study

4.2.1.2 Study Results

Data were collected both on performance and on abnormality location identification which were analysed as follows:

Performance Results

A Chi-square test was carried out with the null hypothesis being that the performance of each feature was divided equally among the two categories (i.e. correct and incorrect) with α (α = 0.0074) adjusted for multiple comparisons to keep the overall α level at .05. The results showed that the ability to detect the Architecture Distortion, Ill-Defined Mass, Asymmetry, Calcification, Spiculate mass was significantly greater than chance (IDM: $\chi^2(1) = 16.2$, $p<.0074$; Calc: $\chi^2(1) = 7.2$, $p<.0074$; ASYM: $\chi^2(1) = 8.89$, $p<.0074$; SPIC: $\chi^2(1) = 3.2$, $p<.0074$). However, on the two normal cases, performance in determining no abnormality was present (i.e. true negatives) fell below chance (both equal 25%). Details are shown in figure 4-2. For the abnormal cases participants correctly identified at least 70% (Spiculate Mass) with

Architectural Distortion successfully identified by everyone. The overall mean correct feature identification was 85.8%.

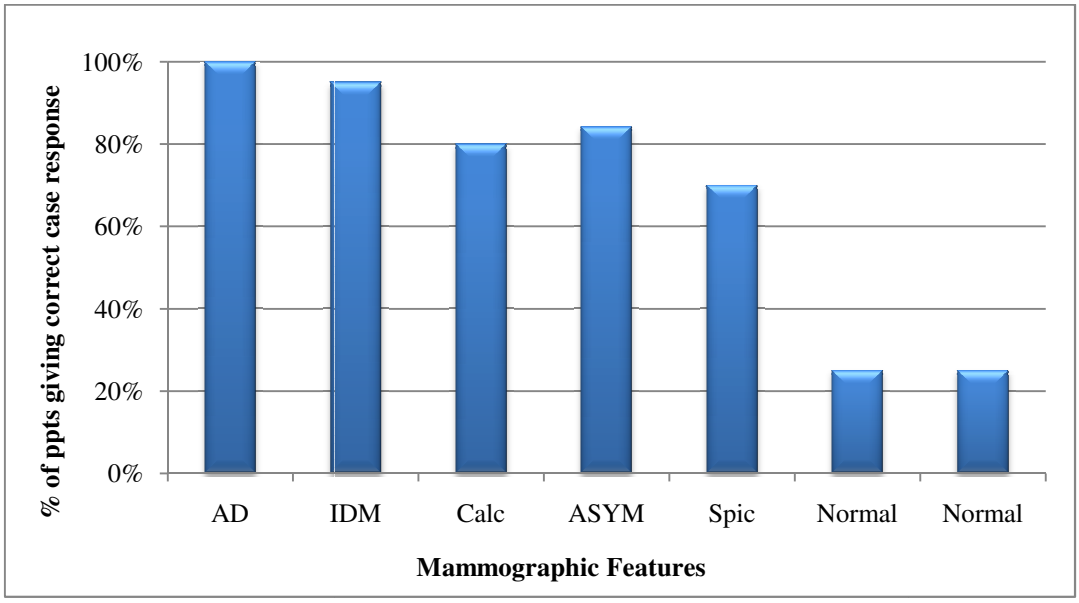


Figure 4-2**.** The percentage of participants giving the correct case response per type of mammographic feature

For image pairs featuring abnormalities, the percentage of participants who specified the correct location is shown in figure 4-3. In terms of performance across participants in identifying the correct location of an abnormality, performance varied from 30% for asymmetries to 80% for ill defined masses.

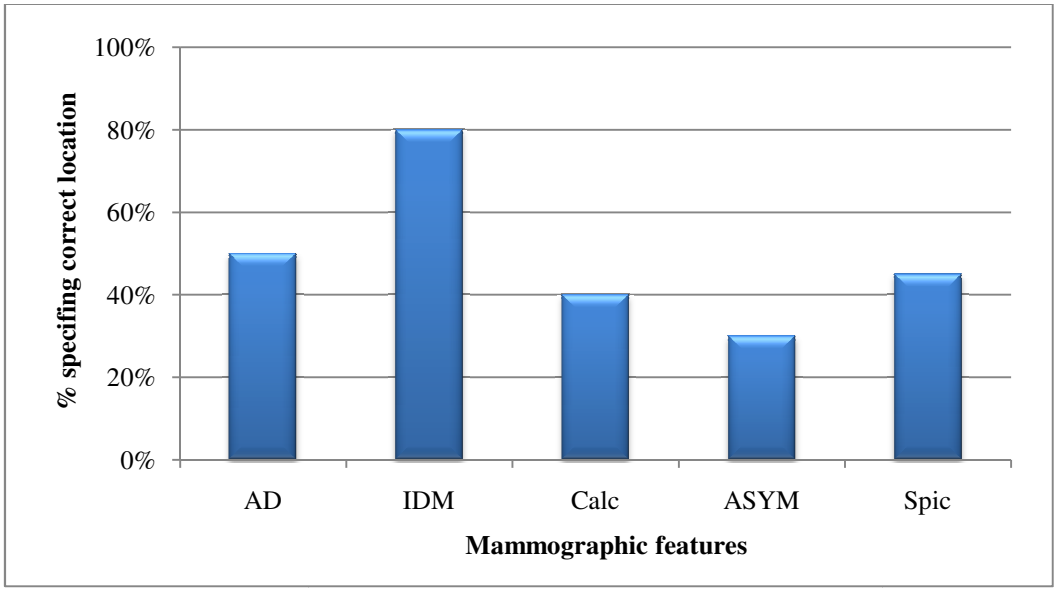


Figure 4-3. The percentage of participants specifying correct location by type of mammographic feature

A Chi-square test was carried out with the null hypothesis being that the performance of each feature was divided equally among the two categories (i.e. correct and incorrect) with α (α = 0.0074) adjusted for multiple comparisons to keep the overall α level at .05. The results showed that the ability to detect the ill-defined mass case and identify its correct location was significantly greater than chance $\chi^2(1) = 8.895$, $p<.0074$. However, performance on the other abnormal cases in determining the abnormality and also specifying correct location was not significantly greater than chance. This possibly reflects the size of the ill-defined mass which was slightly larger (obvious radiological appearance) than the lesions in the other cases.

In general, participants performed better at identifying the abnormal cases than the normal cases. However, their performance on specifying the location of the abnormality was poor (AD case was 100% recalled but only 50% of the participants specified the correct location). The potential explanation could be that with the decrease of image quality, examining an image becomes more difficult (i.e. harder to specify the right location). At the same time, readers increase their decision threshold (i.e. higher sensitivity) when image quality is poorer.

Feedback

Overall, the participants' feedback on viewing these mammogram cases on the PDA can be summarized as follows. Most participants were inherently initially negative about trying to view mammograms in this manner and strongly felt that actually to take part in the study was a waste of their time. However, once they tried to look at the first image and realised that they could zoom and pan around the image then they were much more positive. Having viewed the image set then all participants were more positive about using such a device. Not surprisingly, the PDA screen resolution was judged to be too poor to make a proper diagnosis. For instance, it was commonly reported that it was virtually impossible to identify calcifications confidently; it was also difficult to identify architectural distortion. However, it was felt to be relatively easier to identify masses. It was possible to lose the sense of asymmetry on

the small screen. Whether a case was ultimately benign or malignant was commented upon as being difficult as one would have to zoom in using the single zoom level and then pan around the image, trying to ensure that the whole image was examined – it was felt that it was easy to forget where an individual was in the image when doing this. Overall it was felt that it was difficult to have decent analysis and perception of such a small image.

With regard to zooming; this was seen favourably, although comments were made that when zooming in, more screen pixels would be needed to see detail, which then helps to dismiss potential suspicious areas. Finger touch zooming was mentioned as a way forward. The use of two views (MLO & CC) would be better to help identify the abnormality; performing contrast adjustment and the ambient illumination (role of reflections on the PDA screen) were also commented upon. It was felt that it would be hard to use images shown on a PDA for collecting a second opinion (as happens in routine clinical reporting) if an individual was doubtful about a case.

From the PDA performance data and the participants' feedback it was apparent that it was possible to correctly identify abnormalities using such small and low-resolution displays. However it was not easy to do so.
However, some participants felt that nationally a lot of effort had been made over the years to improve the quality of the mammographic images to the current very high national standard in order to make a proper diagnosis and so using any device which would inherently affect the perceived image would need to be carefully thought through, particularly so as not to degrade the image quality.

A lot of the participants' feedback was positive towards the possibility of using such a portable device as a mammographic interpretation training tool. However, some commented that perhaps in training then the minimum image size should possibly be that of a laptop computer screen.

4.2.2 Pilot Study Two

A second pilot study was undertaken based on the outcomes from the first study. This was designed to collect comments and feedback from experienced screening radiologists about making screening judgments using a laptop computer instead of a PDA. This would provide an indication of the feasibility of inspecting mammograms on normal portable computer devices and potentially indicate directions for further research.

4.2.2.1 Methods and Materials

Participants

Three experienced film-readers (two breast screening radiologists and one breast symptomatic radiologist) were recruited from an opportunity sample at the Symposium Mammographicum scientific meeting (2008) at Lille, France.

Materials

Visual Stimuli: Ten pairs of mammographic images were selected from the PERFORMS archive of previously categorised cases. Five of the pairs featured a specific abnormality (i.e. ill-defined mass (IDM); spiculate mass (Spic); architectural distortion (AD); calcification (Calc); asymmetry (ASYM)) and five of the pairs featured no abnormality (i.e. Normal case with three years follow on that prove to be normal). Each image pair comprised the mediolateral oblique (MLO) and the Cranio Caudal (CC) view of both breasts. These were combined into one large image using Adobe Photoshop and saved as JPEG images with a resolution of 1,400 x 1,050.

Viewing Device: A TOSHIBA TECRA M5 laptop with 1,024 MB of memory was used. The size of the screen display was 14.1 "; screen type: SXGA+ TFT display; internal resolution: 1,400 x 1,050 pixels (see figure 4-4).

Design

All of the ten mammogram pairs were presented to each participant, using the laptop. Their answers to each case were recorded on paper by the experimenter.

Procedure

Each participant viewed a mammogram pair, presented full screen size on the laptop, without using any image manipulation tools, and made a decision as to whether there was an abnormality present. Where an abnormality was identified, its location was indicated and this was noted by the experimenter. Also, his/her feedback from the examination of mammogram images on the laptop was collected after viewing all of the cases. The study was performed in normal room lighting conditions.

Figure 4-4. The laptop used to display mammograms shown in use in parts 2 & 3 of the study

4.2.2.2 Study Results

Data demonstrated that these participants could identify all the abnormal cases in the set, although their reported confidence in their decisions was low.

With regard to the normal cases the participants had difficulty in deciding that these cases contained no abnormality. The feedback and comments from the participants was summarized as follows. The participants thought that a typical laptop PC could be useful as a device for mammographic interpretation. The use of image manipulation tools was suggested to be very important for reading screening cases, especially for visualizing subtle mammographic features, such as micro-calcifications.

4.2.3 Pilot Study Three

On the basis of the first two parts of the study, the third part aimed to further explore the possibility of experienced screening radiologists making screening judgments using a laptop computer. Specifically their cancer detection performance was measured on a laptop using an appropriate DICOM viewer which facilitated image manipulation.

4.2.3.1 Methods and Materials

Participants

Three experienced film-readers (one breast clinician and two breast advanced practitioners) from two different breast screening units in the UK took part.

Materials

Visual Stimuli: Two sets of ten pairs of mammographic images were selected from the PERFORMS archive of previously categorised cases. These were matched by difficulty and feature type. In each set, five of the pairs featured a specific abnormality (i.e. IDM, Spiculate mass, Architecture Distortion, Calcification and Asymmetry) and five of the pairs featured no abnormality (i.e. a normal case that had had a three years follow on screening that had proved also to be normal). Each image pair comprised the Medio-Lateral Oblique (MLO) and the Cranio Caudal (CC) view of both breasts. Each pair of images were combined into one large DICOM image with a resolution of 7,200 x 4,800 and viewed using a DICOM viewer (ImageJ: http://rsbweb.nih.gov/ij) in the study.

Viewing Device: the same as used in Pilot Study Two (see section 4.2.2)

Design

Each of the twenty mammogram pairs was presented to the participants using the laptop. The answers to each case were recorded on paper.

Procedure

Each participant viewed one set of images using the image manipulation tools (i.e., using the HCI tools of zoom in/out, pan, and window/level) and the other set without using these image manipulation tools. For each mammogram pair, in sequence, the participant was required to make a decision as to whether there was an abnormality present. Where an abnormality was identified then the participant also indicated its location which was noted by the experimenter. Also, his/her feedback concerning the examination of mammographic images on the laptop was collected after they had viewed all of the cases.

4.2.3.2 Study Results

The number of cases for which each participant had given the right answer (i.e. specify no abnormality if the case was normal; specify the correct location of each lesion if it was an abnormal case) were examined and compared between the with/without (HCI/non-HCI respectively) image manipulation sets. These results are shown in table 4 -1.

Table 4-1. The number of cases correctly reported with/without the support of HCI (for each column, total cases examined was five cases).

Participant	***HCI***		***Non-HCI***	
	ABNORMAL	NORMAL	ABNORMAL	**NORMAL**
1	4	2	3	2
2	4	2	4	2
3	4	3	3	2

A two-way t-test showed that on average, observers performed not significantly better with HCI (M = 0.63, SE = 0.08) than not supported by HCI (M = 0.53, SE = 0.06, t (5) = 2.236, p >.05, r =.83) The scale of this study is relatively small, which is not enough observers' data to reach any statistical power to reliably infer any difference in performance with/without the support of HCI. However, data were used to estimate the sample size required.

The mean performance difference between HCI and non-HCI was divided by the standard deviation of all data to obtain Cohen's d = .91 (Equation 2.4). For these data the experimental conditions yielded a large correlation coefficient (r=.83) and was also significantly correlated p <.05. These figures were used to derive estimates of the required sample sizes based on designing a study with power of .80 at an alpha level of .05 by consulting the appropriate look-up table (Maxwell, 1990. pp 570, table 13.10). The look-up table showed that a minimum of eight observers are required while d = .75 and r =.83; six observers are required while d = 1.00 and r =.83. Accordingly, it was estimated that seven (between eight and six) observers are required to compare performance differences between image examinations with/without image interpretation tools per viewing condition in order to achieve a study with 80% power at an α -level of .05.

4.3 Discussion

This series of pilot studies provided a clear indication that it is possible to detect mammographic abnormalities (at least the specific exemplars cases used here) using a small PDA. However, the results indicate that there may be specific difficulties to overcome, particularly with respect to confidently determining that an abnormality is absent and also, although to a lesser degree, with respect to accurately locating abnormalities. There is a suggestion that these difficulties may be more pronounced for some kinds of abnormality than others, which may relate to the abnormality size. However, the present indicative data do not allow firm conclusions to be drawn on this matter. Such difficulties may be overcome with a zoom-able HCI interface to allow closer inspection of certain image areas.

With respect to the comments from participants, although most were initially doubtful about the ability to make any decision on such a small image size, afterwards they were amenable to the idea of the potential use in training; particularly the idea of being able to zoom in to an area, or only presenting small areas of interest of mammograms, using such a device.

Displaying mammograms on a PDA as a training tool is in many ways attractive but needs to be very carefully considered to maintain image quality. For viewing the images, zooming in/out with no effective pixel loss was suggested, which helps to see more details and dismiss suspicious areas. Even if zooming in/out is allowed, some mammographic features might still be expected to have difficulty to identify confidently, e.g. calcification is expected to be very difficult to identify and masses are relatively easier.

Despite participants' reticence about the PDA, the data indicated that 85.8% of them correctly identified the features for these particular exemplars of abnormal appearances. The normal cases were only correctly identified by 25% of participants. This indicates that they were over-reading the cases which may well reflect the difficulty in judging a case to be normal in this situation. Participants' suggested the laptop would probably be more suitable for delivering training. Therefore, another two small scale pilot studies were completed using a laptop computer with a typical screen.

4.4 Conclusion

The strong demand of using this as an image interpretation tool became very clear from the first pilot study on the laptop. Accordingly, the second pilot was carried out. Although there were only three participants involved in pilot study three, hence, there was too few people data to reach any statistical power. Data collected were used to calculate the minimum number of participants required in future studies. It was estimated by using the formal Cohen's d calculation that a minimum of 7 participants are required to achieve a study with .80 power, given α level of .05 to compare performance differences

between image examinations with/without image interpretation tools per viewing condition (modality).

On the basis of the pilot studies described in this chapter, the next study was designed to investigate the influence of different viewing devices to investigate further the issues associated with interpreting mammographic images on portable devices with relatively smaller size and a lower resolution (compared with digital mammography workstation) display (Chapter 5). Also, it was aimed to recruit seven observers to achieve a study with 80% power, and with an alpha-level of .05.

CHAPTER 5

Visual Search and Mammographic Interpretation Training

5.1 Introduction

The pilot studies in Chapter four simply investigated whether experienced screeners could identify a limited range of abnormal features on a commonly available small PDA with a relatively low resolution screen. Performance results and feedback collected from the participants implied that viewing mammograms on a small sized screen was feasible. However, to fully explore this then a more in-depth study was required to assess the influence of small screen size and resolution on performance in mammographic interpretation.

This study was an investigation into aspects of the potential of viewing breast screening cases on a series of screen sizes (although all of them are relatively small) for the purpose ultimately of providing some form of training in mammographic examination. Ostensibly a PDA is not a technology of choice as the typical PDA screen resolution can only accommodate about 1/10th of a single mammogram – and in screening each woman must be represented for inspection by four such images. However the purpose here was to determine whether the general approach has merit and then to investigate HCI techniques fully to use such a small display to best effect.

Accordingly, the study was conducted in order to compare the diagnostic efficacy associated with different sizes and resolution of some small displays which, for experimental purposes, were simulated on an LCD computer monitor. The eye movements of the participants were recorded in this task. Of interest here was whether increasing the physical size of the displayed image or increasing the resolution affected the observers' behaviour. In the study, it was hypothesized that those viewing conditions which represent larger image size/higher pixel resolution would give rise to more correctly reported cases by the participants. Additionally it was hypothesized that the more gross mammographic features would be easier to identify on a PDA.

As a starting point to research involving screeners, which would require NHS ethics approval as well as consideration of their extensive time commitments it was decided to utilise a number of radiography students here. Whilst they

had some radiography-related training they had no experience in mammography. This choice of participants aimed to involve individuals with some radiology knowledge and so simulate potential trainees who might use a future mammography training system.

5.2 Study

5.2.1 Methods and Materials

Participants

Eight volunteer radiography students participated who were invited from a UK Bachelor of Science diagnostic radiography training course provided by the University Of Derby School Of Health Sciences. This comprised six second-year diagnostic radiography students aged between 19 and 35 years old with up to two years of radiography-related education and two diagnostic radiography tutors who had more than 10 years of general X-ray film reading experience. None of them had any experience in mammography reading and all the participants had normal vision or corrected to normal vision.

Materials

Visual stimuli: Twenty-four pairs of mammographic images were selected from the PERFORMS archive. Each image pair comprised the Medio-Lateral Oblique (MLO) view of both breasts. The Cranio-Caudal (CC) views were not included as there was a limited available time that each participant could provide for the study. Twelve of the pairs featured one of three specific abnormality appearances;namely: Mass, architecture distortion (AD) / asymmetry (ASYM) and Calcification (Cal)) and twelve of the pairs featured no abnormality (i.e. Normal).

Hardware: The experiment was operated on a Lenovo 8922 laptop with 1,024 MB of memory with an additional monitor for displaying images. The size of the additional screen display was 20" (517 x 72 x 334 mm (W x D x H)); internal resolution: 1,920 x 1,080 pixels (see figure 5-1). The viewing distance was 60cm from the computer monitor.

Eye Tracking Device: A Tobii X50 stand alone remote oculometer eye tracker was used. (Accuracy: 0.5-0.7 degrees visual angle).

Design

Viewing conditions: Viewing mammograms on a small device whilst maintaining ease of recording eye movements was mimicked by using the display monitor for all of the different experimental conditions. Two different image display sizes and two different pixel resolutions were used in the study (table 5-1).

In terms of image size, 3.5" (the most common screen size of PDA at the time of the experiment) and 6" (the largest screen on the then current general market in the UK) were employed. On the monitor these sizes equated to pixel resolutions of 239x179 and 410x308 respectively. In terms of the resolution, the two pixel groups investigated were: 320x240 (the most common screen pixel for a PDA) and 800x600 (the highest pixel level on the current general market at that time in the UK).

Table 5-1. Image size and resolution of the four viewing conditions.

Viewing condition	*Image size (inch)*	*Resolution(pixels)*
1	*3.5"*	*239x179pixels*
2	*4.69"*	*320x240pixels*
3	*6"*	*410x308pixels*
4	*11.72"*	*800x600pixels*

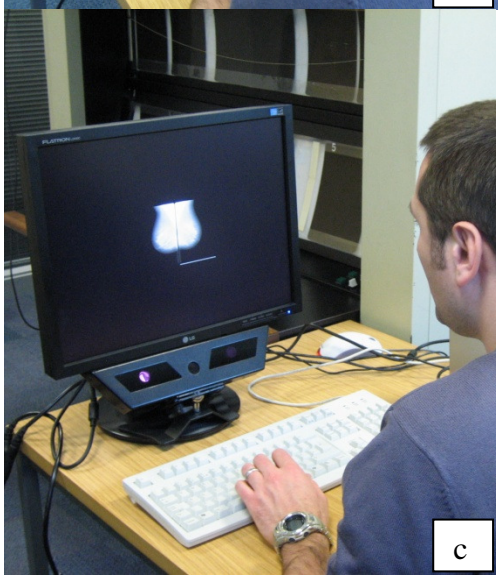
Procedure

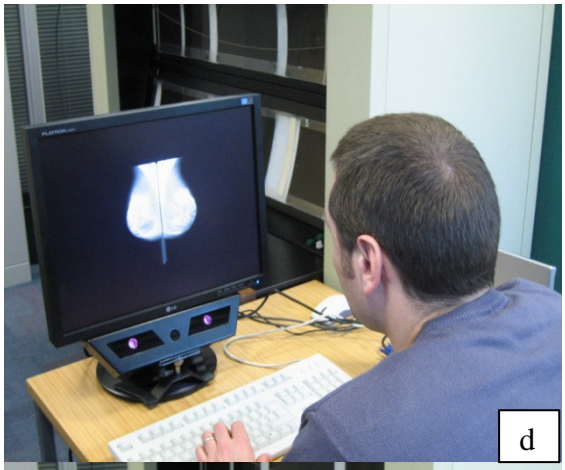
Ethical approval was granted by the Loughborough University ethics committee. All participants were given a participant information sheet (see Appendix C) describing the investigation and informed consent (see Appendix B) was gained before each experiment started.

The experiment took place in a darkened reading room at Derby Breast Unit with a room ambient light level of approx. 22 *lux*. Each participant was first carefully calibrated on the Tobii eye tracker. After that, each observer was given a short standardised training session on the appearance of the mammographic features used in the study (i.e. Mass, Architecture Distortion / Asymmetry and Calcification) using a Powerpoint presentation. Participants were then asked if they had any questions about the task.

They then completed the computer-based image examination task whilst their eye movements were discretely recorded (figure 5-1). In this, the observers were asked to examine each case, identify if there was any abnormality present, and then rate their confidence in whether the image was normal or contained an abnormality. If the latter, they also had to specify the abnormality and its location. Also, they were asked to rate the quality of the images.

Each participant viewed three abnormal and three normal mammogram pairs in each of the four viewing conditions. The six mammogram pairs for each viewing condition were shown in blocks and the order of the viewing condition blocks was counterbalanced across participants using a Latin Square design. Mammogram pairs were pseudo-randomly assigned to viewing condition with the condition that three normal cases and three abnormal cases must appear in the viewing conditions and that the three abnormal cases must be made up of one Architectural Distortion (AD) / asymmetry (Asym); one Calcification (Calc); and one mass.





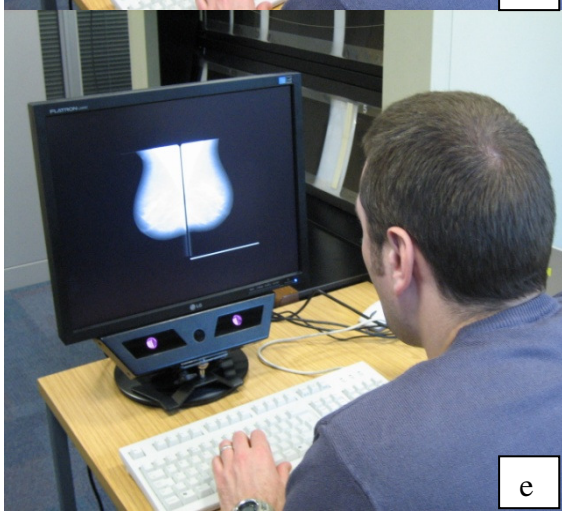


Figure 5-1. Examples of participant: viewing training slides (a); and examining images displayed as; 3.5" (b); as 320x240 pixels (c); as 6" (d); and as 800x600 pixels (e) sizes. The room lights of the viewing room were turned on for photographic purposes.

5.2.2 Study Results

Performance Results

Performance, was measured by the simplest measure of diagnostic decision, i.e. the percentage of cases for which the film reader gives the correct answer. The resultant data were subject to a two-way, repeated-measures ANOVA: viewing condition (3.5", 320x240pixels, 6.0", 800x600pixels) x outcome (abnormality present, abnormality absent).

There was no main effect of viewing condition, F (3, 21) = 1.272, p = .310. However, there was a significant main effect of normality, F (1, 7) = 10.162, p <.05. There was no interaction between viewing condition and normality, F (3, 21) = 0.132, p = .940. For more details see table 5-2.

Table 5-2. Mean performance proportion of correct response by case, known pathology and viewing condition

	3.5"	*6"*	*320x240p*	*800x600p*	***Total***
Abnormal	M=.750 SD=.295	M=.500 SD=.178	M=.646 SD=.187	M=.667 SD=.308	M=.640 SD=.254
Normal	M=.417 SD=.345	M=.250 SD=.295	M=.375 SD=.278	M=.417 SD=.295	M=.364 SD=.298

The mean percentage of correct responses across all the participants was compared by four viewing conditions. Although there were no significant effects of viewing condition on participants' performance, it shows that there were slight differences amongst these conditions. Details are shown in figure 5-2.

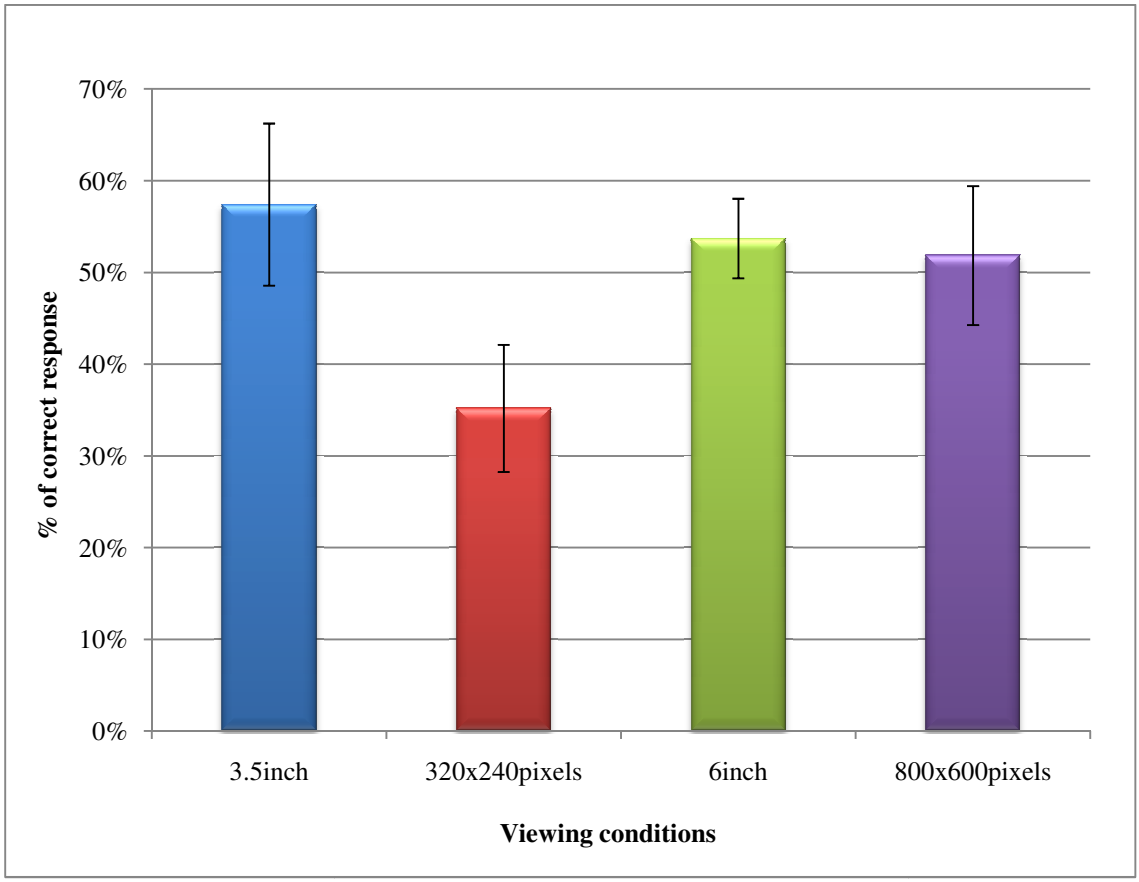


Figure 5-2. Mean percentage of correct responses by viewing conditions

There was a significant difference between performance on abnormal and normal cases (p<.05). The percentage of cases correctly classified by the participants (n=8) indicated that they made significantly less errors on abnormal (64% correct response) as compared to the normal cases (36.4% correct response). Similar results were also observed within the PERFORMS scheme using normal sized mammographic film images and this may be exacerbated here by the small screen size. Even so, this problem is not peculiar to PDAs and tends to result in false positive responses. Details are shown in figure 5-3.

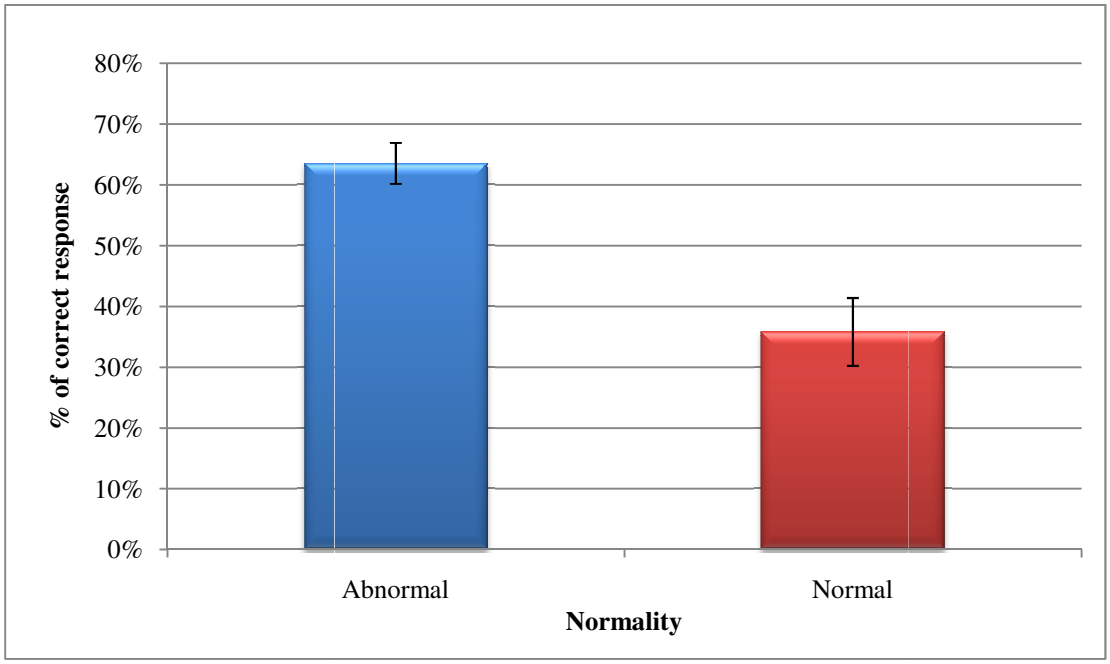


Figure 5-3. Mean percentage of correct responses by normality

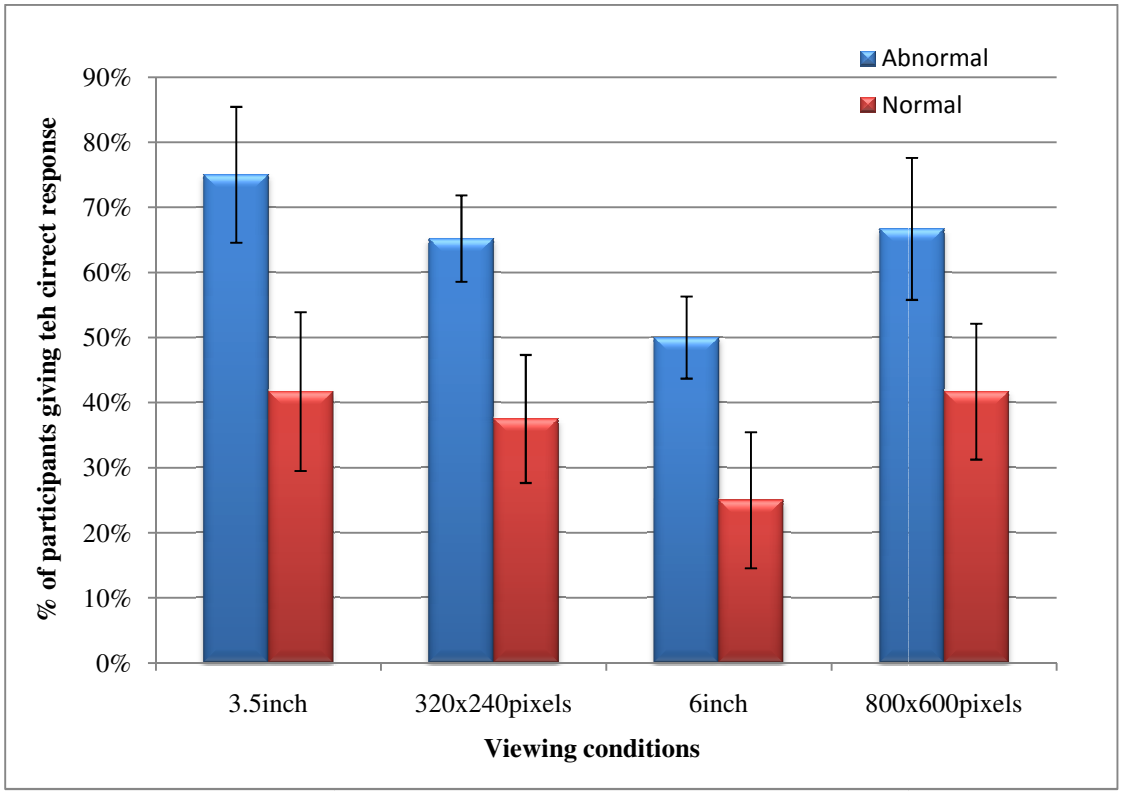


Figure 5-4. Mean percentages of correct response on both abnormal and normal cases by viewing condition.

Figure 5-4 shows the percentage of correct responses on both abnormal and normal cases by viewing condition. There were specific difficulties with confidently determining no abnormality was present. Although there were no significant effects of viewing condition on performance, there were slight descriptive differences between these conditions. The 3.5" screen, the smallest size of all the viewing conditions, appeared to support the highest level of cancer detection. This suggests that, within the range of screen sizes and resolutions reported here, simply increasing the PDA screen size or its pixel resolution might not be a straightforward solution for implementing PDAs for use as a mammographic interpretation training tool.

Although there was a main effect of outcome on participants' performance, the performance difference on the different feature groups was not statistically significant. For image pairs featuring abnormalities, the percentage of participants who specified the correct abnormal area is shown in figure 5-5. In terms of performance across participants on identifying the correct location of an abnormality, performance varied between 47.6% for calcification and 60% for masses.

In terms of the hypothesis that some mammographic features would be easier to identify on a small size/low resolution screen, there was no significant difference between these mammographic features on the percentage of correct responses and the percentage of correct location responses.

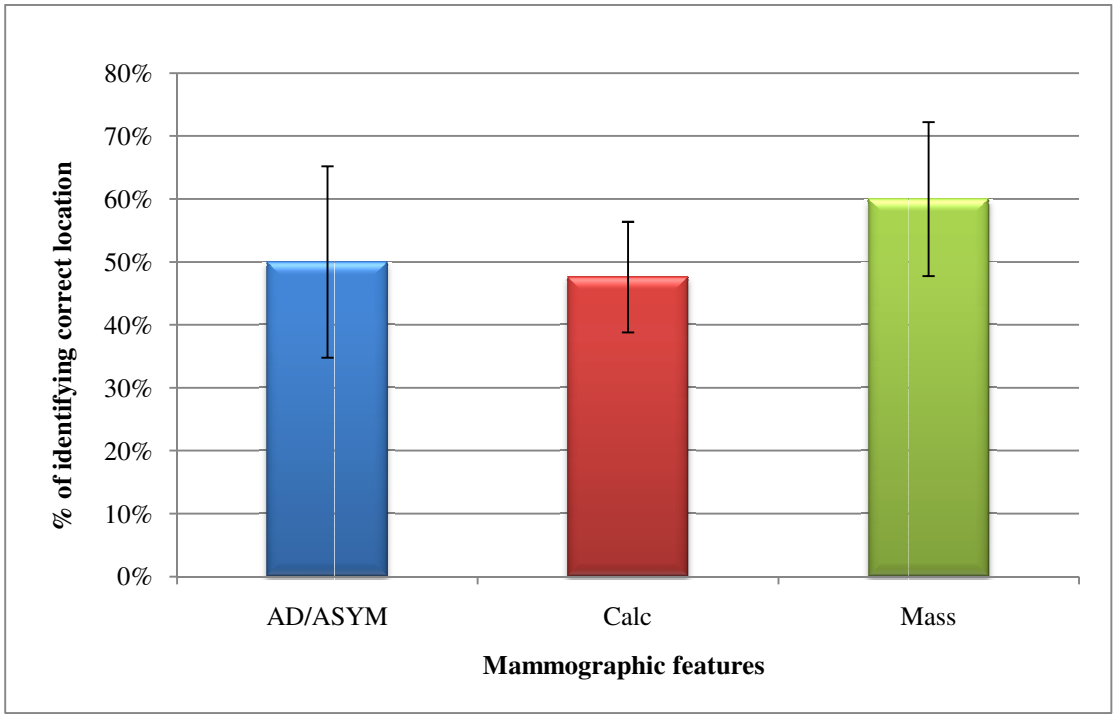


Figure 5-5. Participants: the percentage of giving correct location response by abnormality features.

Eye Movement Data Results

The observers' eye movements were recorded in the study, which allowed examination of each individual's visual search behaviour and, therefore, enabled errors to be studied in more detail (e.g. the false negative errors were classified into search error, detection error, and interpretation error).

The ClearView analysis software is the software that was used to perform the data analysis for the study. The fixation filter implemented in ClearView is based on an algorithm that has two settings; fixation radius and the minimum fixation duration. For such data analyses it is recommended by ClearView that a fixation radius is used which is equivalent to 30 pixels and a minimum fixation duration of 100 ms for viewing mixed content (in the study, participants were invited to read the mixture of images and words).

According to these settings, the eye position data (x,y) were calculated to be sampled fifty times a second. These sampled co-ordinates were then grouped into eye fixations that required at least six data points (100ms).

The Area of Interest (AOI) around a mammographic feature was recommended to be defined with a spatial proximity of within 1^{o} visual angle (subtended at the observer's eye) according to the ClearView's guidance. Participants viewed the monitor at a viewing distance of 60cm (23.6"), consequently this distance was used to calculate the visual angle subtended by the image. A 'hit' (true positive) was scored when the measured point of fixation fell within 1^{o} of the edge of the abnormality; i.e. fell within the Area of Interest (AOI). As the resolution of the display screen that was 1,920 x 1080 pixels, each AOI was therefore defined to approximate 35 pixels around the lesion using the AOI definition tool provided by ClearView.

Additionally, because of the complex nature of the abnormal features, and the relatively small display, the cumulated dwell time chosen to differentiate between detection and interpretation errors was taken as being 1,000 ms (with a detection error being scored when a participant's fixations fell within the AOI for less than this value and an interpretation error being taken as when the fixations fell within the AOI for equal to or greater than this value) (Kundel, 1978).

There were two experimental conditions (i.e. viewing conditions, mammographic features) and the same participants were used in both. Additionally, the data collected were ordinal and so a non-parametric test, Friedman's ANOVA, was chosen to analyse the data. For each type of false negative error, Friedman's ANOVAs were conducted on the proportion of times a given type of error was made on each type of feature (Calcification, Mass, Asymmetry/Asymmetrical Distortion), pooled across display types, and also for the proportion of times a given type of error was made on each type of display (i.e. 3.5inch, 320x240, 6 inch, 800x600), pooled across feature types.

The data indicated the percentage of search errors varied significantly by feature type ($\chi^2(2) = 8.00$, $p < .05$), however, neither the percentage of detection errors nor interpretation errors had a significant effect by the mammographic feature type (Detection error: ($\chi^2(2) = 0.26$, $p = .87$; Interpretation error: ($\chi^2(2) = 4.30$, $p = .12$)). Pairwise Wilcoxon Signed-Ranks tests were conducted between each unique combination of feature type, for the proportion of search errors, with the α-level adjusted to .018 to control for the overall level of type I error. The differences between the proportion of search errors on Calcifications (*Mdn* = 1.00) and Masses (*Mdn* = 0.00), and Calcifications and Asymmetries/Asymmetrical Distortions (*Mdn* = 0.00) approached significance ($Z = -1.89$. $p = .125$, $r = -.51$, for each comparison). Given the small sample size ($n = 8$) and the medium effects sizes ($r = -.51$), the tests may have been underpowered and may well have reached significance in a larger sample.

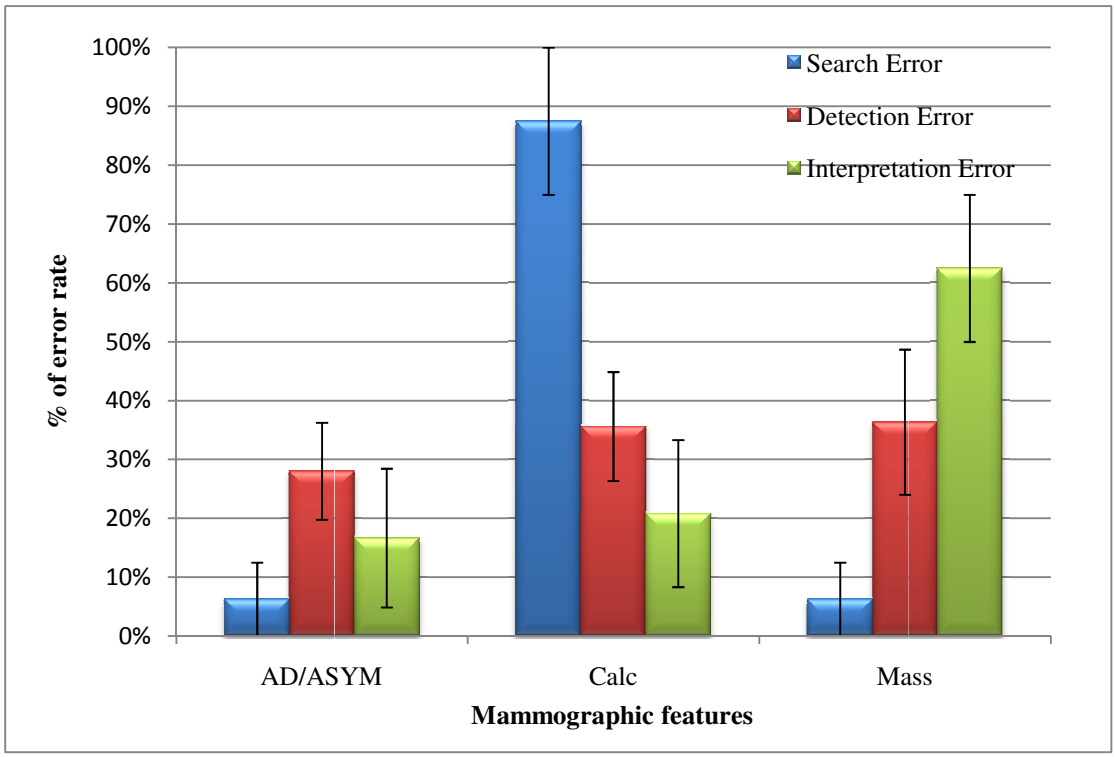


Figure 5-6. Mean percentage of errors by mammographic features.

The difference between the proportions of search errors on Masses and Asymmetries/Architectural Distortions did not approach significance; however,

search errors made on Calcification are significant different from the other two (p<.05). Other statistical tests implied that the sample effect was expected to be significant if there was a larger sample size. Details of the percentage of each type of error within each mammographic feature type are shown in figure 5-6. For illustration, examples of one observer's eye movement data on different mammographic features are shown in figure 5-7. In this figure (A) shows calcification which was missed (a search error) by the observer; (B) shows architectural distortion which was not detected; (C) shows a mass which was fixated and detected but then misinterpreted and (D) shows a normal case.

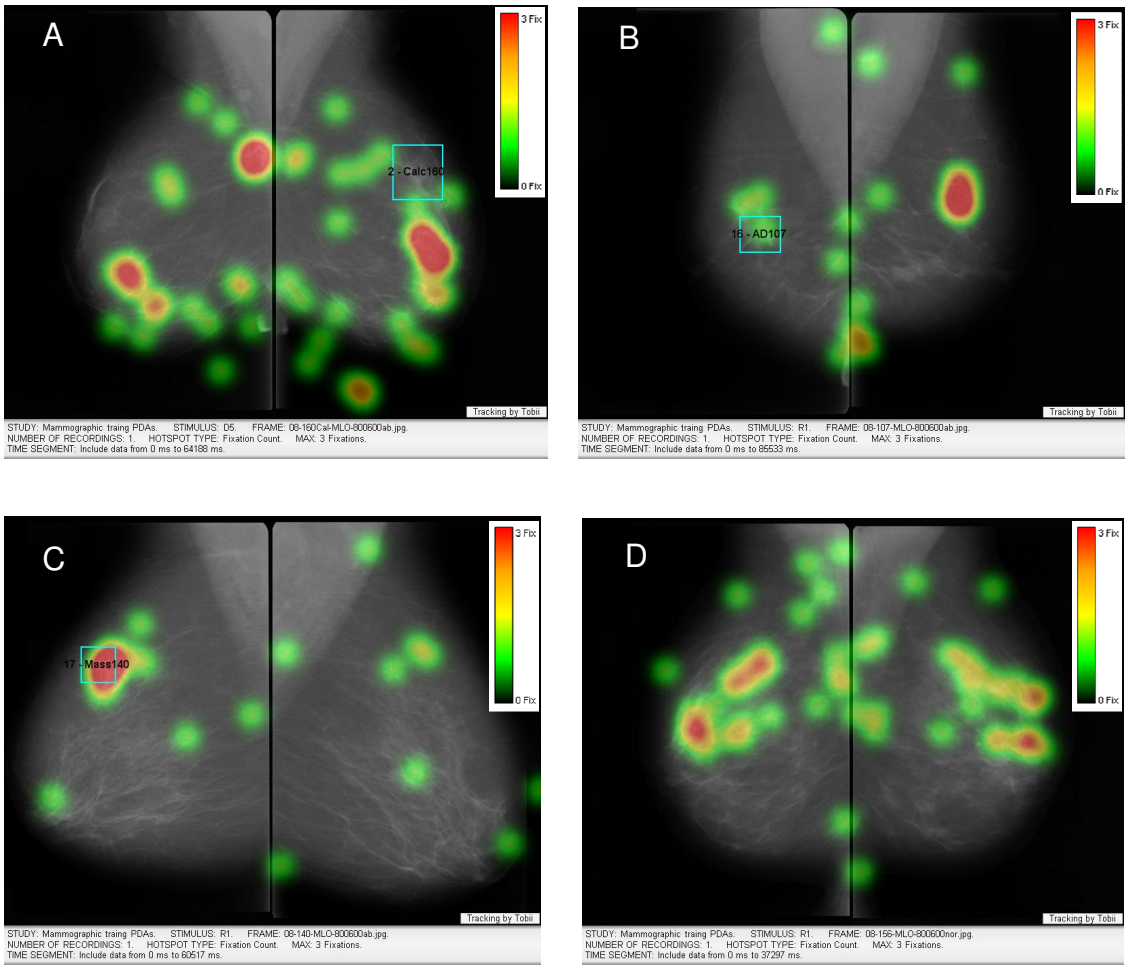


Figure 5-7. Examples of mammograms with different features used in the study with overlaid eye tracking data from one observer. See text for explanation.

The data also indicated that under neither viewing condition, error type varied non-significantly (3.5": $\chi^2(2) = 3.43$, $p = .18$; 320x240pixels: $\chi^2(2) = 2.33$, $p = .31$; 6inch: $\chi^2(2) = 4.46$, $p = .10$; 800x600pixels: $\chi^2(2) = 3.90$, $p = .14$). Details are shown in figure 5-8. Portions of detection errors made vary slightly from 20% to 30%. For the search error, a higher percentage of such error was made on the 3.5" screen, the smallest display; however, the least portion of search error was made on the 6" screen. Highest portion (45%) of interpretation error was made on the 6" screen, compare to lowest, less than 10% on the 3.5" screen.

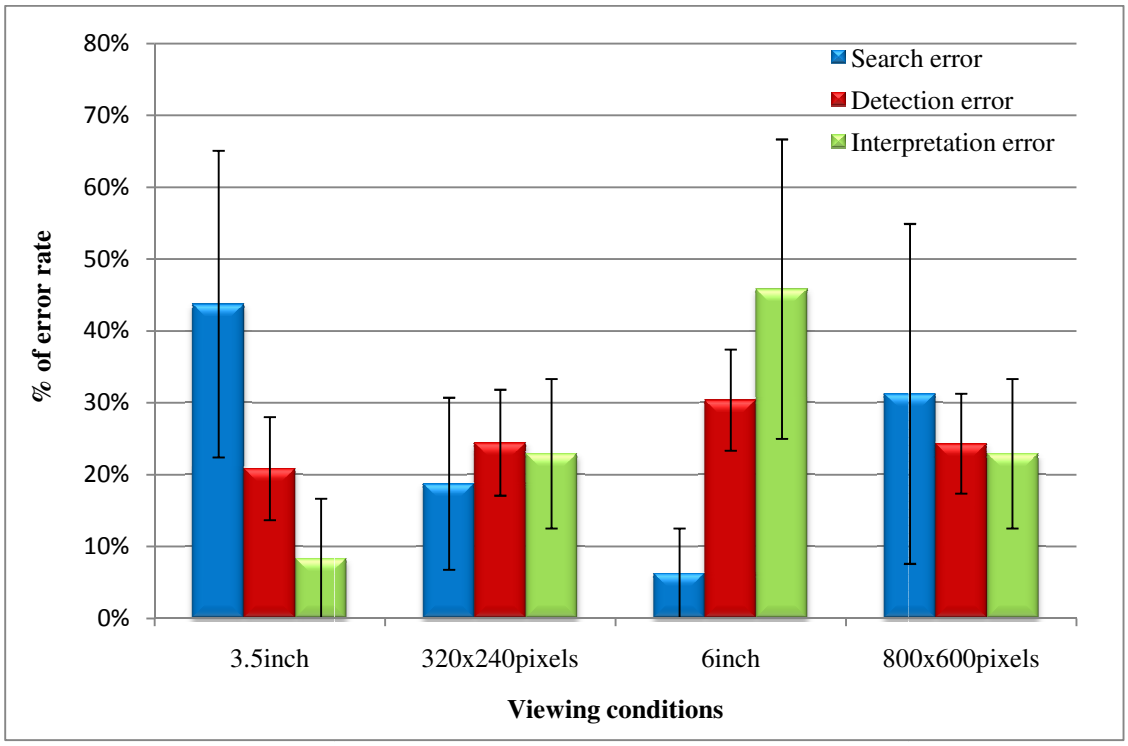


Figure 5-8. Mean percentage of errors of each viewing condition

In addition, with the small image sizes it becomes difficult to differentiate eye fixations to different image areas. This is demonstrated in figure 5-9 which shows examples of eye movement data from the experiment. It is clearly apparent that the eye movement information on the smaller image sizes (e.g. 3.5inch) is insufficient for the implementation of detailed visual search behaviour analysis as is seen through the very large overlapping of adjoining measured fixation positions. Consequently, it is hard to use such eye

movement data to inform recommending tailored training approaches on such small displays. It is argued that HCI techniques (e.g. zoom-in, zoom out, pan) are required to support any PDA-delivered training with such mammographic images.

5.3 Discussion

With a promising result from the pilot studies using a small PDA, described in Chapter 4 (see section 4.2.1), along with the feedback and comments from experienced mammographers, this study set out to collect more detailed information from a group of individuals who were familiar with radiographic appearances but not with mammography per se.

The result from this study suggested the possibility of viewing mammograms on different relatively smaller sized screens (typical PDA screen sizes), although the performance on normal cases (with no key mammographic features present) was poor. This could be due to the nature of the study which caused observers to over-read cases or that the small display actually affects their decision criterion. However, here, a significant main effect of normality was found with significantly less errors on the abnormal cases being made as compared to the normal cases, which is similar with the result of the pilot study and previous PERFORMS data (Scott & Gale, 2005; Scott & Gale, 2006).

There was no significant difference between mammographic features on the percentage of correct responses and the percentage of correct location responses. Although the slight difference within features did not reach statistical significance, calcification appeared to be slightly more problematic, which is to be expected in accordance with the feedback from the pilot study and also from previous reported results from the PERFORMS scheme (Scott & Gale, 2006). The observers' overall performance was relatively poor which implies the need to use participants with knowledge of mammographic image appearances in future studies.

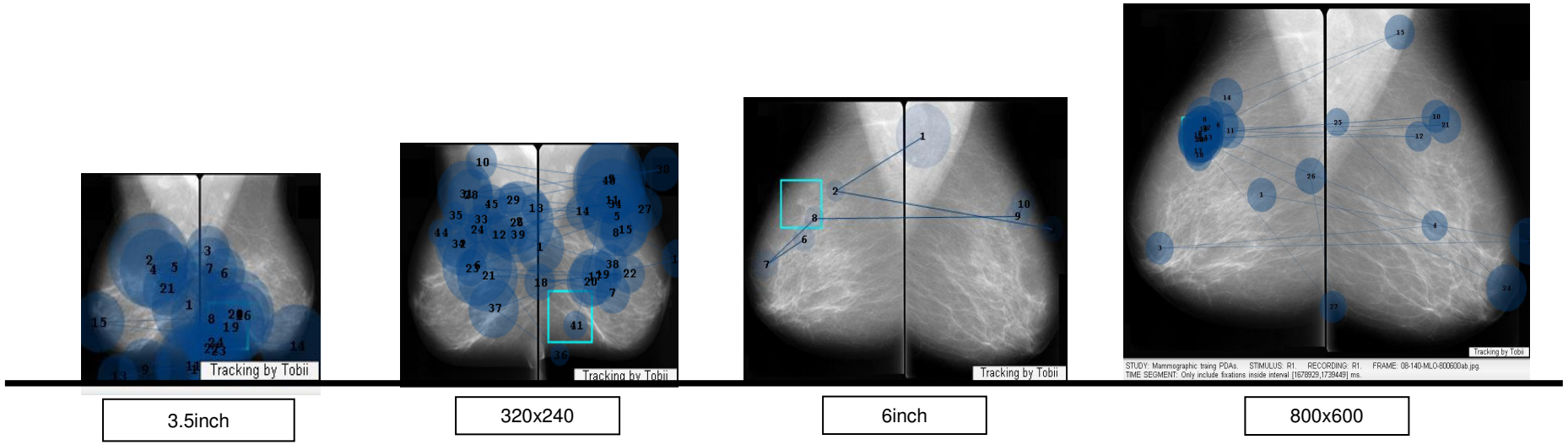


Figure 5-9. Examples of eye movement data on different size mammograms. Note the overlaying of fixations as the image size decreases.

Apart from the performance data, the observers' eye movements were also recorded in the study (examples are shown in figure 5-7), which allowed examination of each individual's visual search behaviour and therefore enabled the errors to be studied in more detail (Kundel, 1978).

The eye movement data indicated the percentage of search errors varied significantly by feature type. Participants made significantly more search errors on the relatively smaller features (i.e. calcifications) as compared with the other two types. It is reasonable to conclude that participants had more difficulty searching for the smaller features on the small/low pixel resolution screen. This highlights the potential difficulty of including calcification cases in potential PDA delivered mammographic interpretation training. However such small features may well be capable of visualisation on such small displays if suitable HCI techniques (zoom pan) are facilitated.

It was hypothesised that those viewing conditions which represent a larger image size/higher pixel resolution would give rise to more correctly reported cases by the participants. However, according to the ANOVA results, the different viewing conditions had no main effect on the performance of examining these images. Although there was no significant difference between the performances in different viewing conditions, the 3.5" screen was shown to be the best viewing condition – this may reflect a matching of the screen's spatial resolution when displaying these abnormalities to that of the human visual system and is a matter for further investigation.

Furthermore, there was a non-significant difference found among the three different visual errors made under the different viewing conditions. This suggested that further research needs to be carried out on how best to employ suitable HCI techniques to increase the feasibility of mammographic interpretation training on PDAs rather than overly, and simply, focusing on increasing the screen size and resolution.

Overall, results from the study indicate the, somewhat surprising, possibility of viewing mammograms, or parts of mammograms, on PDAs and to use these devices to deliver targeted training as, and when, required. However, the fairly poor overall performance data also implied how difficult it was for participants with little knowledge of mammography, even though they were familiar with radiographic appearances, to examine mammograms. This suggested the importance to involve specialized mammogram readers in future studies.

5.4 Conclusions

In general, the results indicate the technical potential of using PDAs as a training technology for examining mammograms. Whilst this is not an obvious choice for such high resolution images and would not be used as the sole training technology, it does demonstrate that PDAs can help deliver training to an individual, as and when they demand it.

The common difficulty in classifying normal cases which could be exacerbated by small screen size with a low resolution suggests the possibility of delivering very specific training on PDAs, which takes the small size and pixel resolution into consideration. Further research needs to investigate improving the usability of mammographic training system by facilitating the engagement of human-computer interaction (HCI) techniques. Also, professional mammogram film readers' participation is essential for future studies investigating performance issues.

CHAPTER 6

Mammographic Interpretation and Naïve Observers

6.1 Introduction

The previous chapters described the current mammographic interpretation training situation in the UK Breast Screening Programme and highlighted the importance of digital mammography training and the opportunity offered by a range of sub-clinical display devices to deliver aspects of that training. This chapter describes a series of studies to investigate different ways of delivering such training.

6.1.1 Mammographic Interpretation Training

With the age range of women who are invited for screening in the UK recently being extended from 47 to 73 years, this has significantly increased the number of women screened annually. Consequently, there is a need to train a growing number of breast screening personnel to be qualified to examine and report breast screening cases within the national screening programme.

Current mammographic interpretation training needs to be undertaken where there is a multi-viewer on which to view the film mammograms, or else on high resolution digital workstations where digital mammography is now available. Consequently, both of these limit the time and locations where training can take place. Ideally, such training would be tailored to the specific needs of the individual (what they require) and on-demand (i.e.: 'whenever' and 'wherever' an individual decides to undertake it). The gradual introduction of digital mammography in the UK is providing a wider range of different training opportunities without such time/location restrictions. Also, it offers the potential to use images more interactively to suit the specific needs of individuals.

The purpose of these studies was to determine whether training could be offered on lower resolution monitors (or laptop computers), which are more widely available and cheaper than high resolution workstations, thereby allowing training to be undertaken at a broader range of times and locations, and how best to achieve this.

6.1.2 Visual Inspection

The particular over-arching research interest is in whether other displays (e.g. desktop computers, laptops, PDAs) can be used to offer training in this domain. Here a typical office PC monitor was employed. Examination of radiological images inevitably produces errors; in particular false negative errors are of interest and whether using such a display affects the types of errors. These false negative errors have been classified into three types; namely errors due to visual search, detection and interpretation (Kundel, 1978).

6.1.3 Research Aims

This research investigates the utility of employing low-cost devices to provide individualised training.

Specifically four different potential approaches to training were developed which all aimed to encourage a user to visually search the mammographic images and identify known key early signs of breast cancer presence. The overall approach taken was firstly to present the two MLO views of a specific case and then follow these with the similar presentation of both of the CC views – this being the most common approach used in breast screening practice where radiologists often tend to find more useful radiological information on the MLO views than on the CC views; using the CC views largely to confirm an opinion gleaned from first examining the MLO views.

By monitoring the eye movements of the participants as they examined the cases then the types of errors they made could be classified into the above three types. It was hypothesized that the different types of training would produce different visual inspection strategies and thus different percentages of these error types.

Two key mammographic features were targeted in this study; masses and calcifications. These primarily differ in size as well as other appearance attributes. Based on the studies in the previous chapters it was hypothesized that masses would be easier to detect than calcifications using a standard monitor display. A key aim was to determine whether one particular form of training was superior to the others and thus should be followed up in future studies at breast screening training centres using actual screeners.

6.2 Studies

An experimental investigation was carried out to examine different kinds of training for breast screening interpretation. Initially the visual inspection behaviour of an expert breast screening radiologist was recorded whilst he examined a series of recent screening mammographic cases which were presented on the same monitor as used in the subsequent study. Analysis of these data allowed identification of regions of interest around particular cancers and also around other image areas which attracted his visual attention but were not judged to be abnormal.

These results were then used to provide training to naïve participants in order to determine if utilising such an expert's knowledge leads to an improvement in cancer detection performance.

6.2.1 Study One

This study investigated the visual inspection behaviour of the experienced breast screening radiologist expert in examining digital mammogram cases. These results were then compared with other participants' visual inspection data collected from the previous experiment (see section 5.2). The results were then used to provide possible training to naïve participants to determine if utilising such knowledge leads to an improvement in cancer detection performance in the later experiments.

6.2.1.1 Methods and Materials

Participant

An expert breast screening radiologist from the Nottingham Breast Institute, Nottingham City Hospital was invited to participate in the study after suitable University Ethics Approval was obtained; the study was deemed to be audit by the chair of the local NHS ethics committee and so Central Office for Research and Ethics Committees (COREC) approval was not required. He had more than 17 years of breast screening experience along with four years of digital mammography experience, has widely published research in breast screening and is internationally well regarded.

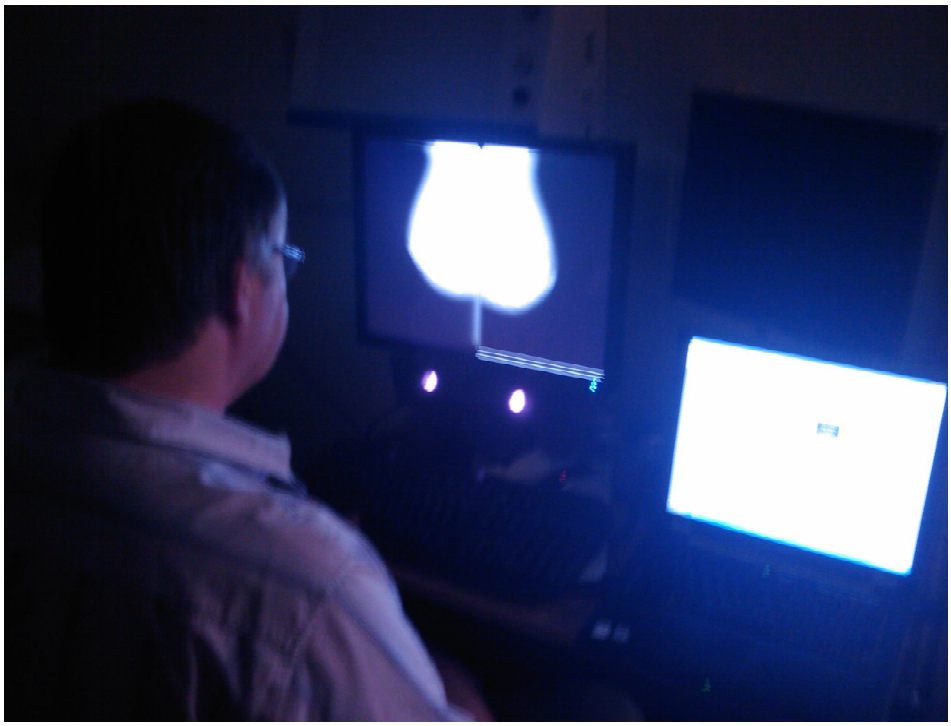


Figure 6-1. An experienced breast screening radiologist examining digital mammogram cases in a darkened clinical room

Materials

Visual Stimuli: twenty-one pairs of mammographic images were selected from the PERFORMS archive of previously categorised cases. Eleven of the pairs featured a specific type of Mass (i.e. ill-defined Mass (IDM), Spiculate mass

(Spic)); ten of the pairs featured Calcifications. Each image pair comprised the Medio-Lateral Oblique (MLO) and the Cranio Caudal (CC) view of both breasts. Each view was combined into one large image of each case using Microsoft Photoshop and saved as JPEG images with a resolution of 1,920 x 1,080 pixels. Seven of these cases had been used previously in the study (see section 5.2) described in the previous chapter.

Viewing Device: A TOSHIBA TECRA M5 laptop with 1024 MB of memory was used with an additional viewing monitor. The size of the screen display was 21.5 " (517 x 72 x 334 mm (W x D x H)); internal resolution: 1,920 x 1,080 pixels (see Figure 6-1).

Eye Tracking Device: A Tobii X50 stand alone remote oculometer eye tracker (accuracy: 0.5-0.7 degrees visual angle) was used which had been used in the previous study (see section 5.2.1)

Procedure

To familiarise the expert with the eye tracking equipment and the procedure, he was first calibrated on the system and then he examined some practice images. He then inspected the 21 images whilst his eye movements were recorded and his radiological comments on each case were tape recorded. The clinical room lighting was dimmed to approximately 21.5 *lux.*

6.2.1.2 Study Results

Not surprisingly, the expert correctly detected all the abnormal lesions and specified their locations accordingly. As he examined seven cases which had been used in the previous study (see section 5.2) then his data (based on visual dwell measures), including key fixation locations, were compared with the eye movement data collected from the radiography student participants from the last chapter. Details are shown in figure 6-1.

The expert viewed all seven images the same size on the monitor whereas the students viewed images presented at different sizes as indicated in figure 6-1. Detailed recording of the expert's visual search behaviour over a number of MLO view screening cases has elucidated somewhat different search behaviour (e.g. Figure 6-2) to that as advocated by Tabar (Tabar & Dean, 2001) and also quantifiably different from the radiography students.

In figure 6-2, each column represents one observer's eye movements on the set of images (the first one is the expert's eye movement), each row represents one image examined by different observers. It is evident that the expert essentially used the same search pattern for each case. He examined each breast in detail before making comparisons between the two breasts. His fixations are relatively short compared to the radiography students in the previous study.

Clearly, differences in saccadic eye movement patterns can be found between experienced radiologists; however it would be expected that there is similarity in the actual image areas foveally examined which should represent (1) known general high probability areas for abnormality presence (e.g. just behind the nipple) and (2) areas which, per image, suggest potential abnormal appearance based wholly on the individual appearance of that image.

All the eye movement data in the twenty-one cases and the audio description of each case were then used to provide training to naïve participants in the following studies.

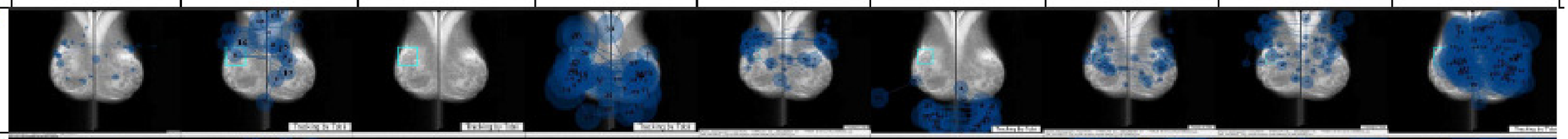
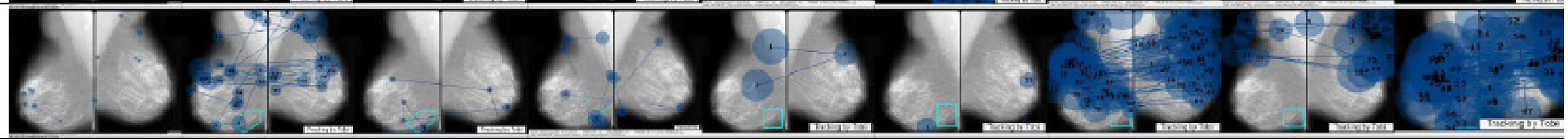
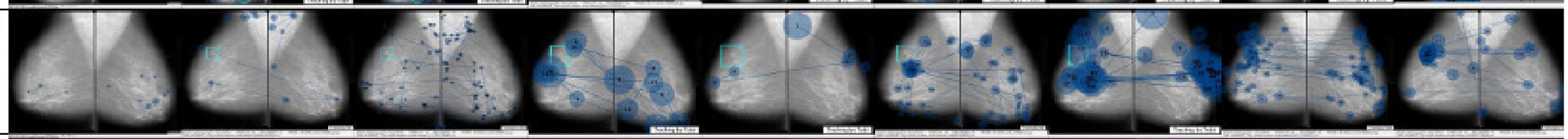
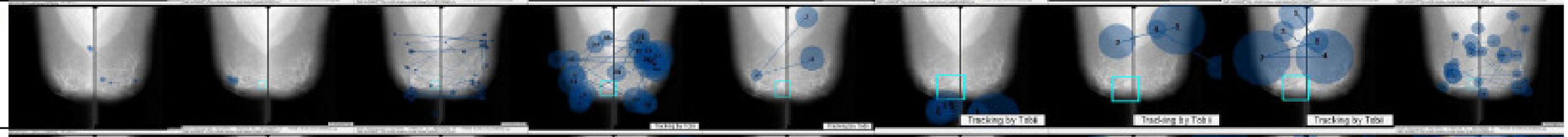
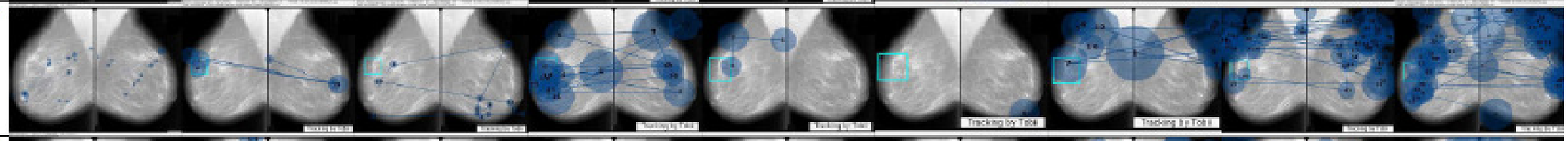
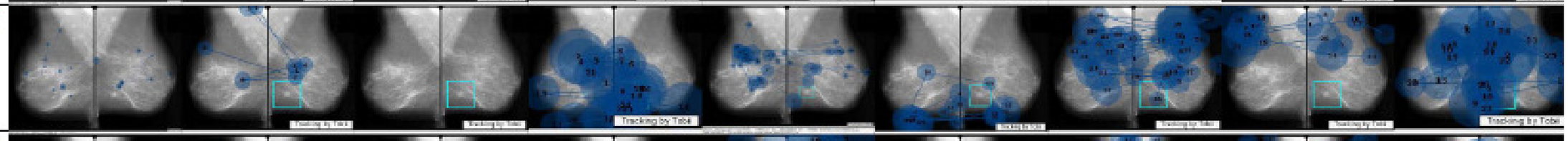
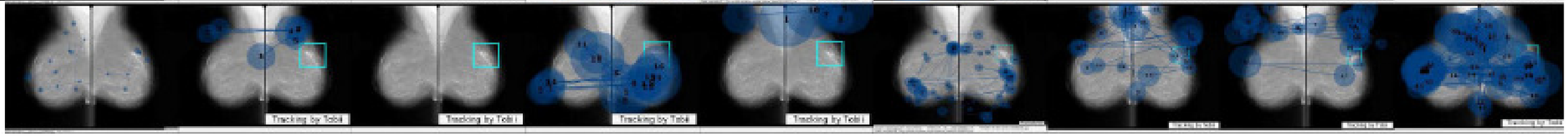
	Expert	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Observer 6	Observer 7	Observer 8
1									
2									
3									
4									
5									
6									
7									

Figure 6-2. Eye movement comparison between the expert and the other eight inexperienced film readers on seven mammogram cases (MLO view)

6.2.2 Study Two

In order to investigate the different types of computer based training for mammographic interpretation, a group of naïve participants were invited to undertake one of four different types of training and two sessions of image examination exercise before and after the training. The objective was not so much to try and specifically train naive participants to successfully identify cancer as much as to determine whether each training approach was feasible and whether there were any performance differences between the approaches.

6.2.2.1 Methods and Materials

The training methods investigated arose from feedback on the questionnaire investigation (Chapter 3) as well as from various discussions with radiologists (Chapter 4) and with the expert radiologist used here in Study One. The initial approach was to present the two MLO views full size on a monitor coupled with the abnormality highlighted in some way, together with some related descriptive text. This would then be followed by the CC views of the same case and related similar information. This mimics a common training approach taken in mammographic textbooks as well as simulating viewing images on a monitor with no HCI functions.

Secondly a similar method was employed but after identifying the feature/location then this area was magnified to increase the perceptibility of the features being described. This is commonly used both in some textbooks as well as in various conference presentations by experts as a means of conveying information concerning feature descriptions easily. It also represents image examination on monitors where HCI manipulations are employed.

The third approach was based around using the expert's visual search behaviour from Study One. Consequently the MLO views were shown and these were then animated with an annotated version of the expert's visual

search behaviour. Then the CC views were similarly shown. The idea here was to see if participants could understand the broad elements of the search behaviour.

Finally the MLO images were shown together with the expert's audio description of what he thought about these images. This was then followed with the appropriate CC views and again the audio comments. This approach mimics aspects of an expert describing a case to a trainee.

Participants

Twenty naive observers participated (seventeen research students in various subjects, and three university employees) in the study. There were nine female participants and eleven male participants aged between 22 and 35 years old. None of them had any experience in mammography reading and all had normal vision or corrected to normal vision. Participants were split into four experimental groups and a control group (which did not undertake any training) with the experimental groups undertaking different forms of computer based training.

Materials

Visual stimuli: Twenty-one sets of recent digital mammographic images were used in the test set. Each image set comprised both the Medio Lateral Oblique (MLO) views and Cranio Caudal (CC) views of both breasts. Fourteen of the pairs featured a specific abnormality which had been grouped into two types (namely: Mass and Calcification) with the abnormality visible on either one or both views. These two features were deliberately chosen as whilst both can be difficult to detect on a mammogram, masses are generally relatively large and calcifications are fairly small irregular abnormalities which can appear singly or in clumps and of various sizes. Seven of the sets featured no abnormality.

Each training set included the 21 cases (11 of the pairs featured a specific type of Mass (i.e. IDM, Spiculate mass); 10 of the pairs featured Calcification) comprising both MLO and CC views of each case but were presented in four different formats (figure 6-3).

Viewing Device: this was the same as the equipment which was used in the previous study (section 6.2.1)

Eye Tracking Device: A Tobii X50 stand alone remote oculometer eye tracker. (Accuracy: 0.5-0.7 degrees visual angle) was used as in the previous study (section 5.2.1).

Design

Training sets: each training set included the 21 cases but were presented in four different formats (figure 6-3). Figure 6-4 demonstrates the design using the actual images.

These formats were:

T1 – whole image: the MLO views of both breasts were presented then, where appropriate, the feature area was highlighted by a circle with descriptive text, followed similarly by the CC views.

T2 – magnified area of interest: the MLO views were presented (as in format T1) but then the area of interest around the abnormality was highlighted followed by this area being magnified and shown alone (size: 8”x 8”) on the monitor screen. This was followed similarly by the CC views.

T3 - eye movements: the MLO views were shown and then overlaid with annotated fixation locations of where the initial expert radiologist had looked and in the order in which his visual search had been performed. Also, the area of abnormality was highlighted. This was followed similarly by the CC views.

T4 - comments: This was similar to format T1 above but with the addition of the expert's audio descriptions concerning the case.

Control Group - a control activity (45 minutes of book reading) was undertaken which took the same length of time as the other training sessions.

Procedure

Ethical approval was granted by the Loughborough University ethics committee. All participants were given a participant information sheet (see Appendix D) describing the study and informed consent (see Appendix B) was obtained before each experiment started.

The experiment took place in a darkened laboratory with a viewing room ambient light level of 23 *lux*. For each participant the eye tracker was first calibrated. Each was then given a short standardised introduction on the computer concerning breast cancer, mammogram images, and were familiarised with the appearance of the two different key breast cancer features in such images (Masses and Calcification). They then visually examined two practice cases and any queries they had were answered.

Subsequently, they completed a computer-based image examination task whilst their eye movements were discretely recorded (figure 6-5). During this task the participants first fixated on a small centrally presented fixation cross and then this was replaced with the MLO views of the case which in turn was followed by the CC views. Participants were asked visually to examine each case view, identify whether the case were normal or contained an abnormality. If the latter, they also had to specify the feature type (i.e. Mass or Calcification) and its location.

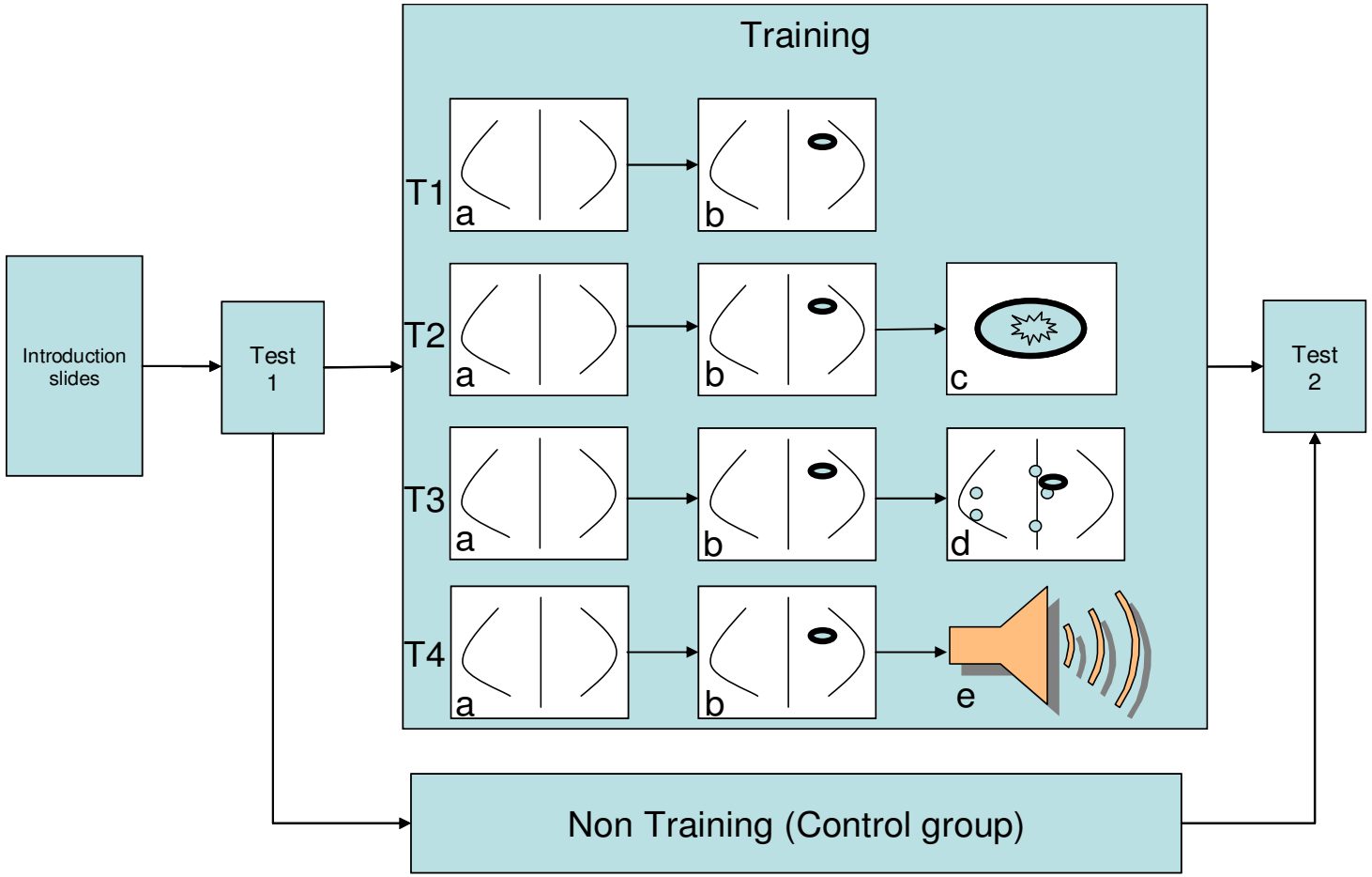


Figure 6-3. Schematic of the different training and control approaches.

a: both MLO views of a case; b. both MLO views with the area of interest highlighted; c. the portion of the mammographic image around a potential abnormality site was presented; d. both MLO views presented with expert's eye fixation path information overlaid; e: playback of the expert's verbal instruction (both with concomitant MLO image presentation). Each training approach was then repeated for the CC views of each case.

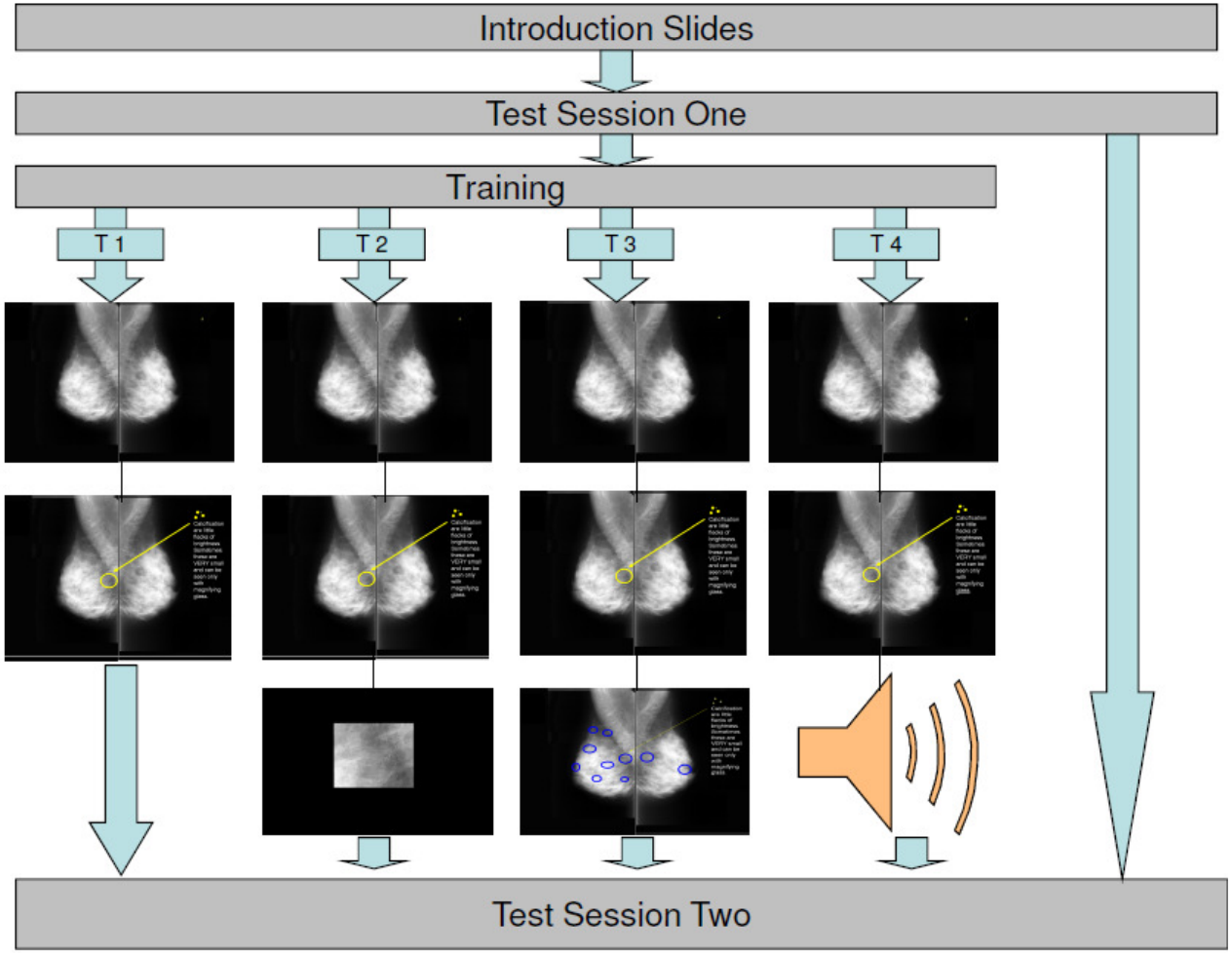
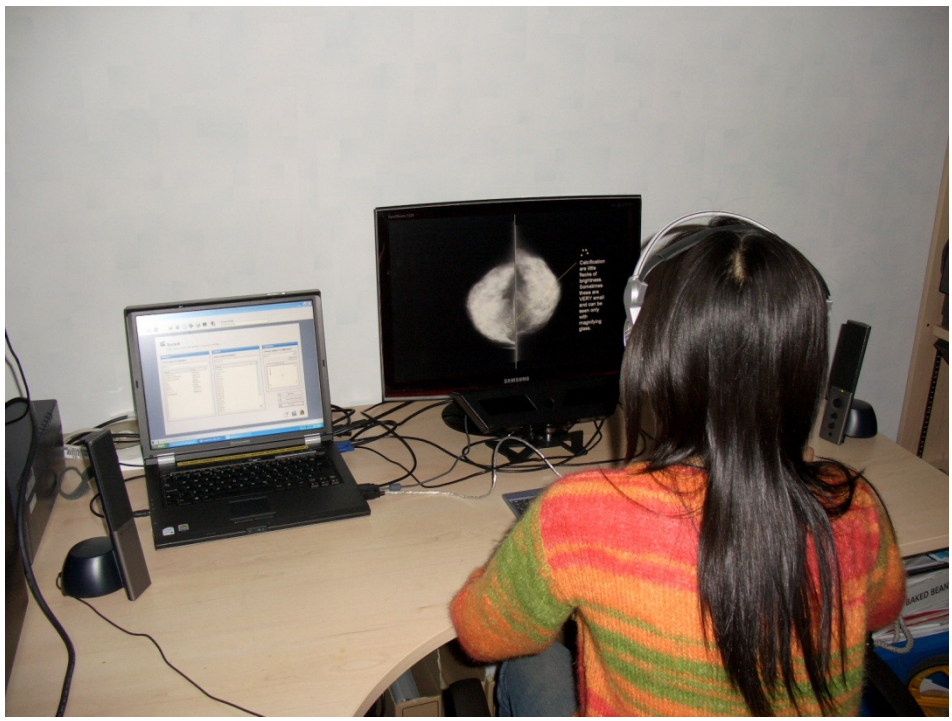


Figure 6-4. Example of schematic of the different training and control approaches.

a)



b)

Figure 6-5. a) shows a participant taking the test viewing the MLO mammograms of a case; b) shows a participant undertaking the audio training whilst viewing the CC images of a case. The photographs were taken under normal room lighting level for illustration purposes.

Both before and after training, each participant was tested by being presented with a series of 21 digital mammography cases (both MLO and CC views) in a random order. Seven cases featured Masses with the abnormality visible on either one or both views; seven cases featured Calcifications with the abnormality visible on either one or both views; the other seven cases were normal and showed no abnormality. Each participant was required to identify whether a case demonstrated either the presence of a Mass, Calcifications or was normal. If they thought some abnormality was present then they indicated its location and rated their confidence in their decision. Participants' eye-movements were recorded throughout.

6.2.2.2 Study Results

Performance Results

In this experimental study of naïve observers some 880 responses were collected and of these only 124 (14.09%) correctly identified both features and locations on both views of a case. Initially, a 2 x 5 mixed design ANOVA revealed that there was a non-significant main effect of test session on the performance before and after the training, F (1, 15) =.252, $p > .05$, $r = .13$. There was a non-significant main effect of training types on the test performance before and after the training, F (4, 15) =.814, $p > .05$, $r = .23$. Also, there was no significant interaction between different training type and test sessions: F (4, 15) = 1.37, $p > .05$, $r = .29$. The percentage of each training type correctly identifying both the feature and location before and after the training are shown in figure 6-6.

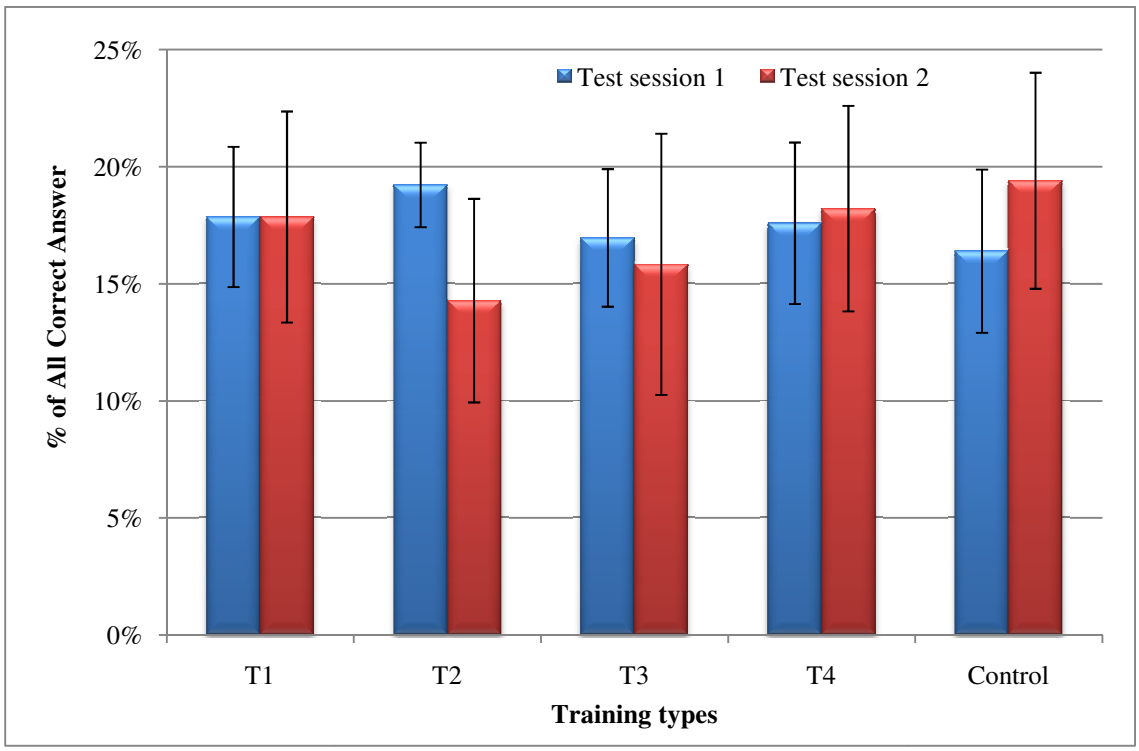


Figure 6-6. Percentage of both correct answers (feature & location) of each training type (both test sessions).

As no statistically significant main effect of training was found, the training types were then grouped as visual (including; whole image, magnified area of interest, or eye movements), audio (comments) or the control group - details are shown in figure 6-7. ANOVA indicated that there was no significant effect of training type: F (2, 437) = 1.4, $p > .05$, $r = .06$; However, in the audio group, on average, a t test revealed that participants performed significantly ($p < .05$) worse in test session 2, than in test session 1 with performance decreased (from 20.4% correctly identify both feature and location down to 9.5% correct answers) after training. For both the visual and control conditions a t-test showed that there was a slight, non significant ($p > .05$), increase (details see figure 6-7).

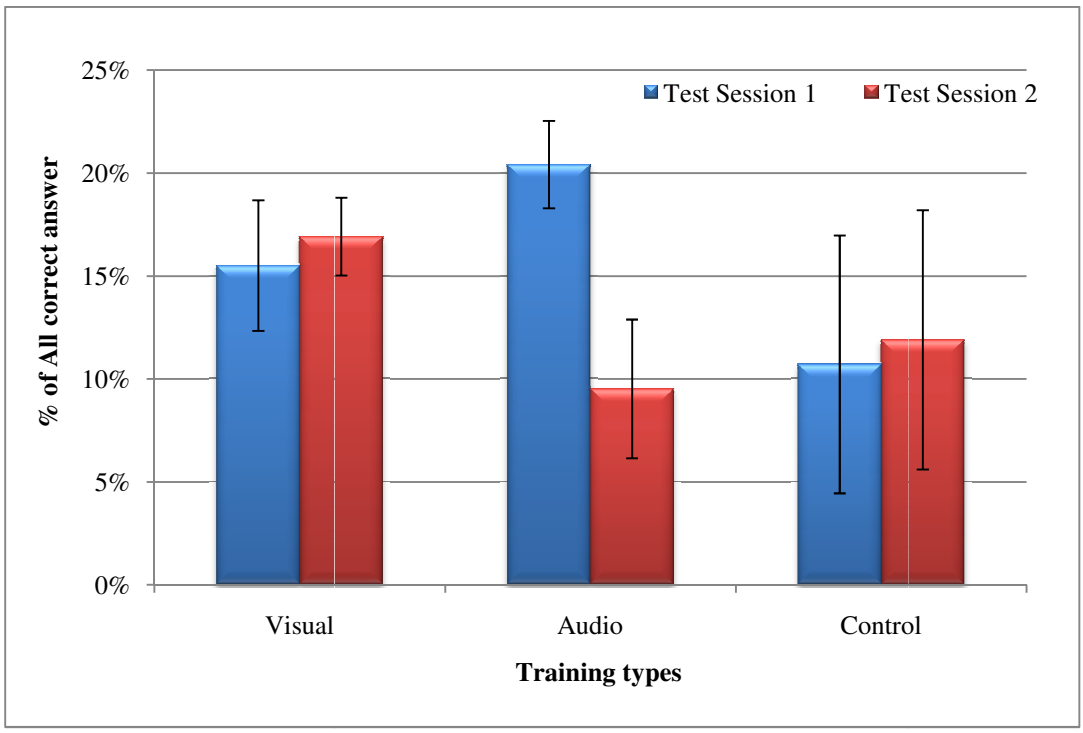


Figure 6-7. Percentage of both correct answers (feature & location) of the training groups (both test sessions)

Given the difficulty of correctly identifying both the abnormal feature and feature location for such naive participants, the data appeared to be insensitive to detecting any increase/decrease in performance before or after the training session. Therefore, the data were analysed by separating the answers on feature and location of each view. To do this the participants' performance in each test session was attributed points concerning whether they correctly identified location and feature on both views. This is shown in figure 6-8 where 1 point is attributed for correctly identifying feature or location in either mammographic view. Thus a score of 4 indicates correctly having identified both the location and feature on both MLO and CC views.

Treating the data in this fashion then most participants scored 1-1.5 indicating that they only correctly scored either location or feature in either view. In the three visual conditions there was a slight increase in performance in test 2, in the audio condition performance decreased and in the control condition there was little variation. For details see table 6-1.

Table 6-1. Examples of the scoring system for five hypothetical cases.

Case Number	*MLO feature*	*MLO location*	*CC feature*	*CC location*	***Score***
1	1 point	1 point	1 point	1 point	**4 points**
2	1 point	1 point	1 point	0 point	**3 points**
3	1 point	1 point	0 point	0 point	**2 points**
4	1 point	0 point	0 point	0 point	**1 point**
5	0 point	0 point	0 point	0 point	**0 point**

A 2 x 5 mixed ANOVA showed that In all conditions participants were significantly better in identifying features correctly than in identifying correct location: $F(1, 435)=121.68$, $p < .05$, $r = .47$; the performance in specifying features was significantly different between training groups: $F(4, 435) = 2.43$, $p < .05$, $r = .07$; the performance in specifying location between training groups was not significantly different : $F(4, 435) = 1.08$, $p > .05$, $r = .05$. Further post hoc t tests showed that Masses were significantly ($p <.05$) better identified after any visual training and only slightly increased after audio training ($p >.05$). Calcifications were detected worse after all three types of visual training as well as the audio training ($p <.05$). In the control condition calcification identification improved. Normal cases were reported worse on every second trial ($p <.05$).

In the control group the correct identification of normal cases dropped on the second test but mass and calcification detection increased ($p >.05$). In the audio condition calcification and normal identification fell on the second test but mass identification improved slightly ($p >.05$). In the three visual conditions only mass identification improved ($p >.05$).

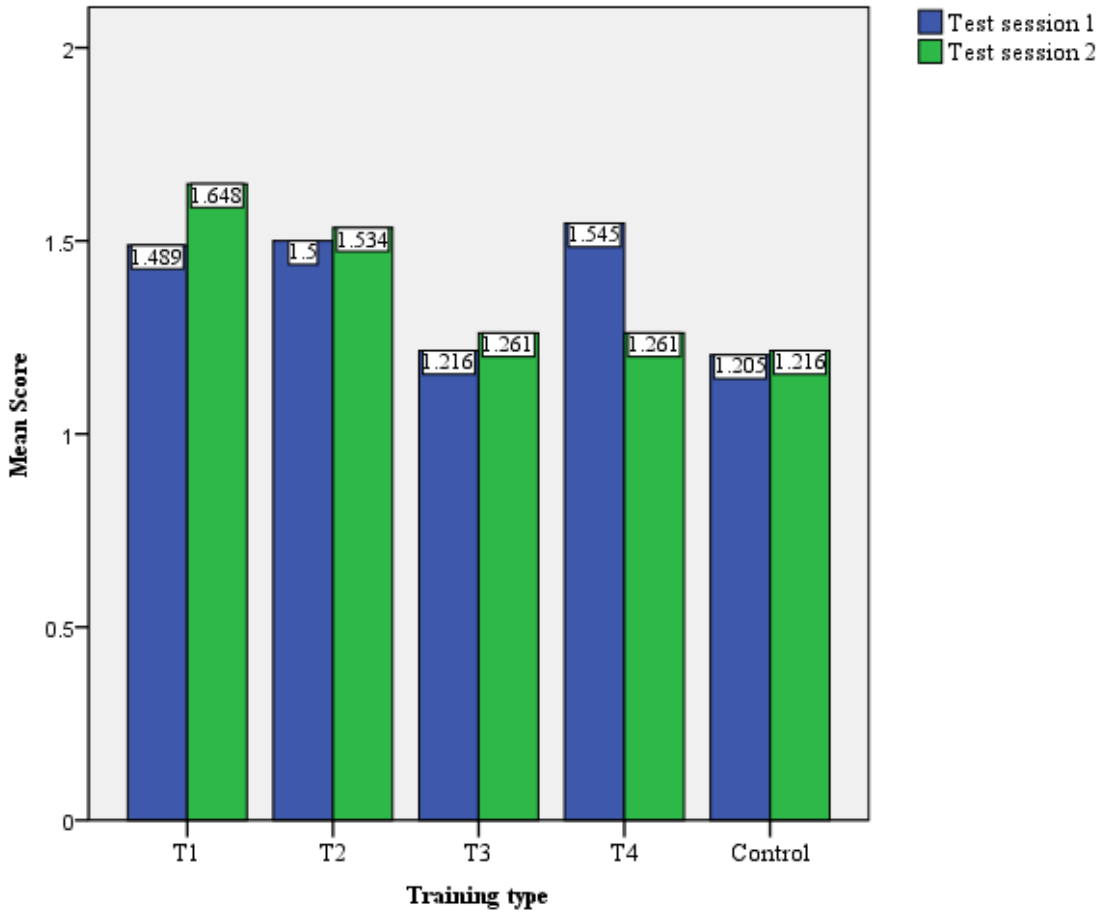


Figure 6-8. Mean performance values of each participant group (both test sessions).

Visual Search Data Analysis

A key interest here was in the participants' visual search behaviour and whether this was affected by the different training approaches. This was analysed in detail and where participants made errors then this was broken down into whether they made search, detection or interpretation errors based on whether they failed to actually fixate on or near (i.e. within the area of interest: AOI) the abnormality (a search error) or whether they did fixate near the abnormality but only for a short period of time (*<1000ms*; detection error) or a longer period of time *(= >1000ms*: interpretation error) (Kundel, 1978).

Examples of the pattern of visual exploration of a number of MLO views of cases by the same observer are shown in figure 6-9: images A1-B4. In A1 the

observer's scan path is shown before training where the abnormality (emphasized for illustration here by a square) was missed because the observer did not look at or near it (a visual search error). After training (A2), when examining the same case the abnormality was fixated and correctly identified. Another example of a search error is shown in B1. B2 shows a detection error where the abnormality was fixated for a short time period but not detected. In B3 the abnormality was fixated for a longer period of time (i.e. detected) but not reported (an interpretation error). B4 shows a case with two abnormalities, one of which was missed (search error) and the other was correctly fixated and reported.

A one-way MANOVA revealed a non-significant multivariate main effect for training type, Pillai's Trace = .300, F (16, 52) = 1.199, p >.05. However, the number of search errors made was significantly more than the other two types of visual errors made in both test sessions (p <.05). The overall pattern of errors is shown in figure 6-10 for each test session.

These data are shown in detail in figures 6-11 and 6-12 for the MLO and CC view of each case respectively. In both figures each error type is plotted for the first and second test phase.

Overall the data indicated that most errors were due to search – participants simply missed abnormalities because they failed to look at or near them (i.e. in examining the case the abnormality failed to fall within their useful field of view) with little variation between the test sessions before and after training. It would not be expected that such a short training session, designed to explore whether each approach was at all useful, would engender major improvements in abnormality identification. On average, figure 6-10 shows that after training errors tended to decrease but not markedly so.

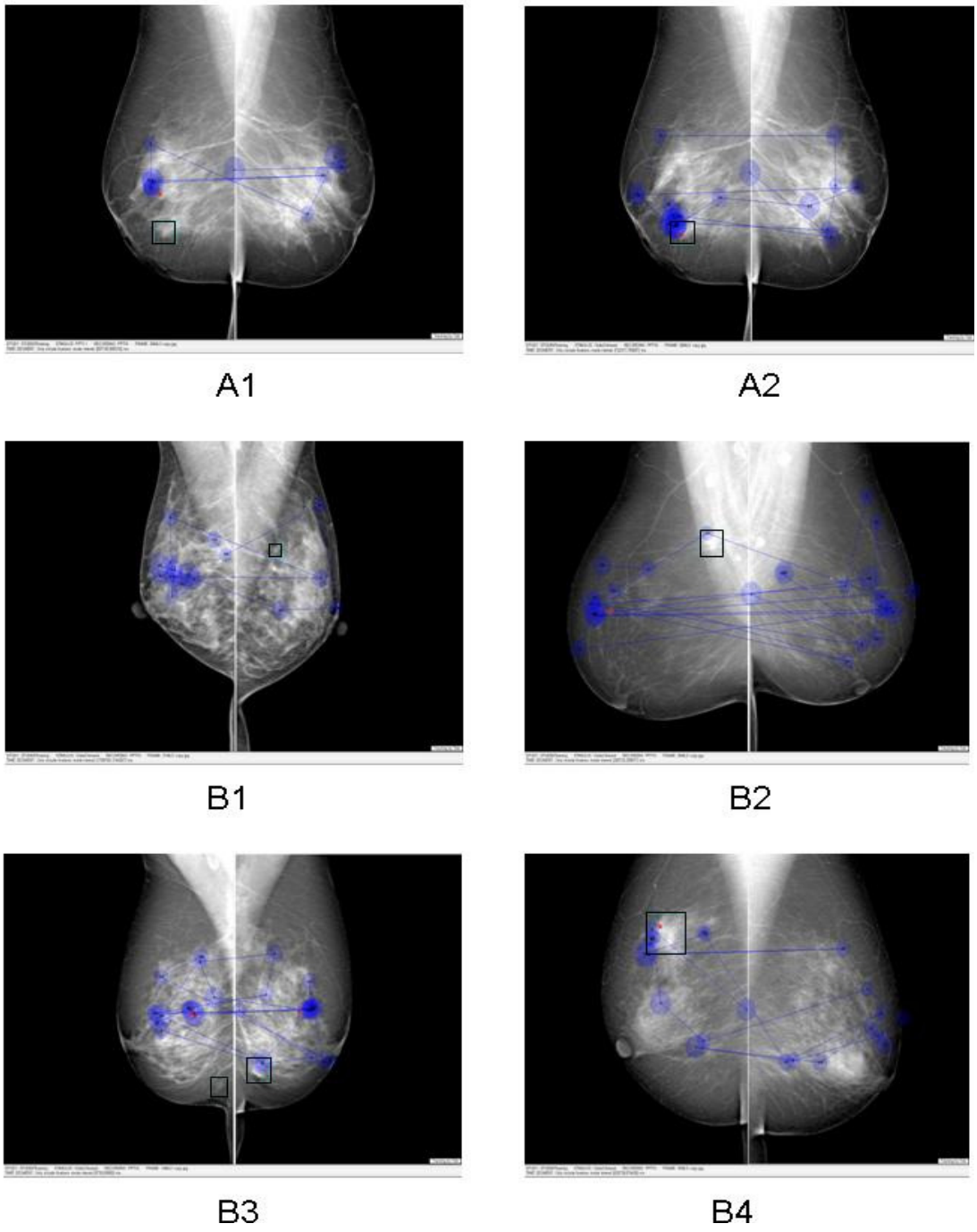


Figure 6-9. Examples of the pattern of visual exploration of six MLO views

As would be expected, the expert's visual search patterns for each case showed an overall planned search behaviour which was modified by the particular appearance and features present in each case. Replaying this for participants was an attempt to see if they would, or could, then follow this planned search over different cases. The data demonstrate that this was very

difficult for them to do, which agrees with prior research on structured search in radiology (Gale & Worthington, 1983).

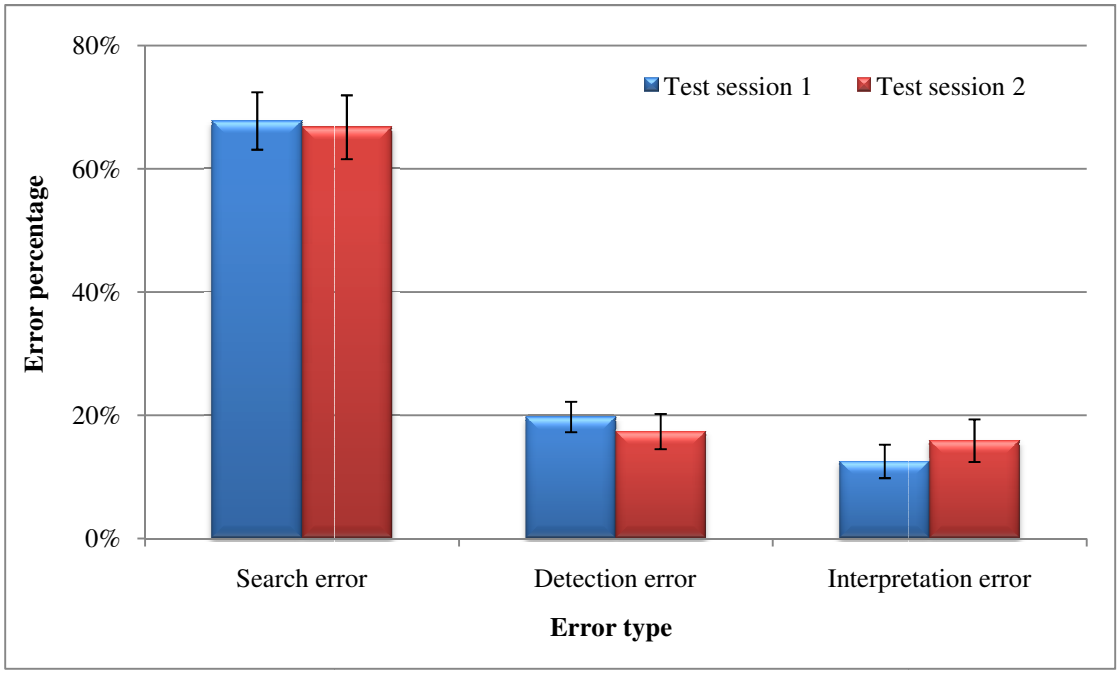


Figure 6-10. visual inspection errors in test session one and test session two

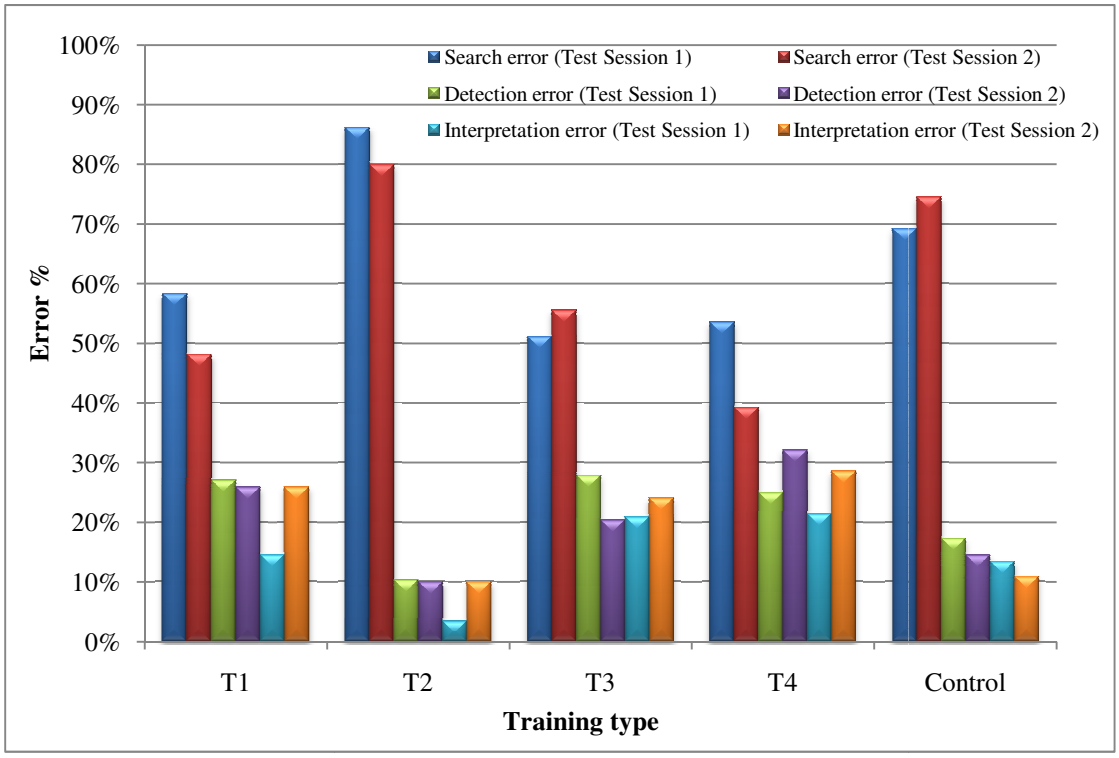


Figure 6-11. Different types of visual inspection errors x training type (MLO view)

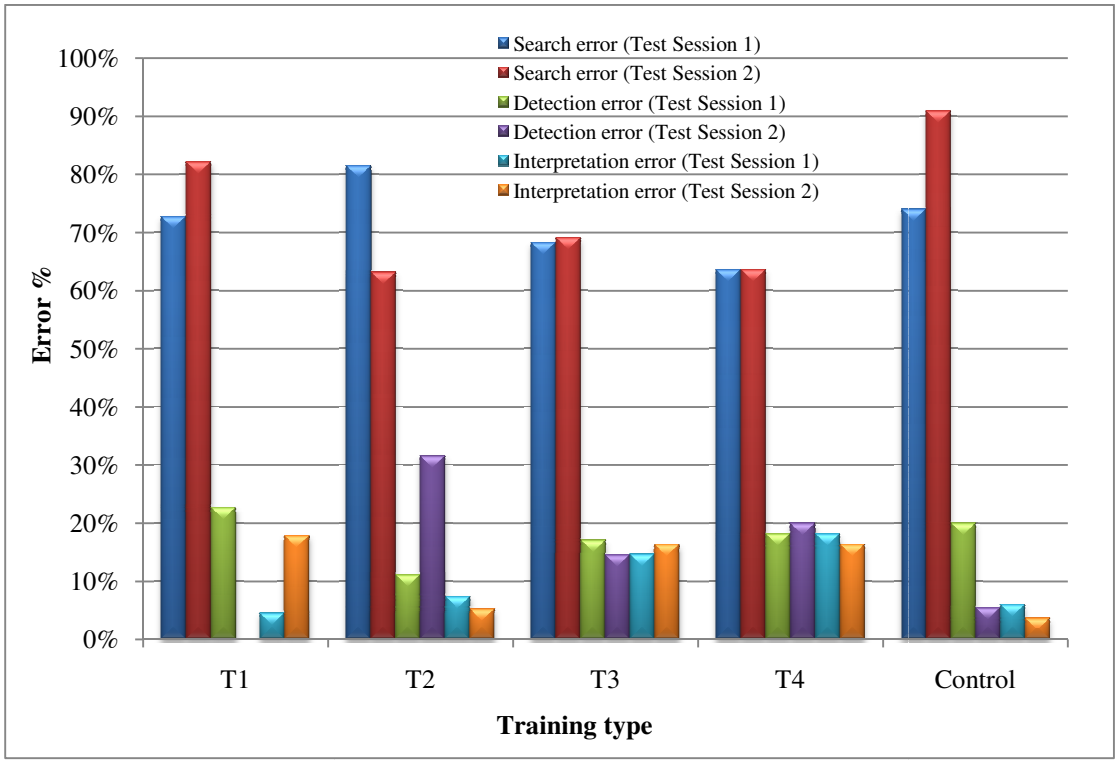


Figure 6-12. Different types of visual inspection errors x Training type (CC view)

The dwell time (the total duration that a participant fixates his/her eyes within the area of interest) was also compared between test session 1 and 2 and between training group and control group (see figure 6-13). However, a one-way repeated measure ANOVA indicated no significant difference between test sessions F (1, 18) = .381, p > .05, effect size Cohen's d = .36, correlation r = .16 (small to medium effect).

These figures were used to derive estimates of the required sample sizes based on designing a study with power of .80 at an alpha level of .05 by consulting the appropriate look-up table (Maxwell, 1990. pp 570, table 13.10). The look-up table showed that a minimum of 59 observers are required while d = .50 and r =.16; and 228 observers are required while d = .25 and r =.16. It was estimated therefore that between 59 and 228 observers are required to compare dwell time differences between test sessions in order to achieve a study with 80% power and an α-level of .05. The study was therefore underpowered to detect an effect of training type: this could be because of the

small sample size; they were not actual screeners, or because only a short training time was employed. Using practicing screeners this would mean that the study requires between 10% and 38% of the screeners reported in Chapter 3 which is practically not feasible to achieve.

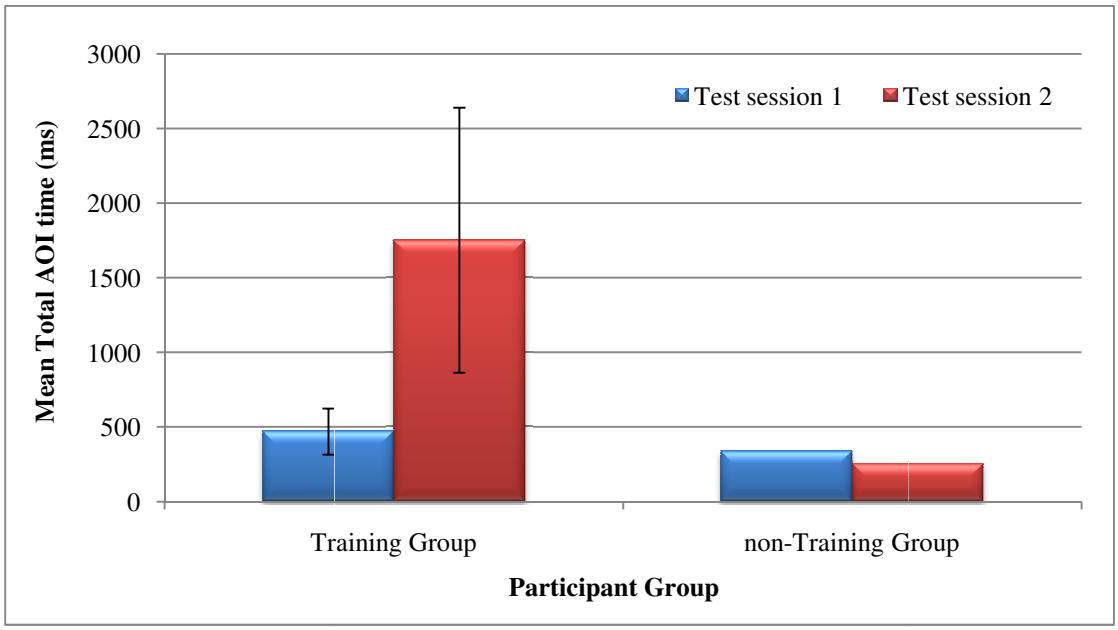


Figure 6-13. Mean of the total times spent within the AOIs for the abnormal images by training group/non-training group

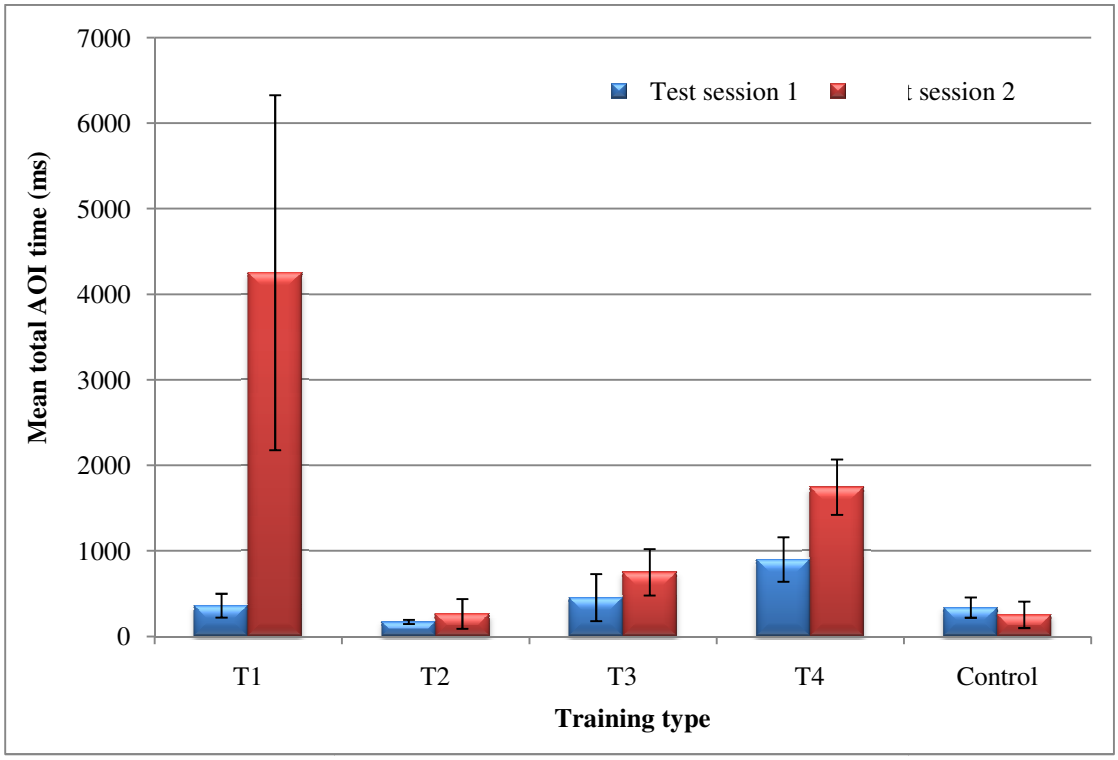


Figure 6-14. Mean of the total times spent within the AOIs for the abnormal images by the different training groups

Figure 6-14 shows, for the abnormal images, the mean of the total fixation time within the AOI for each training group on both test sessions. This shows that after training the participants, in groups T1, T2, T3, T4, spent more time looking within the AOI at potential abnormalities; however, for the control group fixations inside the AOI decreased.

6.2.2.3 Discussion

The experimental study reported here was an investigation utilising a single low cost monitor to deliver mammographic interpretation training as compared to using one or two high resolution workstation monitors. Different training regimes were developed which presented observers with images enhanced using different HCI approaches which were considered suitable in this domain. The approaches used were firstly simply showing the full views of both breast simultaneously which mimicked the display on a workstation but at a much lower resolution. Secondly the full view was shown followed by a

magnified view of the area of interest to mimic basic image manipulation on a workstation. The third approach taken was to utilise areas of interest which attracted the visual attention of an expert together with the scanning path which were overlaid on the image. The fourth approach tried to replicate an expert describing how they examined a case for abnormalities which is a commonly adopted approach in real screening as shown in the questionnaire survey.

The study set out to use short training approaches to examine their effects on naïve observers. This demonstrated that such observers can be trained to recognise certain key breast cancer appearances using a low cost display monitor along with a range of HCI techniques. Two mammographic appearances were studied; small calcifications because these can be difficult to detect in routine breast screening and larger masses. Calcifications were not detected well presumably due to the shortage of image manipulation techniques used here. Naïve observers were used in this study to see how they responded to the different training types.

6.2.3 Study Three

After the training experiment was carried out with the naïve users, experienced breast screeners' opinions on the different training methods used here were elicited. A focus group was undertaken with 15 breast screeners who were shown examples of each of the four training methods and asked to give their evaluations and suggestions concerning each method.

Their feedback included:

1) if they thought they would find each approach helpful as a training tool;
2) what they considered to be the advantages and disadvantages of each method;
3) how further to improve each training method;

4) when and where they considered could be most suitable for the use of each training method.

6.2.3.1 Study Results

From the focus group it was apparent that experienced breast screeners felt that hearing an expert's description about a case (c.f. T4 approach), whilst being presented with that case to examine themselves, would be a very useful training approach. Simply replaying an expert's visual search behaviour overlaying the case itself (T3) was not felt to impart much information. Highlighting an area around a potential abnormality and then presenting this area in a magnified fashion (T2) for closer inspection was also judged to be a useful approach.

6.3 Overall Discussion

Our previous research (Chen, Gale & Scott, 2008) has elicited that users would like to have additional training opportunities to view digital mammographic images over and above simply viewing them in a breast screening centre and using a high resolution diagnostic workstation. There are clearly limits to using other types of display devices (e.g. display size, resolution and contrast issues) for these particular images, where an early sign of cancer may be not much larger than the head of a pin, and it is not argued here that other poorer quality displays should be advocated for detecting possible cancer presence in breast screening. However, it is argued that the use of other display devices for certain aspects of training could be a useful facility which would expand the availability of training opportunities.

This work set out to determine whether any of the proposed training approaches were feasible using a typical office computer monitor and naïve observers. It was not designed as an exhaustive test of these approaches. The four training methods were devised as variants on what breast screeners

do when they interact with a digital workstation coupled with potentially utilizing aspects of the visual search behaviour of an experienced radiologist.

The empirical data demonstrate that these users, after a short familiarisation with the appearance of abnormalities and mammographic images, were often unable to perceive the abnormality and consequently made search errors. When they were able to identify abnormal image areas (figure 6-11) then even though they spent considerable time looking at the area of the abnormality they still made errors. The auditory approach performed worse than the others, however this approach is the one indicated by the focus group with actual screeners that they would prefer – possibly because it is somewhat similar to existing training approaches in screening where images are viewed whilst a tutor describes the image content. Being able to zoom in to an area of interest in an image which allows smaller potential abnormal appearances to be viewed in detail and which simulated one of the HCI manipulations on a digital mammography workstation was envisaged as being particularly useful, however the data indicate. Utilising the visual scan path of the expert radiologist, as implemented here was not found to be as useful as had been supposed. Approaches which result in these naïve participants spending more time looking at the AOI (T1 and T4) are the most promising as this is the beginning of the learning process of identifying abnormal appearances.

6.4 Conclusions

All of the training approaches were found to be feasible to implement but of variable usefulness. Overall, it is argued that these findings taken together indicate that low cost devices can be used for training purposes in digital breast screening with appropriate HCI techniques. These then extend the opportunity for training beyond the clinical workplace.

However, the failure of achieving statistically significant results among different training types from the second study could due to two separate reasons:

1) The participants that were invited to take part in the study were not personnel who have a breast image interpretation background;

2) Mammographic interpretation training has been established to be a long and time-consuming process. For example, in the UK, current radiographers are trained to become advanced practitioners via a year's Masters Course to reach the minimum standard for mammographic examination. This implies that a future study should look into the possible long term effects of the different types of training provided to different groups of participants (i.e. the radiographers who are in mammographic interpretation training to become advanced practitioners). However, this was not the purpose of the current investigation.

This then leads on to a further investigation of practicing screeners' performance on potential training devices and how they utilise HCI techniques.

CHAPTER 7

Mammographic Interpretation Training: Suitability of Displays for Delivering Mammographic Interpretation Training

7.1 Introduction

From the research detailed in the foregoing chapters it is argued that displays less sophisticated than clinical workstations could be used to deliver some aspects of training in breast screening. What is not known is how such displays (e.g. the iPhone representing PDAs and a standard office monitor driven by a laptop computer) fare in comparison to clinical workstations. Earlier it has also been argued that using suitable HCI techniques with such poorer displays should make them more practical and useful, however, exactly how useful is unknown. Figure 7.1 clearly shows the difference in size of these relative displays and their resolutions.

Consequently here, experienced breast screening radiologists and advanced practitioners who have familiarity with both digital and analogue mammograms were used in a study to examine how they performed using such different displays. A known set of difficult cases were examined by participants on three separate occasions using the three displays either with or without HCI aids. The image set comprised difficult recent screening cases containing a range of appearances and key mammographic features ranging from small calcifications to larger masses and architectural distortions. It was hypothesised that whilst all features would be visible on the clinical workstation, on the non clinical displays then the smaller features (calcifications) would probably not be visible without using HCI tools. The participants' visual search behaviour and image interactions were recorded and then later analysed in considerable detail.

It was hypothesised that performance on the workstation would be best, particularly when image manipulations tools were used. This was also predicted to be the fastest condition. The standard monitor, with or without HCI help, was predicted to elicit poorer performance and image examination would take much longer than the workstation. Use of HCI tools was hypothesised to produce better performance than not using such tools. The iPhone was predicted to produce poorer performance even with HCI usage. It

was predicted that participants would take far longer with this display and most likely lose where they were within the displayed image.

7.2 Study

7.2.1 Methods and Materials

Participants

As mentioned in the previous chapters, 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. In doing this they have purchased digital mammography capture and viewing equipment from various manufactures as there is no overall central NHS approach to purchasing equipment – only a requirement to meet certain standards. Thus, whilst all such equipment uses DICOM mammography images and so should be able to display images produced from other manufacturers' systems there are some differences in displays from different manufacturers. This means that images captured by one manufacturer do not easily display on another's display; this has largely come about because different algorithms are used by each manufacturer to maximise their systems. Currently in the UK digital mammography systems are used by the following manufacturers: GE, SECTRA, FUJI, HOLOGIC and SIEMENS. In this study, centres were selected which used both GE capture and display equipment. GE is the major market leader in digital mammography in the UK and so results should be extensible to many other screening centres.

Three centres were selected (Nottingham, Derby and Brighton) which each had GE digital mammography equipment for at least four years and all screeners were familiar with the appearance of images from GE digital mammography equipment. Therefore, nineteen screeners from these three major breast screening centres were approached with detailed study-related information. For logistical reasons (the large amount of experimental

equipment required to be transported from centre to centre), it eventually turned out that to conduct the study at Brighton was not feasible. Consequently, there were fourteen participants (nine consultant radiologists and five advanced practitioners from Nottingham and Derby - two of the major UK breast screening centres - who volunteered to undertake the experiments.

The participants were then divided into two groups according to the different screening centres where they primarily worked.

Materials

Visual stimuli: An expert breast radiologist selected two sets of 20 challenging to 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.

Hardware: This comprised three sets of viewing equipment: (1) GE digital mammography workstations (with 5 megapixel dual monitors; resolution 2,04 x 2,560 pixels each); (2) a standard LCD monitor (images were shown using a DICOM viewer running on a laptop, screen size: 21.5", resolution: 1,050 x1,680); and (3) 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 (figure 7-1).

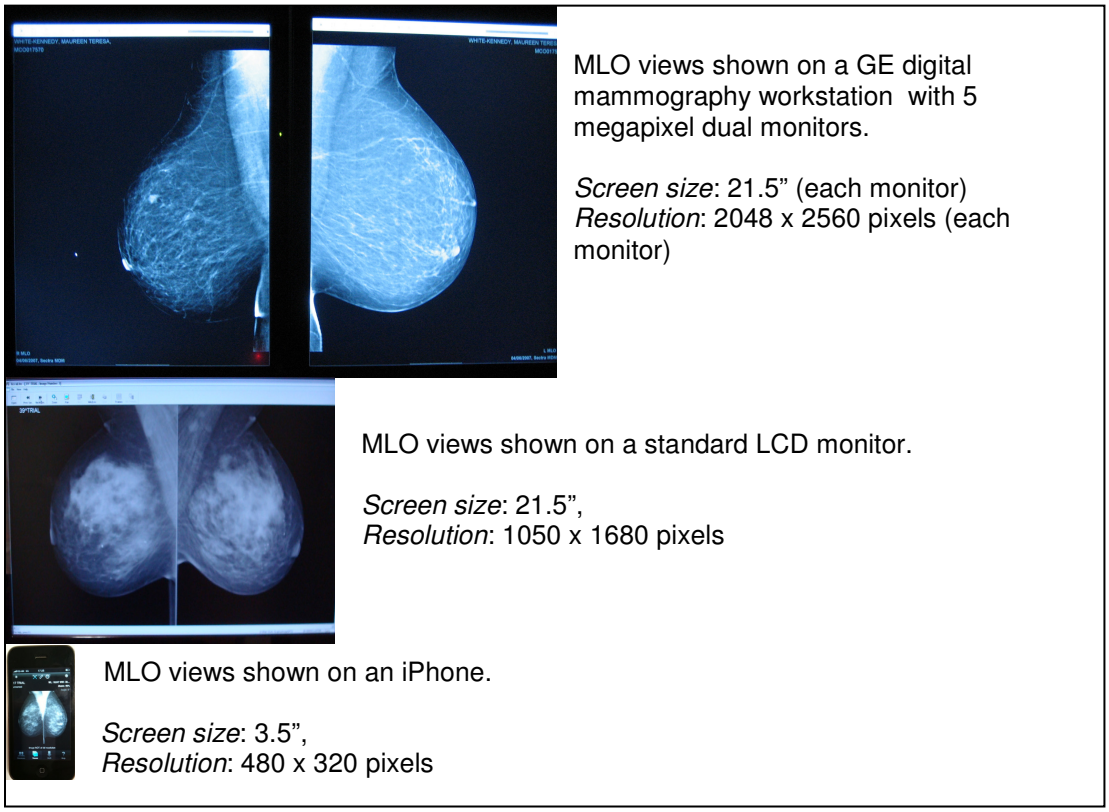


Figure 7-1. Illustration of the respective sizes of the three display systems

Eye Tracking Device: A head mounted eye tracker (ASL 504) was used to monitor the visual search behaviour of the experienced screeners in this study. Because of the size of the workstation monitors the head mounted system was used instead of a remote eye tracking device, which would sit beneath the displays unconnected to the observer. This is because remote systems 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. It also enabled the individuals to perform the task just as they would do in normal everyday screening, rather than performing the task in a prescribed way to suit the experimental situation i.e. they were free to move their heads about at will and to move back and forth to the monitors – such actions have to be somewhat restricted when a remote eye tracking system is used.

Procedure

Appropriate ethical approval was granted by the Loughborough University ethics committee. In addition, an application was prepared for appropriate NHS COREC ethics approval. At this point, the expert radiologist at the Nottingham Breast Institute suggested contacting the chair of the Nottingham Research Ethics Committee who determined that the proposed study came under 'service evaluation' and therefore there was no COREC ethics approval needed. Written confirmation of this was received from the local research ethics committee.

Pilot Study

A pilot study was first conducted. The equipment was set up in a radiological reporting room. A head mounted eye tracker (ASL 504) was used to monitor the visual search behaviour of the experienced radiologist as he examined a series of recent screening digital cases on a GE mammography workstation.

The study was conducted to determine how an experienced observer examined images on the workstation and what interactions/manipulations he made to the images in order to identify abnormalities. Additionally, the investigation provided an estimate of the time required for fitting the ASL eye tracker on the participant and calibrating the eyes. Also, the pilot study was used as an example to estimate the time required for each experiment.

Figure 7-2 shows examples of the expert examining a single case for illustrative purposes, together with corresponding times, which has been taken directly from the initial eye movement record (the large white cross hairs indicate the fixation location at that particular point in the recording). This figure clearly demonstrates that for this mammographic image the overall examination time was less than 30 seconds. Initially the two medio-lateral views (MLOs) were examined in full (for 5s), followed by zooming in to examine the corresponding upper MLO quadrants (for 3s) and then the lower MLO quadrants (4s). This was followed by full MLO viewing again (2s) then

switching to the cranio-caudal (CC) view for both breasts (6s) followed by upper (2s) and lower CC (2s) quadrants then the full CC view (1s) again. This record of visual search behaviour serves to illustrate how experienced observers readily utilise the different digital controls of the workstation to examine in detail the image areas of interest.

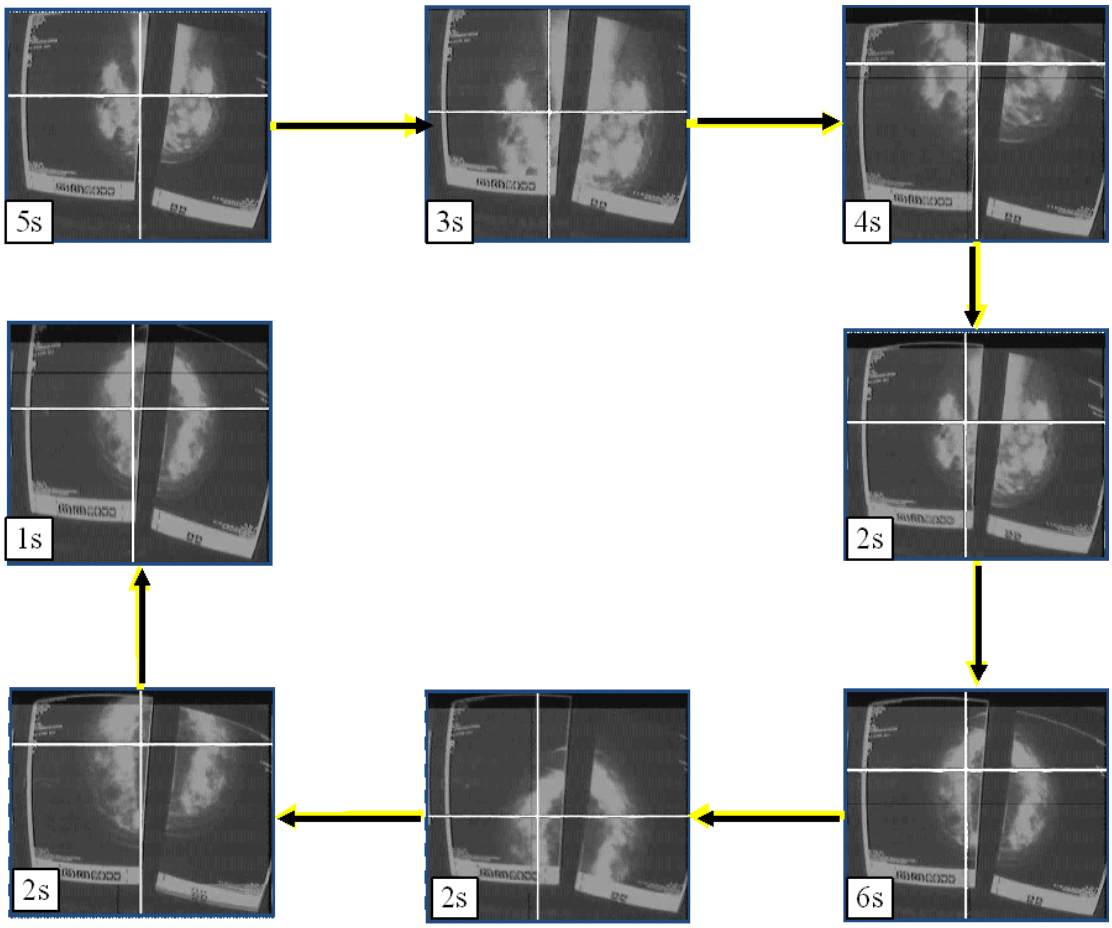


Figure 7-2. Example of the expert radiologist examining one case showing the image manipulation sequence together with his eye movement fixations.

Main Study

Subsequently, over a period of eight months each group undertook three rounds of trials. All participants were initially given the participant information sheet (see Appendix E) and informed consent (see Appendix B) was given before each round of experiment started.

For each round, the experiments took place in a dark image reporting room in the breast screening centre where each participant primarily worked. Every participant examined these DICOM cases on one of three different display devices.

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. For example, all participants from one breast screening 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. Between trials then each participant would have examined approximately over 1,000 routine screening digital cases. Also, there was no case feedback given to participants between each round of the experiment.

The experiment was carried out in darkened radiological reporting rooms with controlled ambient lighting levels of approx. 22 *lux*. For the monitor and iPhone conditions an offset desk lamp was used to provide some additional low level ambient illumination.

For each round of the experiment, each participant read two sets of cases with two viewing conditions. Individuals were either only allowed firstly to view one case set 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 or in the reverse order. The order of viewing the two image sets and the order of viewing conditions were counterbalanced across participants using a Latin Square design. 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 7-3); using two fingers then it was possible to zoom into the image or using one finger to pan or adjust window levels.

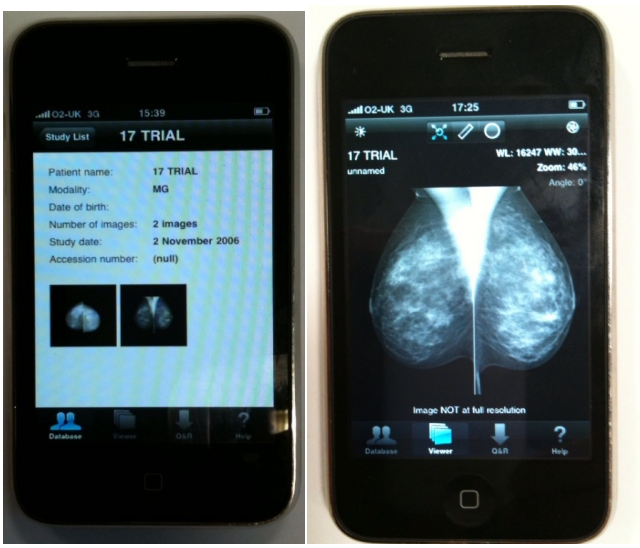


Figure 7-3. iPhone running Osirix DICOM viewing software

Figure 7-4 shows details of the experimental layout when a participant examined images on the workstation. Participants were videotaped using a fixed camera to monitor their behaviour in interacting with the displays. Additionally, they wore an ASL 504 system to record their visual search behaviour throughout. The head mounted eye tracker works along with the eye monitor which was used to monitor the visual data capture. The scene monitor shows the scene video data capture. The eye movement analysis system comprises a Dell computer running the Gaze-Tracker software which records the eye movement data together with the scene data as digital data files.

In figure 7-5 examples of the experimental set up for all three viewing conditions are shown. The left column 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 column illustrates the standard LCD monitor task and an associated eye movement record; the right column 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 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. In the monitor viewing task interaction with the DICOM viewing software was by mouse. Interaction with the GE workstations was by means of the standard GE workstation interaction keyboard. The viewing distance was 55-65 cm depending on each individual in each round.

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 (Normal, Benign, Probably Benign, Indeterminate, Probably Malignant or 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 the individual.

The performance of each participant 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 recent real life performance data from everyday clinical screening.

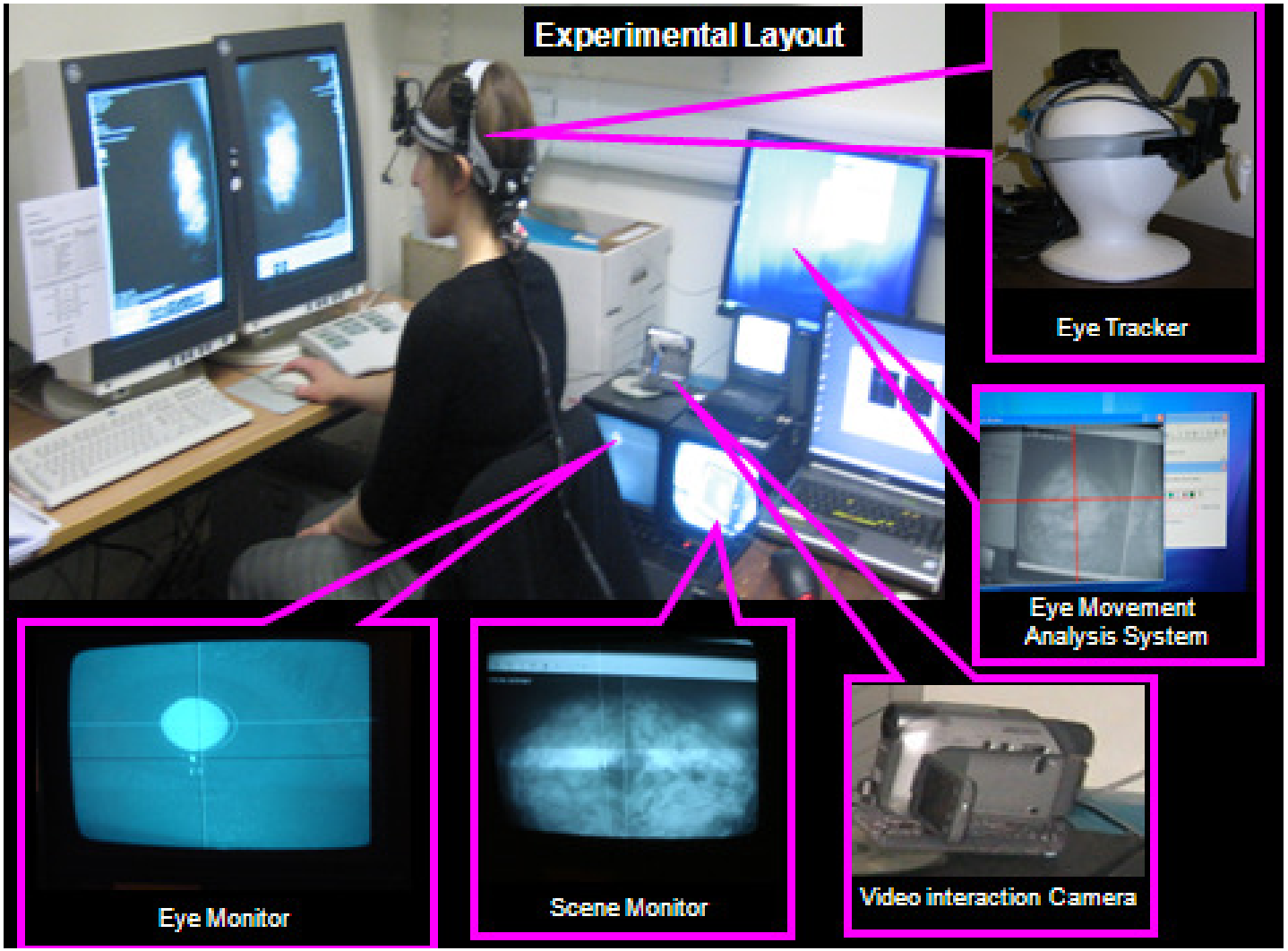


Figure 7-4. Example of the experimental set-up on the workstation

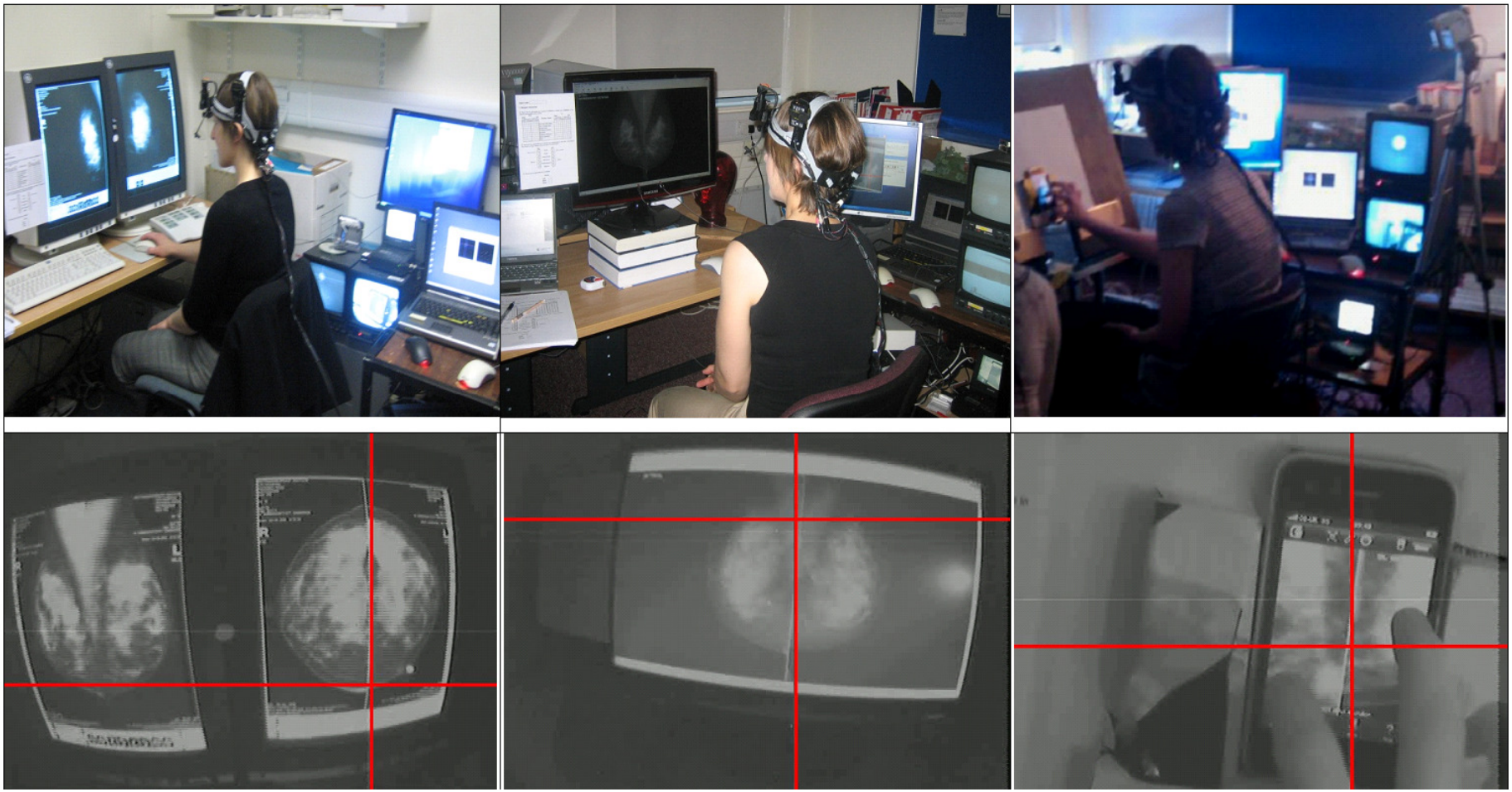


Figure 7-5. Examples of participants examining images on the three different displays. The ambient lighting levels were altered for photographic purposes.

7.3 Study Results

7.3.1 Performance Results

Participants' performance was compared on both levels of digital mammographic image 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 [$F(1, 12)=5.803$, $p<.05$] 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 7-6 below).

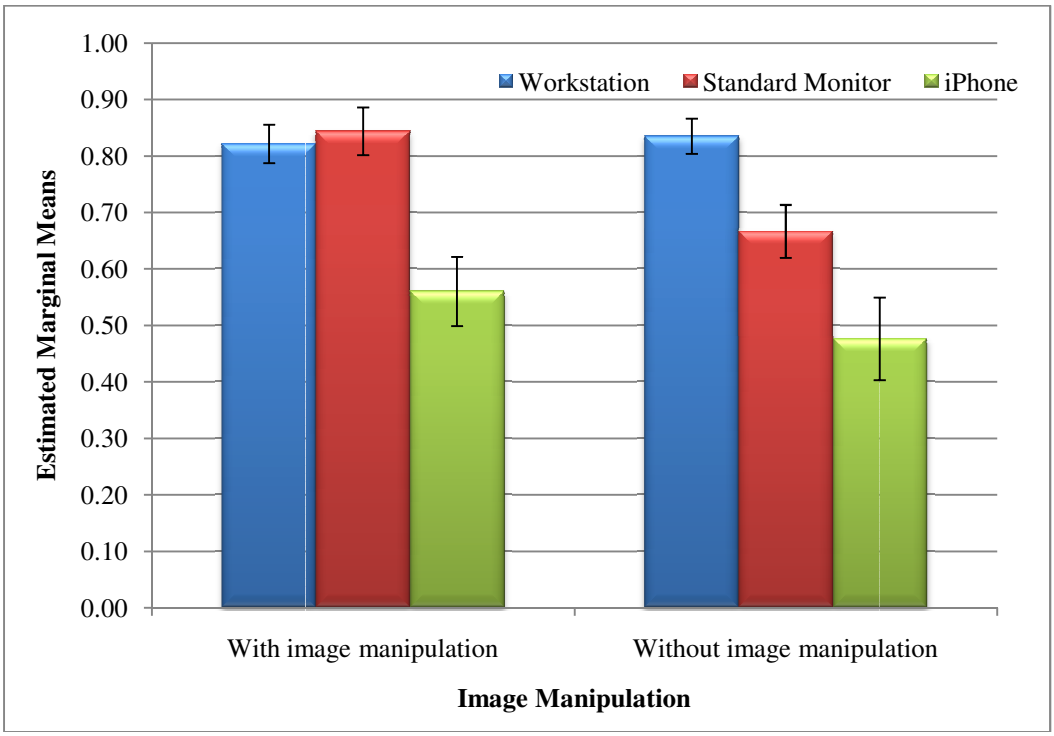


Figure 7-6. Performance on the three modalities with and without HCI

7.3.1.1 Experience

Participants were separated into two groups: those with over 10 years experience and those with less than 10 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 measure 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 7-7) 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 7-8) were 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, r=.43] whereby those in the more experienced group scored significantly higher (~ 93.9%) than those in the less experienced group (~ 86.4%).

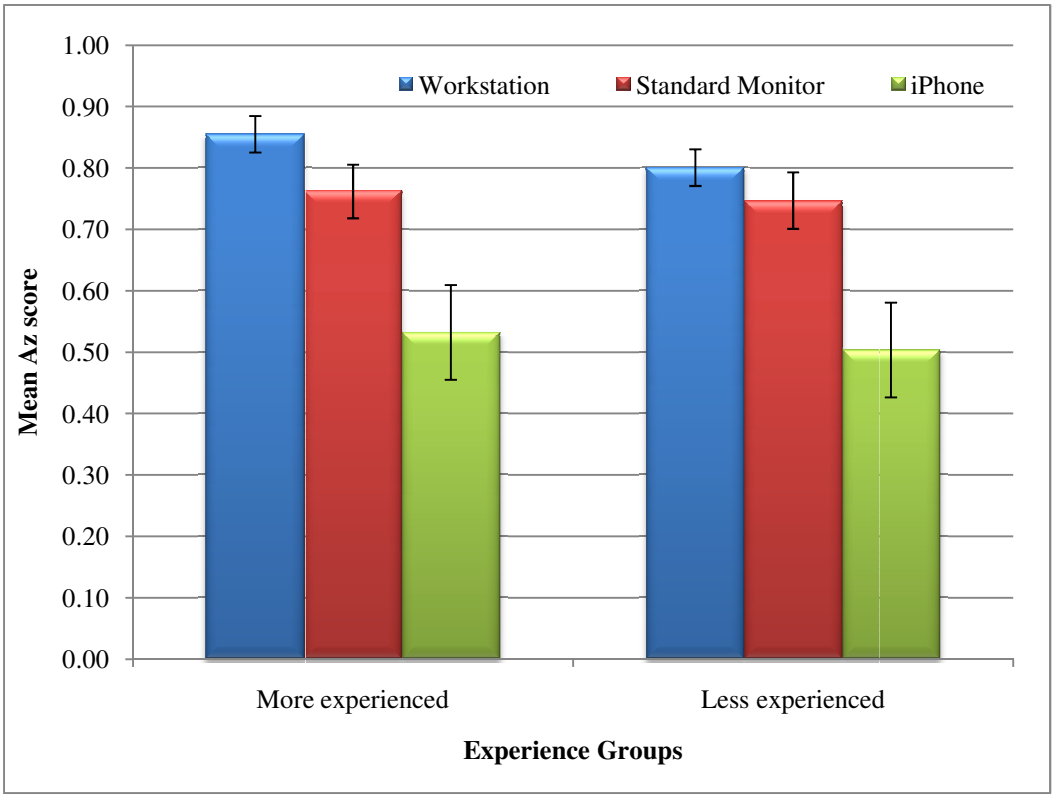


Figure 7-7. Experience Groups by Modality Type

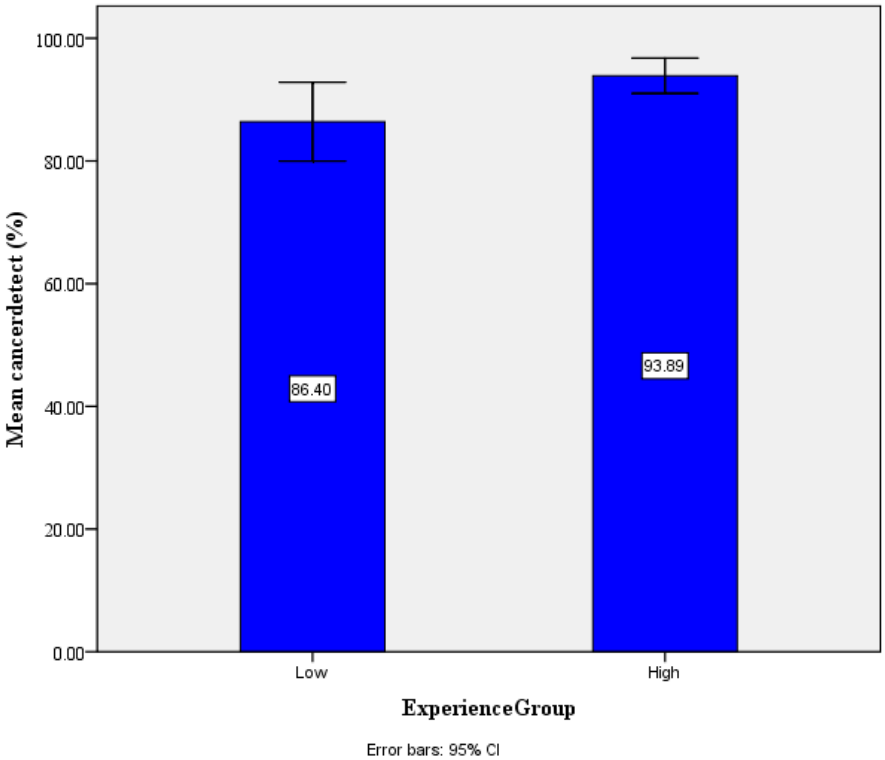


Figure 7-8. Mean cancer detection for the low and high experience groups on the PERFORMS scheme

7.3.1.2 Modalities

Participants' overall performance data were firstly examined by performing a JAFROC analysis for each individual. However, it was difficult to perform this analysis for some individual's data because they had a lack of false positive responses. Consequently, to perform the JAFROC analysis, two participants' data sets were dropped. A repeated measures ANOVA was then carried out which revealed a significant main effect of modality [$F(2, 20)=27.489$, $p<.001$, $r=.76$]. 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 Figure-of-Merit (FOM).

Furthermore, each individual's overall data were pooled together by condition groups in order to include all 14 participants' data. Figure 7-9 illustrates the empirical ROC curve for each modality. JAFROC analysis showed that the mean (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.

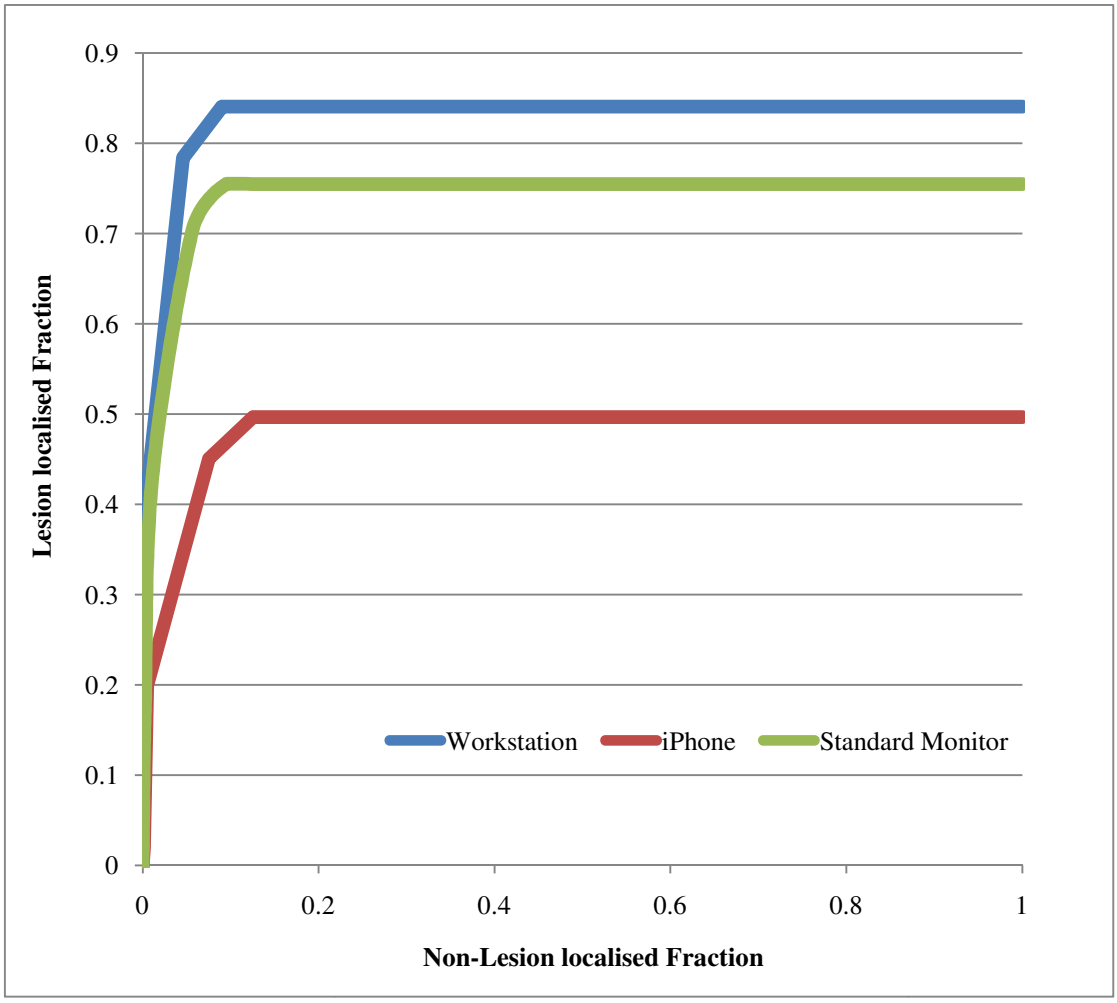


Figure 7-9. Empirical curve of performance in digital workstation, standard LCD monitor, and an iPhone

7.3.1.3 Image Manipulation

As mentioned above, there was a significant difference found overall between performance with/without the support of image manipulation $F(1, 12)=5.803$, $p<.05$, $r = .57$. When the data were analysed by considering modality/image manipulation, further post-hoc analysis (t-tests) elicited that there was little difference (*p=n.s.*) 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 (*p=n.s.*) 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 7-10. The empirical ROC curve was plotted below (figures 7-10) and Figure-of-merit (FOM) was calculated using the trapezoidal rule. Figure 7-9 shows the mean value of Figure-of-merit (FOM) on the workstation, standard LCD monitor and iPhone with/without HCI (non-HCI).

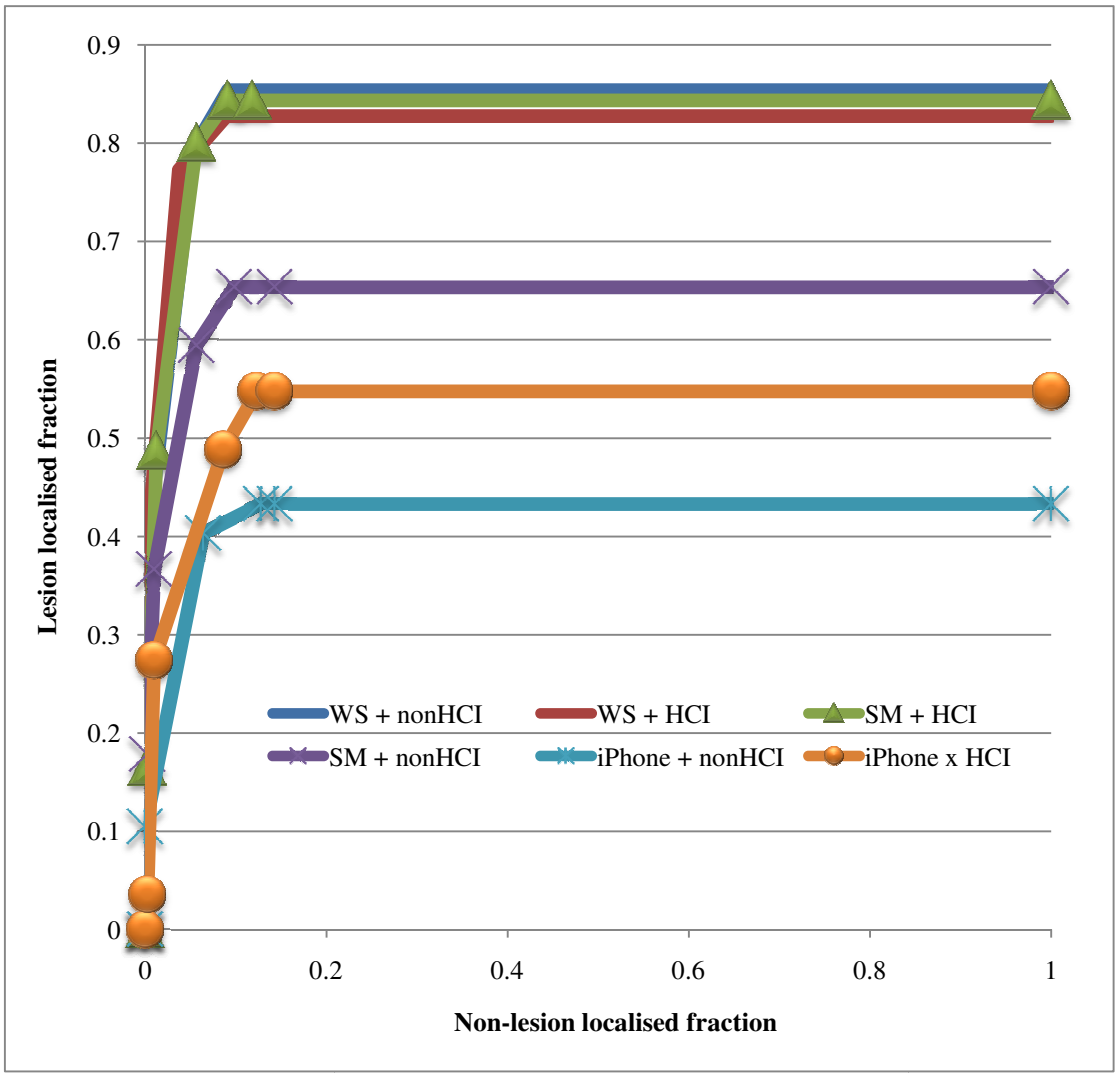


Figure 7-10. FROC curves of performance on workstation (W/S), monitor (SM) and iPhone with/without HCI (nonHCI).

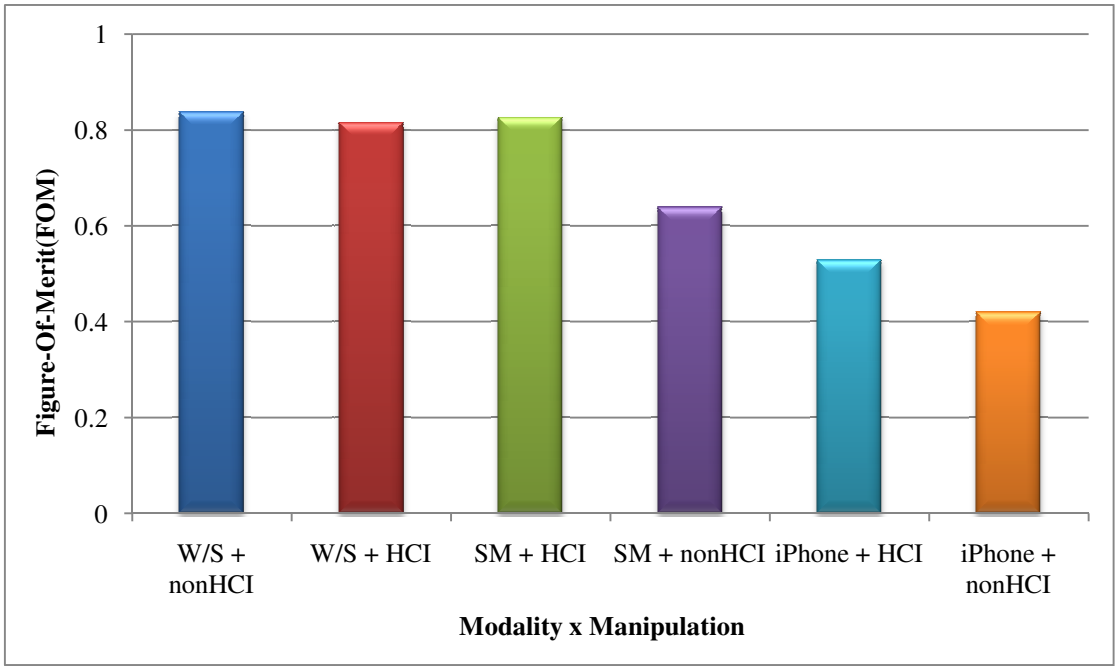


Figure 7-11. Figure-of-merit (FOM) on workstation (W/S), monitor (SM) and iPhone with/without HCI (nonHCI).

A two-way repeated measure ANOVA was performed to compare the performance data between the different feature groups (namely: Mass - Architectural Distortion [AD], ill-defined Masses, and spiculate mass; Calcification and Normal - cases containing no mammography features) in the three viewing conditions (see figure 7-12 for details).

Mauchly's test indicated that the assumption of sphericity had been violated for the main effects of viewing conditions, $\chi^2(2) = 8.74$, $p < .05$. Therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε=.671). There was a significant main effect of feature type on performance $F(2, 28)=8.226$, $p<.01$. Contrast revealed that performance on Mass, $F(1, 14)=20.61$, $p<.05$, $r = .77$, and normal, $F(1, 14)=7.00$, $p<.05$, $r = .58$, were significantly better than calcification cases.

Interaction contrast also revealed that when comparing the normal cases with the calcification cases both for the workstation with HCI to the standard monitor with HCI, $F(1, 14)=7.30$, $p<.05$, $r = .58$, and workstation without HCI to

standard monitor with HCI, $F(1, 14)=5.65$, $p<.05$, $r = .53$. These effects reflect that performance on normal cases (as compared to performance on calcification cases) was lowered significantly less as compared to examining images on the workstation with HCI than examining images on the standard monitor with HCI. This was also true when examining images on the workstation without HCI was compared to examining images on the standard monitor with HCI. The remaining contrasts revealed no significant interaction term when comparing mass cases with calcification cases both for workstation with HCI to standard monitor with HCI, $F(1, 14)=2.04$, $p=n.s.$, $r = .35$, and workstation without HCI to standard monitor with HCI, $F(1, 14)<1$, $p=n.s.$, $r = .23$.

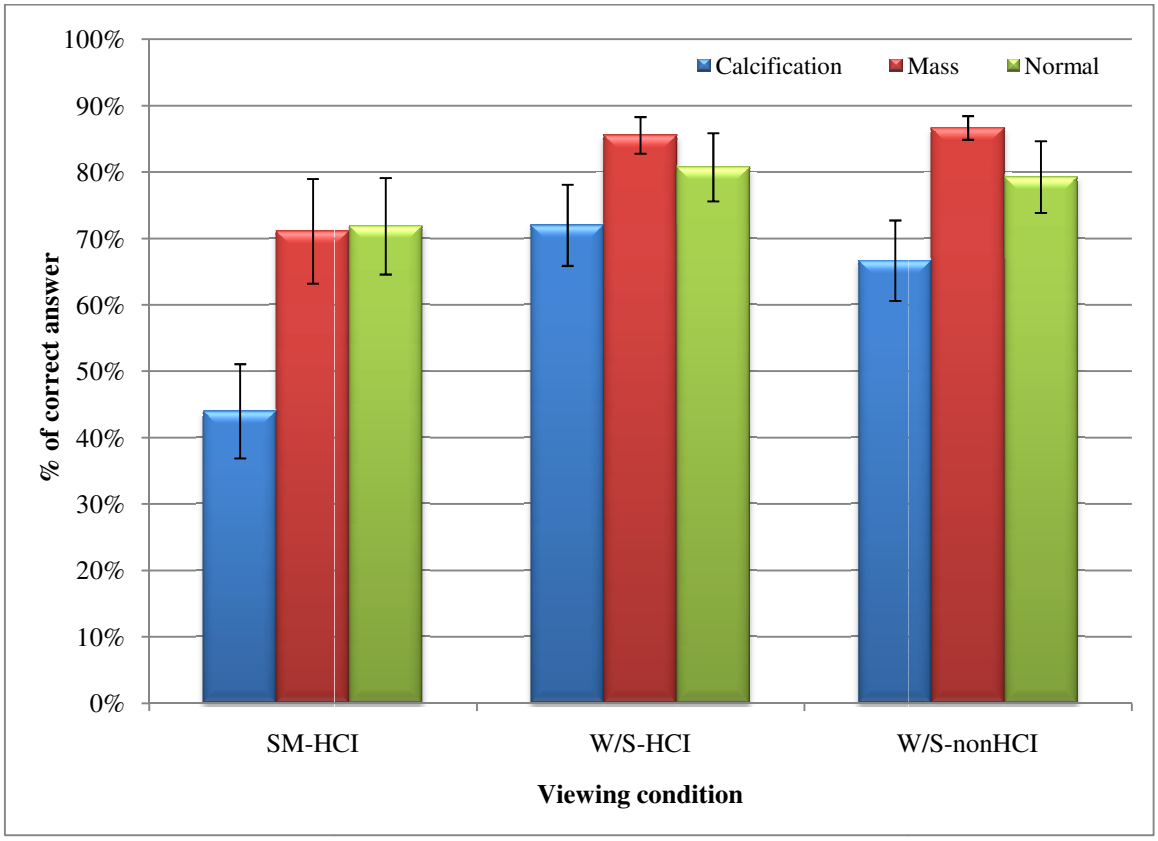


Figure 7-12. Performance in correctly identifying key features, with and without HCI usage on the monitor and workstation (w/s).

7.3.2 Image Viewing Behaviour Analysis

To thoroughly examine the participants' imaging interaction behaviour and relate this to their performance a detailed examination was carried out of the sequence of image manipulation data as recorded by the video interaction camera.

Therefore, according to these performance results, iPhone supported cancer detection performance was very poor; therefore, it is suggested that this should not to be used as a training delivery tool – at least as examined here running this version of the Osirix software. However, the standard monitor was shown to be a potential alternative for the digital mammography workstation due to its good performance. In view of this, it is important to analyse and compare the users' interactive screen behaviour while examining images on the digital workstation and the standard monitor.

7.3.2.1 Pilot Data Analysis

Firstly, some representative sections of videotapes were selected as pilot data which were then reviewed and discussed with other two colleagues to decide: (1) what recorded participant information was necessary to be coded; (2) how to code such behaviour efficiently; and (3) what software was required to facilitate these data analyses. These sections were then converted into .avi computer data files.

Considering that the main purpose of the study was to investigate screeners' imaging interpretation behaviour and how this related to his/her cancer detection performance then it was essential to record any relevant image examination related events. To analyse the behaviour data efficiently, there were two analysis software suites which were considered: 1) Captiv L-2100 software from the TEA group; 2) Observer XT software. Given the comparable functionality and level of easy-to-use, Captiv L-2100 was chosen (see Chapter 2 for further details).

Using this software to analyse the pilot data it became evident that it was necessary to record ten types of actions (i.e. 'events' – see figure 7-13) in order to fully identify all of the observers' interaction behaviour. These were grouped into three different classes of events, as follows:

1. *General behaviour events (shown in figure 7-13 in Pink):*
2. *Image manipulation events (shown in figure 7-13 in Blue):*
3. *Mammographic Image View events (shown in figure 7-13 in Purple).*

These were entered into the software so that when the recorded data were replayed, then the computer screen looked like figure 7-14.

The software allowed the computer data file to be replayed in real time, or slowly, stopped, rewound, fast forwarded etc. When an action occurred on the recording then the appropriate data logging key was simply 'pressed' (using the mouse). Each of these actions was recorded when and if it was performed by a participant and which modified its current state. Replaying the data file also replayed the recorded audio during the experiment which enabled the experimenter to identify when the participants began to verbalise their responses to each image.

Codes	*Actions/Events*	*Meaning of the Codes*
B	**Session begins**	*The exact time when the whole digital mammography case appears*
E	**Session ends**	*The exact time when the whole digital mammography case disappears*
G	**Giving answers**	*The exact time when the participant starts to give answers*
WL	**Windowing on the Left**	*The exact time when the participant starts to change the window/level on the left images (MLO or CC)*
WR	**Windowing on the Right**	*The exact time when the participant starts to change the window/level on the left images (MLO or CC)*
ZPL	**ZP on the Left**	*The exact time when the participant starts to zoom/pan on the left images (MLO or CC)*
ZPR	**ZP on the Right**	*The exact time when the participant starts to zoom/pan on the right images (MLO or CC)*
O	**OFF**	*On image manipulation behaviour (i.e. the time participants spent on visually examining the image while no actual action)*
MLO	**MLO**	*The exact time when the MLO view appears on the screen*
CC	**CC**	*The exact time when the CC view appears on the screen*

Figure 7-13. Types of actions which were recorded in the behaviour analysis and the corresponding codes.

7.3.2.2 Behaviour Data Analysis

Detailed examination of the sequence of the image manipulation data was then carried out. For each participant, their behavioural data for each study round was recorded on two separate video-tapes; one for the case set viewed with the support of the image manipulation tools and the other for the case set viewed without the support of such tools. Each tape was transferred to computer disk and saved as an *.avi* file for to enable data analysis. Each of the 84 *.avi* files (comprising over 100 hours of recorded information) was then reviewed and coded using the Captiv L-2100 software. Figure 7-15 illustrates the video analysis process.

Also, to check on any potential subjective interpretation of the video records, one video recording was first analysed using the Captiv software and then the same video was reviewed again independently by another viewer using the software. The results showed no individual differences in identifying the time sequencing of the actions/events and consequently it was decided that all of the tapes would be analysed by one individual.

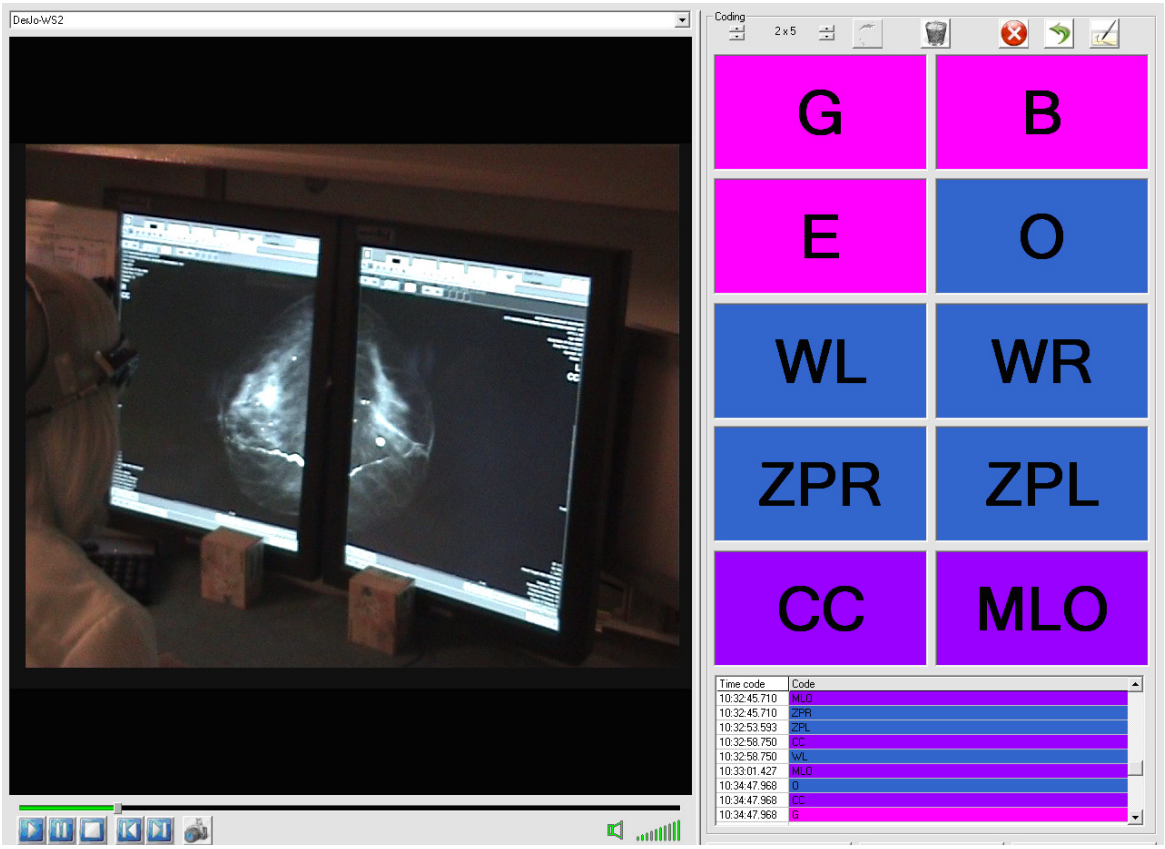


Figure 7-14. Screen shot of analysis of the imaging interpretation behaviour data using Captive L-2100 software.

7.3.2.3 Behaviour Data Analysis Results

Figure 7-15 shows an example of the output from the software of the behavioural sequence for one person examining 20 cases using image manipulation. The figure shows the sequence of overall viewing (pink plot – for each case: when they started and ended viewing and began giving their verbal reports), image manipulation (blue plot – no image manipulation, or windowing or zoom/pan [zp]), and the particular mammographic view (purple – MLO or CC) are shown plotted against time on the *X* axis.

Data were examined across all participants with regard to the average examination time spent on each. T-tests showed non-significant differences (*p=n.s.*) between the time spent examining cases on the standard monitor,

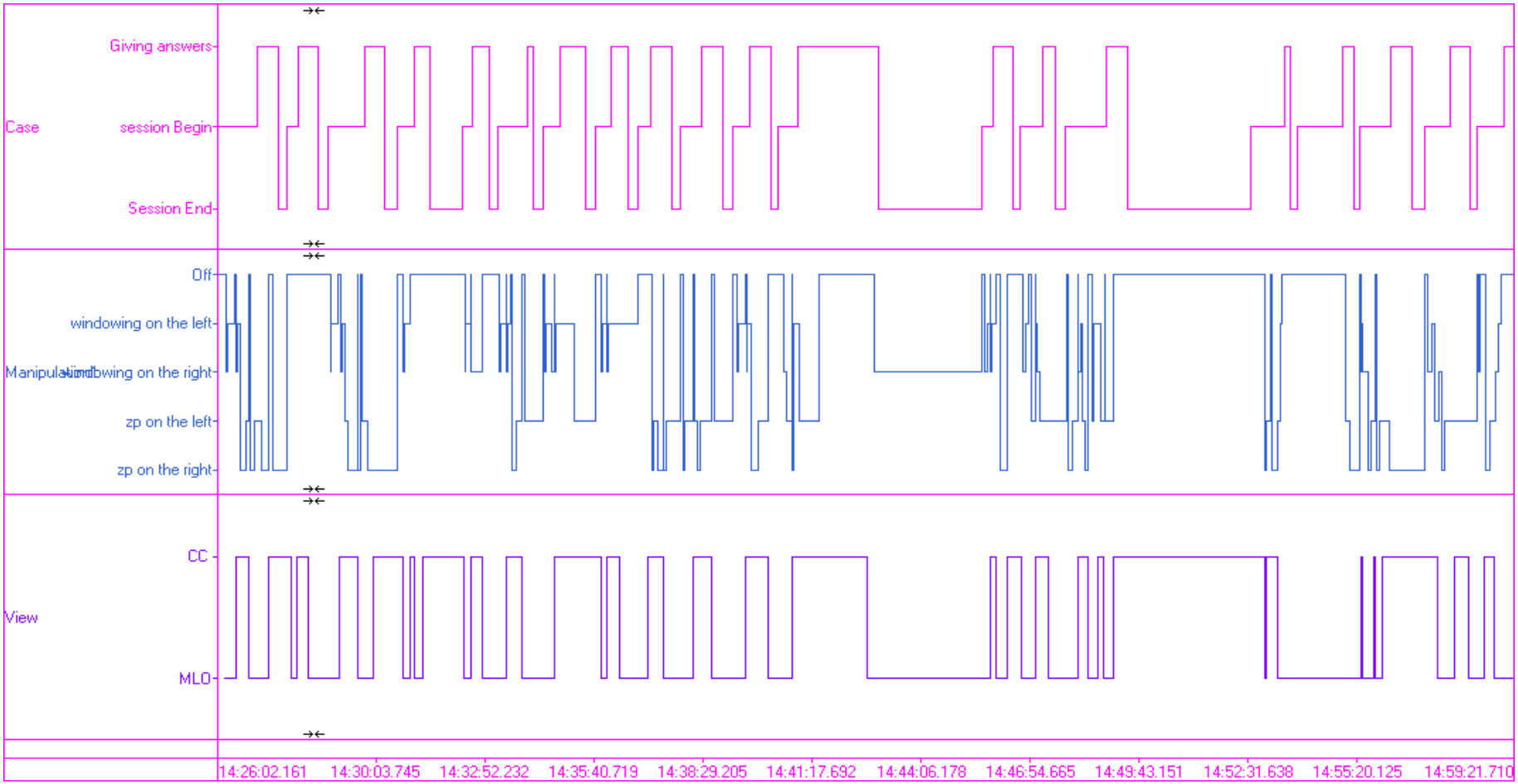


Figure 7-15. Example of an individual's behaviour data result using the Captive L-2100 software.

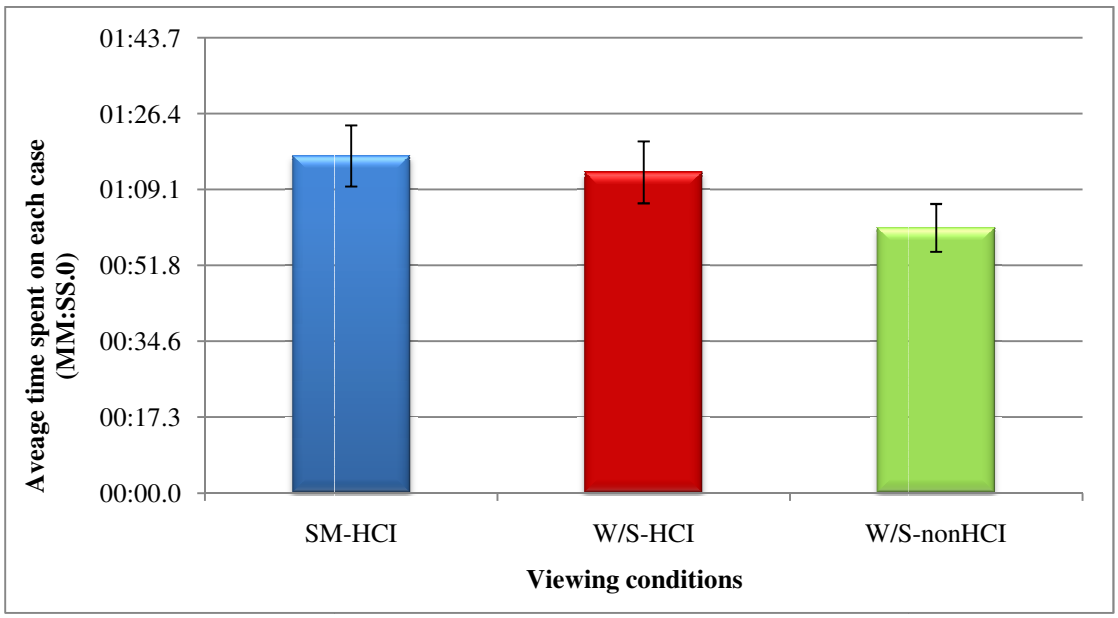


Figure 7-16. Average time spent per case comparison between the three viewing conditions

using image manipulation tools, and on the workstation using, or not using, the image manipulation tools. See figure 7-16 for details.

A 2 x 3 mixed design ANOVA revealed that there was a non-significant main effect of mammography experience on the average time spent on each case [$F(1, 10)=4.17$, $p=n.s.$, $r = .54$]; a significant main effect of viewing conditions on the average time spent on each case, [$F(2, 20)=3.9$, $p < .05$, $r = .40$]. Contrast revealed that the average time spent on each case while examining on the standard monitor with image manipulation tools was significantly longer than if examining on the workstation without using the image manipulation tools [$F(2, 20)=6.39$, $p < .05$, $r = .49$].

There was a significant interaction effect between viewing conditions and participant's mammography experience level, [$F(2, 20)=3.9$, $p < .05$, $r = .40$]. This indicates that the average time on each case in the different viewing conditions differed significantly between the high experienced readers and low experienced readers. To break down the interaction, contrast was performed

and revealed significant interactions when comparing high and low experienced observers' average time on each case:

- while examining images on the standard monitor with image manipulation tools as compared to examining images on the workstation without image manipulation tools, [$F(1, 10)=7.84$, $p < .05$, $r = .66$]
- and whilst examining images on the workstation with image manipulation tools compared to examining images on workstation without image manipulation tools [$F(1, 10)=14.937$, $p < .05$, $r = .77$]. Details are shown in figure 7-17.

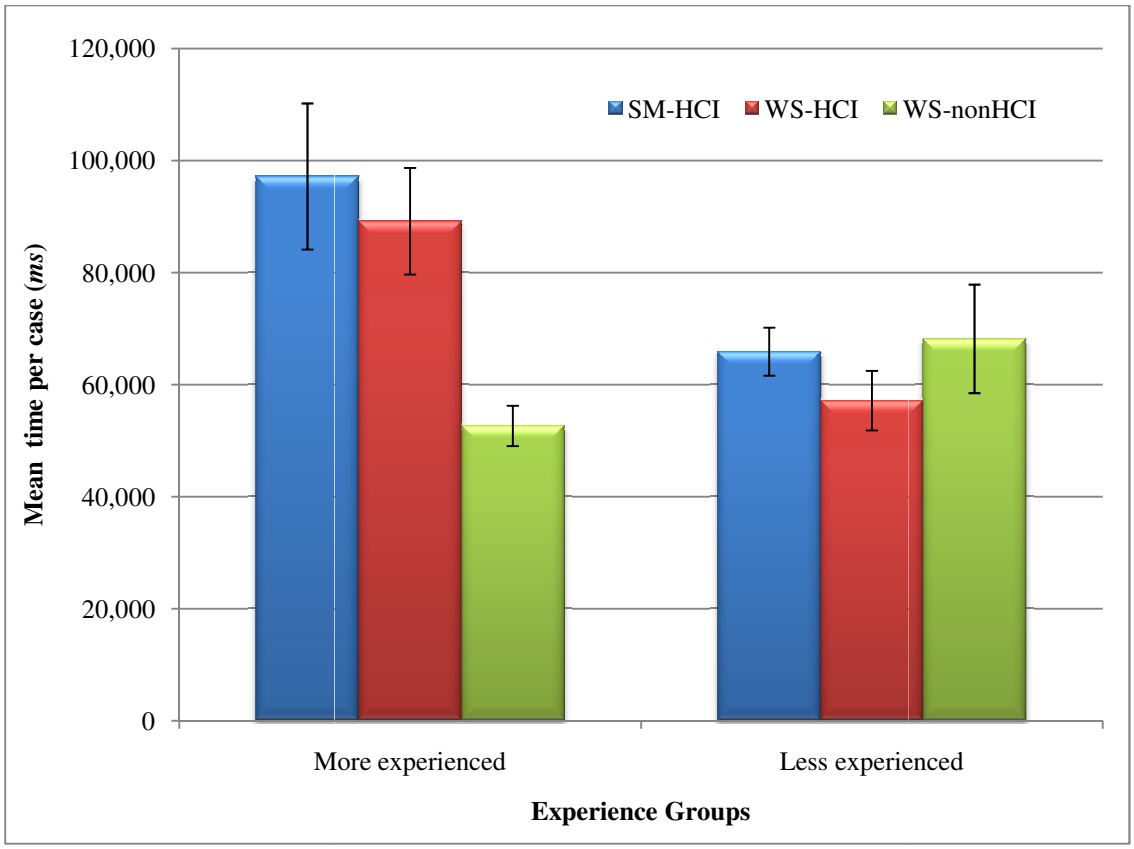


Figure 7-17. Average time spent per case comparison between mammography experience groups x three viewing conditions

Performance vs Image manipulation

Multiple regressions were performed to investigate how two factors; participant's mammography experience (years) and portion of time spent using image manipulation (%), affect their performance accuracy (A_z) on the diagnostic workstation and the standard monitor respectively.

On the Diagnostic Workstation

Workstation image manipulations were not necessarily a pre-requisite for reporting these screening cases appropriately. However, these tools did affect reporting confidence and mammographic case visual examination. The normal clinical digital display has sufficient resolution to visualise most early abnormal signs adequately for detection without using such tools. However, using them improved calcification detection, although this was at the cost of potentially increased recalls. It must be noted that the effect sizes here are small; reflecting the number of participants and further work is required to expand upon this.

Table 7-1. Multiple regressions reporting table

	B	***SE B***	***Beta***
Step 1			
Constant	.948	.016	
yearOfexperience	.002	.002	.370*
Step 2			
Constant	.997	.025	
yearOfexperience	.002	.001	.277*
manipulation time%	-.080	.034	-.583**

Note: R^2 =.137 for Step 1; ΔR^2 =.332 for Step 2. * *p=n.s*; ** *p*<.05

Data indicated that the percentage of time spent on image manipulation exhibited a significant negative relationship to cancer detection performance (R=-.628, *p*<.05). However, there was no significant relationship between participants' years of mammography experience and the percentage of time

they spent on image manipulation (R=-.159, *p=n.s.*) while examining images on the workstation. Table 7-1 showed that participants' years of mammography experience accounted for 13.7% (R^2) of the variation in performance accuracy. The percentage of time spent on image manipulation accounted for an additional 33.2% (ΔR^2) of the variation in cancer detection performance. The relationships between performance, manipulation time (%) and years of mammography experience are shown in figure 7-18.

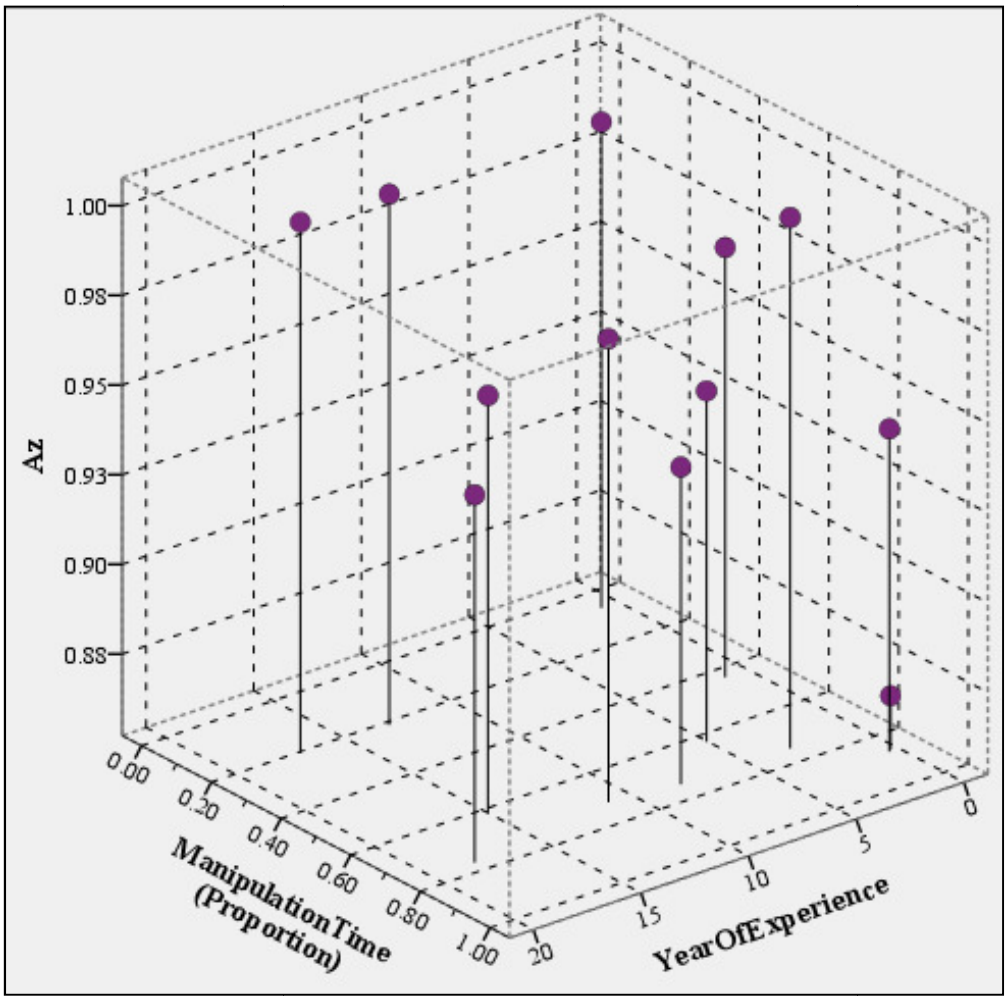


Figure 7-18. 3-D diagram showing the relationship amongst years of experience, % of time spent on manipulation and image interpretation performance on the workstation (A_z)

On the Standard Monitor

Table 7-2. Multiple regressions reporting table

	B	***SE B***	***Beta***
Step 1			
Constant	.712	.124	
Manipulation time%	.227	.218	.313*
Step 2			
Constant	.619	.122	
yearOfexperience	.261	.196	.360*
Manipulation time%	-.010	.005	.499**

Note: R^2 =.098 for Step 1; ΔR^2 =.247 for Step 2. * *p=n.s*; ** *p=n.s*

Data indicated that the percentage of time spent on image manipulation exhibited a non-significant negative relationship to cancer detection performance (R=.313, *p=n.s.*). However, there was no significant relationship between participants' years of mammography experience and the percentage of time they spent on image manipulation (R=-.466, *p=n.s.*) while examining images on the standard monitor. Table 7-2 showed that participants' Years of mammography experience accounted for 9.8% (R^2) of the variation in cancer detection performance. Percentage of time spent on image manipulation accounted for an additional 24.7% (ΔR^2) of the variation in performance accuracy on the standard monitor. The relationships between performance, manipulation time (%) and years of mammography experience are shown in figure 7-19.

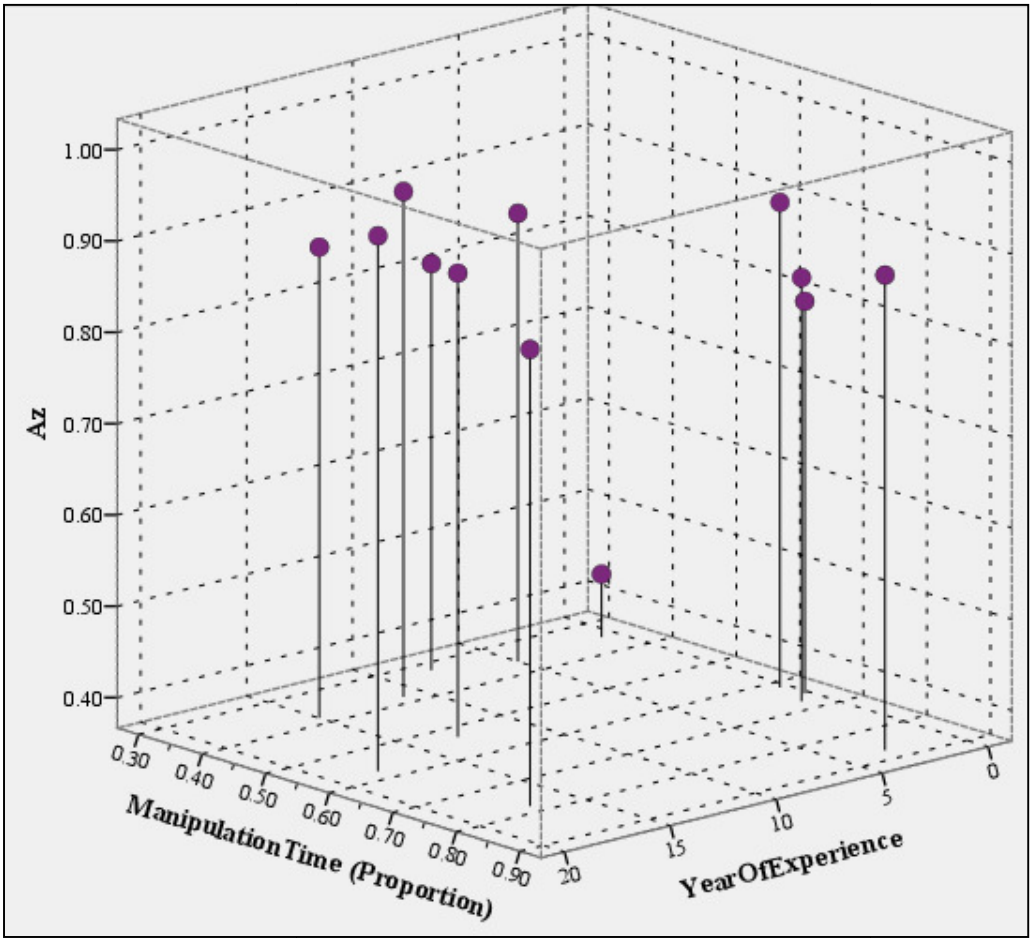


Figure 7-19. 3-D diagram showing the relationship amongst years of experience, % of time spent on manipulation and image interpretation performance on the standard monitor (A_z)

7.3.3 Visual Search Data Analysis

Participants' eye movements were recorded throughout all three parts of the study. An initial simple examination of these data records showed that, as would be expected, the eye movement data on the iPhone was not robust enough for analysis. Whilst it was possible to note that observers' fixation locations were located on the iPhone where one would expect (e.g. on button selection when making a selection) it really was too small a display, given the recording accuracy of the eye movement recording technique, to analyse with any accuracy. This was expected to be the case before the experiment started but the recording of eye movements in this part of the investigation

was still performed so that participants performed all three sections of the study under the same experimental conditions.

A key part of the interaction with the images was how individuals visually examined and interacted with the cases when examining images on the diagnostic workstation and the standard monitor when the image manipulation tools were used. 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.

The raw eye movement data was clustered into fixations using the criteria that have been discussed in the methods chapter (section 2.1.4.1). Fixation data on all the false negative cases made while the participants were examining cases on the workstation while using image manipulation and also examining cases on the standard monitor again while using image manipulation. Data were used to classify the false negative responses into the following three categories:

Search error: neither the location of the abnormality, nor the area of interest, attracted any visual attention;
Detection error: the location of the abnormality was not reported and the area of interest only attracted visual attention for a short *(<1000ms*) period of time;
Interpretation error: the location of the abnormality was not reported even though the area of interest attracted visual attention for longer than 1000ms;

Other information was also collected and analysed, such as:

- 'Time to first hit', which is how long and also how many fixations it took from when the image appeared to first hit the area-of-interest (AOI);

- 'Dwell time' corresponding to the total amount of time spent in the area-of-interest.

7.3.3.1 Eye Movement Data Analysis Results

A two-way repeated-measures ANOVA revealed that there was a non-significant effect of viewing modality on false negative error percentage, $F(1, 14) = 3.5$, p= *n.s.*, $r = .13$; there was also a non-significant effect of visual error classification on false negative error percentage, $F(2, 28) = 1.05$, p= *n.s*, $r = .19$.

Visual search error was the main false negative error (48% of false negative errors were search errors on the standard monitor; 31% of false negative errors were search errors on the digital workstation). The percentages of detection and interpretation errors were similar (32.2% of detection error and 19.6% of interpretation error on the standard monitor; 15% of detection error and 33.3% of interpretation error on the workstation). There was a non-significant interaction effect between the types of viewing modality and the types of visual errors, $F(2, 28) = 1.11$, p= *n.s*, $r = .19$. The percentages of each false negative error type on the two different viewing modalities are shown in figures 7-20 & figure 7-21.

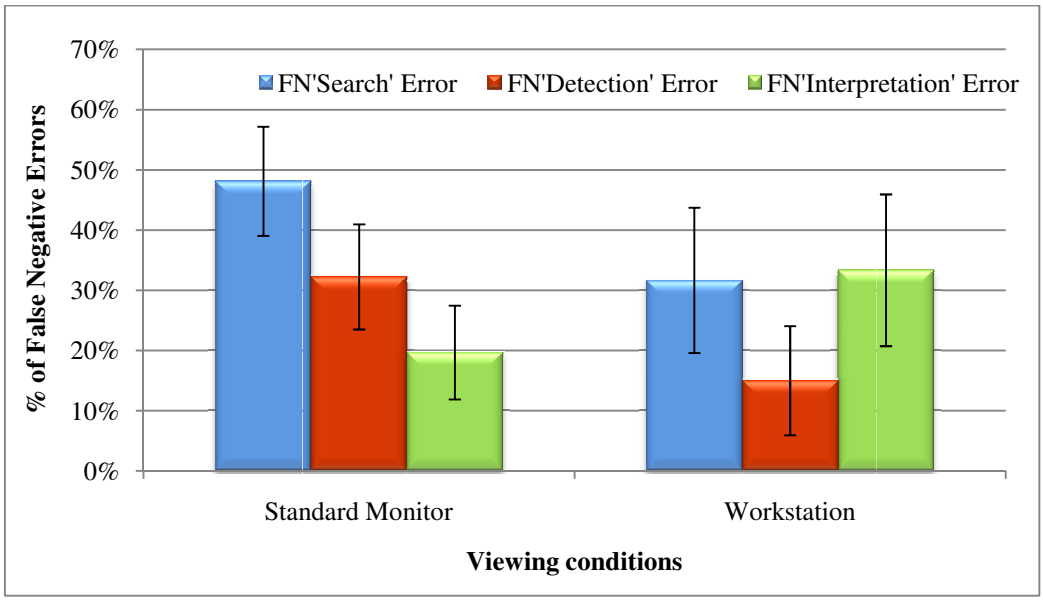


Figure 7-20. Visual errors x viewing conditions

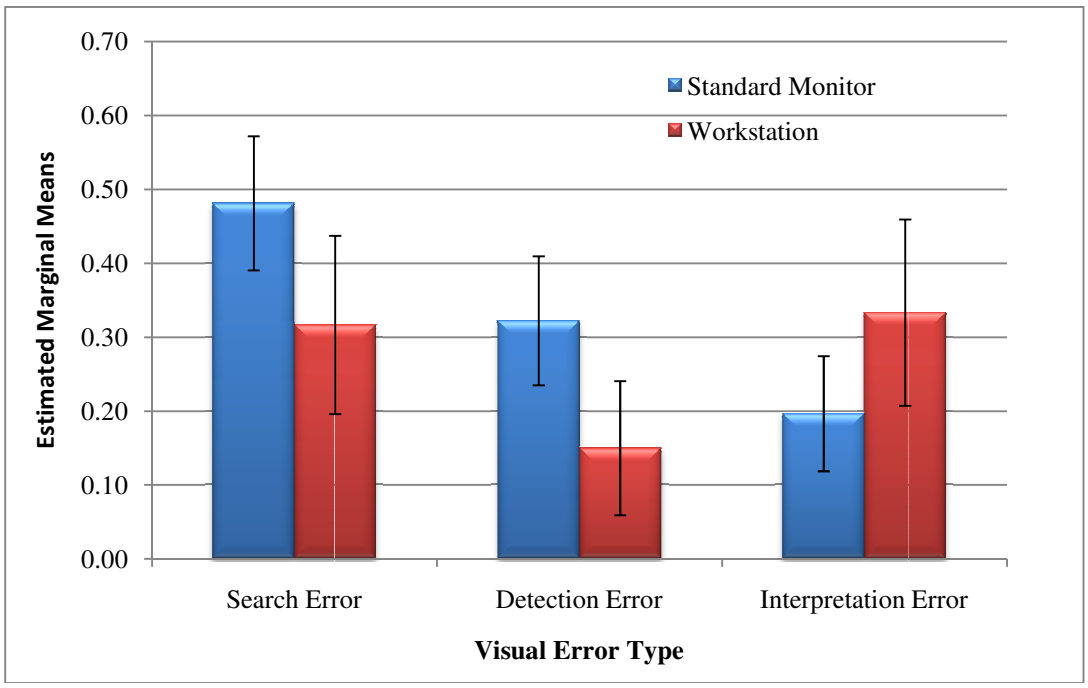


Figure 7-21. Visual errors x viewing conditions

On average, the number of times (approximately 4-5 times) that participants' fixated within the AOI while examining images on a standard monitor (M = 4.6, SE = 1.5) was not significantly different from examining on the workstation (M = 5.2, SE = 3.1, $t(12)= -.162$, *p=n.s.*, $r = .05$). Similarly, on average, participants' eye dwell time within the AOI while examining images on the standard monitor (M = 4.82, SE=2.88) was not significantly different from examining on the workstation (M = 3.27, SE=1.86, $t(12)= .448$, *p=n.s.*, $r = .13$) – although in terms of mean values they 'dwelt' within the AOI for about 2s longer on the monitor than the workstation. Details are shown in figure 7-22 & figure 7-23.

Figure 7-22. Number of times participants fixated within the AOI while examining images on the standard monitor and the workstation

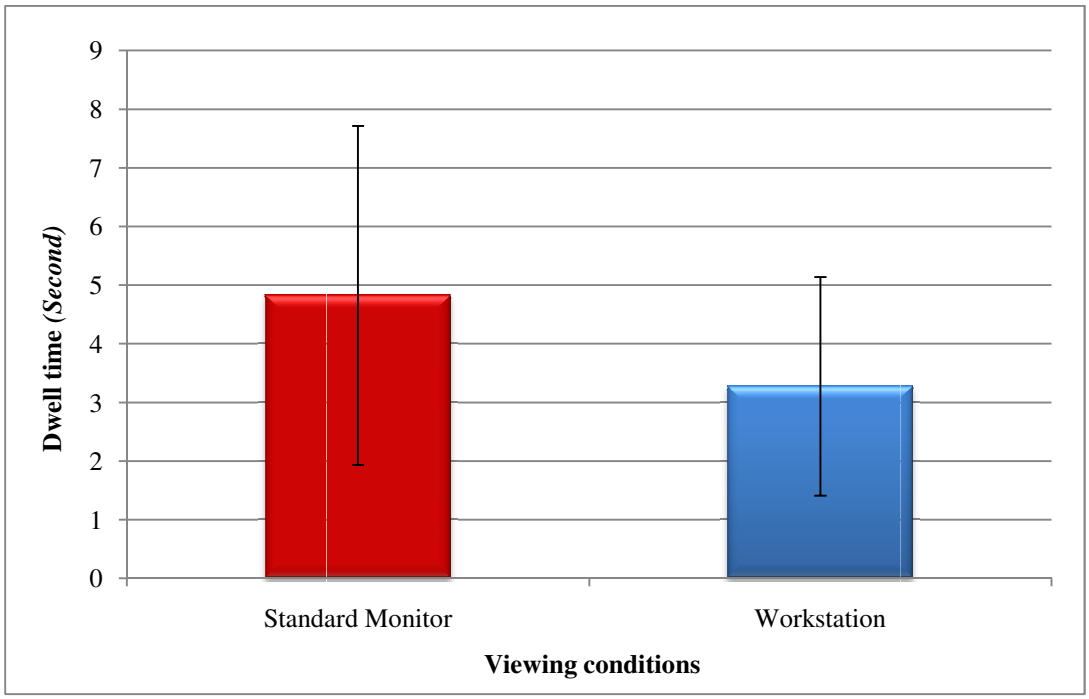


Figure 7-23. Dwell time comparison between examining images on the standard monitor and the workstation

On average, the number of fixations before the eye first fixated within the AOI while examining images on the standard monitor (M = 87.09, SE = 26.01) was not significantly different from examining on the workstation (M = 175.57, SE= 79.08, t(6)=-1.023, p=n.s., r=.39). Similarly, on average, the participants' average duration between image onset to first hit the AOI while examining images on the standard monitor (M = 38.46, SE = 13.20) was not significantly different from examining on the workstation (M = 95.33, SE = 59.4, t(6)=-.897, p=n.s., r=.34). Details are shown in figure 7-24 & 7-25.

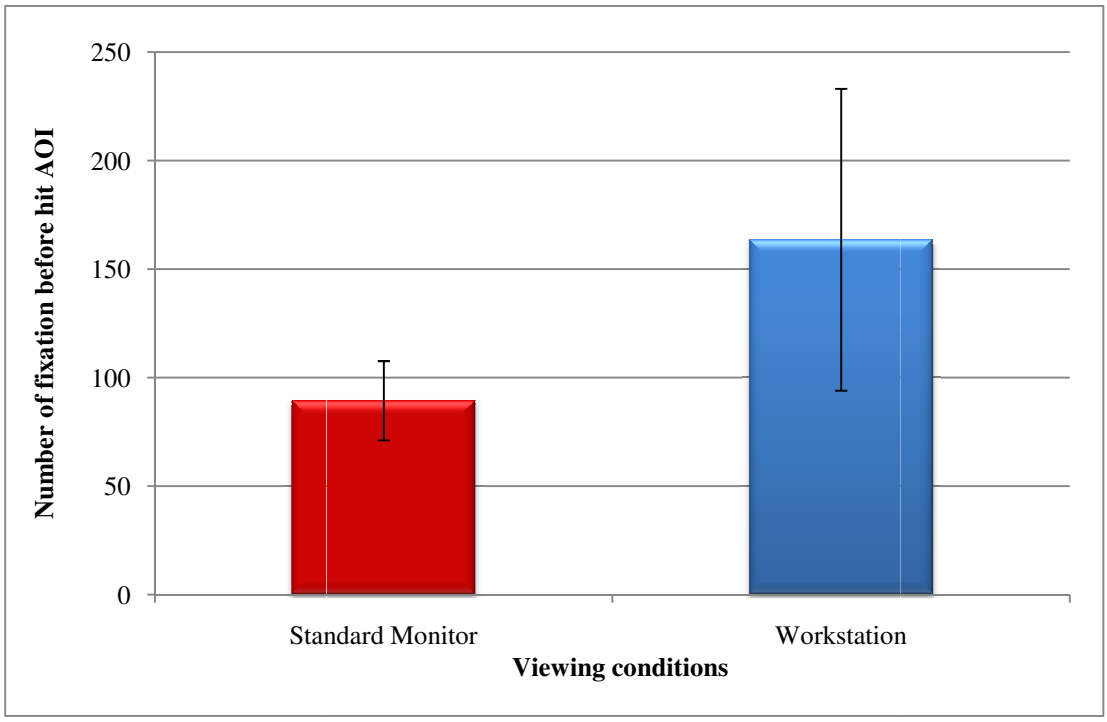


Figure 7-24. Comparison of number of fixations before the eye first fixated within the AOI while examining images on the standard monitor and workstation

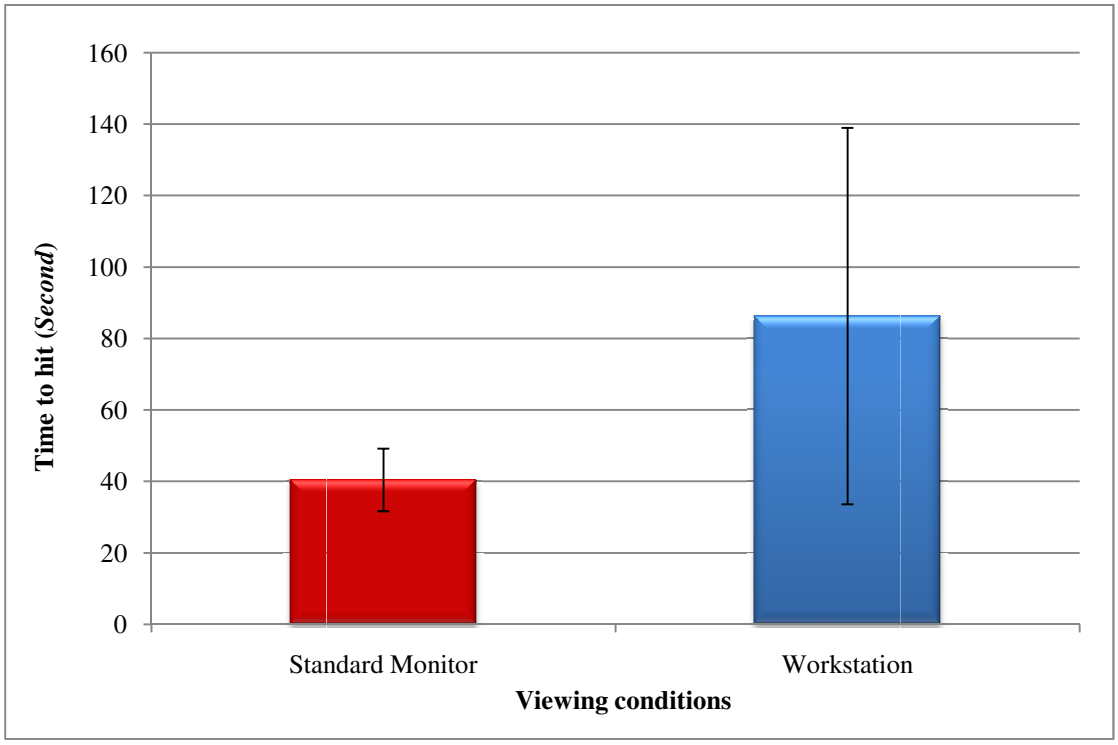


Figure 7-25. Comparison of time to hit between examining images on standard monitor and workstation

7.4 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 it is not proposed that monitors with less resolution, or physical size, than workstations should be used for clinical screening.

A key question 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 detailed 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 at any stage 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 as further elaborated by JAFROC analysis. 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 (*n.s.* difference in performance). This implies that using such a monitor with HCI would be useful for training purposes. Here, a readily available standard DICOM viewer was used. 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, examination of the two groups (>10 years and < than 10 years) data from the PERFORMS scheme showed that the more experienced group were significantly better at detecting cancer. However, in the experiment here no significant difference was found between the two experience groups. This may possible be due to the low number of participants or cases.

More regression analysis revealed that participants' years of mammography experience accounted for little (13.7% on the workstation, 9.8% on the standard monitor) of the variation in performance accuracy. However, the percentage of time spent on image manipulation accounted for more of the variation in cancer detection performance (33.2% on the workstation, 24.7% on the standard monitor). The percentage of time spent on image manipulation had a negative relationship with performance on both modalities, which was significant on the workstation but non-significant on the standard monitor.

To further investigate the potential of using the standard monitor to deliver mammography interpretation training, the participants' image manipulation behaviour on the workstation and the standard monitor was compared. The important finding was that the average time on each case in the different viewing conditions differed significantly between the high experienced readers and low experienced readers. Compared with the low experienced participants, high experienced ones spent less time on the standard monitor and workstation while the image manipulation tool was used. However, high experienced participants spent more time than the inexperienced participants

while the image manipulation tool was not used. This could be because the experienced readers became more careful and therefore spent a longer time examining images while the normal image manipulation tool was not available.

Visual data analysis revealed that the participants made a similar overall pattern of errors on both modalities. There was no significant difference between modalities in terms of visual error types. However, they made more search errors than detection errors on either modality although not significantly. On average, the number of times participants fixated within the AOI and participants' eye dwell time within the AOI while examining images on a standard monitor was not significantly different from examining on the workstation. Similarly, the number of fixations before the eye first fixated within the AOI and the average duration to first hit the AOI while examining images on the standard monitor was not significantly different from examining images on the workstation. These findings are somewhat surprising given the differences in monitor and workstation display sizes and resolutions.

This points to some similarity in how they inspected images on the two displays. Participants were overall faster on the monitor to hit the AOI and made fewer eye fixations before they hit the AOI. However, once within the AOI then they spent more time dwelling there on the monitor. This may well reflect that the abnormalities were able to be seen on either display type but that once identified then it took longer to examine in detail on the monitor.

7.5 Conclusion

Whilst superior performance was attained using the clinical workstations, participants were able to identify abnormal features on both the standard LCD monitor and the 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 Improved mammographic interpretation training software for the iPhone may render it more useful. It is argued that lower resolution displays are useful for training purposes only.

'Common sense' would imply that manipulating difficult images on a workstation should support better cancer detection performance. This did not prove to be the case for these particular cases, except for calcifications, where image manipulation had little overall effect. Accounting for this is quite complex. The more time spent manipulating images actually led to decreased cancer detection, with variation in performance related to both experience and image manipulation. It is suggested that individuals have their own particular way of working (some using a lot of image manipulation and others not) which has implications for training as well as general guidance on image examination.

A standard monitor with the support of image manipulation achieved similar performance as on the digital workstation. Visual data analysis on the false negative errors suggested that participants' visual behaviour on the standard monitor was very similar to their behaviour on the digital workstation. Also, their screen behaviour on these two modalities was shown to be comparable. Although there was some difference in identifying calcification, the standard monitor is considered to be a good alternative to deliver mammographic interpretation training, if image manipulation tools are made available and training is carefully planned.

CHAPTER 8

The Potential for Mammographic Interpretation Training in China

8.1 Introduction

The incidence of breast cancer is being addressed in many western countries by the implementation of breast screening programmes which are aiding in the early detection and suitable treatment of this disease. However, its incidence is increasing in other countries, where traditionally this disease has had a low occurrence, of which a good example is China which has a population of approximately 1.3 billion people. If widespread breast screening were to be undertaken in China then very large numbers of clinical staff would first have to be trained to interpret mammographic images to a consistently high standard and also ongoing training would need to be implemented. Using a range of low cost computing devices to help deliver some aspects of such training would be very beneficial, both financially and logistically. For instance in terms of the huge distances involved across China then their use in an e-learning system could prevent unnecessary travel to potential training centres.

8.1.1 Breast Cancer in China

Breast cancer is the most common type of cancer amongst women around the world (World Health Organization, 2006). China, as a country, has a fifth of the world's female population. However, compared with western countries, Chinese women have been typically considered to have a relatively low breast cancer disease incidence. For instance, Rong (2008) reported that, compared to the top-ranking USA, which has 101.1 breast cancers per 100,000 people of all races and ethnicities, the breast cancer incidence in China is only 18.7 per 100,000 people, ranking the country 142nd in the world for this disease.

In the past the incidence of breast cancer in the Far East has been low. In 2002 the World Health Organization (WHO) reported that China had a breast cancer incidence rate less than 0.02% which was the lowest breast cancer

incidence, and also mortality rate, when compared with fifteen other countries studied. For details see Chapter 1, figure 1-1.

However, over the past twenty years, the rate of breast cancer amongst urban Chinese women has increased sharply. The breast cancer death rate among urban Chinese women has increased 38.9% over the past 10 years. With 3% of disease incidence each year, China has become one of the countries with the fastest growth of breast cancer (Ma *et al.*, 2008). Dramatic rises in breast cancer incidence have been reported in some large cities with a possible sharp increase in the number of breast cancer cases being predicted if these trends spread to the rest of the country (Linos *et al.*, 2008; Ziegler *et al.*, 2008).

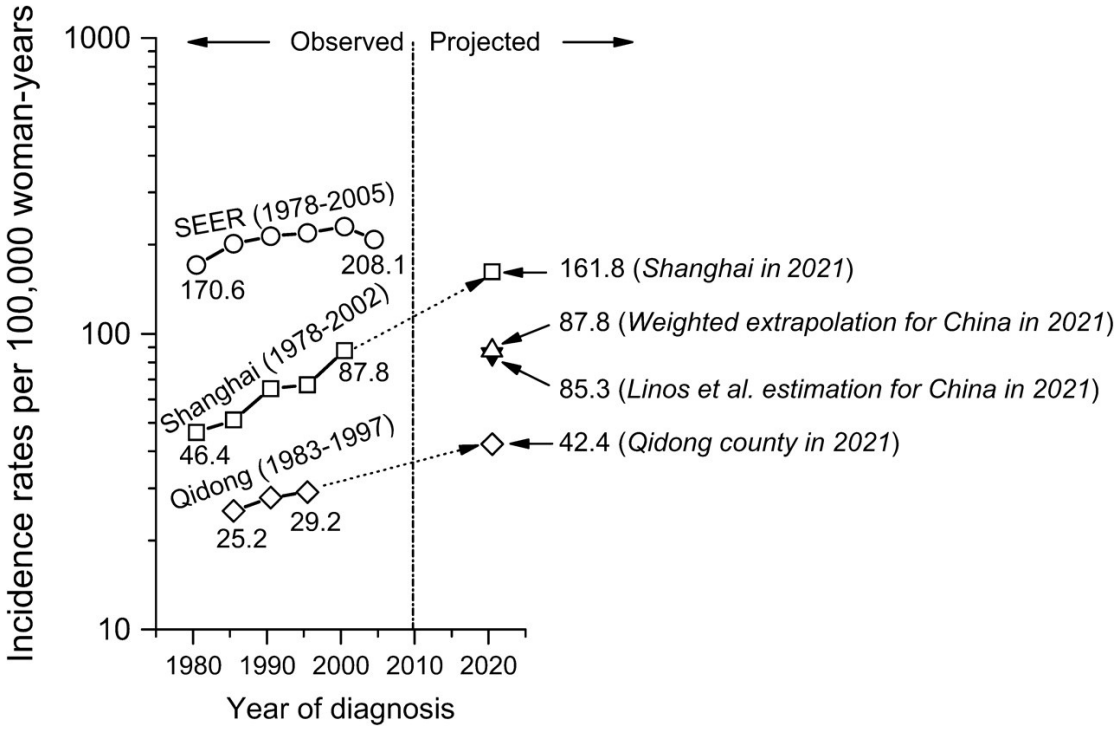


Figure 8-1. Age-standardized breast cancer incidence rates, based on age-specific rates for women aged 35–69 years and adjusted to the world standard population, were plotted on a logarithmic – linear scale by calendar year so that a slope of 10° represented a rate change of 1% per year (Devesa *et al.*, 1995). Surveillance, Epidemiology, and End Results (SEER) data (**open circles**) for breast cancer cases newly diagnosed from 1978 through 2005 were obtained from the National Cancer Institute's SEER database (SEER, 2008). Data for Shanghai from 1978 through 2002

(**open squares**) and Qidong County from 1983 through 1987 (**open diamonds)** were provided by the CANCER Mondial Statistical Information System (Parkin *et al*., 2005; Curado *et al*., 2008). Regression lines were fitted by weighted least squares to the data from Shanghai and Qidong County. The weights were $r^2/\mathrm{Var}(r)$, where r is the direct standardized rate for a given time period; these weights represent inverse variance weighting for the logarithm of the incidence rate. The lines used for extrapolation were 3.8120 + 0.0315(year _1980.5) for Shanghai and 3.1673 + 0.0143(year _ 1980.5) for Qidong County. By substituting 2021 for year, calculating the estimated logarithm of the rate, and exponentiating, we obtained extrapolated breast cancer incidence rates (per 100 000 woman-years) of 161.8 for Shanghai and 42.4 for Qidong. A weighted average of these results with weights for Shanghai and Qidong of 0.38 and 0.62, respectively, which correspond to Chinese National Bureau of Statistics estimates of the proportions of urban and rural populations in China in 2001 (*China Statistical Yearbook 2001, 2002*), yielded the weighted extrapolated estimate of 87.8 per 100 000 woman-years (**open triangle**). The standardized rate from the age-specific rates in table 2 of Linos *et al.* (2008) was 85.3 per 100 000 woman-years (**solid triangle**). (Ziegler *et al.*, 2008)

According to available figures, the death rate from breast cancer in China has been increasing by 3% annually in recent years. It has replaced lung cancer as the most rapidly growing cancer in the country. As the China Daily reported (2007), in China's commercial centre of Shanghai, 55 out of every 100,000 women have been reported to be diagnosed with breast cancer; this represents a 31% increase since 1997. Whilst this rate is still less than that of (for example) the UK where the similar rate is 0.08% it has been predicted (figure 8-2) that by 2021 extrapolated breast cancer incidence rates would be 161.8 for Shanghai and 42.4 for Qidong county per 100,000 woman-years (Ziegler *et al.*, 2008). "Unhealthy lifestyles are mostly to blame for the growing numbers," Professor Qiao Youlin of the Cancer Institute and Hospital of the Chinese Academy of Medical Sciences is quoted as pointing out. He also cites poor diets and environmental pollution along with increased living stress are the top provoking factors (China Daily, 2007).

8.1.2 Breast Cancer Screening in China

Consequently, the growing demand for potential breast screening coupled with a relatively limited capacity to provide such screening services are

placing a huge pressure on breast cancer detection within the Chinese health organization. To improve the situation, the China National Health Service and the China Cancer Research Institute co-operated with the US Cancer Research Institute and in April 2005 started the 'Million Women Breast Screening Project'. This was planned to involve 100 qualified hospitals to perform breast screening for one million women; with each woman being invited for screening four times within six years (Ministry of Health, China, 2003). Available data show that to the end of March 2006, there had been 37 hospitals (from 19 provinces) involved in the project, which had screened 19,642 women with a 0.21% cancer detection rate (Hu *et al.,* 2007); thus the project has not been as successful as planned. More recently, the Union for International Cancer Control reported on an effort to screen 5,000 women in China; whilst this is admirable this is far below the number of screened women China needs to address (UICC, 2010).

8.1.3 Potential Difficulties in China for Breast Cancer Screening

Overall then, breast cancer screening is still in its infancy in China. Despite various initiatives, the under-development of screening (e.g. the shortage of mammogram readers and the inadequate capacity of breast screening) limits the current potential to detect breast cancer nationally. Therefore, it can immediately benefit from moving straight to employing FFDM rather than conventional analogue imaging using X-ray film. A difficulty for any country is training a sufficient number of breast screening radiologists and for these individuals rapidly to gain sufficient expertise to perform well in a screening situation where the everyday presentation of breast cancer cases is extremely low yet their vigilance in being able to identify early signs of cancer must remain high. The volume of cases read per year relates to expertise in interpreting screening cases (Scott & Gale, 2007). Therefore when screening is introduced more widely in China it is unlikely that there will be much expertise in identifying early cancer signs until a large number of radiologists

have gained considerable experience. Consequently, there is a crucial demand to standardise the breast screening procedure and to train a large number of mammogram readers to examine and report screening cases. Being able to deliver aspects of such training using non-clinical workstations would be very important.

8.2 Studies

The opportunity arose to investigate aspects of current mammographic interpretation in parts of China. Consequently the following studies describe this work. In China most radiologists carrying out mammography perform symptomatic work – thus they are more used to seeing abnormalities at a later stage than would typically present in screening (as in the UK for instance). Furthermore some Chinese radiologists are very familiar with digital mammography unlike many UK radiologists.

8.2.1 Study One

The study was designed to compare the performance of experienced radiologists in China making screening judgments on digital mammogram images viewed on a low-cost computer display with their performance when examining mammogram films on a multi-viewer. Whilst it was predicted that examining film images would be better, the investigation would also provide an indication of the feasibility of presenting mammograms on a low-cost computer display in another country. Therefore the investigation could suggest directions for further research to enable training on such devices. Comparison of participants' film reading performance with comparable data from the UK would also give an overall indication of what type of training these individuals may require. The data from examining images on the Tablet PC would also provide some information about whether they would find such small devices acceptable for training purposes. The study was conducted in two rounds, one year apart.

8.2.1.1 Methods and Materials

Participants

Breast cancer screening within China is at its very early stage, consequently there are shortages of experienced radiologists who are specially trained to examine breast images. Therefore, it was particularly difficult and challenging to recruit participants for the study. Initially numerous hospitals and radiologists were identified from contacts within China and then they were contacted from the UK. Subsequently they were contacted again by the author when in China. Six hospital sites from two provinces in the central south part of China were visited in person and radiologists there were invited to participate in the study.

However, in the event, only three experienced mammogram readers finally agreed to take part voluntarily from three different hospitals in these two Chinese provinces. These were: the Hunan XiangYa No.3 Hospital, Changsha, Hunan Province; the HuNan People's Hospital, Changsha, Hunan Province, and the Wuhan Xiehe Hospital, Wuhan, Hubei Province. Out of these three readers from the first round of the study, there were only two available for the second round of the study which was carried out a year later.

Two of the three participants had more than five years of mammogram reading experience (including three years of digital mammography experience), however, their main experience was of symptomatic mammography rather than breast screening mammography. The other participant had less than three years of mammogram reading experience (with no digital mammography experience). Each of them read an average of 20-25 patient cases per day, which is estimated to be around 6,000-7,500 cases per year, with most patients being symptomatic cases and a relatively smaller number of screening cases.

Materials

Visual Stimuli: Two sets of sixty pairs of mammographic images were selected from the PERFORMS archive of previously categorised cases. These were matched by difficulty and feature type in each set; each case comprised both the Medio-lateral Oblique (MLO) and Cranio-Caudal (CC) view of both breasts. Abnormal cases (approximately half of the set) covered several mammographic features [i.e. well-defined mass (WDM); ill-defined mass (IDM); spiculate mass (Spic); architectural distortion (AD); suspicious calcification (Calc); benign calcification; and asymmetry (ASYM)] and half of the cases featured no abnormality (i.e. a normal case that had had a three years follow on screening that had proved also to be normal).

Viewing Device: In the first round one set of film-based cases was viewed on a standard radiological viewing box (see figure 8-2: a) and the other digital-based set of sixty cases were viewed directly on the Tablet PC (XPlore technology iX104, 10.4" XGA transmissive LCD screen - see figure 8-2: b). The size of each digital image was 3.5". On the second round the digital set was also viewed on a typical office monitor with screen display size of 21.5" (517 x 72 x 334 mm (W x D x H)); and internal resolution of 1,920 x 1,080 pixels (figure 8-2: c).

Procedure

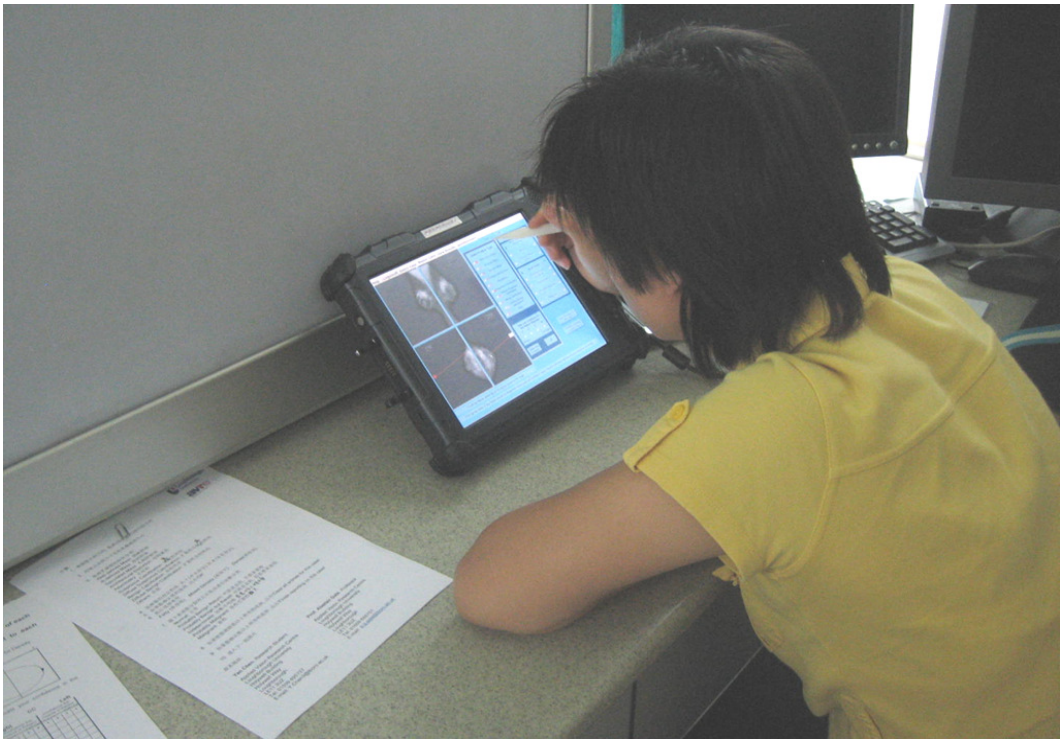
Round One: Initially, sixty screening mammogram film cases were examined on a standard multi-viewer and decisions were recorded on to a Tablet PC (See Figure 8-2: a); the other sixty cases were viewed directly (Figure 8-2: b) only as small digital images (3.5" each) on the Tablet PC (without any image interaction functions supported).

Round Two: Subsequently, after a gap of a year, two participants were revisited and invited to examine the digitised film set which had been viewed in the first round (sixty cases) again on the office monitor. Participants

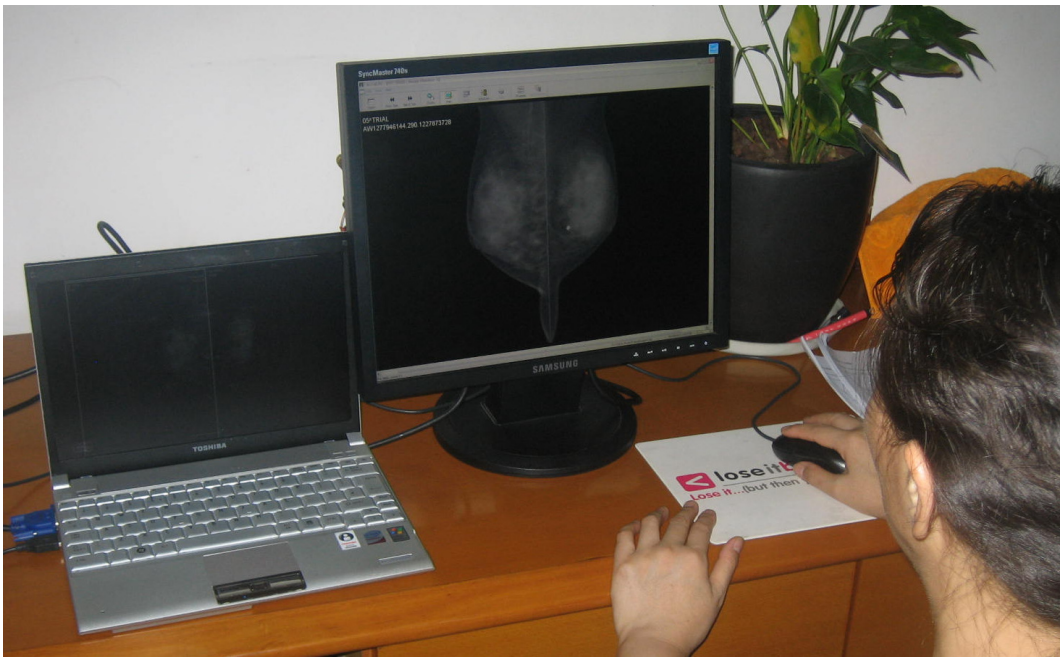
examined these images using the DICOM viewer which supported some essential image manipulation functions (i.e. zoom, pan and window/level)

For each case in both rounds, each participant was asked to decide whether or not abnormalities were present and to rate their confidence in that decision on a 6 point scale: 1. Normal; 2. Benign; 3. Probably benign; 4. Indeterminate; 5. Probably malignant; 6. Malignant. If an abnormality was identified, then the participant was asked to specify the feature type and to indicate the abnormality location. Finally they were asked to judge the density of the breasts. At the end of each experimental session, feedback was sought from the participants regarding the examination of digital mammographic images on the small display.

a)



b)



c)

Figure 8-2. a): Example of the set up for examining the film-based images on the viewing box and reporting decisions onto the Tablet PC (during the experiment the additional light from the viewing box was masked; it is shown here illuminated for photographic purposes); b): Viewing the digital images only directly on the Tablet PC; c): Example of the set up for examining the digital mammograms on the office monitor in Round Two. Ambient illumination increased for all photographs.

8.2.1.2 Results

First Round

The data from the first round were analysed firstly in terms of the correct detection of cancer and correct return to screen decisions (i.e. the basic breast screening decisions). These individual data were then compared with the similar UK national data (from radiologists, advanced practitioners and others) which had been collected from participants who had taken part in the PERFORMS SA07 case set in 2007. Figure 8-3 shows the distribution of the UK data from all 506 participants who undertook both parts of the SA07 scheme with the mean of the UK performance for correct recall (CR) and correct return to screen (CS) decisions respectively shown. Overlaid on these graphs are the data of the three Chinese radiologists (labelled P1, P2 and P3) in the figure. Although the graphs depict UK data from both parts of this SA07 scheme and the Chinese data represents only either SA07 part 1 (film) and SA07 part 2 (Tablet) the figures give a representation of where the Chinese radiologists' performance sits as compared to the UK screening radiologists on these images. As can be seen the three participants did not fare as well as their UK counterparts. P1 and P2 did not recall enough cases, although P3's performance matched the UK mean performance. In terms of return to screen decisions then P3 under-judged these whilst P1 and P2 matched or exceeded the UK mean value. Essentially P2 and P1 under-read the cases and P3 over-read the cases.

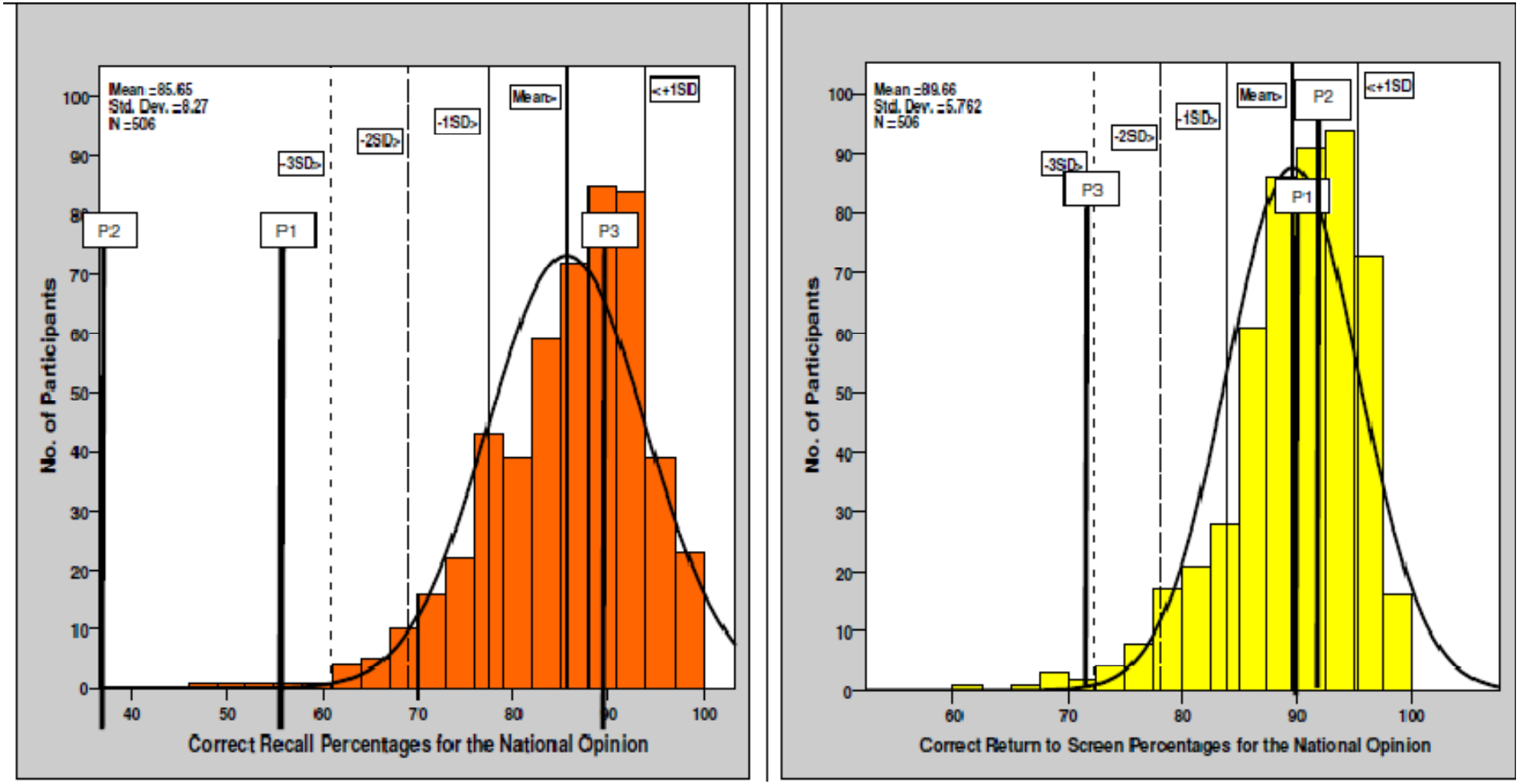


Figure 8-3. The distribution of the UK PERFORMS data performance for correct recall (CR) and correct return to screen (CS) decisions

Subsequently, participants' sensitivity (CR) and specificity (CS) were compared between the two viewing conditions. A one-tailed paired t-test showed a significant difference in sensitivity (p <.05) and specificity (p <.05) between using the film and viewing images directly on the Tablet PC. The details of the measures of sensitivity and specificity of each participant are shown in figures 8-4 and 8-5.

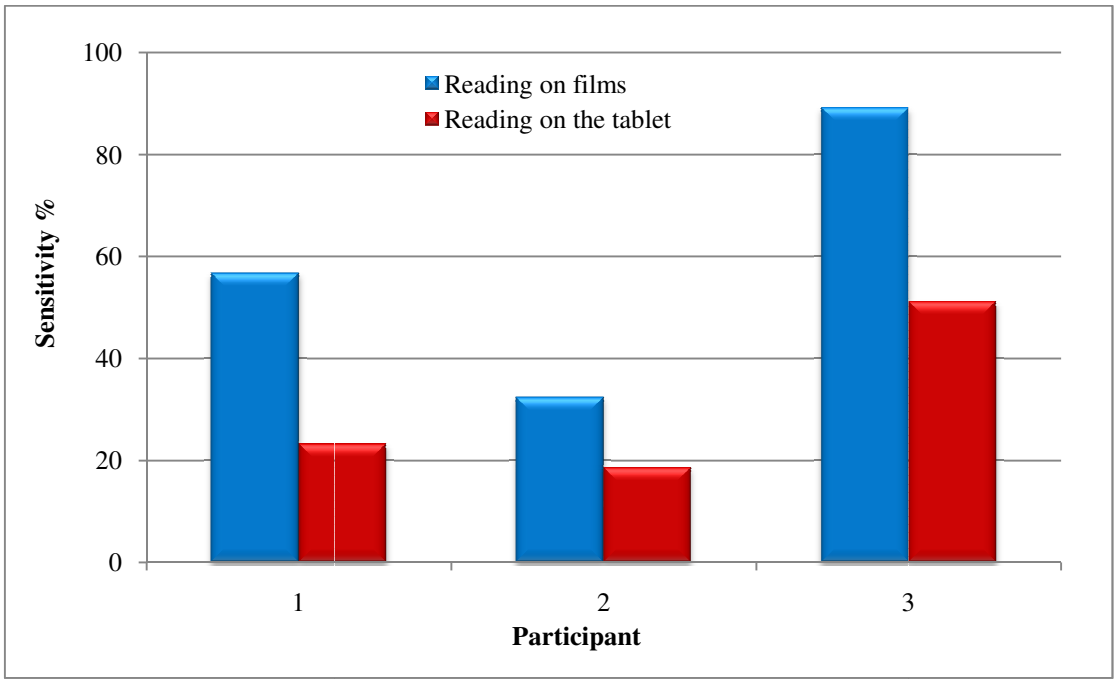


Figure 8-4. Participants' sensitivity measures - examining images on films and on the tablet PC

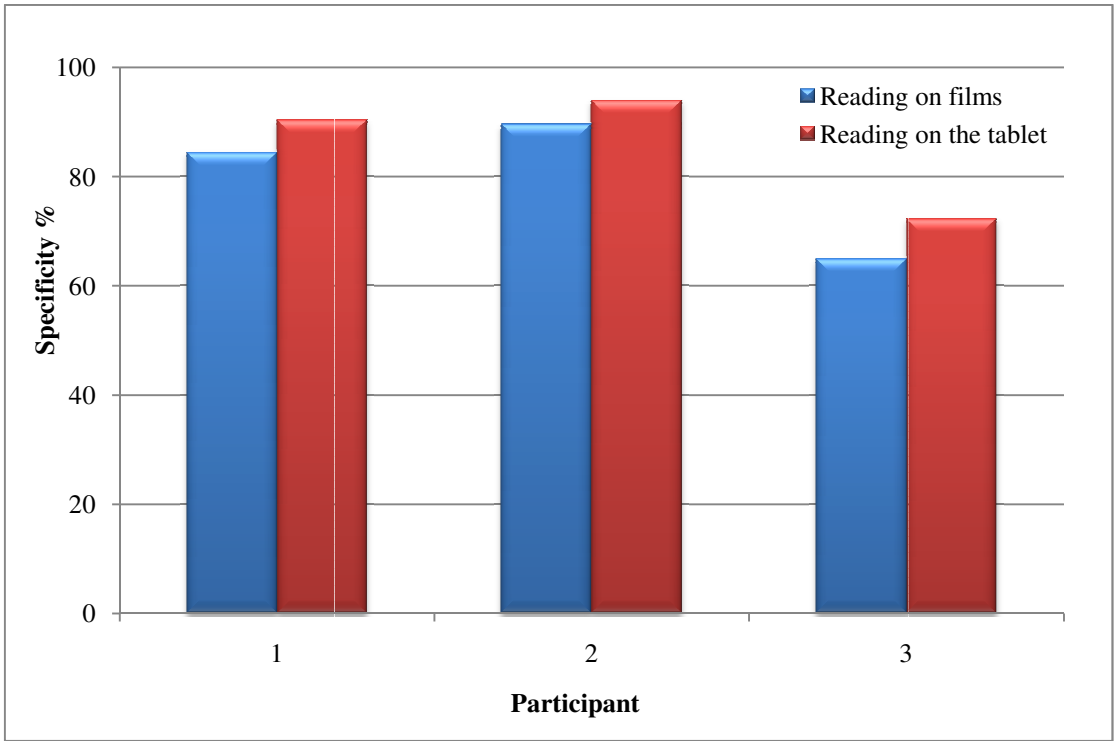


Figure 8-5. Participants' specificity measures - examining images on films and on the tablet PC

The participants' reporting performance was also evaluated using Receiver Operating Characteristic (ROC) analysis (ROCKIT) to compare their average performance between examining film-based images with digital images displayed on the tablet PC. More details are shown in Figure 8-6, 8-7, 8-8.

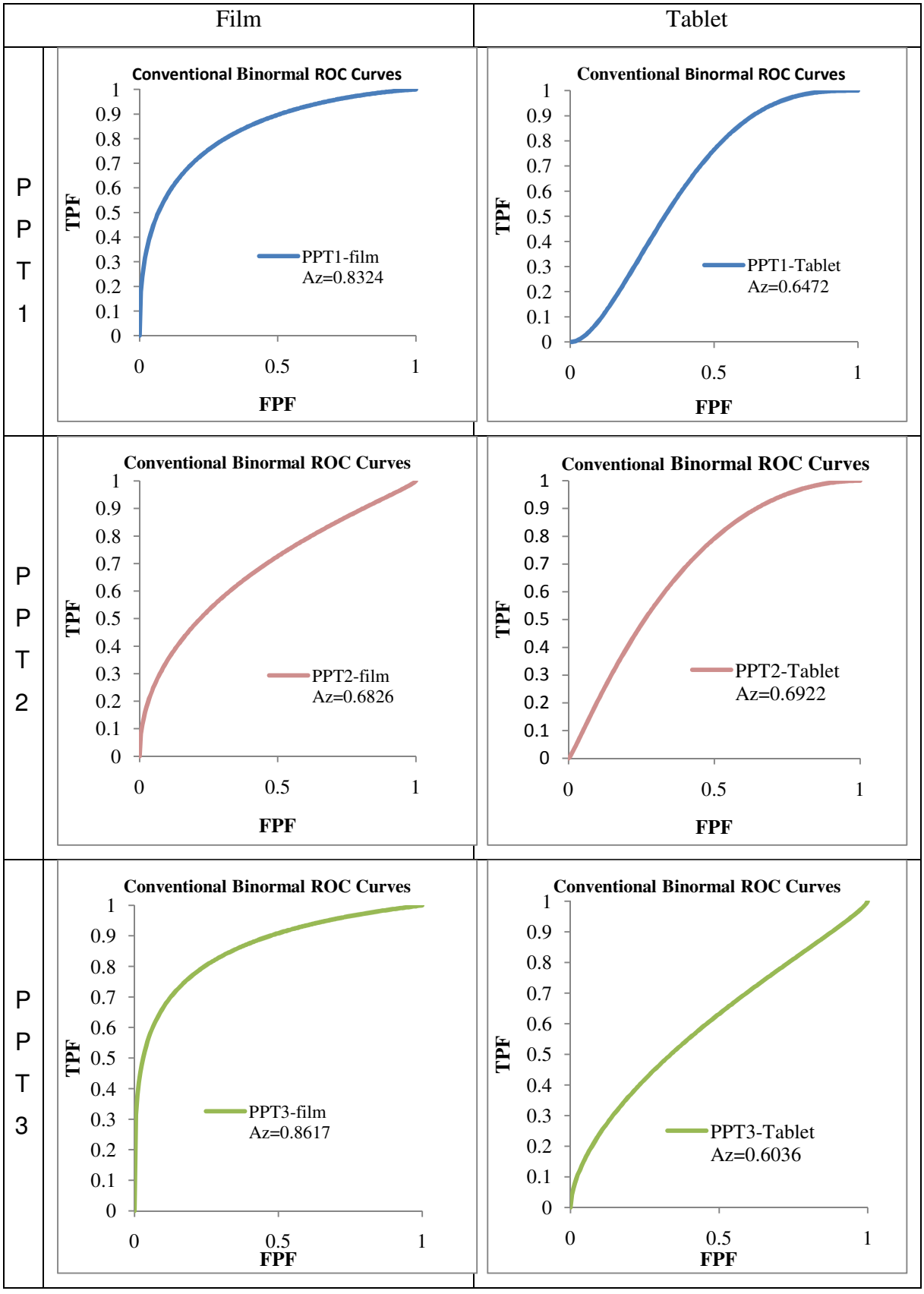


Figure 8-6. ROC plots of participants examining film-based mammograms *vs.* ROC plots of participants examining digital mammograms on the tablet PC

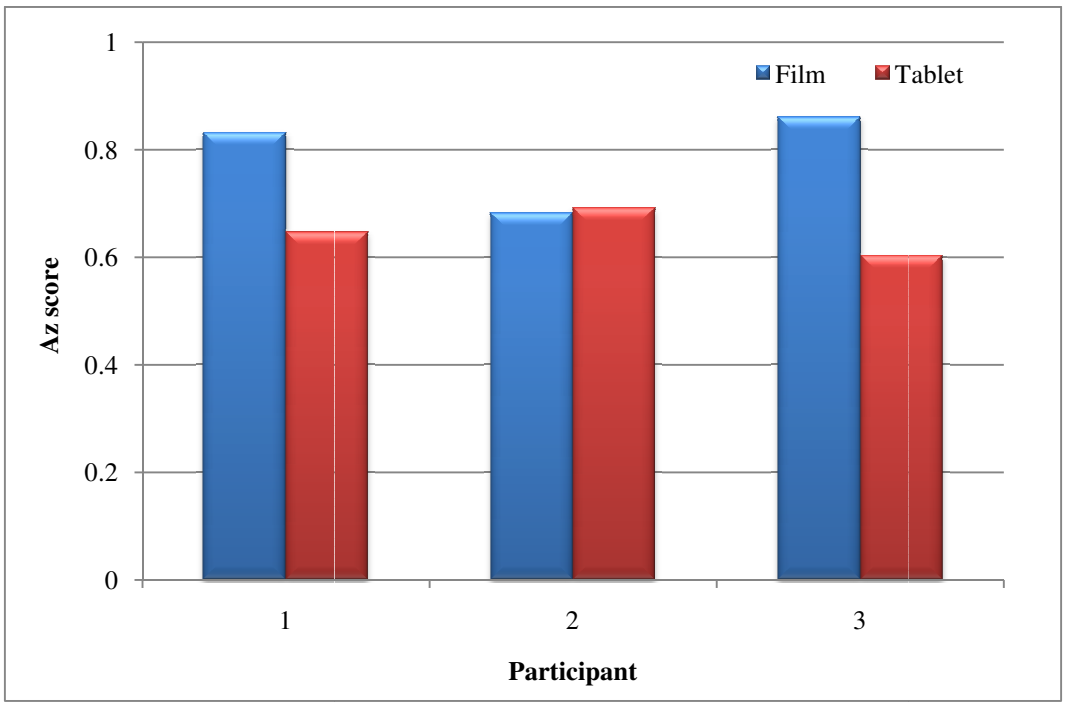


Figure 8-7. The ROC area (A_z) of each participant examining film-based mammograms & examining digital mammograms on the tablet PC

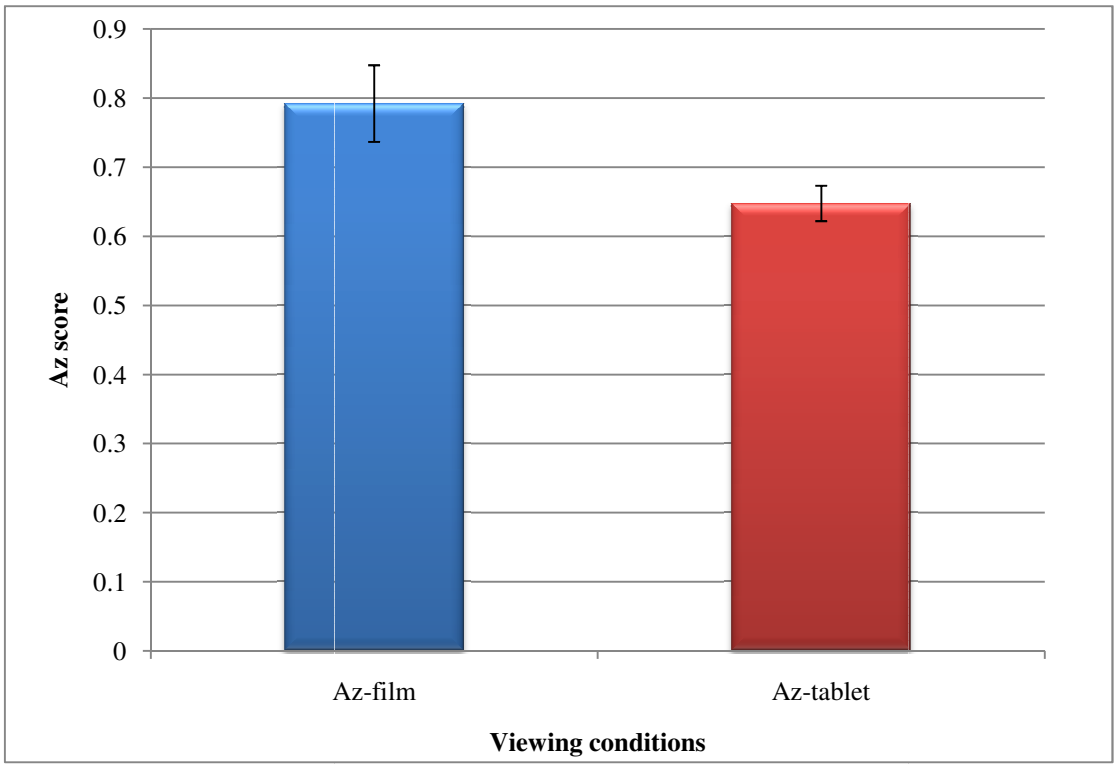
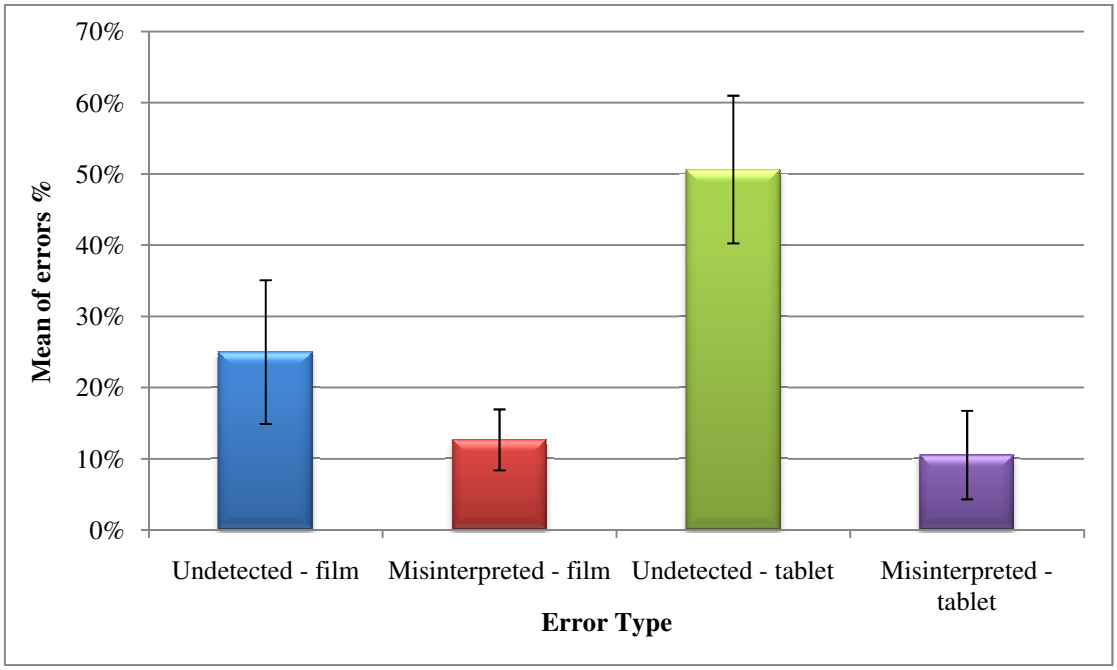


Figure 8-8. The mean of the ROC area (A_z) for each viewing condition.

As shown in the figures above, the trend of the results shows that participants performed better on film with the area under the curve for film-based images interpretation (A_z = 0.792) being greater than for tablet PC-based images interpretation (A_z = 0.647). Although, a one-tailed paired t-test indicated that there was a non-significant group difference between A_z values for the two conditions. Not surprisingly, most participants performed better when examining images on the films comparing to the Tablet in detecting cancer. However, their performance on the Tablet PC images relates to them examining fixed size small images with no image manipulation functions (e.g. zoom, pan, contrast adjustments).

False negative errors are typically subdivided into those due to search, detection and interpretation. Participants' eye movements were not recorded in the study, because of the logistics of transporting eye movement equipment to China and so the number of search errors could not be investigated. However, the other two categories could be investigated and so errors were analyzed by being separated into two groups: detection errors (i.e. an abnormal appearance was undetected) and interpretation errors (i.e. an abnormal appearance was detected but mistakenly classified).

One-tailed paired t-tests showed that the participants made significantly more detection errors than interpretation errors ($p<.05$). Also, there was a significant difference ($p<.05$) between detection errors made on the digital set viewed on the Tablet and such errors made on the film-based set. However, there was little difference found between other group comparisons, i.e. There was a non-significant difference ($p>.05$) between interpretation errors made on both modalities; and a non-significant difference ($p>.05$) between the two types of errors made on the Tablet or film. Details are shown in figure 8-9.



.

Figure 8-9. Mean percentages of both types of errors in each image set

To further analyze the errors made on both the film and the Tablet, the means of errors that were made on the different mammographic features were examined as shown in Figures 8-10, 8-11, and 8-12.

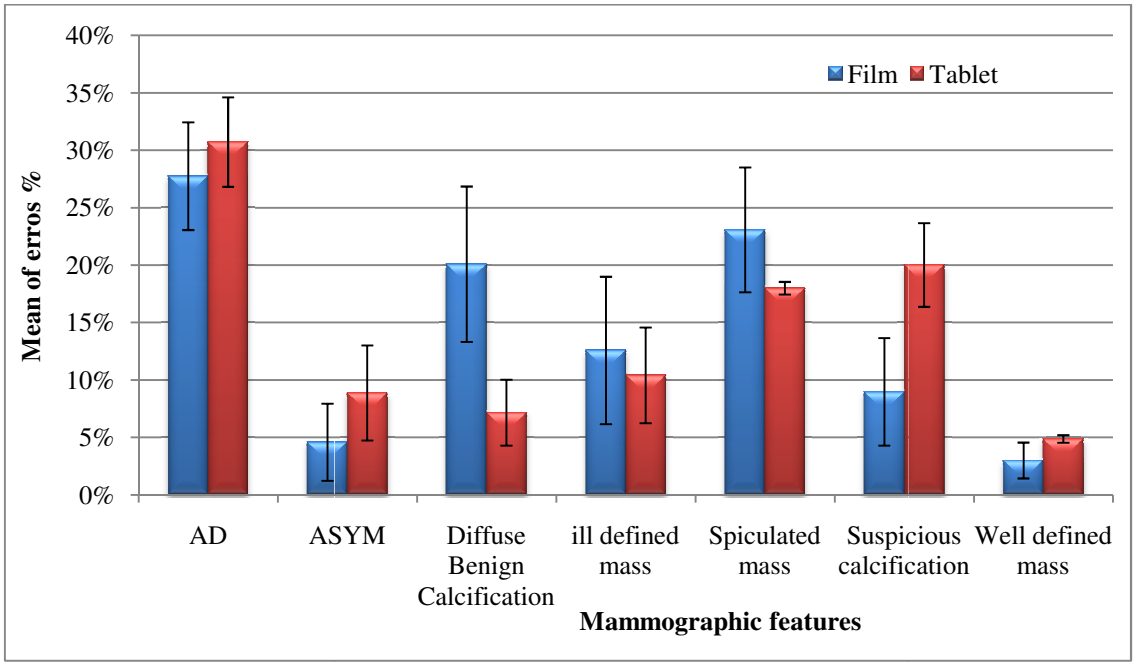


Figure 8-10. Mean error (%) of both errors **X** Mammographic features

Architectural Distortion (AD) was shown to be the most mistaken feature, however the cases were viewed. This follows the general pattern found for the same case set when it was examined by UK screening radiologists in the PERFORMS scheme. Suspicious calcification (arguably the smallest feature) elicited more detection errors on the Tablet as compared to other features as might be expected. Although this difference is not statistically significant, it highlights the difficulty of detecting such small features on the small display.

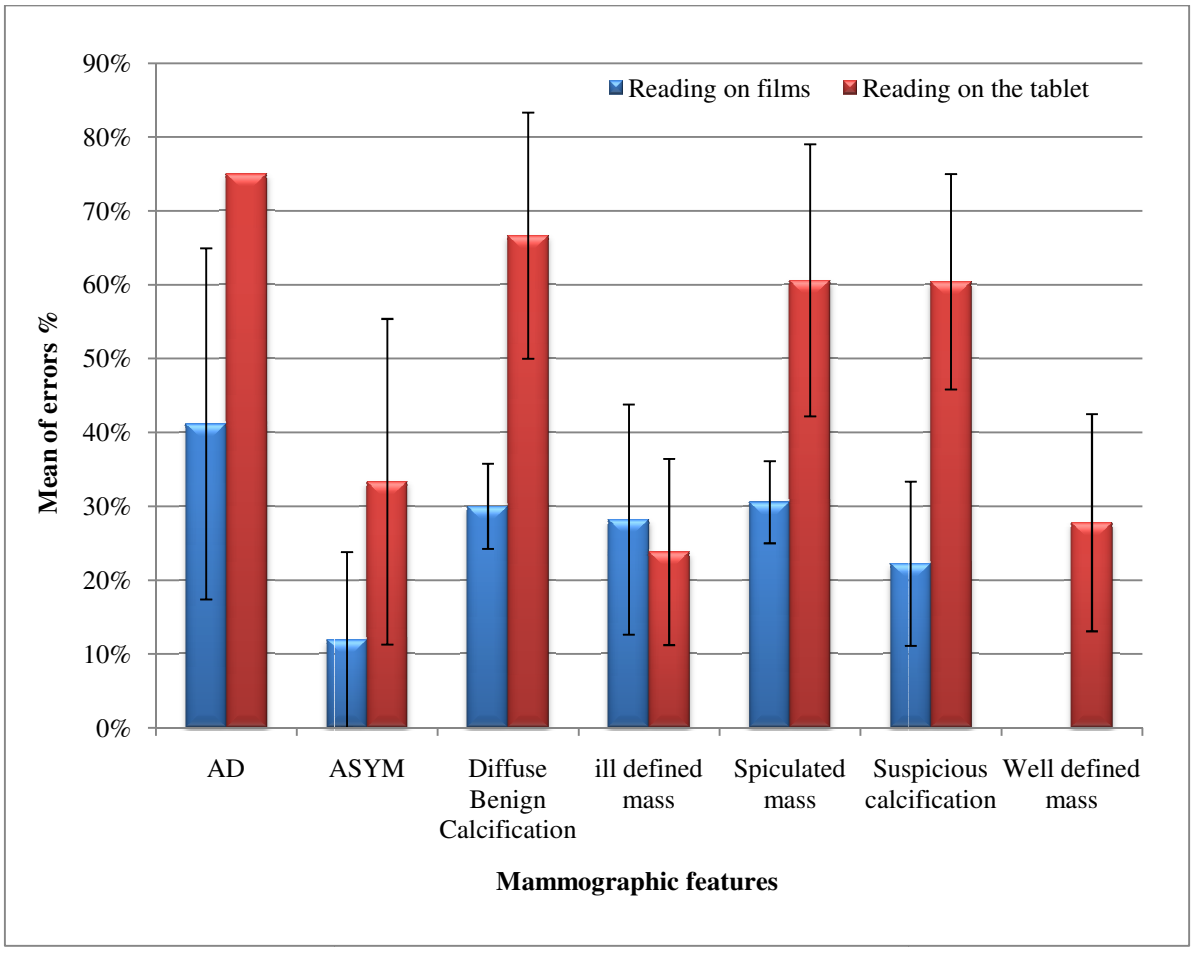


Figure 8-11. Mean of % of detection errors **X** Mammographic features

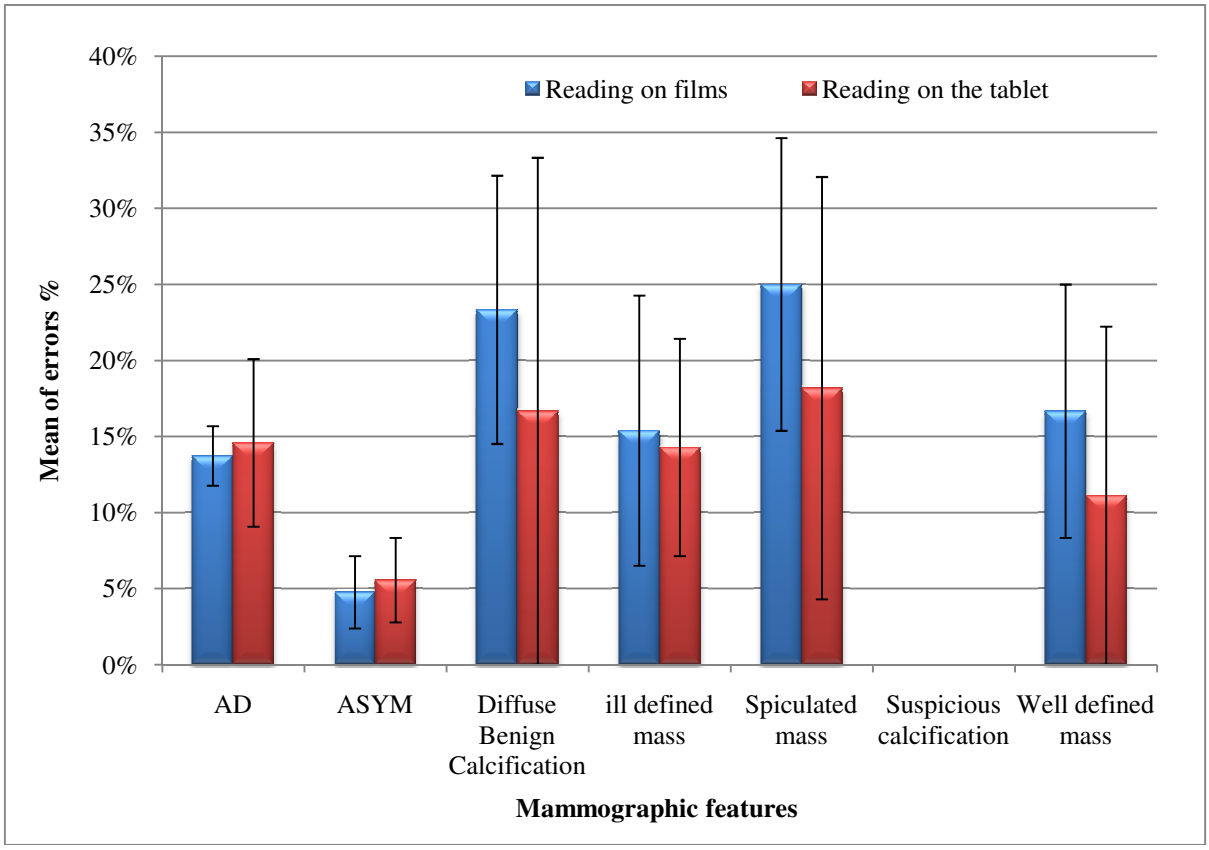


Figure 8-12. Mean of % of misinterpretation errors **X** Mammographic features

The film set used here was carefully constructed from difficult exemplars specifically to stretch participants' skills. In the UK this case set has been shown to elicit a good number of errors, across several hundred participants, when they examine the film images. The data from these Chinese radiologists is then not too dissimilar, especially considering that they were unfamiliar with the task.

Interpreting small sized images was problematic which suggests that adding image manipulation functions to such images would improve performance. This is the subject of the second round of research and it is argued that indeed this does render using small display devices useful training adjuncts.

Second Round

Considering the small number of participants that were involved in the study, the data of each participant was analysed separately. Receiver Operating Characteristic (ROC) analysis was used to compare participants' reporting performance between examining film-based images (in the previous round) with digital images displayed on an office monitor. More details are shown in figure 8-13.

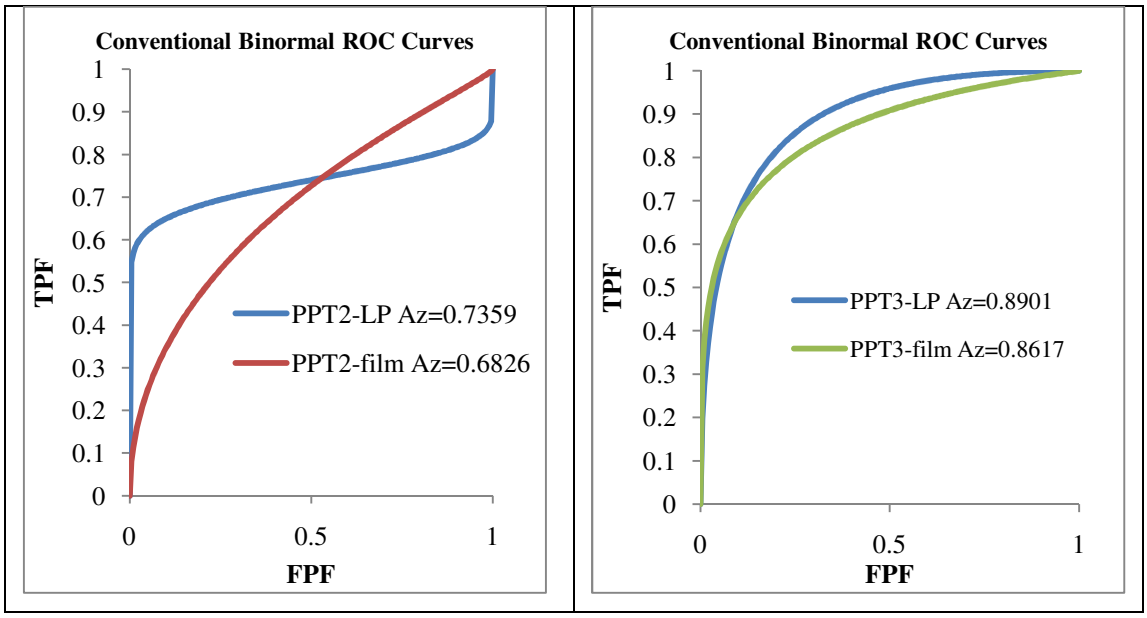


Figure 8-13. ROC plots of participants examining digital mammograms on an office monitor and on film

As shown in the figures above, the results show that both participants performed better on the digital image based task (with the area under the curve for participant 2: A_z = 0.7359; participant 3: A_z = 0.8901) being numerically greater than for film-based image interpretation (with the area under the curve for participant 2: A_z = 0.6826; participant 3: A_z = 0.8617)

Spearman's non parametric correlation test revealed a significant relationship between the rating scale given by both participants for each case viewed on the film-based images and on the digital images (P2: r=.37, p<.01; PPT3: r=.57, p<.01).

ROC analysis for each feature and each participant was also considered as an analysis approach. This was tried to plot each mammographic feature type against 'normal' within each participant. However, this proved to be unsuccessful within this round of the study due to the shortage of the rating data for some features. Nonetheless, the percentage of correct answers for each feature was compared between the two viewing modalities. Details are shown in figures 8-14 and 8-15.

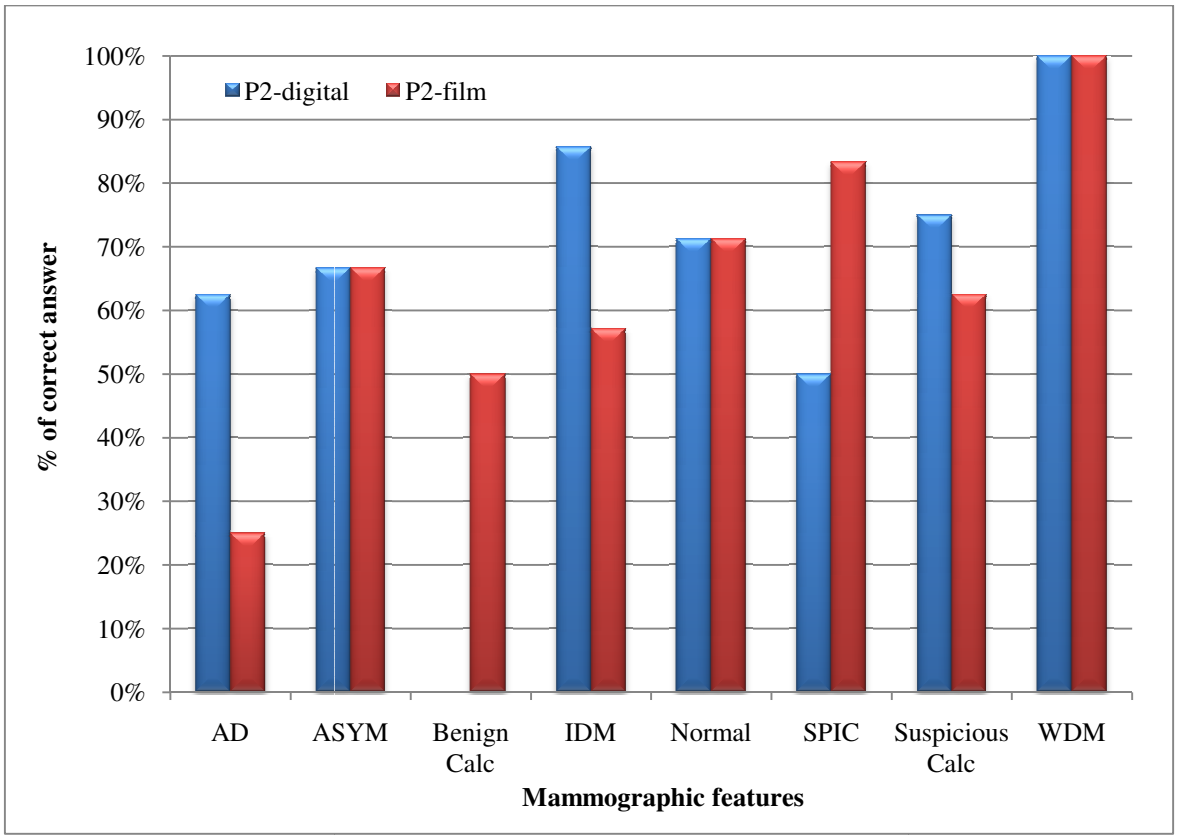


Figure 8-14. Participant 2's correct answer on both modalities x mammographic features (Normal: no mammographic feature was present)

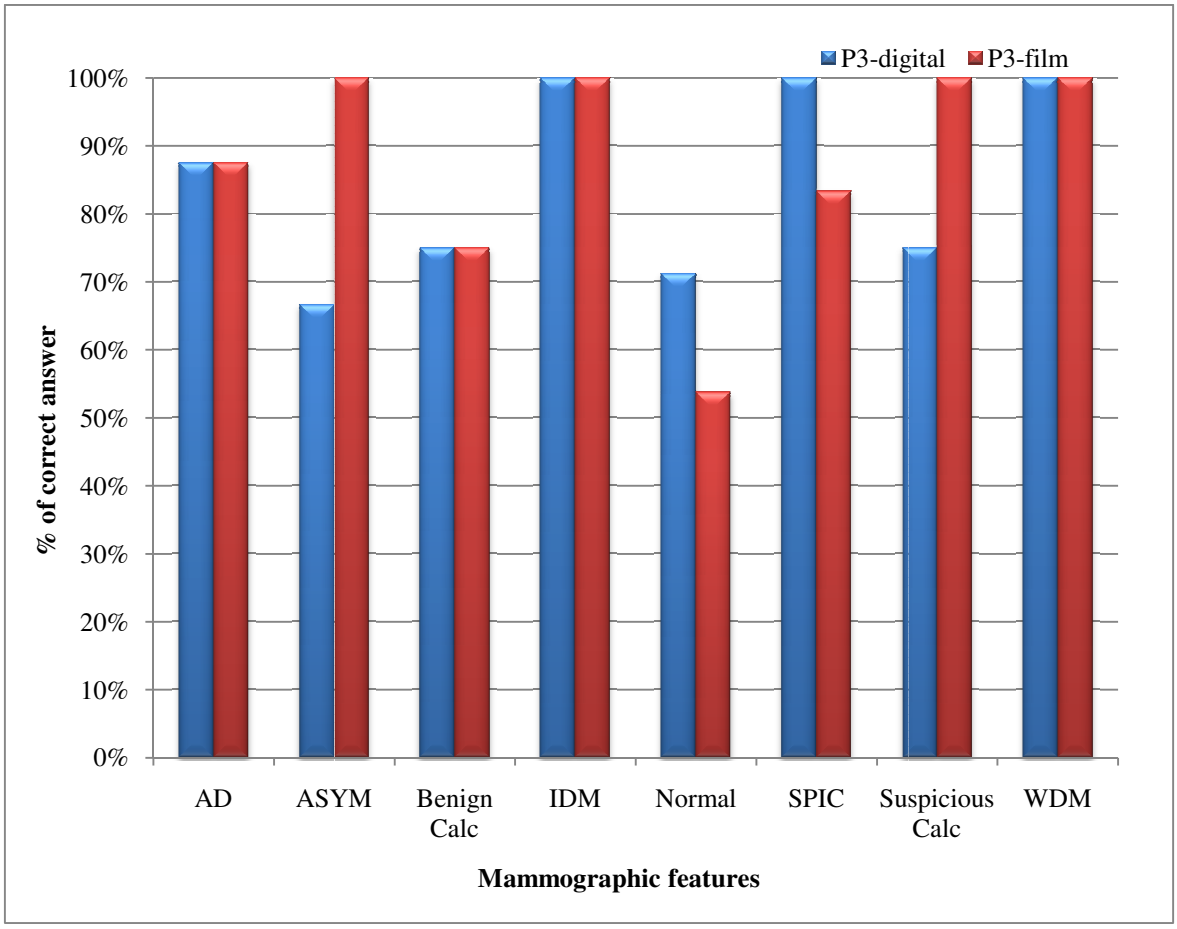


Figure 8-15. Participant 3's correct answer on both modalities X Mammographic features (Normal: no mammographic feature was present)

These figures illustrate performance differences in identifying appropriate features between these two radiologists as well as individual differences in correctly identifying key mammographic features.

8.2.1 Study Two

A small scaled study was also designed and carried out in China to collect feedback for a possible training prototype.

8.2.1.1 Methods and Materials

Participants

In the study, the same group of radiologists who were involved in Study one were invited. However, only two mammogram readers were available to take part.

Both participants had more than five years of mammogram reading experience (including three years of digital mammography experience), however, their main experience was of symptomatic mammography, which is estimated to be around 6,000-7,500 symptomatic cases per year.

Materials

Visual Stimuli: Two sets of ten pairs of digital mammographic images were selected by an expert radiologist from Nottingham Breast Institute. These cases were all pathology proved malignant cases and were grouped by two mammographic feature type, i.e. micro-calcification, and mass. In each set, each case comprised both the Medio-lateral Oblique (MLO) and Cranio-Caudal (CC) view of both breasts.

Training Device: the image set was viewed using a DICOM viewer on the same office monitor which was used in Study One. The detailed information of each case was shown using PowerPoint on the laptop (see figure 8-2:c).

Procedure

Participants were invited to undertake a possible approach to training in digital mammography using a laptop computer. This involved examining two sets of ten digital cases on the office monitor by means of the DICOM viewer which allowed image manipulation. Then, they could check the answer for each case (including the location of the abnormality on both views and other case related information, i.e. expert's comments on each case) by referring to the appropriate image on the laptop screen. After viewing all the images they were asked the following questions:

1. Do you think that this display or approach could be useful for training?

2. What do you think the training approach would be good for (or what features would it be good for) ..And… not good for ?

3. If this was developed into a full training programme – what would you want it to be able to do? E.g. would you use zoom, window, etc.

4. Do you think a training approach like this could be useful in helping new radiologists to screening to gain a rapid understanding of the main features to be looking for?

8.2.1.2 Results

Very positive attitudes towards possible laptop delivered digital mammography training were reported from both participants. Both training sets were considered to be very helpful despite the feature type. Also, it was agreed that the training would be helpful for both new and experienced radiologist groups.

The suggestions collected from participants' comments concentrated on three main points: 1) it would be helpful to be able to examine four views of each case, even if the size of each view would be relatively small; 2) it is essential to be able to use the image manipulation tools, 3) access to prior images of each case would also be useful.

8.3 Discussion

Previous chapters have proposed that effective training can be undertaken using less complex displays than high resolution clinical digital mammographic displays. As well as this being useful in the UK a potential key usage for such an approach would be in countries where large scale

training needs to be undertaken and/or where there are limited funds to support extensive use of clinical workstations for training. An example of such a country is China where breast screening has been trialled and is gradually being introduced (Li & Zhang, 2010). Here, three Chinese radiologists, familiar with mammographic appearance through their symptomatic work, were studied as examples of Chinese mammographic skill level. Whilst accepting that statistically they cannot be taken as indicative of a general screening skill level in China they were the only participants that could be encouraged to take part and they were senior radiologists with considerable mammography experience.

Compared to UK screeners on these PERFORMS cases they did not fare as well. This is not surprising as they would not be used to the early appearances of the difficult abnormalities in the PERFORMS case set, nor would they be used to using the tablet computer for reporting cases.

Their performance on reporting films was compared to that of just reporting directly from the small mammographic images on the tablet PC (with no HCI tools). In general reporting films was better, producing fewer false negative detection errors; there was no significant difference between the two in terms of false negative interpretation errors. More detection errors were found for calcifications on the tablet. This is taken as supportive for presenting screeners from other countries with images on a small computer display – even when no image interactions were possible; the poor performances are taken as support for the need for HCI tools with such displays. When two of the radiologists, a year later, examined the same case set but on a monitor with HCI tools then their performance was improved. This may well reflect their prior experience with examining digital mammograms.

When these two radiologists were subsequently shown digital mammograms which they could view on a laptop using a DICOM viewer with HCI tools and also see feedback on each case as an example of a possible training approach then their views were very encouraging.

8.4 Conclusions

The three radiologists who voluntarily took part here were experienced in examining breast images and in particular symptomatic images. The case sets which they examined were based on known difficult UK screening cases and so they would have found these to be challenging as disease presentation would typically be at an earlier stage than what they were used to. Furthermore, with the rapid implement of digital mammography in some major Chinese breast units, the radiologists who were invited in the study were more familiar with the digital mammograms than with the film-based mammograms. Thus it is not surprising that their performance was poorer than that of UK radiologists in examining the film cases. Additionally the UK radiologists annually read such difficult film sets and consequently are quite used to the potential make up of abnormal cases within the sets. This is indicative of the need for widespread training in screening case appearances to be undertaken in China to facilitate the introduction of breast screening.

The data showed that they performed comparably well on examining digital mammograms on the normal office monitor with the support of image interpretation tools, however, relatively poorly on the small tablet. The data also suggested that they could benefit from further training that could be offered using the widely available low-cost office monitors. In order to utilise low-cost computer devices for mammographic interpretation training, suitable image interaction techniques need to be employed appropriately which would then aid the introduction of widespread screening into countries such as China. The displays (monitor, tablet PC) were acceptable and usable by these participants which shows the way forward for offering training in other countries using HCI-enabled low cost displays. Not only can such devices be used within the UK screening programme but these are also applicable in other countries where there exist different levels of screening skills.

CHAPTER 9

Summary, Implications and Future Work

9.1 Summary

Radiology has undergone massive technological developments in recent years which have seen the shift from using X-ray film to digital images. In 2010 the Radiological Society of North America (RSNA)'s annual meeting – the world's leading radiological conference - was entitled "Personalised Medicine" to reflect the use of computing devices to offer truly personalised image inspection. Virtually every vendor of radiological imaging equipment now has applications for smart mobile devices, such as the iPad, iPhone and other PDAs. Such devices are perceived as offering the physician the opportunity to share images with patients whilst on hospital wards or for teams of medics to discuss cases. Some researchers have shown that reporting CT cases on such devices is possible.

Despite this expansive use of mobile devices, the FDA has ruled recently (Diagnostic Imaging, 2010) that 'Mobile MIM', an iPhone application that allows medical images to be examined by iPhone users "has new technological characteristics that could adversely affect safety and effectiveness and raise new types of safety and effectiveness questions…Therefore, this device is classified by statute into class III (Premarket Approval)." This decision virtually ruled that any PDA-type devices should not in fact be used for diagnostic reporting purposes. However, despite this, the ongoing mass of applications for mobile devices being developed by manufacturers demonstrates that there is a strong market desire to use such devices, where possible, for numerous purposes within radiology.

Mammography is the last domain within radiology to make the changeover from X-ray film to soft copy reporting. The advent of digital imaging in mammography raises the potential of using a range of display devices, such as PDAs, for image inspection purposes. In this thesis the potential of using such a range of sub-clinical display devices for training purposes in breast screening has been investigated from the HCI viewpoint. High resolution clinical displays cannot be replaced for screening reporting or diagnostic

purposes but such equipment is both expensive and has limited availability for non-clinical uses. Thus if other cheaper, less high quality displays, can be shown to be useful for training purposes then there could be a widespread need for such displays. The domain concentrated upon is breast screening, where the Loughborough Applied Vision Research Centre already has a major research investment in assessing the annual performance of all UK screeners. The research presented here asks whether there is a need for training to be undertaken on portable, and other, devices and if so then how should this be accomplished? What type of displays could be used and how might training using such displays be undertaken?

At the current time the UK breast screening programme is being expanded which means that more individuals need to be trained both to participate in interpreting screening mammograms in the national programme and also to maintain existing screeners' high skill levels. Additionally, breast screening is being introduced in many other countries and China is taken as an exemplar country where there is a rapidly growing need for screening which demands that large numbers of health professionals be trained to adequately inspect screening mammograms. To accomplish these using clinical workstations would be financially and logistically impossible hence the use of other display devices is a necessity as well as being highly beneficial.

From the outset it was assumed that over the course of the time of this research technological progress would have addressed any technical limitations of such portable devices evident at the outset – this is confirmed, for instance, by the release of the iPad and the retina display of the iPhone 4 which are major capability changes from the PDAs that were available at the start of the research.

Initially (Chapter 3) a UK national survey of all the mammography film readers within the NHSBSP was undertaken which showed that the current main difficulties of mammographic (film) interpretation training were: inflexible training time and a lack of designated training delivery equipment. With the

gradual introduction of digital mammography, it is argued that the training situation resulting from the high UK breast screening workload could potentially be improved by implementing some low-cost mobile viewing devices as an alternative to the current image viewing devices, namely X-ray film multi-viewer and digital workstation. Whilst a positive attitude towards using such displays as a technology for delivering 3W ('whatever required, whenever, wherever') training employing digital images was found, demonstrating the consumer pull for 3W, the potential for mammographic interpretation training on different viewing devices had not been examined.

Consequently, empirical research was undertaken to investigate this, firstly (Chapter 4) by trialling experienced breast screening radiologists with some mammography images on a small PDA. Their initial reticence on examining very large mammographic images on such small devices was expected. However, their fair cancer detection performance and their positive views having tried to use the PDA suggested the potential of using such devices for delivering aspects of mammographic interpretation training whilst highlighting the need for appropriate HCI (image manipulation and interaction).

On the basis of the results and participants' comments of this trial, another two small-scale pilot studies were completed (Chapter 4) using a laptop computer. Results again indicated the promising possibility of using a laptop as an image interpretation training tool. Although there were too few participants involved in the pilots to reach any statistical power, the data were used to estimate minimum number of participants for the subsequent studies.

Having elicited that using small display devices could be acceptable to end users (breast screening personnel) a more detailed study (Chapter 5) was then undertaken of radiography students who had no mammography experience but some radiographic knowledge (i.e. potential users of a 3W training system). This investigated how they examined mammograms on four different small screens with various sizes and resolutions (representing typical PDA screen sizes/resolution).

Participants were eye tracked which allowed examination of each individual's visual search behaviour and therefore enabled their errors to be studied in more detail. In general, the results from this study demonstrated that it was possible to view mammograms on different small sized screens (typical PDA screen sizes), although the performance on calcification and normal cases (with no key mammographic features present) was poor. Also, It had been predicted that those viewing conditions which represented a larger image size/higher pixel resolution would give rise to more correctly reported cases. However, the data showed no significant differences between performances in the four different viewing conditions. Hence, it implied the possibility of delivering very specific training on PDAs, which takes the small size and pixel resolution into consideration. This suggested that further research needed to be focused on how best to employ suitable HCI techniques to increase the feasibility of mammographic interpretation training on PDAs rather than overly, and simply, focusing on increasing the screen size and resolution.

Additionally, participants' eye movement data on some cases were compared with a mammography experts' data (Chapter 6). Clear differences in saccadic eye movement patterns and fixation locations were found between the expert and the participants. Taking into account the participants' relatively poor performance data along with their eye movement data, it demonstrated it was very important to have readers with mammography knowledge involved in future studies. However, such personnel are not always readily available to participate in experimental studies.

Another study (Chapter 6) firstly recorded an expert's eye movements and his radiological comments on a set of screening cases. Subsequently, a series of four potential training approaches were designed based on utilising aspects of his inspection behaviour (i.e. his comments, visual inspection patterns and eye movements). A group of naive observers were then trained using these approaches to investigate whether any approach was potentially useful. Subsequently, a focus group was undertaken with 15 breast screeners who

were shown examples of each of the four training methods and asked to give their evaluations and suggestions concerning each.

Empirical data, along with results from the focus group, showed that the four training approaches were all feasible to implement but of variable usefulness for delivering mammographic interpretation training. Although the expert audio instruction (i.e. the traditional approach in radiology) was recommended by the focus group, this was not supported by the experimental data with these participants. It was shown that naive participants could be trained by these visual methods which resulted in them spending more time looking at the abnormality location. This is most promising as this is the beginning of the learning process of identifying abnormal appearances. The importance of this is shown in Chapter 7 when experienced screeners' performance was examined and it was found that most errors (circa 60%) were attributable to visual search - they simply failed to look at the abnormality.

These results also again called attention to the importance of involving participants who have a breast imaging interpretation background. Consequently, a large scale study (Chapter 7) was undertaken employing breast screening radiologists and advanced practitioners. Three major breast screening centres took part with 15 mammogram film readers involved in the study over a period of eight months. Each participant undertook three rounds of reading the same set of cases on three different modalities (digital mammography workstation, standard monitor and iPhone) with at least a two months gap in between. Participants either examined images on the displays or were additionally allowed to interact with the images and manipulate them.

JAFROC analysis revealed that overall performance on the standard monitor while the image manipulation tool was used was almost as good as their performance on the clinical workstation. Such a promising result strongly implies that a standard monitor can be used successfully to deliver mammographic interpretation training. However, the iPhone performance, with

or without image manipulation tool use, was significantly lower than either workstation or standard monitor performance.

Participants' image manipulation behaviour on each modality was recorded and analysed directed by their performance data results. The important finding was that their screen interaction behaviour on the digital workstation and the standard monitor was shown to be comparable. However, a significant difference was found when participants were separated by their mammography experience into two groups. Compared with the low experienced participants, high experienced ones spent less time on the standard monitor and workstation while the image manipulation tool was used. However, high experienced participants spent more time than the inexperienced participants while the image manipulation tool was not used. This suggests that high experience readers adjust their way of reading, hence, increasing the image manipulation time, if the normal image manipulation tool was not available.

Furthermore, the results of the participants' visual inspection data indicated that the participants made similar overall pattern of errors on both modalities: 1) the average number of times participants fixated within the AOI; 2) participants' eye dwell time within the AOI while examining images; 3) the number of fixations before the eye first fixated within the AOI; 4) and the average duration to first hit the AOI while examining images on a standard monitor was not significantly different from examining on the workstation.

Overall results from participants' data on performance, image manipulation behaviour and visual search indicated that a standard monitor could be employed as an alternative for the digital workstation for delivering on-demand mammographic interpretation training which uses the full mammographic case images.

Having shown that non-clinical displays can be used for training within the UK it would be valuable for such an approach to be implemented in countries

where large scale training needs to be undertaken and/or where there are limited funds to support extensive use of clinical workstations for training. Therefore, a small study (Chapter 8) was undertaken in China. Firstly, the performance of some, primarily breast symptomatic, Chinese radiologists in reading PERFORMS cases were compared with UK screeners. A year later, the same set of images were viewed again on a laptop. Furthermore, the potential method of delivering training was tried by the Chinese radiologists. Overall the data showed that the Chinese radiologists were not as good as the UK radiologists were, as was to be expected for various reasons. However, much more importantly it showed that delivering training using small display devices was not only very applicable in the UK but also in other countries with different levels of screening skills.

9.2 Contributions to Knowledge

Changes to the UK NHS Breast Screening Programme to use digital imaging have been taken as pivotal in investigating the introduction of a range of potential training delivery tools.

The research has identified that there is a user demand for training approaches which are personalised, accessible and remote (i.e. 3W – Whatever required, Whenever, and Wherever) from clinical workstations, not only in the UK, but also in developing countries.

This thesis presents the first known study to investigate potential alternatives to the digital mammographic workstation as a training delivery tool which uses the full mammographic case images. This is very important as visual search errors are the most common occurring type of error and so approaches which only utilise part of the mammographic image in training, whilst useful, immediately ignore the main error factor in screening. How such training could be delivered has been extensively explored in a number of studies here which have shown that a standard office monitor can be used very successfully. Other displays ranging in size down to a PDA have been

investigated and found to be poor - although future developments of displays and software may well make them more feasible.

Methods of training using aspects of an expert's visual inspection behaviour have been shown to be a useful and fruitful approach.

Furthermore this research presents the first known study tentatively to investigate screening performance in China as compared to the UK and demonstrate that training can be offered in China using small devices.

The various experimental studies carried out in the course of this research have been disseminated at international scientific conferences and also been published in various edited publications. These are detailed at the end of this chapter.

9.3 Future work

The research presented here has generated several ideas for future research, some of which have already begun to be investigated.

Reducing False Negative Errors in Breast Screening

First of all, further research is needed both in evaluating mammogram readers' performance and also in training them to better their performance. This is evidenced by the latest data from the PERFORMS scheme (2010) which shows that on a test set of difficult cancer cases which were examined in 2009/2010 then some 15% of these cancer cases were missed when examined by all UK screening personnel.

From the current experiments visual search errors are the major false negative (miss) errors that have been found. Thus, approaches which seek to train individuals to improve their visual search inspection behaviour are important. Employing aspects of an expert's image inspection behaviour to develop training regimes is one novel approach which was found in this

research to be useful. Here, the potential of these different methods was examined. Future research needs to expand on these and examine the effects of extensive different training techniques and how they impact upon performance.

Improvement to Mammographic Interpretation Performance Evaluation

Radiological CADe and CADx systems are evaluated using different techniques. Free-response ROC (FROC) analysis is often employed to evaluate CADe systems. Currently, jack-knife (JAFROC) is the most commonly used methods to compare two CADe system (i.e. compare two FROCs). On a mammogram case, two views of the same lesion (on both MLO and CC view) are marked separately by the CADe system, although two images' CAD scores might be correlated as these are the same anatomy. Current FROC analysis using re-sampling methods (including bootstrap, jackknife, and permutation tests) to handle the possible correlation of multiple marks has been used to compare the performance accuracy between CAD systems (Samuelson & Petrick, 2006; Samuelson, *et al*., 2007).

However, radiologists examine and mark cases as a whole in their clinical practice:

1) For each breast, the lesion on MLO and CC views (if the lesion can be seen on both views) are always considered together. The classification/rating on the lesion on each view (i.e. each individual data point analysed by the FROC) potentially influences each other;

2) Comparing both breasts can aid in detecting several types of breast cancer features, this applies for every mammography feature except micro-calcification. Therefore, for these feature types, the classification/rating on each breast (i.e. each individual data point analysed by the FROC) can potentially influence one-another.

Hence, it is important to decide that participants are scored by each view or by each case. Also, re-sampling methods that are used to handle the possible correlation of the multiple marks have not been evaluated using human readers' data. Consequently, ongoing collaborative work by the author with Prof. Dev Chakraborty (the author of JAFROC) is applying such methods to evaluate and compare performance accuracy between mammogram readers. This will develop a new ROC analysis technique for widespread implementation by all medical imaging researchers.

Visual Research with New Breast Screening Modalities

Furthermore, visual performance research is currently not undertaken in other imaging modalities within the area of breast cancer detection. A considerable amount of work has been undertaken in investigating mammogram film readers' eye movements while they are examining images (such as with SFM and digital mammograms). However, there has been some dramatic changes with imaging modalities within the area of breast cancer detection. Radiologists have started to examine stacks of images of each patient as compared to the traditional four mammographic views per woman. With the development and wider usage of new image types, it is very important to extend current visual performance research to the related medical imaging techniques in breast screening. Consequently, discussions are currently under way with key UK screening centres and the Director of the NHS Cancer Screening Programmes to expand the current research to encompass both breast MRI and breast tomosynthesis as well as new breast ultrasound Doppler techniques.

Visual Search & CAD

Another aspect that needs further investigation is the potential of using CAD to assist mammographic interpretation training. CAD systems were developed with the aim of helping the reader in identifying abnormality presence (CADe) and classify it appropriately as either benign or malignant (CADx). CADe has been well researched in order to aid radiological imaging interpretation by

suitably applying image processing algorithms in order to identify known key image features. Whilst CADe is used to aid in detection, it is proposed and argued here that CADe can also very usefully be employed in training; this is a wholly new proposition.

Recording saccadic eye movements whilst observers are examining displays is an area which has grown enormously in recent years. Because the saccadic eye movement system is not always consciously driven then often the user is unaware of precisely where in an image they have looked (i.e. which covert cues may have attracted their attention). Gaze location in a display is related to aspects of the image in that location which have attracted visual attention and consequently a useful approach is to record the user's eye movements whilst they examine an image and then show them the image again overlaid with areas highlighted where they have looked. Such an intelligent computing approach offers the observer the opportunity to examine in detail those image areas which have attracted their attention but of which they are probably not aware. In medical imaging, eye tracking has been used experimentally to investigate certain performance aspects such as expertise development although it is not used in routine clinical practice.

It is proposed here that eye gaze can be used in association with CADe to assist mammographic interpretation training. For instance, the training approach would be for a radiologist to examine an image while recording his eye gaze, with CADe applied in the background. This will generate image areas which the radiologist spends considerable time examining (using a fixation time criterion cut off value which can be empirically derived) as well as CADe prompted image areas. Where only CADe prompted areas exist, the area will be shown afterwards to the film reader as the training clue.

To explore the potential of the above proposal, an initial experimental investigation has been undertaken in co-operation with Hologic (a major manufacturer of breast imaging equipment), from which an example is shown in figure 9-1. A digital mammogram (figure 9-1a), containing abnormal

features, was presented to an experienced breast scientist on a 17" monitor and her visual search behaviour recorded using a Tobii eye tracker. This plots as an output the observers' eye movements and eye fixation locations. These were analysed and a useful fixation time criterion (500 ms) employed to identify image areas which were visually attended to (shown as squares in figure 9-1c). The potential abnormalities that were identified by a CAD system (Hologic R2) are highlighted by circles in figure 9-1b. By comparing both figures 9-1b and 9-1c, there are two areas which are agreed by both the observer and the CAD (highlighted by hexagon areas) and which are argued here demonstrate a very high probability of abnormality presence. Other areas separately prompted by the CAD and eye movement data are also shown. By feeding back to the observer the areas which they have examined, coupled with the CAD prompted areas, is proposed could be used to train individuals to recognise which areas they have ignored as well as which areas they have examined. For instance in figure 9-1d the red circle identifies a CAD prompted area which the observer did not look at. Feeding such information back to the observer as part of their interactive training is therefore proposed as a new training technique.

Radiological inspection performance is subject to errors, like all human inspection situations. New intelligent computing applications based on monitoring the radiologist's eye gaze as s/he examines medical images, coupled with CADe, are further proposed as a new mammographic interpretation approach.

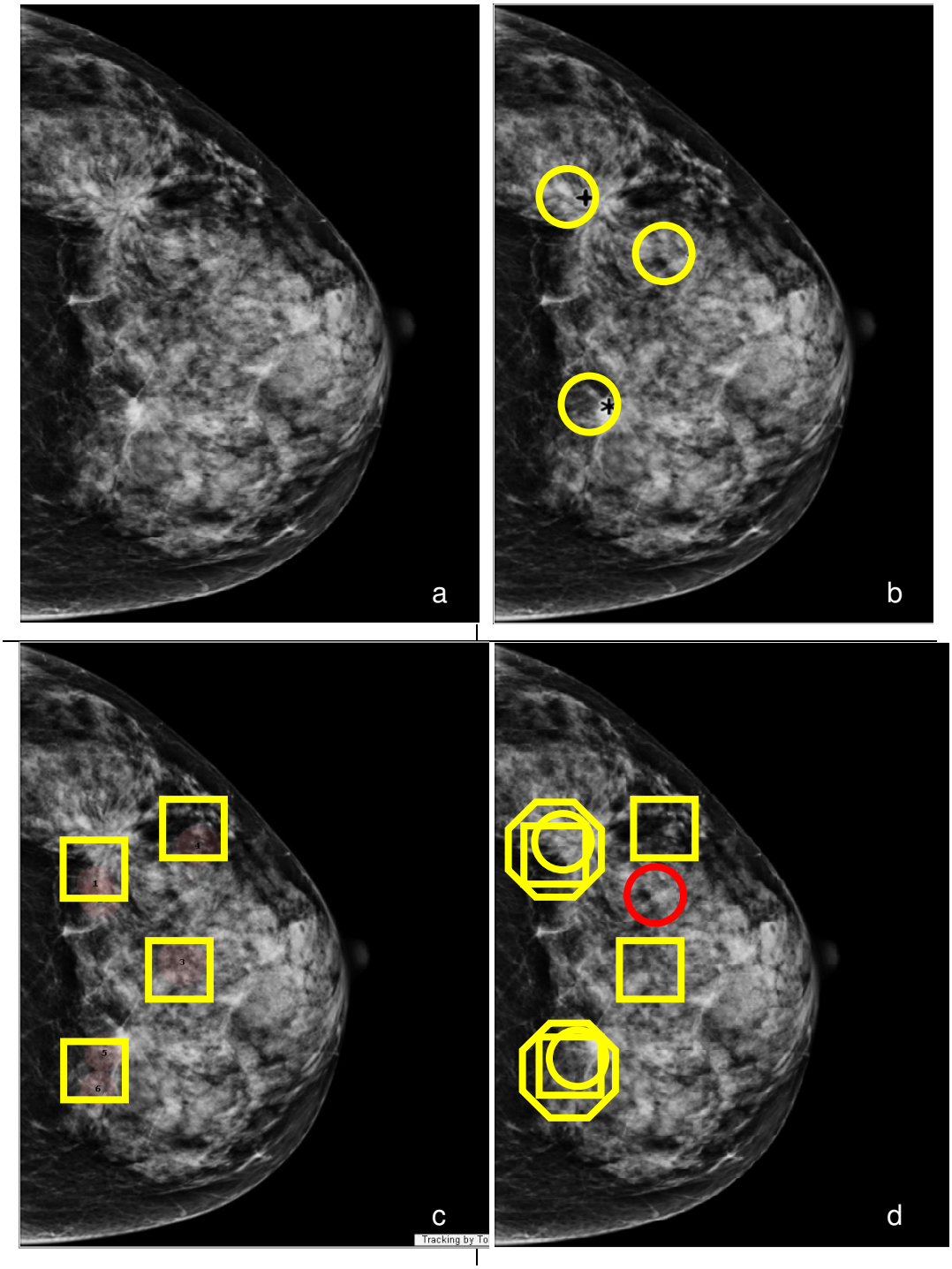


Figure 9-1. Digital mammogram with prompted areas. See text for explanation. Images were kindly supplied by Hologic.

9.4 Publications from This Research

1. Chen, Y., James, J., Turnbull, A., Gale, A.G., Is image manipulation necessary to interpret digital mammographic images efficiently?. In *Proceeding of SPIE Medical Imaging Conference 2011*.

2. Darker, I.T., Gale, A. G., Chen, Y., Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme, In *Proceedings of SPIE Medical Imaging Conference 2011*.

3. Chen, Y., James, J., Evans, A., Turnbull, A. and Gale, A., 2010. An investigation of workstation image manipulation usage when examining FFDM images. *Breast Cancer Research*, 12 (Suppl 3), P9.

4. Darker, I.T., Gale, A. G., Chen, Y., 2010. Variability in film reader estimates of breast density in the PERFORMS scheme, *Breast Cancer Research*, 12(Suppl 3):P42

5. Chen, Y. & Gale, A., 2010. Intelligent computing applications based on eye gaze: their role in medical image interpretation. *Advanced Intelligent Computing. Theories and Applications: 6th International Conference on Intelligent Computing,* Springer, P320.

6. Chen, Y., Gale, A., James, J. & Turnbull, A., 2010. The effects of image manipulation on digital mammographic interpretation, In *Proceedings of Symposium Mammographicum 2010, British J. Radiology online*.

7. Chen Y., & Gale A.G., 2010. Mammographic interpretation training: how could low cost display devices help? In *Digital Mammography. J.* Marti, A. Oliver, J. Freixenet, & R. Marti (eds). Springer.

8. Chen Y., Gale A.G., 2010. Using Eye Gaze in Intelligent Interactive Imaging Training, In *Proceedings of the International Conference on User Interfaces*.

9. Chen, Y., James, J., Turnbull, A., Scott, H. & Gale, A.G., 2010. Is image manipulation necessary in digital mammographic interpretation? In *Proceedings of the European Conference on Radiology 2010*

10. Chen, Y., Gale, A.G., James, J., Evans, A. & Turnbull, A., 2010. Breast Screening: visual search as an aid for digital mammographic interpretation training. In B. Sahiner & D.J. Manning (Eds.) *Proceedings* of *SPIE Medical Imaging 2010:* Image Perception, Observer Performance, and Technology Assessment, Vol 7627.

11. Chen, Y., Gale, A.G., & Evans, A., 2009. The feasibility of vision-supported computer-based training in digital mammography, *Breast Cancer Research 2009*, 11(Suppl 2):P29.

12. Chen, Y., James, J., & Gale, A.G., 2009. How important is mammographic image manipulation when examining digital screening cases?, *Breast Cancer Research 2009*, 11(Suppl 2):P7.

13. Chen, Y., Gale, A.G. & Scott, H.J., 2009. Breast screening: a vision supported healthcare training approach. In *Proceedings of the IEA 17th World Congress on Ergonomics*.

14. Chen, Y., Gale, A., Scott, H., Evans, A. & James, J., 2009. Computer-Based Learning to Improve Breast Cancer Detection Skills. *In Proceedings of the 13th International Conference on Human-Computer Interaction*. Part IV: Interacting in Various Application Domains; Springer; JA Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2009, LNCS 5613, pp. 49–57.

15. Chen, Y., Gale, A.G., Scott, H.J., Wang, X.Y., Liu, H.H. & Xie, A.M., 2009. Delivering mammographic interpretation training in China: can low-cost computer displays help?, *BJR Congress Series -Proceedings of the UK Radiological Congress* 2009, p106.

16. Chen, Y., Gale, A.G. & Scott, H.J., 2009. Mammographic interpretation training in the UK: current difficulties and future outlook. In B. Sahiner & D.J. Manning (Eds.) *Proceedings of SPIE Medical Imaging 2009*: Image Perception, Observer Performance, and Technology Assessment, Vol. 7263, pp C1 – C10.

17. Chen, Y., Gale, A.G. & Scott, H.J., 2008. Mammographic interpretation training: what exactly do film-readers want?, *Breast Cancer Research*, Vol. 10, p S3, 2008.

18. Chen, Y., Gale, A.G., Scott, H.J., 2008. The possibility of 3W (Whatever required, Whenever, Wherever) Mammographic Interpretation Training. *BJR Congress Series - Proceedings of the UK Radiological Congress* 2008, p63.

19. Chen, Y., Gale, A.G. & Scott, H.J., 2008. Anytime, Anywhere Mammographic Interpretation Training. In P.D. Bust (Ed.) *Contemporary Ergonomics* 2008, 375-380. Taylor & Francis, London, 2008.

20. Chen, Y., Gale, A. & Scott, H.J., 2008. Mammographic interpretation training: how useful is handheld technology? In D. Manning and B Sahiner (Eds.). In *Proceedings of SPIE Medical Imaging 2008:* Image Perception, Observer Performance, and Technology Assessment. Vol. 6917;12, 1-10, 2008.

9.5 Conference Presentations

1. Chen Y., James, J., Turnbull, A. & Gale A.G.: Is image manipulation necessary to interpret digital mammographic images efficiently? Presented at the SPIE Medical Imaging Conference, Florida USA, February 2011.

2. Darker I.T., Gale A. G. & Chen Y.: Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme. Presented at the SPIE Medical Imaging Conference, Florida USA, February 2011.

3. Chen Y., James J., Evans A., Turnbull A. & Gale A.G.: An investigation of workstation image manipulation usage when examining FFDM images. Presented at the Royal College of Radiology Breast Group Annual Meeting, Brighton UK, 31st Oct- 2nd November 2010.

4. Darker I.T., Gale A. G. & Chen Y.: Variability in film reader estimates of breast density in the PERFORMS scheme Introduction. Presented at the Royal College of Radiology Breast Group Annual Meeting, Brighton UK, 31st Oct- 2nd November 2010.

5. Chen Y. & Gale, A.G.: Intelligent Computing Applications based on Eye Gaze: their role in Medical Image Interpretation. Presented at the Sixth International Conference on Intelligent Computing, Changsha, China, August 2010.

6. Chen Y., James, J., Turnbull, A. & Gale A.G.: Mammographic interpretation training: how could low cost display devices help?, Presented at the International Workshop on Digital Mammography, Girona, Spain, June 2010

7. Chen Y., James J., Turnbull A., Scott H.J. & Gale A.G.: Is image manipulation necessary in digital mammographic interpretation? Presented at the European Congress of Radiology, Vienna, Austria, March 2010.

8. Chen Y. & Gale A.G.: Using Eye Gaze in Intelligent Interactive Imaging Training, Presented at the International Conference on User Interfaces, Hong Kong, China, February 2010.

9. Chen Y., James, J., Turnbull, A., Evans, A. & Gale, A.G.: Breast Screening: visual search as an aid for digital mammographic interpretation training. Presented at the SPIE Medical Imaging Conference, San Diego USA, February 2010.

10. Chen Y., Gale A.G., James J., Evans A. & Scott H.J.: Breast Screening: visual search as an aid for digital mammographic interpretation training. Presented at the Radiology Society of North America (RSNA) Annual Meeting, Chicago USA, November 2009.

11. Chen Y., Gale A.G. & Evans A.: The feasibility of vision-supported computer-based training in digital mammography. Presented at the Royal College of Radiology Breast Group Meeting, Belfast UK, November 2009.

12. Chen Y., James J. & Gale A.G.: How important is mammographic image manipulation when examining digital screening cases? Presented at the Royal College of Radiology Breast Group Meeting, Belfast UK, November 2009.

13. Chen Y., & Gale A.G.: Eye movements and computer-based mammographic interpretation training, Presented at the Medical Imaging Perception Society conference, Santa Barbra, USA, October 2009.

14. Chen Y., & Gale A.G.: A vision supported healthcare training approach in breast screening, Presented at the15th European Conference on Eye Movements, Southampton UK , August 2009.

15. Chen Y. & Gale A.G.: Breast screening: a vision supported healthcare training approach. Presented at the International Ergonomic Association conference, Beijing China, August 2009.

16. Chen Y., Gale A.G., Scott H.J., Evans A. & James J.: Computer-Based Learning to Improve Breast Cancer Detection Skills. Presented at the 13th International Conference on Human-Computer Interaction, San Diego USA, July 2009.

17. Chen Y., Gale A.G., Scott H.J., Wang X.Y., Liu, H.H. & Xie A.M.: Delivering mammographic interpretation training in China: can low-cost

computer displays help? Presented at the UK Radiological Congress 2009, Manchester UK. June 2009.

18. Chen, Y., Gale A.G. & Scott H.J.: Mammographic interpretation training in the UK: current difficulties and future outlook. Presented at the SPIE Medical Imaging 2009 conference, San Diego, USA, February 2009.

19. Chen Y., Gale A.G. & Scott H.J.: Mammographic interpretation training: what exactly do film-readers want? Presented at Symposium Mammographicum, Lille, France, July 2008.

20. Chen Y., Gale A.G. & Scott H.J.: The possibility of 3W (Whatever required, Whenever, Wherever) Mammographic Interpretation Training. Presented at the UK Radiological Congress 2008, Birmingham UK, June 2008.

21. Chen Y., Gale A.G. & Scott H.J.: Anytime, Anywhere Mammographic Interpretation Training, Presented at the Annual Conference of the Ergonomics Society, Nottingham UK, April 2008.

22. Chen Y., Gale A.G. & Scott H.J.: Mammographic interpretation training: how useful is handheld technology?, Presented at the SPIE Medical Imaging Conference, San Diego USA, February 2008.

,

REFERENCES

American Cancer Society, 2006. *What Are the Key Statistics for Breast Cancer?* Cancer Reference Information.

American Cancer Society, 2008, *Cancer Facts and Figures*.

American College of Radiology (ACR), 1998. *Illustrated Breast Imaging Reporting and Data System (BI-RADS TM)* Reston, Virginia.

American College of Radiology (ACR), 1999. *ACR mammography quality control*. Reston, Virginia

American College of Radiology (ACR), 2007a. *Practice Guidelines for Determinants of Image Quality in Digital Mammography.* Reston, Virginia

American College of Radiology (ACR), 2007b. *Practice guideline for Digital Radiography.* Reston, Virginia

Anastasio, M.A., Kupinski, M.A. & Nishikawa, R.N., 1998. Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach. *IEEE Trans Med Imaging*, 17:1089.

Anderson, E.D.C., Muir, B.B., Walsh, J.S. & Kirkpatrick, A.E., 1994. The efficacy of double reading mammograms in breast screening. *Clinical Radiology*, 49(4), 248-251.

Anderson, B.O., Yip, C., Smith, R.A., Shyyan R., Sener, S.F., Eniu, A., Carlson R.W., Azavedo, E., Harford, J., 2007. Guideline implementation for breast healthcare in low-income and middle-income countries: overview of the Breast Health Global Initiative Global Summit. Cancer 113:2221–43.

Astley, S.M. & Gilbert, F.J., 2004. Computer-aided detection in mammography. *Clinical Radiology*, 59(5) 390-9.

Atchison, D.A. & Smith, G., 2000. Optics of the human eye. *Butterworth-Heinemann Medical*.

Balleyguier, C., Ayadi, S., Van Nguyen, K., Vanel, D., Dromain, C. & Sigal, R., 2007. BIRADS™ classification in mammography. *European Journal of Radiology*, 61(2), pp. 192-194.

Bertling C.J. Simpson, D.E., Hayes, A.M., Torre, D., Brown, D.L. & Schubot, D.B., 2003. Personal digital assistants herald new approaches to teaching and evaluation in medical education. *WMJ*;102:46-50.

Bick, U. & Balleyguier, C., 2010. *Digital Mammography: With 16 Tables*. Springer.

Bird, R.E., Wallace, T.W., & Yankaskas B.C. Analysis of cancers missed at screening mammography. *Radiology* 1992; 184: 613–617.

Blanks, R.G., Wallis, M.G., Given-Wilson, R.M. 1999. Observer variability in cancer detection during routine repeat (incident) mammographic screening in a study of two versus one view mammography. *J Med Screen*, 6:152-158.

Boonn, W.W. & Flanders, A.E. 2005. Survey of Personal Digital Assistant Use in Radiology1. *Radiographics*;25(2):537.

Boonn, W. & Flanders A.E. 2005. Informatics in radiology (infoRAD): survey of personal digital assistant use in radiology. *Radiographics*. Mar-Apr;25(2):537-41.

Breast Cancer Screening. 1987. *Report to the Health Ministers of England, Wales, Scotland and Northern Ireland by a Working Group chaired by Sir Patrick Forrest*. Her Majesty's Stationery Office

Brennan, P.C. McEntee, M., Evanoff, M., Phillips, P., O'Connor, W.T., & Manning, D.J., 2007. Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist. *American Journal of Roentgenology*, 188(2), pp. W177.

Boetes, C., Mus, R.D., Holland, R., Barentsz, J.O., Strijk, S.P., Wobbes, T., Hendriks, J.H., & Ruys, S.H., 1995. Breast tumors: comparative accuracy of MR imaging relative to mammography and US for demonstrating extent. *Radiology*, 197(3), pp. 743.

Bordas, P., Jonsson, H., Nystrom, L. & Lenner, P., 2009. Interval cancer incidence and episode sensitivity in the Norrbotten Mammography Screening Programme, Sweden. *Journal of medical screening*, 16(1), pp. 39.

Bunch, P.C., Hamilton, J.F., Sanderson, G.K. & Simmons, A.H. 1977. A free response approach to the measurement and characterization of radiographic observer performance. *Proc SPIE* 127: 124.

Busch, J.M. Barbaras, L., Wei, J., Nishino, M., Yam, C. & Hatabu., H., 2004. A mobile solution: PDA-based platform for radiology information management. *Am J Roentgenol*183(1):237-42.

Carmody, D.P., Nodine, C.F. & Kundel, H.L., 1980. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, 9(3), 339-344.

Castledine, J., 2002. *Background Paper: Estimated Training Requirements for the Expansion of the NHS Breast Screening Programme in England* [Unpublished report]. NHS Cancer Screening Programmes.

Chakraborty, D. P. & Berbaum, K. S., 2004. Observer studies involving detection and localization: Modeling, analysis and validation. *Medical Physics* 31 (8), 2313-2330.

Chakraborty, D.P., Jackknife Free-Response Receiver Operating Characteristic Analysis Software, [Computer software], Available at: www.devchakraborty.com [Accessed 15th January 2010].

Chakraborty D.P., 1989. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med Phys*, 16: 561.

Chakraborty D.P. & Winter L.H.L., 1990. Free-response methodology: alternate analysis and a new observer performance experiment. *Radiology*, 174: 873.

Chakraborty, D.P., 2008. Validation and statistical power comparison of methods for analyzing free-response observer performance studies. *Academic Radiology*, 15(12), pp. 1554-1566.

Chen, Y. & Gale, A.G., 2008. 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* 6917, Volume 6917, pp. 691712-10.

Chen, Y., Gale, A.G. & Scott H., 2008. *Anytime, Anywhere Mammographic Interpretation Training*. In P.D. Bust (Ed.) Contemporary Ergonomics. Taylor & Francis, London, 2008.

Chen Y, Gale A, Scott H, Evans A, James J., 2009. Computer-Based Learning to Improve Breast Cancer Detection Skills. *Proceedings of the 13th International Conference on Human-Computer Interaction*. Part IV: Interacting in Various Application Domains; Springer; JA Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2009, LNCS 5613, pp. 49–57, 2009.

Chen Y. & Gale A.G., 2009. Breast screening: a vision supported healthcare training approach. In *Proceedings of the 17th World Congress on Ergonomics:* IEA 2009

Chen, Y., Gale, A.G. & Scott, H.J., 2008. Mammographic interpretation training profile in the UK: current difficulties and future outlook. In Image Perception, Observer Performance, and Technology Assessment. D. Manning and B Sahiner (Eds.). *Proceedings of SPIE* 6917;12, 1-10, 2008

Chen Y., Scott H.J. & Gale A.G.: Breast Screening: visual search as an aid for digital mammographic interpretation training, In Image Perception, Observer Performance, and Technology Assessment. D.J. Manning & C. Abbey (Eds.), *Proceedings of SPIE* 7627, Volume 7627, pp. 7627-0C.

China Daily, 2007. More Chinese women getting breast cancer. *China Daily,* October 30, 2007.

China Statistical Yearbook 2001, 2002, People's Republic of China: China Statistics Press, Beijing.

Ciatto, S.*,* Ambrogetti, D., Bonardi, R., Catarzi, S., Risso, G., M Rosselli Del Turco & Mantellini, P*.,* 2005. Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme*. Journal of medical screening,* 12(2), pp. 103.

Cornford, E., Reed, J., Murphy,A., Evans, A. & Bennett, R., 2009. Optimal mammography reading volumes: evidence from real life. Breast Cancer Research, 11(Suppl 2), pp. O2.

Cornsweet, T. & Crane, H., 1973. Accurate two-dimensional eye tracker using first and fourth Purkinje images. J. Opt. Soc. Am. 63, 921-928.

Cowley, H., Gale, A. & Wilson, R., 1996. Mammographic training sets for improving breast cancer detection. in Medical Imaging: Image Perception and Performance edited by Miguel P. Eckstein & D.P. Chakraborty, Proceedings of SPIE Vol. 2712, pp. 102-112.

Curcio, C.A., Sloan, K.R., Kalina, R.E. & Hendrickson, A.E., 1990. Human photoreceptor topography. The Journal of comparative neurology, 292(4), pp. 497-523.

Curado M.P., Edwards B., Shin H.R., *et al, 2007, Cancer Incidence in Five Continents, Vol. IX* . Lyon , France: World Health Organization Press.

Day N.E. 1985, Estimating the sensitivity of a screening test. *J Epidemiol Community Health*, 39:364-66.

Darker, I.T., Kuo, P., Yang, M-Y., Blechko, A., Grecos, C., Makris, D., Nebel, J-C., Gale, A.G. 2009. Automation of the CCTV-mediated detection of individuals illegally carrying firearms: combining psychological and technological approaches. In Z. Rahman, S. E. Reichenbach, M. A. Neifeld (Eds.) *Proceedings of SPIE Defense, Security, and Sensing* Vol. 7341, 73410P.

Department of Health, 1986. *Breast Cancer Screening Report to the Health Ministers of England*, *Wales*, *Scotland and Northern Ireland by a Working Group Chaired by Sir Patrick Forrest*. London, HMSO.

Department of Health, 2007. *Cancer 2012 Visions*. London.

Department of Health, 2007. *Breast Cancer Screening Annual Report*. London.

Department of Health, 2007. *Cancer Reform Strategy*. London .

Department of Health, 2004. *Manual for cancer services*. London.

Department of Health, 2000. *The NHS Cancer plan: a plan for investment, a plan for reform.* London.

Diagnostic imaging, 2010. *FDA ruling on mobile image viewers could shake iphone app market* [online] (Updated 30 March 2010) Available at: < http://www.fiercemobilehealthcare.com/story/fda-ruling-mobile-image-viewers-could-shake-iphone-app-market/ > [Accessed 7 December 2010].

Dilorio C., Hockenberry-Eaton M., Maibach E., & Rivero T. 1994. Focus groups: an interview method for nursing research. *Journal of Neuroscience Nursing* 26(3, 175–180).

Dorfman, D. D., Berbaum, K. S., & Metz, C. E., 1992. Receiver Operating Characteristic Rating Analysis: Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology*, 27, 723-731.

Dorfman, D.D., Berbaum, K. S., Lenth, R.V., Chen, Y. & Donaghy, B.A., 1998. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. *Academic Radiology*, 5:591– 602.

Duchowski, A.T., 2007. *Eye tracking methodology: Theory and practice*. Springer-Verlag, New York Inc.

Eberl, M.M., Fox, C.H., Edge, S.B., Carter, C.A. & Mahoney, M.C., 2006. BI-RADS classification for management of abnormal mammograms. *The Journal of the American Board of Family Medicine*, 19(2), pp. 161.

Edwards, D.C., Kupinski M.A., Metz C.E. & Nishikawa R.N., 2002. Maximum-likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. *Med Phys*, 29: 2861.

Elmore, J.G., Jackson S.L., Abraham, *L.,* Miglioretti, D.L., Carney, P.A., Geller, B.M., Yankaskas, B.C., Kerlikowske, K., Onega, T., Rosenberg, R.D., Sickles, E.A. & Buist, D.S., 2009. Variability in Interpretive Performance at

Screening Mammography and Radiologists' Characteristics Associated with Accuracy. *Radiology*, 253(3), pp. 641.

Elmore, J.G., Wells, C.K., Lee, C.H., Howard, D.H. & Feinstein, A.R., 1994. Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine*, 331(22), pp. 1493.

Eng. J., 2003, Sample Size Estimation: How Many Individuals Should Be Studied? *Radiology*, 227:309-313.

Espey, D. K., Wu, X.C., Swan, J., Wiggins, C., Jim, M.A., Ward, E., Wingo, P.A., Howe, H.L., Ries, L.A., Miller, B.A., Jemal, A., Ahmed, F., Cobb, N., Kaur, J.S. & Edwards, B.K., 2007. Annual report to the nation on the status of cancer, 1975-2004, featuring cancer in American Indians and Alaska Natives. Cancer 110 (10), 2119-52.

Europa Donna, 2007. A Short Guide to the European guidelines for quality assurance in breast cancer screening and diagnosis. [online] Available at: <http://www.europadonna.org/europadonna/getStaticModFile.aspx?id=2844>.

Evans, A., 2010. UK screening: are the age extensions the best way to improve it? *Royal College of Radiology Breast Group Annual scientific Meeting*, Brighton, United Kingdom, 31st October-2nd Novemeber,

Ferlay, J., Parkin, D.M. & Steliarova-Foucher, E. G., 2008. Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer. [online] Available at: http://globocan.iarc.fr

Ferlay, J., Parkin, D.M. & Steliarova-Foucher, E., 2010. Estimates of cancer incidence and mortality in Europe in 2008. *Eur J Cancer*. 46(4):765-81. Epub 2010 Jan 29.

Findlay, J. M., 1974. A simple apparatus for recording microsaccades during visual fixation. *Quarterly Journal of Experimental Psychology*, 26, 167-170.

Flanders, A.E., Wiggins, R.H. III & Gozum, M.E., 2003. Handheld computers in radiology. *RadioGraphics,* 23:1035-47.

Fleissig, A., Jenkins, V., Catt, S. & Fallowfield, L., 2006. Multidisciplinary teams in cancer care: are they effective in the UK? *The Llancet Oncology*, 7(11), pp. 935-943.

Fornvik, D., Andersson, I., Svahn, T., Timberg, P., Zackrisson, S., & Tingberg, A., 2010. The effect of reduced breast compression in breast tomosynthesis: human observer study using clinical cases. *Radiation Protection Dosimetry*, 139(1-3), pp. 118.

Fryback, D.G. & Thornbury, J.R., 1991. The efficacy of diagnostic imaging. *Medical Decision Making*, 11(2), 88.

Gale, A.G., 1997. Human response to visual stimuli. In Hendee W. & Wells P. (Eds.) *Perception of Visual Information - Second Edition*, (New York) Springer Verlag,.

Gale, A.G., Murray, D., Millar, K. & Worthington, B.S., 1984. Circadian variation in radiology, Theoretical and applied aspects of eye movement research. Proceedings of the Second European Conference on Eye Movements, Nottingham, England, North Holland, pp. 313.

Gale, A.G. & Scott, H. 2010. Measuring Radiology Performance in Breast Screening. In M. Michell Contemporary Issues in Cancer Imaging – Breast Cancer, Cambridge UP, Cambridge.

Gale, A.G. & Walke, G.E. 1991. *Design for performance: Quality assessment in a national breast screening programme*. In Lovesay E. (Ed.) Ergonomics - design for performance, Taylor & Francis, London.

Gale, A., 2010. Maintaining quality in the UK breast screening program, *Proceedings of SPIE, 2010*, pp. 762702.

Gale, A.G., Roebuck, E.J., Riley, P. & Worthington, B.S. Computer aids to mammographic diagnosis. *British Journal of Radiology*, 1987, 60, 887-891

Gale, A.G. 1997. *Human response to visual stimuli*. Perception of Visual Information - second edition, New York: Springer Verlag,.

Gale, A.G., 2003. PERFORMS-a self assessment scheme for radiologists in breast screening. *Seminars in Breast Disease* 6(3), 148-52.

Gale, A.G. & Scott, H.J., 2003. Measuring Radiology Performance in M. Michell (ed.) Breast Screening. In *Contemporary Issues in Cancer Imaging – Breast Cancer*, Cambridge University Press.

Gennaro, G. & Di Maggio, C., 2006. Dose comparison between screen/film and full-field digital mammography. *European Radiology*, 16(11), 2559-2566.

Gilbert, F.J., Astley, S.M., Gillan, M.G., Agbaje, O.F., Wallis, M.G., James, J., Boggis, C.R., Duffy, S.W., 2008. CADET II Group. Single reading with computer-aided detection for screening mammography. *New. England Journal of Medicine*, 359(16):1675-84.

GIMI, [online] Available at: <http://www.gimi.ox.ac.uk/>. [Accessed 18 Jan, 2008].

Goodale, M.A. & Milner, A.D., 1992. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1), pp. 20-25.

He, J., Gu, D., Wu, X., Reynolds, K., Duan X., Yao, C., Wang, J., Chen, C., Chen, J., Wildman, R. P., Klag, M.J. & Whelton, P.K., 2005. Major causes of death among men and women in China. *New England Journal of Medicine*, 353(11):1124-34.

Hellinger J., Ramen, R., Ramen, B., Ramen, L. & Beaulieu, C., 2004. Mobile Image Interpretation Using Handheld Personal Digital Assistants: Preliminary Clinical Validation Using Head and Chest CT. RSNA scientific poster 2004, Chicago.

Hesketh, T., Lu, L. & Xing, Z.W., 2005. The effect of China's one-child family policy after 25 years. *N Engl J Med 2005; 353:1171-6 Diagnostic Imaging, China chooses CR for breast screening*, [online] Available at : <http://www.diagnosticimaging.com/digital-x-ray/content/article/113619 /1242697> [Accessed 23rd Oct, 2010].

Heywang, S.H., Wolf, A., Pruss, E., Hilbertz, T., Eiermann W. & Permanetter, W., 1989. MR imaging of the breast with Gd-DTPA: use and limitations. *Radiology*, 171(1), pp. 95.

Hogan, M.J., Alvarado, J.A. & Weddell, J.E., 1971. Histology of the human eye: an atlas and textbook. *Saunders Philadelphia*.

Horii, S.C., 1989. Environmental designs for reading from imaging work stations: ergonomic and architectural features. *J Digit Imaging*, 2:156-162.

Hu, C. & Fang, Z., 2007. The current situation of breast screenining in China, *China J Breast Dis* (Electronic Version), Vol.1, No.5.

International Agency for Research on Cancer, 2007. *Cancer screening: Breast.* [online] Available at: <http://eu-cancer.iarc.fr/cancer-13-breast-screening.html,en#block-20-50> [Accessed 10 October 2009].

Istepanian, R.S.H., Pattichis, C.S. & Laxminarayan, S., 2006. *M-Health: Emerging Mobile Health Systems.* Springer.

James, J.J., Gilbert, F.J., Wallis, M.G., Gillan, M.G., Astley, S.M., Boggis, C.R., Agbaje, O.F., Brentnall, A.R. & Duffy, S.W., 2010. Mammographic Features of Breast Cancers at Single Reading with Computer-aided Detection and at Double Reading in a Large Multicenter Prospective Trial of Computer-aided Detection: CADET II. *Radiology*, 256(2), pp. 379.

Jemal, A., Ward, E. & Thun, M., 2007. Recent trends in breast cancer incidence rates by age and tumor characteristics among US women*. Breast Cancer Research*, 9(3), R28.

Kalser, W. & Zeitler, E., 1989. MR imaging of the breast: fast imaging sequences with and without Gd-DTPA. *Preliminary observations. Radiology*, 170(3), pp. 681.

Kong, L.(ed.) 2003. *CHINA Overview: National Cancer Control Plan*. Ministry of Health.

Kimme-Smith, C., Haus, A., Debruhl, N. and Bassett, L., 1997. Effects of ambient light and view box luminance on the detection of calcifications in mammography. *American Journal of Roentgenology*, 168(3), pp. 775.

Klabunde, C.N., Sancho-Garnier, H., Taplin, S., Thoresen, S., Ohuchi, N., & Ballard-Barbash, R., 2002. International Breast Cancer Screening Network. Quality assurance in follow-up and initial treatment for screening mammography programs in 22 countries. *Int J Qual Health Care,* 14(6):449-61.

Krupinski, E.A., Nodine, C.F., Kundel, H.L., 1998. Enhancing recognition of lesions inradiographic images using perceptual feedback. *Opt Eng*, 37: 813–818.

Krupinski, E.A. & Nodine, C.F., 1994. *Gaze duration predicts the location of missedlesions in mammography.* In Gale AG editor. Digital mammography. Amsterdam: Elsevier Science.

Krueger, R.A., 1995. The future of focus groups. *Qualitative Research* 5(4), 524-530

Kundel, H.L., Nodine, C.F. & Carmody. D., 1978. Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. *Investigative Radiology*, 13(3):175-181.

Kundel, H.L., Nodine, C.F. & Krupinski, E.A., 1990. Computer displayed eye position as a visual aid to pulmonary nodule interpretation. *Investigative Radiology*, 25:890–896.

Kuhl, C.K., Schrading, S., Leutner, C.C., Morakkabati-Spitz, N., Wardelmann, E., Fimmers, R. & Kuhn, W., 2005. Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *Journal of clinical oncology,* 23(33), pp. 8469.

Lewin, J.M., D'Orsi, C.J., Hendrick, R.E., Moss, L.J., Isaacs, P.K., Karellas, A, & Cutter G.R., 2002. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. *American Journal of Roentgenology*, 179(3), 671.

Liddell, H., 1919. Eye-movement during fluctuation of attention. *The American Journal of Psychology,* 30(3), pp. 241-252.

Linos, E., Spanos, D., Rosner, B.A., Linos, K., Hesketh, T., Qu, J.D., Gao, Y.T., Zheng, W., & Colditz, G.A., 2008. Effects of Reproductive and Demographic Changes on Breast Cancer Incidence in China: A Modelling Analysis. *Journal of the National Cancer Institute,* 100:1352-60-6

Li, H. J., 2009. The Plight of China's Breast Cancer Screening Program. *China Today*, [online] Available at: <http://www.chinatoday.com.cn/ctenglish/se/txt/2009-03/30/content_188455.htm> [Accessed 25 Oct 2010].

Li, H.P., Ji, J.F., Hou, K.Y., Lei, Y.T., Zhao, H.M., Wang, J., Zheng, J., Liu, J.Y., Wang, M.P., Xiao, Y., Wang, Y.F. & Lü, Y.Y., 2010. Prediction of recurrence risk in early breast cancer using human epidermal growth factor 2 and cyclin A2. *Chin Med J,* 123: 431-437.

Liu X. & Gale A.G., 2007. Searching for possible threat items to safe air travel: human error and training. Stephanidis C. (Ed.). *In Lecture Notes in Computer Science series, Springer*.

Ma, X.J., Lin, C. & Zhen, W., 2008. Cancer care in China: A general review. Biomedical Imaging and Intervention Journal, 4(3), e39.

Machworth, N.H., 1976. *Stimulus density limits the useful field of view*. In Monty RA, Senders JW(ed), Eye movement and psychological processes. Hillsdale, NJ, Erlbaum.

Mackworth, N. F. & Mackworth, H. H., 1958. Eye fixations recorded on changing visual scenes by the television eye-marker. *Journal of the Optical Society of America*, 48, 439.

Mao, L.J., 2010. Fighting breast cancer in the Chinese countryside, *China Daily* Online, [online] Available at: <http://www2.chinadaily.com.cn/cndy/2010-03/15/content_9588636.htm>. [Accessed 25 Oct 2010].

Mann, R.M., Kuhl, C.K., Kinkel, K., Boetes, C., 2008. Breast MRI: Guidelines from the European Society of Breast Imaging. *Eur Radiol,* 18(7): 1307-18.

Mann, R.M., Veltman, J., Barentsz, J.O., Wobbes, T., Blickman, J.G., Boetes, C., 2008. The value of MRI compared to mammography in the assessment of tumour extent in invasive lobular carcinoma of the breast. *European Journal of Surgical Oncology (EJSO)*, 34(2), pp. 135-142.

Manning, D.J., Ethell, S.C. & Donovan, T., 2004. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *British Journal of Radiology*, 77(915), 231.

Maxwell, A.J., Ridley, N.T., Rubin, G., Wallis, M.G., Gilbert, F.J., Michell, M.J., 2009. The Royal College of Radiologists Breast Group breast imaging classification. *Clinical Radiology* Volume 64, Issue 6, Pages 624-627.

Maxwell, S.E., Delaney, & Harold D., 1990. *Designing experiments and analyzing data: a model comparison perspective.* Belmont, Calif.:Wadsworth Pub. Co.

McNicol, D., 1972. *A primer of signal detection theory; Chapter 5*. George Allen & Unwin Ltd..

Mello-Thoms, C., 2010. Visual search characteristics in mammogram reading: SFM vs. FFDM*, Proceedings of SPIE*, pp. 76270D.

Mello-Thoms, C., Dunn, S.M., Nodine, C.F. & Kundel, H.L., 2001. An analysis of perceptual errors in reading mammograms using quasi-local spatial frequency spectra. *J Digit Imaging*, 14: 117–123.

Mello-Thoms, C., Dunn, S., Nodinem, C.F., Kundel, H.L. & Weinstein, S.P., 2002. The perception of breast cancer: what differentiates missed from reported cancers in mammography? *Acad Radiol,* 9: 1004–1012.

Merchant, J., Morrissette, R. & Porterfield, J.L., 1974. Remote measurement of eye direction allowing subject motion over one cubic foot of space. *Biomedical Engineering, IEEE Transactions on*, (4), 309-317.

Metz, C.E., 1978. Basic principles of ROC analysis, Seminars in nuclear medicine. *Elsevier*, pp. 283-298.

Metz, C.E., 2008. ROC analysis in medical imaging: a tutorial review of the literature. *Radiological Physics and Technology*, 1(1), pp. 2-12.

Metz, C. E., 2003, ROC Analysis: Methods and Practical Applications. *Presentation notes at AAPM 45th of Annual Meeting*, San Diego, USA

Mowrer, O. H., Ruch, R. C., & Miller, N. E. 1936. The corneoretinal potential difference as the basis of the galvanometric method of recording eye movements. *American Journal of Physiology*, 114, 423.

Nakata, N., Kandatsu, S., Suzuki, N. & Fukuda, K., 2005. Informatics in Radiology (infoRAD): Mobile wireless DICOM server system and PDA with high resolution display: feasibility of group work for radiologists. *RadioGraphics,* 25, 273-83.

National Statistics, 2007. *Breast Cancer: Incidence rises while deaths continue to fall*. UK National Statistics Office.

National Institute for Health and Clinical Excellence 2009. *Breast Cancer (Early & Locally Advanced): Diagnosis and Treatment*. London: NICE.

NHSBSP, 2005. *Quality Assurance Guidelines for Breast Cancer Screening Radiology.* NHSBSP Pub. No. 59, (2005).

NHSBSP, 2009. *What happens at an NHS Breast Screening Unit?* [online] Available at: <http://www.cancerscreening.nhs.uk/breastscreen/bsu.html> [Accessed: 23rd Oct, 2010].

Nickerson, C. & Cush, S., 2004. New ways of working in the NHS breast Screening Programme, Second report on Implementation. *NHS Breast Screening Programme Publication.*

Nodine, C.F., Mello-Thoms, C., Weinstein, S.P., Kundel, H.L., Conant, E.F., Heller-Savoy, R.E., Rowlings, S.E. & Birnbaum, J.A., 2001. Blinded Review of Retrospectively Visible Unreported Breast Cancers: An Eye-Position Analysis. *Radiology*, 221,122-129.

Nodine, C.F., Kundel, H.L., Mello-Thoms, C., Weinstein, S.P., Orel, S.G., Sullivan, D.C. & Conant, E.F., 1999. How experience and training influence mammography expertise. *AcadRadiol*, 6: 575–785.

Nodine, C.F. & Kundel, H.L., 1987. Using eye movements to study visual search and to improve tumor detection. *Radiographics,* 7: 1241–1250.

Niklason, L.T., Christian, B.T., Niklason, L.E., Kopans, D.B., Castleberry, D.E., Opsahl-Ong, B.H., Landberg, C.E., Slanetz, P.J., Giardino, A.A., Moore, R., Albagli, D., DeJule, M.C., Fitzgerald, P.F., Fobare, D.F., Giambattista, B.W., Kwasnick, R.F., Liu, J., Lubowski, S.L., Possin, G.E., Richotte, J.F., Wei, C. & Wirth, R.F., 1997. Digital tomosynthesis in breast imaging. *Radiology*, 205(2):399–406.

O'donnell, A.B., Lutfey, K.E., Marceau, L.D. & Mckinlay, J.B., 2007. Using focus groups to improve the validity of cross-national survey research: A study of physician decision making. Qualitative health research, 17(7), pp. 971.

Patnick, J. eds. 2005. One vision. *NHS Breast Screening Programme Annual Review 2005*. NHS Cancer Screening Programmes.

Patnick, J. eds. 2008. *Celebrating 20 years of screening, NHS Cancer Screening Programmes 2008*. NHS Breast Screening Programme.

Patnick, J. eds. 2009. *Expanding our reach, NHS Cancer Screening Programmes Annual Review 2009*. NHS Breast Screening Programme.

Patnick, J. eds. 2005. *One vision. NHS Breast Screening Programme Annual Review 2005*. NHS Cancer Screening Programmes.

Park, J.M., Franken, E.A.Jr, Garg, M., Fajardo, L.L., Niklason, L.T., 2007. Breast Tomosynthesis: Present Considerations and Future Applications. *Radiographics*, 27(suppl 1), pp. S231.

Parkin D.M., Whelan S.L., Ferlay J., Storm H., 2005, *Cancer Incidence in Five Continents, Vol. I to VIII*. Lyon, France: World Health Organization Press.

People's Daily Online, 2005. *One million Chinese women to get breast screening to avoid breast cancer*. [online] Available at:<Http://english .peopledaily.com.cn/200504/21/eng20050421_182197.html#> [Accessed 25 Oct 2010].

Perez, L. Huang, J., Jansky, L., Nowinski, C., Victorson, D., Peterman, A., Cella, D., 2007. Using focus groups to inform the Neuro-QOL measurement tool: exploring patient-centered, health-related quality of life concepts across neurological conditions. *Journal of Neuroscience Nursing*, 39(6), pp. 342.

Perry, N., Broeders, M., de Wolf, C., Törnberg, S., Holland, R., von Karsa, L., 2008. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Annals of Oncology*, 19(4), 614-622.

Performs Report to National Breast Screening QA Radiologists Group 2010. *Personal performance in Mammographic Screening* SA09 Round.

Phillips, P.W. Crawford, T., Higham, S., Manning, D.J. and Donovan, T., 2005. A Software Framework for Diagnostic Medical Image Perception with Feedback, and a Novel Perception Visualisation Technique. *Proc. SPIE* Vol. 5749, 572-580.

Pisano, E.D., Gatsonis, C., Hendrick, R.E., Yaffe, M., Baum, J.K., Acharyya, S., Conant, E.F., Fajardo, L.L., Bassett, L, D'Orsi, L., Jong, R., and Rebner, M., 2005. Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group. Diagnostic performance of digital versus film mammography for breast-cancer screening*. N Engl J Med*, 353(17), 1773-1783.

Pisano, E.D., Hendrick, R.E., Yaffe, M.J., Baum, J.K., Acharyya, S., Cormack, J.B., Hanna, L.A., Conant, E.F., Fajardo, L.L., Bassett, L.W., D'Orsi, C.J., Jong, R.A., Rebner, M., Tosteson, A.N.A. and Gatsonis, C.A. 2008. DMIST Investigators Group. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. *Radiology*, 246(2), 376-383.

Poplack, S.P., Tosteson, T.D., Kogel, C.A. & Nagy, H.M., 2007. Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. *American Journal of Roentgenology*, 189(3), pp. 616.

Radiation Advisory Committee: Subcommittee on quality assurance (The Pritchard Committee), 1998. *The Pritchard Report Guidelines on the establishment of a quality assurance system/or the radiological aspects of mammography used/or breast screening*.

Raman, B., Raman, R., Raman, L. & Beaulie, C.F., 2004. Radiology on handheld devices: image display, manipulation, and PACS integration issues*. Radio- Graphics,* 24:299-310.

Roe, C.A., Metz, C.E., 1997. Dorfman–Berbaum–Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. *Academic Radiology*, 4:587– 600.

Rong, X., 2000. Breast Cancer Rates Higher among Chinese Immigrants: Balanced Diet and Screening Best Prevention. *Sing Tao Daily.* Published Oct. 22,2000.

Ruggiero, C., 1998. Teleradiology: a review. *Journal of Telemedicine and Telecare*, 4, 25-35(11).

Salvucci, D.D. & Goldberg, J.H., 2000. Identifying fixations and saccades in eye-tracking protocols, *Proceedings of the 2000 symposium on Eye tracking research & applications, 2000, ACM* , pp71-78.

Samei, E., Saunders, R.S., Baker, J.A. & Delong, D.M., 2007. Digital mammography: effects of reduced radiation dose on diagnostic performance. *Radiology*, 243:396–404.

Samuelson, F.W. & Petrick, N., 2006. Comparing image detection algorithms using resampling, *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on, 2006, IEEE*, pp. 1312-1315.

Samuelson, F.W. & Petrick, N., Paquerault, S., 2007. Advantages and examples of resampling for cad evaluation, *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium* on, 2007, IEEE, pp. 492-495.

Saslow, D., Boetes, C., Burke, W., Harms, S., Leach, M.O., Lehman, C.D., Morris, E., Pisano, E., Schnall, M., Sener, S., Smith, R.A., Warner, E., Yaffe, M. & Andrews, K.S., 2007. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. *CA: A Cancer Journal for Clinicians*, 57(2), pp. 75.

Schweitzer, T., Engelmann, U., Schröter, A., Borälv, E., Meinzer, H., 2002. Teleradiology on a Personal Digital Assistant. *2nd Conference on Mobile Computing in Medicine 15, workshop of the Project Group MocoMed* 2002, pp73-80.

Scott, H.J. & Gale, A.G., 2005. Breast screening: when is a difficult case truly difficult and for whom?. *Proceedings of SPIE* 2005:5749, 557-65.

Scott H.J. & Gale A.G., 2007. How much is enough: factors affecting the optimal interpretation of breast screening mammograms. In Image Perception, Observer Performance, and Technology Assessment. Y Jiang and B Sahiner (Eds.) *Proceedings of SPIE 2007: 6515, 65150F*

Scott, H.J., Yap, M.H. & Gale, A.G., 2008. Individualised Grid-based Training and Self-assessment in Breast Screening. *Radiological Society of North America (RSNA*), *Scientific Poster Presentation*, 01 Dec 2008.

Scott, H.J., Gale, A.G., 2006. Breast screening: PERFORMS identifies advanced mammographic training needs*. British Journal of Radiology*, 79, S127-S133.

Scott, H.J. & Gale, A.G., 2005. Breast screening: when is a difficult case truly difficult and for whom?. *Proceedings of SPIE 2005* 5749, 557-65.

Shapiro, S., Coleman, E.A., Broeders, M., Codd, M., de Koning, H., Fracheboud, J., Moss, S., Paci, E., Stachenko, S., Ballard-Barbash, R., 1998. Breast cancer screening programmes in 22 countries: current policies, administration and guidelines*. International journal of epidemiology*, 27(5), pp. 735.

Simpson, A., Slaymaker, M., Power, D.J., Russell, D., Yap, M.H. & Gale, A.G., 2009. On the utilisation of a service-oriented infrastructure to support radiologist training. *Proceedings of CBMS 2009.*

Sickles, E.A., 1984. Mammographic features of "early" breast cancer. *American Journal of Roentgenology*, 143(3), 461.

Sickles, E.A., 1986. Mammographic features of 300 consecutive nonpalpable breast cancers. *American Journal of Roentgenology*, 146(4), 661.

Sickles, E.A., Wolverton, D.E. & Dee, K.E., 2002. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*, 224:861-869.

Siddiqui, K.M., Chia, S., Knight, N. & Siegel, E., 2006. Design and ergonomic considerations for the filmless environment. *JACR*,3:456-467.

Skaane, P., Diekmann, F., Balleyguier, C., Diekmann, S., Piguet, J.C., Young, K., Abdelnoor & M., Niklason, L., 2008. Observer variability in screen-film mammography versus full-field digital mammography with soft-copy reading. *European radiology*, 18(6), pp. 1134-1143.

Skaane, P., 2010. Digital Mammography in European Population-Based Screening Programs*. Digital Mammography*, 155-173.

Skaane, P., Young, K. & Skjennald, A., 2003. Population-based Mammography Screening: Comparison of Screen-Film and Full-Field Digital Mammography with Soft-Copy Reading—Oslo I Study1. *Radiology*, 229(3), 877.

Smith, A., 2005. Full-field breast tomosynthesis. *Radiol Manage*,27(5):25–31.

Smith-Bindman, R., Chu, P.W., Miglioretti, D.L., Sickles, E.A., Blanks, R., Ballard-Barbash, R., Bobo, J.K., Lee, N.C., Wallis, M.G., Patnick, J.,

Kerlikowske, K., 2003. Comparison of screening mammography in the United States and the United Kingdom. *JAMA*, 290(16), pp. 2129.

Smith, F.W., Hutchison, J.M., Mallard, J.R., Johnson, G., Redpath, T.W., Selbie, R.D., Reid, A., Smith, C.C., 1981. Oesophageal carcinoma demonstrated by whole-body nuclear magnetic resonance imaging. *British Medical Journal* (Clinical research ed.), 282(6263), pp. 510.

Smith, G., 2006. *Introduction to RIS and PACS, PACS*, Springer pp9-25.

Snell, R.S. and Lemp, M.A., 1998. *Clinical anatomy of the eye*. Wiley-Blackwell.

Society for Women's Health Research, 2005. *Women's Fear of Heart Disease Has Almost Doubled in Three Years, But Breast Cancer Remains Most Feared Disease.* [Press Release.]

Soutter, J., Campos, J., Hartswood, M., Jirotka, M., Procter, R., Slack, R. & Taylor, P., 2003. Grid-based mammography training, *Hospital Radiologist* 5(6).

Straw, R.B. & Smith, M.W., 1995. Potential uses of focus groups in federal policy and program evaluation studies. *Qualitative Health Research* 5(4), 421 427

Stavros, A.T., Rapp, C.L. & Parker, S.H., 2004. *Breast ultrasound*. Lippincott Williams & Wilkins.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer. cancer.gov) SEER*Stat Database: Incidence-SEER 9 Regs Limited-Use, Nov 2007 Sub (1973 – 2005) Katrina/Rita Population Adjustment> - Linked To

County Attributes - Total U.S., 1969 – 2005 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2008, based on November 2007 submission .

Svahn, T., 2007. Dose reduction and its influence on diagnostic accuracy and radiation risk in digital mammography: an observer performance study using an anthropomorphic breast phantom*. Br J Radiol* 80:557–562.

Svahn, T., 2007. In-plane artifacts in breast tomosynthesis quantified with a novel contrast-detail phantom.

Swensson R.G., 1996. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996; 23.

Tabár, L., Tot, T. & Dean, P.B., 2008. *Breast cancer: the art and science of early detection with mammography.* Thieme Medical.

Tabar, L., Gad, A., Holmberg, L. & Ljungquist, U., 1985. Reduction in Mortality From Breast Cancer after Mass Screening with Mammography. Randomised Trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *The Lancet,* 325(8433), 829-832.

Taylor, P., 2010. Scoring Systems in Computer-Based Training for Digital Mammography. *Digital Mammography*, pp. 490-496.

Toomey, R.J., 2007. The Use of Personal Digital Assistants in the Reporting of Radiological Images. *Medical Image Perception Society (MIPS)* Iowa City, Iowa, USA. 17-20th October 2007

Toomey, R. J., Ryan, J. T., McEntee, M. F., McNulty, J., Evanoff, M. G., Cuffe, F., Yoneda, T., Stowe, J. & Brennan, P. C., 2009. 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* 7263:726307-11

UICC (Union for International Cancer Control), 2010. *Breast Cancer Screening Education programme – China* [Online]. Available at: <http://www.uicc.org/programmes/breast-cancer-screening-education-programme-china> [Accessed 8 December 2010].

Uffmann, M., Prokop, M., Kupper, W., Mang, T., Fiedler, V., Schaefer-Prokop, C., 2005. Soft-copy reading of digital chest radiographs: effect of ambient light and automatic optimization of monitor luminance. *Investigative radiology*, 40(3), pp. 180.

Van Dijck, J., Verbeek, A., Hendriks, J. & Holland, R., 1993. The current detectability of breast cancer in a mammographic screening program: a review of the previous mammograms of interval and screen-detected cancers. *Cancer,* 72:1933-1938.

Van Erkel, A.R. & Pattynama, P.M.T., 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology*, 27(2), pp. 88-94.

Van Goethem, M., Tjalma, W., Schelfout, K., Verslegers, I., Biltjes, I. & Parizel, P., 2006. Magnetic resonance imaging in breast cancer. European Journal of Surgical Oncology, 32(9), pp. 901-910.

Vinnicombe, S., Pinto Pereira, S.M., McCormack, V.A., Shiel, S., Perry, N. & Dos Santos Silva, I.M., 2009. Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. *Radiology*, 251(2), 347.

Wald, N.J., Murphy, P., Major, P., Parkes, C., Townsend, J. &Frost, C., 1995. Multicentre randomised controlled trial of one and two view mammography in breast cancer screening BMJ, *UKCCCR*, 311(7014): p. 1189-93

Warren Burhenne, L.J., Wood, S.A., D'Orsi, C.J., Feig, S.A., Kopans, D.B., O'Shaughnessy, K.F., Sickles, E.A., Tabar, L., Vyborny, C.J., Castellino, R.A., 2000. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology*, 215:554-562.

White G.E. & Thomson A.N., 1995. Anonymized focus groups as a research tool for health professionals. *Qualitative Health Research,* 5(2), 256 261

Wiggins, R.H. III., 2003. Handhelds in radiology. *Semin Ultrasound CT MRI*, 24(6):434-41.

World Health Organization, 2006. *Cancer Fact sheet 297*.

World Health Organization, 2005. *Screening for Breast Cancer*. [online] Available at:<http://www.who.int/cancer/detection/breastcancer/en/print.html>. [Accessed 22 Oct 2010].

World Health Organization, 1982. *Quality Assurance in Diagnostic Radiology*.. Geneva, Switzerland.

Xin, H. & Frey, E., 2009. ROC, LROC, FROC, AFROC: An Alphabet soup. *Journal of the American College of Radiology.* 6:652-655.

Yaffe, M.J., 2010. Basic Physics of Digital Mammography. *Digital Mammography*, 1-11.

Yang L., Parkin, D.M., Ferlay, J., Li, L., Chen, Y., 2005. Estimates of cancer incidence in China for 2000 and projections for 2005. *Cancer Epidemiol Biomarkers Prev,* 14: 243-250.

Yap, M.H. and Gale, A.G., 2009. Individualised grid-enabled mammographic training system. *Proceedings of SPIE*, pp. 72640V. In Advanced PACS-based

Imaging Informatics and Therapeutic Applications. Khan M. Siddiqui and Brent J. Liu (Eds.). *Proceedings of SPIE*. 7264,72640V.

Yeh, S.T. & Glaxosmithkline, C., 1991. Using trapezoidal rule for the area under a curve calculation. *SAS Users Group Int*, 27, pp. 227-229.

Yousef, E.L., Duchesneau, R.H., Alfidi, R.J., Haaga, J.R., Bryan, P.J., LiPuma, J.P., 1984. Magnetic resonance imaging of the breast. Work in progress*. Radiology*, 150(3), pp. 761.

Young L.R. & Sheena, D., 1975. Methods & designs: Survey of eye movement recording methods. *Behaviour Research Methods & Instrumentation,* 7:397-429.

Young L.R. & Sheena, D., 1988. Eye movement measurement techniques. In: *Webster J, ed. Encyclopedia of Medical Devices and Instrumentation*. John Wiley & Sons, Inc.; 1259-1269.

Zhang, J., Gao, W., Wang, P. & Wu, Z.H., 2010. Relationships among hope, coping style and social support for breast cancer patients Chinese Medical Journal 2010;123(17):2331-2335 2331.

Zheng, W., 2005. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *American Journal of Epidemiology*. 1;162(11):1123-31.

Ziegler, R. G., Anderson, W.F. & Gail, M.H., 2008. Increasing Breast Cancer Incidence in China: The Numbers Add Up. *Journal of the National Cancer Institute* 100(19): 1339-1341.

APPENDICES

Appendix A: Questionnaire

PERFORMS
PERsonal perFORmance in Mammographic Screening

Questionnaire - Potential digital mammographic training for image interpretation

Currently most training in the interpretation of screening mammograms needs to be undertaken where there is a mammo-alternator (roller viewer) or other suitable light box on which to view the mammograms; consequently limiting the time and places where training can take place. However, the gradual introduction of digital mammography is opening up new opportunities to provide a wider range of training opportunities without such restriction. For instance, as well as being able to view digital breast images on high resolution monitors at the Breast Screening Centre it could be possible also to view training images on a range of computer displays (which have a lower resolution and admittedly would never be used for making a clinical decision) - for instance desktop PC, laptop PC or even small handheld devices, provided that it were possible to zoom into and otherwise interact with such images. We are interested in examining mammographic interpretation training opportunities as an additional possible future part of the PERFORMS scheme. Such an approach would never replace viewing images on high resolution monitors but could act as an adjunct to this which would enable some training to be undertaken on a small computer of some sort whenever and wherever an individual wanted - to suit each individual's needs.

This short questionnaire simply seeks your anonymous views on this proposed approach which would help us in designing our future training software. Thank you very much in advance.

Please feel free to contact us if you have any questions or if you want more information:

Miss Yan Chen, Prof. Alastair Gale
Applied Vision Research Centre, Loughborough University

Tel.: 01509 635737
E-mail: Y.Chen3@lboro.ac.uk

1. Are you _____ (Please tick)
   - ☐ Consultant radiologist
   - ☐ Advanced practitioner
   - ☐ Radiology SPR
   - ☐ Breast clinician
   - ☐ Radiographer
   - ☐ other, please specify: _____

2. What forms of mammographic interpretation training do your current training / CPD involve?
   - ☐ Arbitration/consensus
   - ☐ MDT's
   - ☐ Conferences
   - ☐ Review of individual film reading data
   - ☐ PERFORMS
   - ☐ CPD course
   - ☐ Interval cancer review
   - ☐ others, please specify: _____

   (Please tick as many as apply)

3. How would you rate the amount of opportunity you have for mammographic interpretation training? (Please tick)

   ☐ Highly ☐ More than adequate ☐ Adequate ☐ Could be more ☐ Not enough

4. Do you have any difficulties when undertaking mammographic interpretation training? (Please tick)

   ☐ Yes. ☐ No.

   If **Yes**, were these due to:
   - ☐ Limited time for training
   - ☐ Limited access to roller viewer
   - ☐ Limited access to digital workstation
   - ☐ other, please specify: _____

5. What do you find most useful in your current mammographic interpretation training?

   _____

   _____

Please Turn Over

6. What do you find least useful in your current mammographic interpretation training?

_____

_____

7. Would you like to be able to carry out mammographic interpretation training whenever and wherever suits you? (Please tick)

☐ Yes. Please specify when and/or where you would prefer to perform your mammographic interpretation training:

_____

_____

☐ No. Please specify the reason:

_____

_____

8. Would you like to have tailored mammographic interpretation training offered to you which is based on your latest PERFORMS report? (Please tick)

☐ Yes.

☐ No. Please specify the reason:

_____

9. Have you had any experience of digital mammography? (Please tick)

☐ Yes. Please specify approximately _____ months you have been using digital mammography

☐ No.

10. If acceptable image quality is achieved, do you think you would undertake digital mammographic interpretation training using: (please tick for each device type)

Rating / Device Type	Not at all	Probably Not	Maybe	Probably Yes	Definitely Yes
Digital workstation					
Desktop PC					
Laptop PC					
Small handheld devices					

11. Do you have any other suggestions about mammographic interpretation training using digital images?

_____

_____

Please return in the freepost envelope provided before 23rd June 2008.

Applied Vision Research Centre, Loughborough University, Loughborough, Leicestershire, LE11 3TU
Tel: +44 (0)1509 635735 Fax: +44 (0)1509 635736

Appendix B: Participant Consent Form

CONSENT FORM

(To be completed after Participant Information Sheet has been read)

The purpose and details of the study have been explained to me. I understand that this study is designed to further scientific knowledge and that all procedures have been approved by the Loughborough University Ethical Advisory Committee.

I have read and understood the Participant Information Sheet and this consent form.

I understand the emphasis of the study is on the usability of technology itself. My performance will not be assessed in terms of film reading proficiency.

I have had an opportunity to ask questions about my participation.

I understand that I am under no obligation to take part in the study.

I understand that I have the right to withdraw from this study at any stage for any reason, and that I will not be required to give any explanation for withdrawing.

I understand that all the information I provide will be treated in strict confidence. My personal records are held on a computer that is password protected so that only authorised members of staff may gain access. All people that participate in these trials are allocated a number so that their names do not appear with their results.

I agree to participant in this study.

Your signature:……..

Print name:……

Signature of investigator ……..

Date:……..

Appendix C: Participant Information Sheet

Participant Information Sheet

Mammographic interpretation training: how useful is handheld technology?

The purpose of the study

From examination of several years of PERFORMS data, problematic mammographic features which film-readers have more difficulty with can be identified. Accordingly, further training is suggested to improve their performance. Ideally, such training would be on-demand; that is whenever and wherever an individual decides to undertake it. To use a portable device for such a purpose would be attractive on many levels, although currently it is not known whether such technology could be used effectively for such high resolution mammographic images.

The information gathered from this experiment will be used for three purposes:

1) To indicate the possibility of using handheld device of Mammographic interpretation training;
2) To suggest the potential work on improving the usability of the handheld device of mobile training scheme by engaging the human-computer interaction techniques.

Taking part

Taking part will involve the following steps:

1) Answer a few questions about your background as a film reader;

2) Complete a computer-based image examination, rate your confidence in whether the image was normal or contained an abnormality. If the latter they also had to specify the abnormality and its location. In addition, your eye movements while examining the images will be tracked;
3) Answer a few questions on the images quality.

The whole process should take no longer than 30 minutes. You will examine 24 images in total. The emphasis of the study is on the usability of technology itself. Therefore, your performance will not be assessed in terms of film reading proficiency.

Your data

The only information we wish to keep for the study are the answers you give along with the eye-tracking data. The information will not be associated with your name. It will be stored securely. Where your data is stored in hardcopy, it will be locked in a filing cabinet to which access is restricted to the research group. Where your data is stored electronically, it will be saved on Loughborough University owned PC, in password protected files, to which access is restricted to the research group. Your data

will be stored in accordance with the Data protection Act. You can request your data be destroyed at any time.

Contact

Please feel free to get in touch with us if you want to ask any questions or if you want more information:

Yan Chen, Research Student

Applied Vision Research Centre
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ
Tel.: 01509 695737
E-mail: Y.Chen3@lboro.ac.uk

Prof. Alastair, Gale, Research supervisor

Applied Vision Research Centre
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ
Tel.: 01509 695701
E-mail: a.g.gale@lboro.ac.uk

Appendix D: Participant Information Sheet

Participant Information Sheet

Computer-based learning to improve breast cancer detection rate

The purpose of the study

In breast cancer screening it is important both to improve and maintain cancer detection skills at their highest levels. The introduction of digital imaging enables computer-based learning to be undertaken outside breast screening centres using a range of different devices. The potential for providing computer-based interpretation training using low-cost devices is detailed.

The information gathered from this experiment will be used for three purposes:

1) to indicate the possibility of using low-cost computer display for Mammographic interpretation training;
2) to suggest the potential work on the delivery of mammographic interpretation training

Taking part

Taking part will involve the following steps:

1) Complete a computer-based image examination, make a decision on whether the image was normal or contained an abnormality. If the latter they also had to specify the abnormality and its location. In addition, your eye movements while examining the images will be tracked;

2) Practise examining images in one of four different ways: follow expert's eye movement; follow expert's verbal instruction; read portion of the images with abnormality on; read full images with the region of interest highlighted;

3) Complete a computer-based image examination, make a decision on whether the image was normal or contained an abnormality. If the latter they also had to specify the abnormality and its location. In addition, your eye movements while examining the images will be tracked;

The whole process should take no longer than 60 minutes. You will examine 40 images in total. The emphasis of the study is on the usability of technology itself. Therefore, your performance will not be assessed in terms of film reading proficiency.

Your data

The only information we wish to keep for the study are the answers you give along with the eye-tracking data. The information will not be associated with your name. It will be stored securely. Where your data is stored in hardcopy, it will be locked in a filing cabinet to which access is restricted to the research group. Where your data is

stored electronically, it will be saved on Loughborough University owned PC, in password protected files, to which access is restricted to the research group. Your data will be stored in accordance with the Data protection Act. You can request your data be destroyed at any time.

Contact

Please feel free to get in touch with us if you want to ask any questions or if you want more information:

Yan Chen, Research Student

Applied Vision Research Centre
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ
Tel.: 01509 695737
E-mail: Y.Chen3@lboro.ac.uk

Prof. Alastair, Gale, Research supervisor

Applied Vision Research Centre
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ
Tel.: 01509 695703
E-mail: a.g.gale@lboro.ac.uk

Appendix E: Participant Information Sheet

Participant Information Sheet

Possible devices&HCI of delivering mammographic interpretation training

The purpose of the study

Mammographic interpretation in breast screening is known to be a particularly difficult task, partly due to the rarity of abnormalities within the screening population and partly due to the various subtle ways in which an abnormality can present. The increased use of digital mammography, and its forthcoming widespread adoption in the UK, offers new opportunities for mammographic interpretation training. For instance, as well as being able to view digital breast images on high resolution monitors at a Breast Screening Centre, it would technically be possible to view images on a range of alternative low cost computer displays: e.g. laptop computers or even mobile devices (e.g. large PDA). It is not suggested that such devices would be used for diagnosis but as an additional low-cost training resource. Technically, the limits of display screen size of such smaller computer screens to display very large digital mammogram images can be overcome by using suitable Human-Computer Interaction (HCI) techniques (e.g. zoom, pan, etc.). However, it is not known whether such alternative display devices could be used effectively to deliver mammogram interpretation training or whether these devices would be acceptable to breast screening radiologists.

In other radiological domains research, spanning the past 30 years, has shown the importance of investigating visual search behaviour during radiological inspection. It has been robustly demonstrated that by monitoring the eye movement patterns of individual radiologists as they examine images then false negative errors (where abnormalities are missed) can be classified into search errors, detection errors or interpretation errors. This knowledge can then be used in further mammographic interpretation training as it gives insight into whether the individual requires further training concerning how better to visually examine the image, how better to detect features in the image, or cognitive interpretation training of detected image features.

An empirical study is proposed to aims to compare the ability of different display devices in supporting mammogram film reader's image interpretation performance either with or without the support of HCI techniques.

Taking part

Taking part will involve the following steps:

1) Examine 40 screening cases on a mammographic digital workstation with half of them using manipulative tools (e.g. zoom, pan, change contrast) but half of them without. For each image, please rate their confidence in whether an abnormality was present and specify its location and feature type if abnormal. Time−on−task will be recorded so as is your eye movements which will be

recorded using a standard eye movement recording technique. A video camera, mounted beside the user, will be used to record your manipulation of the images;

2) Examine the same 40 screening cases on a laptop PC with a 20" screen with half of them using manipulative tools (e.g. zoom, pan, change contrast) but half of them without. For each image, please rate their confidence in whether an abnormality was present and specify its location and feature type if abnormal. Time−on−task will be recorded so as is your eye movements which will be recorded using a standard eye movement recording technique. A video camera, mounted beside the user, will be used to record your manipulation of the images;
3) examine the same 40 screening cases on a PDA with a 8" screen with half of them using manipulative tools (e.g. zoom, pan, change contrast) but half of them without. For each image, please rate their confidence in whether an abnormality was present and specify its location and feature type if abnormal. Time−on−task will be recorded so as is your eye movements which will be recorded using a standard eye movement recording technique. A video camera, mounted beside the user, will be used to record your manipulation of the images.

Anonymised data will be examined to determine: (a) if screeners can identify abnormalities when images are displayed either full size (without using HCI) on non-workstations and when HCI is used; (b) how performance on non-workstations compares to using workstation displays; (c) what effect using different display types has on users' eye movement behaviour as they examine the images and interact with them.
The whole process should take no longer than 60 minutes. You will examine 40 images in total. The emphasis of the study is on the usability of technology itself. Therefore, your performance will not be assessed in terms of film reading proficiency.

Your data

The only information we wish to keep for the study are the answers you give along with the eye-tracking data. The information will not be associated with your name. It will be stored securely. Where your data is stored in hardcopy, it will be locked in a filing cabinet to which access is restricted to the research group. Where your data is stored electronically, it will be saved on Loughborough University owned PC, in password protected files, to which access is restricted to the research group. Your data will be stored in accordance with the Data protection Act. You can request your data be destroyed at any time.

Contact

Please feel free to get in touch with us if you want to ask any questions or if you want more information:

Yan Chen, Research Student

Applied Vision Research Centre
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ
Tel.: 01509 695737
E-mail: Y.Chen3@lboro.ac.uk

Prof. Alastair, Gale, Research supervisor

Applied Vision Research Centre
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ
Tel.: 01509 695703
E-mail: a.g.gale@lboro.ac.uk