

**RECALL RATES IN SCREENING MAMMOGRAPHY:
VARIABILITY IN PERFORMANCE AND DECISIONS**

Norhashimah Mohd Norsuddin

**A thesis submitted in fulfilment of the requirement for the degree of
Doctor of Philosophy**

**Faculty of Health Sciences
The University of Sydney**

2017

SUPERVISOR'S STATEMENT

This is to certify that the thesis entitled “**Recall Rates in Screening Mammography: Variability in Performance and Decisions**” submitted by **Norhashimah Mohd Norsuddin** in fulfilment of the requirements for the degree of Doctor of Philosophy is in a form ready for examination.

Signed 

Date 29/06/2017

Asc. Prof Sarah Lewis

Discipline of Medical Radiation Sciences,

Faculty of Health Sciences,

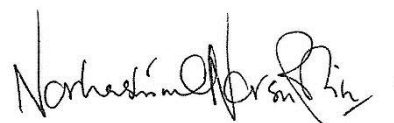
The University of Sydney.

CANDIDATE'S STATEMENT

I, **Norhashimah Mohd Norsuddin**, hereby declare that this submission is my own work and that it contains no material previously published or written by another person except where acknowledged in the text. Nor does it contain material which has been accepted for the award of another degree.

I, Norhashimah Mohd Norsuddin, understand that if I am awarded a higher degree for my thesis entitled “**Recall Rates in Screening Mammography: Variability in Performance and Decisions**” being lodged herewith for examination, the thesis will be lodged in the University Library and be available immediately for use. I agree that the University Librarian (or in case of a document, the Head of Department) may supply a photocopy or microform of the thesis to an individual for research or study or to a library.

Signed :



Name : Norhashimah Mohd Norsuddin

Date : 22 June 2017

ABSTRACT

Introduction and aim

Despite the wide variation in the target recall rates recommended across different breast screening programs, how such variations affect the performance of breast radiologists and their decisions to recall when detecting breast cancer are not well understood. Although having a high recall rate may increase the probability of cancer being detected earlier, a high recall rate also has been related to increased false positive decisions, causing significant psychological and economical costs for both screened women and the mammography screening service. Such negative effects may hamper the success of mammographic screening. Therefore, the purpose of this thesis is to explore the impact of various recall rates on breast radiologists' performance in a laboratory setting, which may lead to a potential improvement in the efficacy of breast cancer screening programs.

Methods

Institutional ethics approval was granted. This study was designed to encompass two aspects. The first aspect investigated the effect of setting varying recall rates on the performance of breast radiologists in screening mammography. The second aspect

examined which types of mammographic appearances of breast cancer are more likely to be missed when breast radiologists read at different recall rates. To achieve these two aspects, five Australian breast radiologists were recruited to read one single test set of 200 de-identified mammographic cases, containing 180 normal and 20 abnormal cases, over three different recall rate conditions: free recall, 15% and 10%. These breast radiologists were tasked with marking the location of suspicious lesions and providing a confidence rating and they could not exceed the recall rate prescribed, as described in detail in the Methods chapter in this thesis. Breast radiologists' performance was analysed by receiver-operating characteristic area under the curve (ROC AUC), Jackknife Free-Response Receiver Operating Characteristic (JAFROC) Figure of Merit (FOM), sensitivity, case sensitivity, lesion sensitivity and specificity. All performance data were analysed using JAFROC Version 4.2 software and statistical analysis was performed using SPSS software version 22.0.

Results

Reading at a lower recall rate had a significant positive effect on specificity ($H=12.704$, $P=0.002$). However, a significant decrease in breast radiologists' performance was observed when reading at lower recall rates (15% and 10%), with lower sensitivity ($H=12.891$, $P=0.002$), case location sensitivity ($H=12.512$, $P=0.002$) and ROC AUC ($H=11.601$, $P=0.003$). No significant changes were evident in lesion location sensitivity ($H=1.982$, $P=0.371$) and JAFROC FOM ($H=1.820$, $P=0.403$). The second study of this thesis showed that breast radiologists demonstrated higher sensitivity and receiver operating characteristic (ROC) area under the curve (AUC) for stellate masses (NSD) ($H=5.36$, $P=0.07$ and $H=7.35$, $P=0.03$ respectively) and mixed features ($H=9.97$, $P=0.01$

and $H=6.50$, $P=0.04$ respectively) when reading at 15% and 10% recall rates. No significant change was observed in sensitivity and ROC AUC on cancer characterized as stellate ($H=3.43$, $P=0.18$ and $H=1.23$, $P=0.54$ respectively) and architectural distortion (AD) ($H=0.00$, $P=1.00$ and $H=2.00$, $P=0.37$ respectively). Across all recall conditions, stellate lesions were likely to be recalled (90.0%) while NSDs were likely to be missed (45.6%).

Conclusion

Albeit a significant improvement in specificity was observed, reducing the number of recalled cases to 10% significantly reduced breast radiologists' performance in this laboratory study. The mammographic appearances of cancers contributed to the breast radiologists' clinical decision-making at low recall rates. Breast lesions characterized as stellate were continuously recalled regardless of any recall conditions, which may be due to the high likelihood of stellate lesions being malignant. On the other hand, cancers with subtle malignancy signs and a high likelihood of being benign, such as non-specific density (NSD), were most likely to be missed at reduced recall rates.

ACKNOWLEDGEMENT

First and foremost, I must acknowledge my limitless and thanks to Allah, the Ever Magnificent, the Ever Thankful, for His help and blessings. I am totally sure that this work would have never become truth, without His guidance.

I would like to express my sincere appreciation to the many people that have supported me to enable the undertaking of this research and the completion of this thesis. First and foremost, I would like to express my heartfelt gratitude to Associate Professor Sarah Lewis, my primary supervisor, who has provided academic and supportive guidance along the journey. Your superb supervision, guidance reviews and counsel provided to me during the course of my PhD candidature have been invaluable. Further appreciation goes to Associate Professor Claudia Mello-Thoms and Dr Warren Reed, especially during the research design and data analysis. Your wealth of experience, advice and encouragement has been welcomed. The opportunity to work with these three amazing supervisors for the last three years and eight months has been a true pleasure and a blessing.

Special thanks to a key individual, Bao Lin Pauline Soh who has provided her test set and Dr Mary Rickard who has shared her valuable insight to this research, without whom this study would not be possible.

I also would like to express my wholehearted thanks to the immeasurable personal support of my amazing family and friends while undertaking this research and completing this thesis. I am indebted to their patience, practical assistance, and providing encouragement and sustenance along this journey of growth and during times that have been challenging for many personal reasons. I am very blessed to have many people in my life and I wish to particularly recognize my incredibly affirming husband Mohd Taher Kamil, the one who stands by me through everything for his immeasurable support, unwavering love, endless support, encouragement, patience and faith.

To my parents, Mohd Norsuddin Tegoh and Rominah Separi thank you for endless love and the million prayers you have sent for my success. To my siblings, Shaifful Bakri, Kadir, Nur Arifah, Muhammad Faiz and Muhammmad Rafiuddin, my in-laws Kak Ain (Nur Aifa), Anga (Nurul Huda) Suharti and Nazmi, thank you for your enormous and numerous help, kindness and compassion in taking care of our family while I was away. No one can ever replace all of you in my life. To my late brothers; Mohaini (2013, the year I began my PhD journey) and Nor Azali (2016, the year I was to ended the PhD journey), although you won't be here to see this, the fond memories of our happy times together makes me feel your presence. You were significantly a part of this journey. All of my thoughts and prayers are with you.

I also would like to extend my gratitude to my extended family in Sydney: Yasmin and Arif, Kak Nora, Abang Zaed and family, Kak Su, Abang Azhar and family, Anis, Abang Syam and family, Shuhada, Kak Ruki, Abang Norman and family, Nina for their enormous helps, endless support physically and emotionally through this roller coaster journey. May Allah reward you for all the good deeds and kindness you have done for me.

I owe profound gratitude to Eliza Yazid, whose constant encouragement, and generous support she and her family provided me throughout this PhD's journey particularly her advice and her friendly assistance during my hardships. Also to my best friend, Qatrunnada Razali, who have been so supportive along the way of doing my study. Because of their unconditional love and prayers, I have the chance to complete this thesis. Last but not least, deepest thanks go to all people who took part in making this thesis real.

PUBLICATIONS AND PRESENTATIONS

The work presented in this thesis has been published and presented in the following forms:

PUBLICATIONS

Journal articles

1. N. Mohd Norsuddin, W. Reed, C. Mello-Thoms, S.J. Lewis. Understanding recall rates in screening mammography: A conceptual framework review of the literature. *Radiography* 21 334-341 (2015).
2. N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, M.Rickard, S. Lewis. An investigation into the mammographic appearances of missed breast cancers when recall rates are reduced. *British Journal of Radiology (BJR)* (In Press) (2017).

Chapters in book

3. N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, P. Brennan, S. Lewis. Lower recall rates reduced readers' sensitivity in screening mammography. In *Breast Imaging* (pp 106-111). Springer International Publishing. (2016).

Submitted papers currently under review

4. N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, B.P.Soh, S. Lewis. Radiologists' performance at reduced recall rates in mammography: A laboratory study. *The Breast Journal (TBJ)*. (2017).

PRESENTATIONS

Oral presentations

1. Forced recall rates in screening mammography. Three Minute Thesis. The University of Sydney, Sydney, Australia. (2014).
2. Forced recall rates in screening mammography. Physics and Perception Meeting, Brain and Mind Research Institute. Sydney, Australia. (2014).
3. Forced recall rates in screening mammography: Laboratory based experiment. Meeting with visiting professor Craig Abbey, Medical Radiation Sciences (MRS), The University of Sydney, Cumberland Campus, Lidcombe, Australia. (2014).
4. Methodology: Forced recall rates in screening mammography. Meeting with visiting professor Steve Hillis, Medical Radiation Sciences (MRS), The University of Sydney, Cumberland Campus, Lidcombe, Australia. (2015).

5. How specific recall rates affect observer performance in screening mammography?
Meeting with visiting Professor Jeremy Wolfe, Medical Radiation Sciences (MRS),
The University of Sydney, Cumberland Campus, Lidcombe, Australia. (2015).

Poster presentations

6. N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, P. Brennan, S. Lewis. Optimising recall rates through observer performance in screening mammography. HDR Conference: Imag!ne.U – Creating the Future. Faculty of Health Sciences, The University of Sydney. Sydney, Australia (2014) (Appendix F: HDR poster presentation, 2014).
7. N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, P. Brennan, S. Lewis. Lower recall rates reduced readers' sensitivity in screening mammography. In *Breast Imaging* (pp 106-111). Malmö, Sweden (2016) (Appendix K). **Awarded Best Research Poster at IWDM 2016 Breast Imaging Workshop**
8. N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, M. Rickard, S. Lewis. Detection for non-specific density lesions is lowered when recall rates are reduced for Australian breast radiologists. Sydney Cancer Conference, Sydney, Australia (2016) (Appendix L).

TABLE OF CONTENT

SUPERVISOR’S STATEMENT	I
CANDIDATE’S STATEMENT	II
ABSTRACT	III
Introduction and aim.....	iii
Methods	iii
Results	iv
ACKNOWLEDGEMENT	VI
PUBLICATIONS AND PRESENTATIONS	IX
PUBLICATIONS.....	ix
Journal articles.....	ix
Chapters in book.....	ix
Submitted papers currently under review.....	x
PRESENTATIONS	x
Oral presentations	x
Poster presentations.....	xi
TABLE OF CONTENT	XII
LIST OF TABLES	XIX

LIST OF FIGURES	XXI
LIST OF ABBREVIATIONS	XXIII
THESIS STRUCTURE	24
CHAPTER 1	26
INTRODUCTION	26
Breast cancer screening mammography	27
Adverse effects of mammographic screening.....	28
Recall rates in breast screening mammography programs	30
Recall rate and mammographic appearances.....	33
Objectives.....	35
References.....	36
CHAPTER 2	40
UNDERSTANDING RECALL RATES IN SCREENING MAMMOGRAPHY: A CONCEPTUAL FRAMEWORK REVIEW OF THE LITERATURE	41
STATEMENT FROM AUTHOR.....	42
CHAPTER 3	51
EXTENDED METHODS	51
Ethical Approval.....	52
Participant	52
Research design	54
Workstation set-up.....	54
Monitor calibration	55

Ambient lighting measurement.....	61
Customised recording software	62
Application set-up.....	65
Test set development	67
Justification for 200 mammographic cases.....	68
Truth (Gold standard)	69
Mammographic appearances of cancers in the test set: detailed classification	70
Randomizing cases in Sectra system	76
Randomizing cases in customised recording software	77
Image presentation (display protocol)	77
Reading task	79
Breast Screen Australian classification system	81
Data Analysis.....	83
Recall rate to assessment	85
ROC analysis	85
JAFROC analysis.....	86
Data extraction.....	88
Lesion localization.....	89
Comparison recording software output to truth (gold standard)	92
Statistical analysis.....	92
Reader Performance	92
Mammographic appearances of missed cancers.....	93
References	95

CHAPTER 4	96
EXTENDED RESULTS	96
Individual performance within the free recall.....	97
ROC curve	98
AFROC curve	98
Lesion detectability at free recall, 15% and 10% recall rates	101
References	104
CHAPTER 5	105
RADIOLOGISTS’ PERFORMANCE AT REDUCED RECALL RATES IN MAMMOGRAPHY: A LABORATORY STUDY	105
STATEMENT FROM AUTHOR.....	106
Abstract	107
Introduction	109
Materials and Methods	110
Participants	110
Experimental protocol	111
Cases	111
Reading environment	112
Reading Task	113
Results	117
Discussion	121
Conclusion	125
References	126

CHAPTER 6 **129**

**AN INVESTIGATION INTO THE MAMMOGRAPHIC APPEARANCES OF
MISSED CANCERS WHEN RECALL RATES ARE REDUCED.** **129**

STATEMENT FROM AUTHOR..... 130

Abstract. 131

Objectives 131

Methods 131

Results 131

Conclusion 132

Advances in knowledge 132

Introduction 133

Materials and methods 134

Sample 134

Cases 134

Reading sessions 137

Reading task 138

Data analysis..... 139

Results 141

Discussion 146

Conclusion 150

References 151

CHAPTER 7	154
DISCUSSION AND CONCLUSION	154
Background	155
Discussion of Study 1: To investigate the effect of reduced recall rates on breast radiologists' performance using receiver operating characteristic (ROC) and Jackknife free response operating characteristic (JAFROC) analysis.....	157
Discussion of Study 2: To assess which types of mammographic appearances of breast cancer are more likely to be missed when breast radiologists read at lower recall rates.	163
Implications of the findings this thesis	167
Strengths of the thesis	171
Limitations of the studies.....	172
Recommendation for future work.....	174
Conclusion	176
Reference	177
APPENDICES	182
Appendix A: Ethics Approval	183
Appendix C: Participant consent form	188
Appendix D: Participant questionnaire.....	190
Appendix E: 3 Minute thesis competition	191
Appendix F: HDR poster presentation, 2014	192
Appendix G: Physics and Perception Meeting, 2014	193
Appendix H: Presentation for Professor Craig Abbey, 2014	194
Appendix I: Presentation for Professor Steve Hillis, 2015.....	195

Appendix J: Presentation for Jeremy Wolfe, 2015	196
Appendix K: IWDM 2016 13th International Workshop on Breast Imaging, Malmö, Sweden	197
Appendix L: Sydney Cancer Conference 2016, Australian Technology Park, Sydney, Australia	198
Appendix M: IWDM 2016 Proceeding paper	199
Appendix N: Performance metrics	204

LIST OF TABLES

Table 1 Demographic details of participating breast radiologists	53
Table 2 Calibrated monitor results.....	58
Table 3 Output results of reporting monitors.....	61
Table 4 Definition of lesion abnormalities according to Australian Synoptic Breast Imaging Report.....	71
Table 5 Mammographic appearances of 20 cancer cases present in the test set.....	75
Table 6 Recall recommendation based on Australian Classification System.....	82
Table 7 The truth location of lesion coordinates of abnormal cases.....	91
Table 8 Distribution of individual recall decisions for each cancer case at three recall conditions (S1: Free recall; S2: 15% recall; S3: 10% recall)	103

CHAPTER 5:

Table 1 Demographic details of participating breast radiologists	111
Table 2 Results for sensitivity, lesion location sensitivity, case location sensitivity, specificity, ROC AUC and JAFROC FOM at free call, 15% and 10% conditions	119
Table 3 Kruskal-Wallis analysis and post hoc Mann-Whitney U test of sensitivity, lesion location sensitivity, case location sensitivity, specificity, ROC AUC and JAFROC FOM ...	120

CHAPTER 6

Table 1 Definition of lesion terms used for classification ¹⁷	136
Table 2 Mean values of sensitivity and receiver operating characteristic (ROC) area under the curve (AUC) of each mammographic feature at free recall, 15% and 10% recall rates	142
Table 3 Distribution of detection and cancer appearances (lesion type, breast density and lesion location) for each cancer in relation to case difficulty at free recall, 15% and 10% recall rates.....	144

LIST OF FIGURES

Figure 1 Workstation setup for participants in MIOPeG laboratory.	55
Figure 2 The TG18-QC comprehensive test pattern used in the study.....	56
Figure 3 Measuring minimum luminance and maximum luminance of the monitors at the centre of the display screen.	57
Figure 4 Konica Minolta Spectroradiometer CS-2000 (Japan)	59
Figure 5 Digital mammography display for user interface in customised recording software	63
Figure 6 Snapshot of the lesion box and confidence score displays on left medio-lateral oblique (MLO) view.....	64
Figure 7 Snapshot of a black window (shell)	65
Figure 8 Example of mammographic images with different breast density A) < 25% glandular tissue, B) 25-50% glandular tissue, C) 51-75% glandular tissue, D) >75% glandular tissue	73

Figure 9 Mammographic images above showed lesion (circled) location to fibroglandular tissue A) overlapping fibroglandular tissue, B) outside fibroglandular tissue and C) at the edge of fibroglandular tissue 74

Figure 10 Image display sequence for reading process 78

Figure 11 Reading process flow for readers when performing the reading task 80

Figure 12 Schematic overview of the steps involved in the analysis process 84

Figure 13 Histogram of individual recall rate of each reader at the free recall condition.. 97

Figure 14 The empirical ROC curves for each reader when performing at three different recall conditions (i. Free recall, ii. 15% recall and iii. 10% recall). 99

Figure 15 The empirical AFROC curves for each reader when performing at three different recall conditions (i. Free recall, ii. 15% recall and iii. 10% recall). 100

CHAPTER 6

Figure 1 Examples of lesion features present in this study a) stellate mass; b) mixed features of calcification and architectural distortion (AD); c) non-specific density (NSD)140

Figure 2 All three cancers above characterized with non-specific density (NSD) but with variability in lesion detectability and level of difficulty at reduced recall rates; **a)** MJCQ: lower difficulty; **b)** MJBK: medium difficulty and **c)** MJCF: higher difficulty..... 145

LIST OF ABBREVIATIONS

- AD (Architectural distortion)
- AUC (Area under the curve)
- BI-RADS (Breast imaging reporting and data systems)
- BSNSW (BreastScreen New South Wales)
- CC (Cranio-caudal)
- FOM (Figure of merit)
- FP (False positive)
- JAFROC (Jack-knife free-response operating characteristic)
- MLO (Medio-lateral oblique)
- NBCC (National Breast Cancer Centre)
- NSD (Non-specific density)
- RANZCR (Royal Australian and New Zealand College of radiologists)
- ROC (Receiver operating characteristic)
- TP (True positive)

THESIS STRUCTURE

The thesis is arranged into eight chapters and structured in the following manner (**Table 1**):

Table 1 Thesis structure

Chapter	Description
1	This chapter provides an overview of the aims addressed by the thesis.
2	This chapter presents a detailed literature review and provides a summary of the findings of studies related to recall rates. This chapter fulfills Objective I.
3	This chapter describes the details of an extended version of the methodology used to perform the experiments in this thesis.
4	This chapter describes a detailed analysis and results from the experiments.

Chapter	Description
----------------	--------------------

5	This chapter presents the main experiment which investigates the impact of different target recall rates on reader performance in a laboratory setting when reading screening mammograms. This chapter fulfills Objective II.
----------	---

6	This chapter presents the extended analysis of the investigation on the types of mammographic appearances of breast cancer that are most likely to be missed when breast radiologists read at lower recall rates. This chapter fulfills Objective III.
----------	--

7	Discussion and conclusion presents a discussion of the work as well as its implications, limitations and the potential future direction of this research.
----------	---

8	Appendices provide an additional published output from this thesis, presentations related to this study and the relevant documents for participating breast radiologists used in the studies.
----------	---

CHAPTER 1

INTRODUCTION

Breast cancer screening mammography

Mammography has been shown to be an effective screening tool for detection of breast cancer. Worldwide, breast cancer is the most frequent cancer among females and is the second most common cancer overall. In 2012, there were an estimated 1.67 million new diagnoses, which contributed to one in four of all cancer diagnosis (1). Breast cancer was also the most frequent cancer occurring among women worldwide, which places it as the fifth most frequent cause of death (522,000 deaths) from cancer overall (1). In Australia, breast cancer is currently the most common cancer among women (2). Over the last three decades, the breast cancer incidence rate has more than doubled, from 5,310 to 13,567 newly diagnosed cases in this country (2). There was a significant increase in the cancer incidence rate between 1990 and 1995, after the national breast cancer screening program, known as BreastScreen Australia, was introduced (3).

Along with breast cancer being most common cancer among Australian women, the percentage of deaths caused by the disease has been decreasing over recent years. Early detection through breast screening mammography has been demonstrated to significantly improve breast cancer survival (4) and it is reported that in 2015, there was a 90% chance of surviving at least five years after diagnosis, particularly for women diagnosed with invasive ductal carcinoma and smaller tumour sizes, as compared to 72% between 1982-1987 for the same diagnosis (2, 3). Along with treatment advances in recent years, screening mammography is contributing to the reduction in breast cancer mortality (4),

from 68 deaths per 100,000 women in 1991 to 42 deaths per 100,000 women in the general population for those aged 50-69 in 2012 (3).

Through early screening, mammography is able to detect particular changes and early signs of cancer before they have developed to an advanced stage. Finding a cancer at early stage might lead to less aggressive treatment for the patient and will improve the prognosis of the disease. However, there are some population health drawbacks from a screening program, such as overdiagnosis (5), false positive and false negative results (6). There are debates around these issues, namely about whether screening programs such as BreastScreen Australia (BSA) do more harm than good for the greater screened population of women who are healthy and cancer free.

Adverse effects of mammographic screening

It is known that some cancers may never progress to become symptomatic or clinically relevant and this is termed “overdiagnosis”. In every 1000 women attending breast cancer screening throughout the last 10 years, 5 healthy women will be overdiagnosed with breast cancer and will likely undergo unnecessary treatment or further imaging (6). It is estimated that overdiagnosis can be as high as 42% in women aged 50-59 (5), prompting great concern in breast screening programs. Another potential adverse effect that may compromise the success of population-based screening mammography programs is the false positive recall rate (7, 8). Women that are recalled for further assessment and do not have breast cancer are referred as having a false positive result for their screening. It has been estimated that the cumulative risk of a false-positive screening

result in women aged 50–69 undergoing 10 biennial screening tests varied from 8% to 21%(7).

Although recalling a high number of women for assessment may have the potential to increase the probability of cancer being detected, an effective screening mammography program must consider achieving an appropriate balance between cancer detection and recall rates. This is because a large proportion of screened women being recalled for additional investigations contributes to unnecessary assessments, patient anxiety and additional financial costs (9). Adverse psychological outcomes resulting from a false positive result have been shown to have a negative impact on women attending subsequent screening in mammography programs (10-12).

Unfortunately, it has been difficult to avoid these potential harms, and with a low incidence of breast cancers in the screened population, which is only 118 per 100,000 women (13, 14), it makes screening a complex task for radiologists and other mammogram breast readers (15-17). Thus, screening programs usually have a stringent quality assurance procedure in place to assure the harms in mammographic screening can be minimized and the performance of screening programs can be optimized.

In Australia, a quality improvement program is in place for BreastScreen Australia to ensure the ongoing success of the population screening mammography program, with high standards quality service in compliance with the National Accreditation Standards (NAS). The use of a series of performance indicators/measures, such as sensitivity,

specificity, positive predictive value (PPV) and recall rates facilitates monitoring and evaluating the program's aims of reducing morbidity and mortality from breast cancer.

Recall rates in breast screening mammography programs

Recall rates within a breast cancer screening mammography program refer to the proportion of screened women who are recalled for further assessments (diagnostic follow-up or biopsy procedure). When a radiologist identifies a suspicious finding on a mammogram, additional diagnostic follow-up imaging, such as ultrasound, magnetic resonance imaging (MRI), digital breast tomosynthesis (DBT) or fine needle biopsy may be required to confirm the presence of the cancer, and the patient will be asked to return for further assessment. Recall rates for first screenings are more likely to be higher than subsequent screenings due to lack of previous or prior mammographic images for comparison (3).

National and regional bodies have recommended a maximum percentage of women to be recalled for further assessment as a clinical guideline for radiologists when reporting screening mammograms. According to the National Accreditation Standards (NAS) for BreastScreen Australia, the recommended maximum percentage of women recalled for their first screen and subsequent screen should be less than 10% and 5% respectively (18). The United Kingdom (UK) has similar recall rates recommendation for the first screen as Australia, while the recall rate for the subsequent screenings has been recommended to be less than 7% (19). Similarly, the American College of Radiology (ACR) also has recommended a target recall rate of less than 10% for women who are attending their first

screening mammogram (20), while recommended recall rates in Europe have suggested target recall rates less than 5% for the first screen and 3% for the subsequent screen (21).

A large variation in recall rates have been reported, ranging from approximately 1.4% in the Netherlands to more than 15.1% in the US (22). A rate that is too low may be associated with decreased sensitivity and a higher number of cancers being missed. Conversely, if the recall rate is too high, it may be associated with increased false positives findings (23, 24). A report on the performance of BreastScreen Australia showed an upward trend in “recall to assessment results”, with a significant increase between 1996–2000 and 2001–2005, from 6.9% to 9.2% in the first screening round and from 3.8% to 4.0% in the subsequent rounds. Despite the “recall for subsequent screening” remaining constant at 4% among screened women in the Australian population, the proportion of women being recalled for further assessment for their first screening through BreastScreen Australia has increased to a high of 12% in 2013(3). Of these recalled women, most (91%) were found to have false positive results, and only 9% had breast cancer (invasive breast cancer and ductal carcinoma in situ (DCIS)).

The variations in recall rates may also be confounded by other factors, including the level of expertise of the readers, population screening demographics, the complexity of interpretation of mammography images or the national health policy (25-27). These factors are generally beyond the radiologist’s control. Additionally, the results of recall rates reported in previous studies may also be influenced by the screening interval (the time interval between screenings), with a significant increase in the likelihood of women being recalled as the screening interval increases (28). Detailed literature on how these factors

affect recall rates results in screening mammography is presented as a published article in Chapter 2.

The relationship between recall rates and sensitivity in screening mammography is unclear, as evidenced by past conflicting research (23, 24, 29, 30). A study by Gur et al (31) reported a linear relationship between recall rates and the cancer detection rate, with a significant correlation ($r = 0.76$; $P = 0.01$), and the study suggested that radiologists may find more cancers when they recalled at higher rates ($>10\%$). However, other studies found that the number of cancers that can be detected does not increase once the recall rates reach approximately 5% (23, 29), and suggested an increase in recall rates greater than 5% only resulted in higher false positive outcomes, suggesting lower specificity (32).

Hence the issue around the large international variation in recalling screened women for further assessment on diagnostic accuracy remains unresolved. It is unclear why some countries such as the Netherland recommend very low recall rates (currently at recall rate less than 2%) but still have similar cancer detection rates as the countries such as the United States with higher target recall rates (recall rate more than 15%) (32-34). There is very little information on how restricting the percentage of women referred for further assessment affects radiologists' performance, in terms of both receiver operating characteristic (ROC) area under the curve (AUC) and Jackknife free response operating characteristic (JAFROC) Figure of Merit (FOM) scores. These questions underscore the need for further research to explore the effect of various recall rates on diagnostic accuracy and radiologists' performance.

Recall rate and mammographic appearances

Many factors influence the difficulty of reaching a correct diagnosis for normal and abnormal cases, especially the characteristics of cancer lesions. A review of previously missed cancers reveals lesions with subtle mammographic signs of malignancy are commonly missed in mammographic screening, with cancers that present with architectural distortion (AD) being the most challenging malignant feature for readers to detect (17).

Fibroglandular tissues and cancer masses appear as radiopaque on mammograms. As mammography produces two dimensional (2D) images which are characterized with superimposition and overlapping of breast tissue, the presence of breast cancer lesions in high mammographic density can be obscured and difficult to visualize. Previous studies have shown that breast cancer detection decreased in women with high mammographic density, from 80-90% sensitivity in fatty breasts to 29-75% in dense breasts (35-37). Although the sample size used for these studies were relatively large (ranging from 576 to 329,495), there was some variation in classifying the breast density into BI-RADS categories, most likely due to human subjectivity. The qualitative approaches used in BI-RADS classification may lead to differences in decision-making regarding breast density information. There is also an increased likelihood of false positive results for women with high mammographic density than for women with low mammographic density (38-40). According to US and Australian population breast screening data, a higher proportion of

women with high breast density, who were predominantly younger (screened women aged 40-49), were recalled for further assessment (3, 39, 41).

However, no study has explored the key features of mammographic images related to the type of cancer lesions and breast densities that are more likely to be missed and affected when radiologists are reading at various recall rates. Identification of the impact of specific mammographic features on diagnostic accuracy would facilitate an understanding of the complex issue of the image interpretation process and recall decision making, which may in turn explain or shed light on variations in radiologists' performance (42).

Objectives

The work presented in this thesis was designed to determine the effect of varying prescribed target recall rates on the breast radiologists' performance when reading mammographic images in a laboratory environment.

The objectives of this research are as follows:

- i. To thoroughly review published studies concerning recall rates in screening mammography.
- ii. To investigate the effect of reduced recall rates on breast radiologists' performance using receiver operating characteristic (ROC) and Jackknife free response operating characteristic (JAFROC) analysis.
- iii. To determine whether specific types of mammographic appearances are more likely to be missed than others when breast radiologists are operating at different recall rates.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 2015;136(5):E359-E86.
2. Australian Institute of Health and Welfare. *Cancer in Australia: an overview 2012*. 2013.
3. Australian Institute of Health and Welfare. *BreastScreen Australia monitoring report 2012–2013*. Canberra: AIHW, 2015 Cancer series no 95 Contract No.: CAN 93.
4. van Schoor G, Moss SM, Otten JDM, Donders R, Paap E, den Heeten GJ, et al. Increasingly strong reduction in breast cancer mortality due to screening. *Br J Cancer*. 2011;104(6):910-4.
5. Morrell S, Barratt A, Irwig L, Howard K, Biesheuvel C, Armstrong B. Estimates of Overdiagnosis of Invasive Breast Cancer Associated with Screening Mammography. *Cancer Causes & Control*. 2010;21(2):275-82.
6. Brodersen J, Jørgensen KJ, Gøtzsche PC. The benefits and harms of screening for cancer with a focus on breast screening. *Polskie Archiwum Medycyny Wewnętrznej*. 2010;120(3):89-94.
7. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *Journal of medical screening*. 2012;19(suppl 1):57-66.
8. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. *Ann Intern Med*. 2007;146(7):502-10.
9. Burnside E, Belkora J, Esserman L. The Impact of Alternative Practices on the Cost and Quality of Mammographic Screening in the United States. *Clinical Breast Cancer*. 2001;2(2):145-52.
10. Alamo-Junquera D, Murta-Nascimento C, Macia F, Bare M, Galceran J, Ascunce N, et al. Effect of false-positive results on reattendance at breast cancer screening programmes in Spain. *Eur J Public Health*. 2012;22(3):404-8.
11. Sim MJ, Siva SP, Ramli IS, Fritschi L, Tresham J, Wylie EJ. Effect of false-positive screening mammograms on rescreening in Western Australia. *The Medical journal of Australia*. 2012;196(11):693-5.

12. Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, et al. Systematic review of the psychological consequences of false-positive screening mammograms. *Health technology assessment (Winchester, England)*. 2013;17(13):1-170, v-vi.
13. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-86.
14. Australian Institute of Health and Welfare. *Breast cancer in Australia: An overview*. Canberra: AIHW, 2012 Cancer series No 71 Contract No.: CAN 67.
15. Craft M, Bicknell AM, Hazan GJ, Flegg KM. Microcalcifications Detected as an Abnormality on Screening Mammography: Outcomes and Followup over a Five-Year Period. *International Journal of Breast Cancer*. 2013;2013:7.
16. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology*. 1992;184(3):613-7.
17. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Brenner RJ, Sickles EA. Analysis of 172 Subtle Findings on Prior Normal Mammograms in Women with Breast Cancer Detected at Follow-up Screening. *Radiology*. 2003;226(2):494-503.
18. BreastScreen Australia. *National Accreditation Standards: BreastScreen Australia Quality 2008* [cited 2014 May 21]. Available from: <http://www.cancerscreening.gov.au>
19. National Health Service Breast Screening Radiologist Quality Assurance Committee. *Quality assurance guidelines for radiologists*. National Health Service Breast Screening Programme publication no 15 Sheffield, England: NHSBSP Publications. 1997.
20. Feig SA, D'Orsi CJ, Hendrick RE, Jackson VP, Kopans DB, Monsees B, et al. American College of Radiology guidelines for breast cancer screening. *American Journal of Roentgenology*. 1998;171(1):29-33.
21. Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*. 2008;19(4):614-22.
22. Yankaskas BC, Klabunde CN, Ancelle-Park R, Rennert G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? *Journal of medical screening*. 2004;11(4):187-93.
23. Otten JDM, Karssemeijer N, Hendriks JHCL, Groenewoud JH, Fracheboud J, Verbeek ALM, et al. Effect of Recall Rate on Earlier Screen Detection of Breast Cancers Based on the Dutch Performance Indicators. *Journal of the National Cancer Institute*. 2005;97(10):748-54.

24. Schell MJ, Yankaskas BC, Ballard-Barbash R, Qaqish BF, Barlow WE, Rosenberg RD, et al. Evidence-based Target Recall Rates for Screening Mammography. *Radiology*. 2007;243(3):681-9.
25. Duijm LEM, Groenewoud JH, Fracheboud J, de Koning HJ. Additional Double Reading of Screening Mammograms by Radiologic Technologists: Impact on Screening Performance Parameters. *Journal of the National Cancer Institute*. 2007;99(15):1162-70.
26. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International Variation in Screening Mammography Interpretations in Community-Based Programs. *Journal of the National Cancer Institute*. 2003;95(18):1384-93.
27. Haneuse S, Buist DSM, Miglioretti DL, Anderson ML, Carney PA, Onega T, et al. Mammographic Interpretive Volume and Diagnostic Mammogram Interpretation Performance in Community Practice. *Radiology*. 2012;262(1):69-79.
28. Hofvind S, Geller BM, Skelly J, Vacek PM. Sensitivity and specificity of mammographic screening as practised in Vermont and Norway. *The British journal of radiology*. 2012;85(1020):e1226-e32.
29. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of Recall Rates with sensitivity and positive predictive values of screening mammography. *American journal of roentgenology*. 2001;177(3):543-9.
30. Hardesty LA, Klym AH, Shindel BE, Chough DM, Sumkin JH, Gur D. Is Maximum Positive Predictive Value a Good Indicator of an Optimal Screening Mammography Practice? *American Journal of Roentgenology*. 2005;184(5):1505-7.
31. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, et al. Recall and detection rates in screening mammography. *Cancer*. 2004;100(8):1590-4.
32. Kemp Jacobsen K, O'Meara ES, Key D, S.M. Buist D, Kerlikowske K, Vejborg I, et al. Comparing sensitivity and specificity of screening mammography in the United States and Denmark. *International Journal of Cancer*. 2015;137(9):2198-207.
33. Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, et al. Comparison of screening mammography in the united states and the united kingdom. *JAMA*. 2003;290(16):2129-37.
34. Hofvind S, Yankaskas BC, Bulliard J-L, Klabunde CN, Fracheboud J. Comparing interval breast cancer rates in Norway and North Carolina: results and challenges. *Journal of medical screening*. 2009;16(3):131-9.
35. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, et al. Individual and combined effects of age, breast density, and hormone

replacement therapy use on the accuracy of screening mammography. *Ann Intern Med.* 2003;138(3):168-75.

36. Cawson JN, Nickson C, Amos A, Hill G, Whan AB, Kavanagh AM. Invasive breast cancers detected by screening mammography: a detailed comparison of computer-aided detection-assisted single reading and double reading. *Journal of medical imaging and radiation oncology.* 2009;53(5):442-9.

37. Buist DSM, Porter PL, Lehman C, Taplin SH, White E. Factors Contributing to Mammography Failure in Women Aged 40–49 Years. *JNCI: Journal of the National Cancer Institute.* 2004;96(19):1432-40.

38. van Dijck JA, Verbeek AL, Hendriks JH, Holland R. The current detectability of breast cancer in a mammographic screening program. A review of the previous mammograms of interval and screen-detected cancers. *Cancer.* 1993;72(6):1933-8.

39. Cook AJ, Elmore JG, Miglioretti DL, Sickles EA, Aiello Bowles EJ, Cutter GR, et al. Decreased accuracy in interpretation of community-based screening mammography for women with multiple clinical risk factors. *J Clin Epidemiol.* 2010;63(4):441-51.

40. Al Mousa DS, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? *Clinical Radiology.* 2014;69(4):333-41.

41. Litherland JC, Evans AJ, Wilson ARM. The effect of hormone replacement therapy on recall rate in the national health service breast screening programme. *Clinical Radiology.* 1997;52(4):276-9.

42. Antonio AL, Crespi CM. Predictors of interobserver agreement in breast imaging using the Breast Imaging Reporting and Data System. *Breast cancer research and treatment.* 2010;120(3):539-46.

CHAPTER 2

Reader variability in recall rates have important clinical and economic implications, such as unnecessary follow-up procedures, additional costs to the health care system and psychological effects for the women themselves associated with false-positive mammogram results. However how varying recall rates affect a reader's performance when recalling women for further assessment in the screening service is not well understood.

The purpose of Chapter 2 (literature review) in this thesis is to provide the reader with a detailed understanding of the multifactorial nature that may affect recall rates in mammography through a conceptual map of the current literature. This conceptual map provides a holistic understanding of recall rates for the health care practitioner in general, and for the breast radiologists in particular, of their decision making processes surrounding recalling women in a screening service. Chapter 2 is a published journal paper titled "Understanding recall rates in screening mammography: A conceptual framework of the literature", which appeared in the journal *Radiography – Special Edition in Breast Imaging* (2015).

**UNDERSTANDING RECALL RATES IN SCREENING
MAMMOGRAPHY: A CONCEPTUAL FRAMEWORK REVIEW OF
THE LITERATURE**

Chapter 2 is published as:

N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, S. Lewis. “*Understanding recall rates in screening mammography: A conceptual framework review of the literature*”. *Radiography – Special Edition in Breast Imaging* (2015); 21(4): pp: 334-41. doi: 10.1016/j.radi.2015.06.003

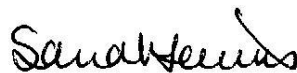
STATEMENT FROM AUTHOR

Statement from authors confirming authorship contribution of the PhD candidate

As co-authors of the paper “Understanding recall rates in screening mammography: A conceptual framework review of the literature”, we confirm that Norhashimah Mohd Norsuddin has made the following contributions:

- Conception and design of the research
- Data collection
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Asc. Professor Sarah Lewis



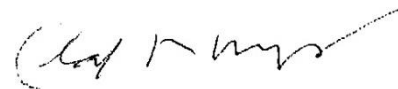
Date: 21/03/2017

Dr Warren Reed

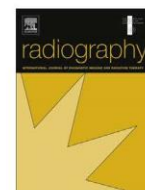


Date: 21/03/2017

Asc. Professor Claudia
Mello-Thoms



Date: 20/03/2017



Understanding recall rates in screening mammography: A conceptual framework review of the literature



N. Mohd Norsuddin^{a, b, *}, W. Reed^a, C. Mello-Thoms^a, S.J. Lewis^a

^a Medical Imaging Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences, Faculty of Health Sciences and Brain Mind Research Institute, The University of Sydney, Lidcombe, NSW, Australia

^b Diagnostic Imaging & Radiotherapy Programme, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, 50300 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:

Received 2 April 2015
Received in revised form
25 May 2015
Accepted 15 June 2015
Available online 17 July 2015

Keywords:

Recall rates
False positive results
Cancer detection rate
Screening mammography
Multidisciplinary
Conceptual framework

ABSTRACT

Recall rates are one of the performance measures used to evaluate the effectiveness of mammography screening programs. There is conflicting evidence regarding the link between recall rates and cancer detection rates and a variety of differing recall rates exist between countries and readers. This variability in recall rates may have important clinical and economic implications such as unnecessary follow-up procedures, additional costs to the health care system and psychological effects for the women themselves associated with false-positive mammograms results. In order to reduce the impact of false positive recall rates in screening mammography, it is essential for all multidisciplinary health care providers, especially those in medical imaging, to fully understand the factors that may contribute and affect recall rates. The multifactorial nature of recall rates is explored in this paper through the construction of a conceptual map based on a review of the current literature.

© 2015 The College of Radiographers. Published by Elsevier Ltd. All rights reserved.

Introduction

Recall rates are one of the main performance indicators that play a significant role in determining the overall accuracy of a screening mammography programme.^{1–6} The performance of screening mammography from a medical imaging perspective is generally measured by indicators such as sensitivity, specificity, positive predictive rates (PPV) and cancer detection rates (CDR). To underpin this paper, the performance measures associated with screening mammography as framed from a medical imaging practice perspective have been defined in Table 1 to assist with understanding the conceptual framework showcasing recall rate variability (Fig. 1).

There are benefits to recalling a percentage of women within a screened population as there is a relationship between recall rates and the early detection of breast cancer.⁴ However, it is important to consider what may be an “optimal” rate as false positive recalls have important clinical and economic implications such as unnecessary follow-up procedures, costs and adverse psychological

effects upon the women recalled.^{7–9} Statistics have shown that screening mammography can be successful in reducing breast cancer mortality, with reductions of 21%–26% in women aged 50–69 and a 32% reduction in women aged 70 and above.¹⁰ Inadequate specificity in mammography leads to potential harms such as overdiagnosis, missed cancers and false-positive results.^{11–13}

The risk of a false-positive screening result is positively correlated with the recall rate.⁴ This rate is initially influenced by the mammographic technologies available at the point of screening, such as screen-film mammography (SFM) and full field digital mammography (FFDM).^{14,15} The human physical parameters, that are brought into consideration include the breast reader's expertise and work experiences as well as the woman's presentation (for example clients age, screening history, use of hormone therapy, breast density, previous invasive procedure and familial breast cancer).^{5,16,17} Although the technical factors are often easier to control and can be standardized to some extent the human physical parameters of the readers (generally radiologists but also occasionally non-radiologist readers such as breast physicians and radiographers) along with the population of screened women that may present at any given time to a screening mammography programme present a real challenge and contribute to the variability in recall rates.¹⁸

* Corresponding author. Medical Imaging Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences, Faculty of Health Sciences, The University of Sydney, East Street, Lidcombe, NSW 2141, Australia. Tel.: +61 4 50595504.
E-mail address: nmoh5894@uni.sydney.edu.au (N. Mohd Norsuddin).

Table 1
Definition of terms in screening mammography performance.

Name	Definition
Sensitivity	Measures the percentage or fraction of actual positive cancer cases that are correctly identified. Often described as a decimal. The ability of a test to correctly detect the presence of disease.
Specificity	Measures the percentage or fraction of cancer free cases that are correctly identified. Often described as a decimal. The ability of a test to correctly detect the absence of disease.
Recall rate	The proportion of screened women that are asked to return for further assessment. Often expressed as a percentage.
False positive result	The decision made in error that a case is positive for cancer when the case is actually cancer free.
Positive predictive value (PPV)	The probability of screened women with a positive (malignant) test that do have breast cancer. Often expressed as a percentage.
Negative predictive value (NPV)	The probability of screened women with a negative (normal) test that do have breast cancer. Often expressed as a percentage.
Cancer detection rate (CDR)	The proportion of screened women with breast cancer who test positive for breast cancer. Often expressed as a percentage.

In this review we introduced a novel conceptual mapping of the factors that need to be considered when recall rates are evaluated. In particular, we focus on three main areas that may contribute to a woman being recalled for further assessment, including imaging technologies, differences in practices among breast readers and characteristics of the population screened (patient). These three areas were chosen as focal points of discussion in this article through broad thematic analysis of over 400 past published research papers in the area of recall rates in screening mammography.

A conceptual understanding of recall rates

Fig. 1 illustrates the multifactorial nature of recall rates from the literature. The development process of conceptual mapping in this paper began with identification of the wide variation in recall rates in international screening mammography programmes. The search of the literature was conducted in MEDLINE, CINAHL (EbSCOhost), SPIE library, Web of Science, PubMed, Scopus databases and Google Scholar. No specific year of publication was imposed in this search however, we prioritised studies from 2000 onwards which were likely to capture current imaging modalities in screening mammography. Keywords included in this review were recall rates, false positive results, screening mammography, observer performances, performance measures, screening performance, sensitivity, specificity, radiologists and medical imaging. The aetiology of the recall rates were identified and classified into three main sections; imaging technologies such as digital and analogue image acquisition, differences in practice such as the volume of cases read and the experience of the reader and characteristics of screened population including breast density and geographical location. The emboldened factors in the conceptual framework indicate distinct sections that are further discussed in this review.

Despite the fact that breast cancer diagnosis and care is multi-disciplinary and inter-professional, much of the past literature has been narrow in scope, seeking to understand recall rates as a single entity remote from other influences. Fig. 1 shows the breath of the variables that affect recall rates, providing health care practitioners with a greater understanding of their role in the context of the larger decision making processes surrounding recalling women in a screening service. This way, each practitioner unit can comprehend holistically the importance of optimising their practice, implementing quality control and an appreciation for individual women's scenarios and why they may be concerned about being recalled for further assessment.

Imaging technologies (screen-film – full field digital mammography)

Over the last decade, the evolution of digital technologies has transformed the technical quality of mammograms through improved image receptor systems that allow for more consistent image quality with higher contrast resolution, fewer artifacts^{19,20}

and lower radiation dose^{21,22} than previous screen-film mammography (SFM). The benefits of post processing capabilities in full field digital mammography (FFDM) have also improved cancer detection, especially in dense breast parenchyma.^{14,23} With FFDM, images that have been under-exposed or over-exposed are no longer necessarily repeated, resulting in a lower recall rate among screened women who may have questionable technical images. Although SFM has better spatial resolution than FFDM, which allows better visualization of fine structures that act as biomarkers for breast cancer, such as microcalcification, FFDM has the ability to alter the image contrast and digital information after exposure through magnification, image windowing and panning.

Trials involving comparisons with FFDM and SFM in a screening context have demonstrated conflicting results with regards to recall rates.^{15,24,25} A clinical trial by Lewin et al.¹⁵ in the Colorado–Massachusetts Study found no significant differences between FFDM and SFM in cancer detection but with significantly reduced recall rates for women imaged with FFDM. A prospective trials by Skaane et al. concurred with Lewin et al. for results in cancer detection but found higher recall rates for FFDM (Oslo I, 4.6%; Oslo II, 4.2%) when compared to SFM (Oslo I, 3.5%; Oslo II, 2.5%).^{24,25} Despite these inconclusive findings, other studies have not replicated such a vast difference between SFM and FFDM.^{22,26,27} Results from the Digital Mammographic Imaging Screening Trial (DMIST) report there were no differences between SFM and FFDM for the entire population, with the CDR of 0.4% and 0.44% for SFM and FFDM respectively and the recall rate was exactly the same at 8.6% for SFM and FFDM. A specific finding of DMIST was that FFDM demonstrated a significant advantage for specific cohorts of women such as those under the age of 50 years, women with mammographically dense breasts and premenopausal women.²⁸

Some limitations of the earlier studies when comparing recall rates in SFM and FFDM included workstation designs with limited post-processing capabilities, emerging detector development and the unfamiliarity of the readers with reading digital cases when transitioning from only reading SFM cases which may affect their study findings.¹⁵ However, with the present wide spread use of FFDM, especially in developed countries, these earlier influencing factors have been improved. Interestingly, readers that have experience in reading SFM perform higher for specificity than those who have only read digital cases. A recent study of 129 radiologists by Rawashdeh et al. (2015) has found that readers who had limited experience with screen-film reading were likely to have lower specificity (0.70 versus 0.83; $p < 0.001$) and hence higher recall rates in comparison to readers that had previous hard copy reading experience, even when there was statistical control for age and experience.²⁹

Differences in practice (breast readers)

Reader background

In this section, we explore the characteristics of breast readers, that is, the observer who views the images and arrives at a decision

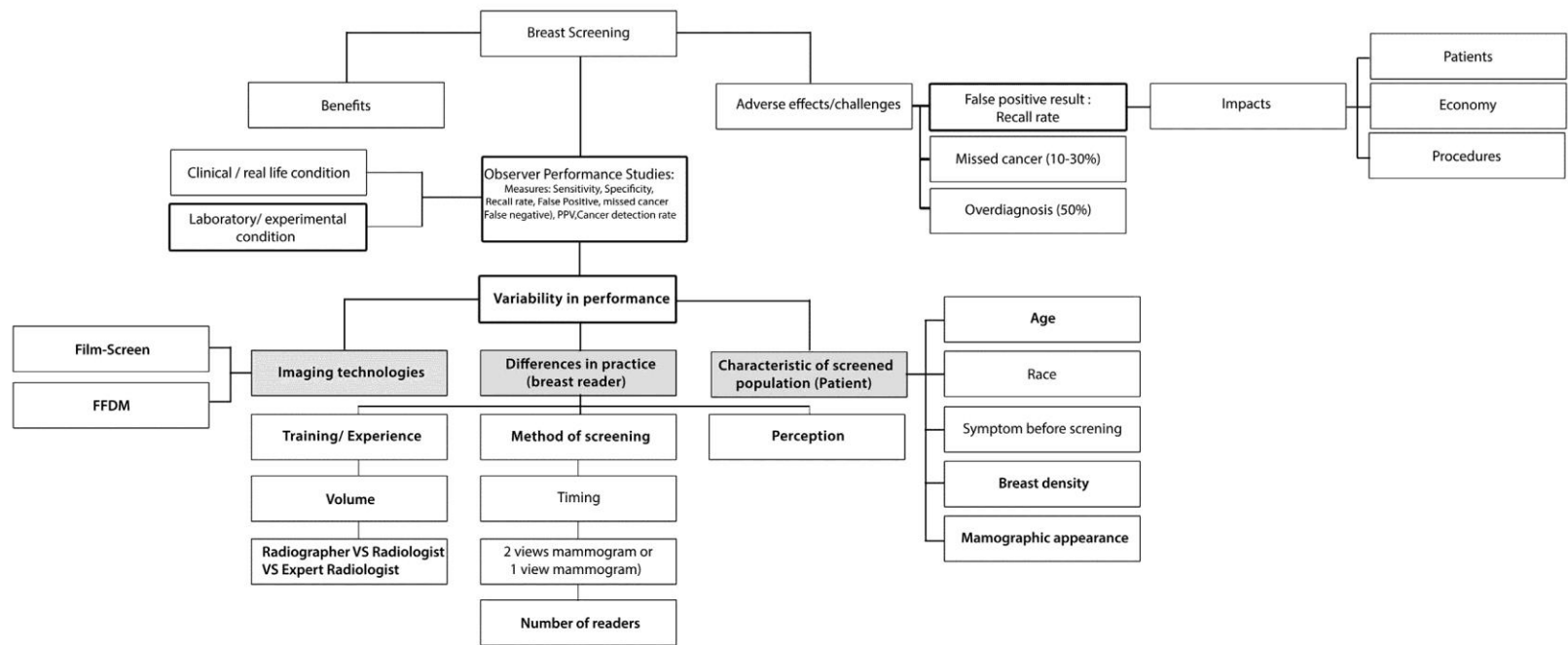


Figure 1. Conceptual diagram of factors that contribute to a false positive-recalled decision outcome.

of recall in a screening programme. In most countries this work is performed by radiologists although some countries have breast readers with similar reporting duties, such as breast-reporting radiographers in the United Kingdom (UK) and breast physicians in Australia. However, most of larger published literature relates to the performance of radiologists as a distinct cohort and here we discuss the research for this group, using the term reader and radiologist interchangeably.

In acknowledging the role of non-radiologists as readers, there has been recent published literature about the role of radiographers acting as screen readers. In the UK, radiographers are part of the workforce that have been formally trained to undertake mammography as well as read screening mammograms.³⁰ Although evidence suggests that the accuracy of radiographer screen-readers is acceptable for practice,³⁰ Bennett et al. reports an increase in recall rates associated with radiographer readers without an improvement in the cancer detection rate. Similarly in a small study of Australian radiographer screen readers ($n = 10$, reading screen-film cases not FFDM), Debono et al. details a variety of accuracy levels with reasonable sensitivity and specificity when compared to radiology-gold standard decision.³¹ Further analysis of recall rates for radiographer readers is necessary to assess their widespread impact.

Dedicated education in mammography

It is anticipated that readers will improve their performance in screening mammography throughout their career. Dedicated training and specialized education in breast imaging has been shown to produce better observer performance, especially in reducing the number of recalled women alongside higher cancer detection rates. A study by Miglioretti et al. (2009) has demonstrated that readers who have undertaken dedicated fellowship training in breast imaging have reduced false positive recall rates and meet an acceptable standard of national performance very early within their working lives, exhibiting a faster learning curve than radiologists that acquire expertise through years of experience alone.³² Despite this finding, it appears the fellowship system as a mechanism for improved post-graduate specialisation has not seen wide spread use internationally, with a paucity of research focusing on potential benefits of formal mammographic education. We could only identify one study by Elmore et al. (2009), who has a similar authorship team to Miglioretti et al., which also addresses the benefits of fellowship training. In both articles, the authors report that fellowship-trained radiologists make a small proportion of breast radiologists (less than 10%) who are available to report on screening mammograms but that the recall rate could be decreased by the fellowship system.

In the last decade, dedicated training or continuing professional development and feedback programmes for readers to improve their clinical performance have been introduced, including Personal Performance in Mammographic Screening (PERFORMS)³³ in the UK and the BreastScreen Reader Assessment Strategy (BREAST) in Australia.³⁴ In these structured programmes breast readers are asked to interpret test sets of mammographic cases for their own performance development. With BREAST, immediate feedback is provided to the reader about their cancer-detection performance allowing for identification of an individual's strengths and weaknesses as well as an analysis of the types of lesions commonly missed or associated with incorrect decisions.³⁴ Feedback in real time screening mammography is unlikely however the value of multidisciplinary breast cancer care and the importance placed upon the feedback loop between radiologists, pathologists and surgeons was highlighted in an article by Alcantara et al. (2013). This study of Australian readers reported that radiologists who attended

multidisciplinary breast cancer team meeting had improved confidence in their screening decisions.³⁵

Reader experience and volume

Recent studies have demonstrated that inter observer variability may be due to differences in readers' characteristics and work practices. Two variables commonly investigated to determine performance are reader experience (the length of service or years in interpreting mammograms) as cumulative experience and annual reading volumes.^{18,36} Experienced readers generally have reduced false positive rates in screening mammography.³⁷ Readers with more years of experience read and interpret images faster and can identify lesions more accurately^{38,39} and eye tracking analysis has also shown that incorrect decisions such as false positives and false negatives actually attract extended visual time, with correctly identified cancers visually located more quickly and efficiently.^{39,40}

Reading high volumes of mammograms is likely to improve the identification of the normal variations in breast tissues allowing readers to be more certain of benign findings and identify a range of cancers presentations.⁴¹ It has been shown that readers who read a larger number of mammograms have lower false positive results, because these readers have developed a better knowledge bank of normal presentations seen on screening mammograms.^{36,42} Various breast screening programmes and organizations have set minimum annual volumes of cases to be read by breast readers. This volume varies across countries; from 960 cases during a 24 months period in the United States,⁴³ 2000 cases per year in Australia⁶ and as high as 5000 mammograms per year in the UK.⁴⁴ It is suggested that a minimum case threshold may allow readers to reduce the number of women recalled for further assessment (abnormal interpretation rate) and to increase the CDR.⁴⁵

In considering if there is a maximum threshold by which over there is saturation or decline in performance, an editorial by Given-Wilson (2011) reviewed two studies conducted in East Midland (England) and Scotland, which sought to determine if reading higher volumes could be detrimental. These two studies, indicated there is no evidence that higher volumes of reading would decrease overall performance, with Concord et al. noting at higher volumes (more than 8333 reads per year) recall rates may slightly increase without any associated increase in the CDR.⁴⁶ A similar result was found among high volume readers in US^{47,48} and Australian studies by Reed et al.⁴⁹ and Rawashdeh et al. have demonstrated a significant correlation between, the CDR and volume of cases ($P \leq 0.03$; with higher specificity).⁵⁰ Furthermore, the findings from Rawashdeh et al.²⁹ suggest that readers who read more than 5000 mammographic cases per year have significantly better results in identifying normal mammograms effectively increasing in the readers' specificity scores and reducing recall rates through false positive decisions. On the reverse side, readers who report on less than 1000 mammographic case per year appear to have lower specificity scores (0.60)⁵⁰ and this in turn has a direct correlation to increased recall rates.

These research findings have important implications for radiologists or other breast readers that infrequently read screening mammograms or for countries where populations and readers may be geographically dispersed and teleradiology is not available. The importance of large and regular volumes of screening cases appears vital to produce higher performance and lower recall rates. Proposals for the wider inclusion of other readers to boost mammography reporting workforce, such as radiographer readers need to consider the effects of infrequent reading or low volume reads.³¹

The variability in the case load of different breast screening programmes may also contribute overall to reader performance and hence recall rates. Reading mammograms is a difficult visual search task for readers and can lead to error (perceptual and

Table 2
Summary of studies investigating the link between recall rates and performance measures.

Authors in reference list	Aims	Method	Population mammography screening	Time period (subsequent screening)	Statistical analysis	Main findings
⁵⁸	To explore the variance in recall rates and the extent to which recall rates can be compared across countries.	Prospective Study: questionnaire		10–29 months	<ul style="list-style-type: none"> Multivariate analysis 	No consistent relationship of initial to subsequent PPV Pattern: increasing recall rate, decreasing PPV
³	To measure the effect on sensitivity and positive predictive value as recall rates.	Prospective Study: 31 radiologists	Population-based mammography database (Carolina Mammography Registry)	12 months	<ul style="list-style-type: none"> Reduced monotonic regression analysis (nonparametric approach). Linear regression analysis 	Practices with recall rates between 4.9% and 5.5% achieve the best trade-off of sensitivity and PPV.
⁴	To determine the effect of changes in recall rate on earlier detection of cancers (in relation to false-positive rates)	Retrospective - 15 blinded radiologists (10 Dutch, 5 non-Dutch) - image interpretations alone	Population-based Dutch Screening Program (biennial screening)	24 months	<ul style="list-style-type: none"> Localization Receiver Operating Characteristic (LROC) 	The effect of increasing the recall rate is most obvious between 1% and 4%. If recall rate more than 4%, cancer detection rates level off, and will increase numbers of false positives
⁵⁴	To investigate the correlation between recall and detection rates in a group of 10 radiologists who had read high volume of screening mammograms	Retrospective - 10 radiologists - film based - BIRADS interpretation systems - with prior images		–	<ul style="list-style-type: none"> Correlation: Parametric Pearson (r) and nonparametric Spearman (rho) 	Linear fit between the recall and detection rates. - significant correlation ($p < 0.05$) - an average of 0.22 additional detections per 1% increase in recall rates - higher recall rates, higher detection rates - significant inter-reader variability
⁷³	To identify target recall rates for screening mammography on the basis of how sensitivity shifts with recall rate.	Retrospective study (1991–2001)	Community-based screening program		<ul style="list-style-type: none"> Isotonic regression analysis Reduced monotonic regression Reduced monotonic regression model Concave fit 	Recall rates of 10.0% for first and 6.7% for subsequent mammograms are recommended targets on the basis of their additional work-ups per additional cancer detected (AW/ACD) rates (less than 100). At a recall rate of 12.3%, the estimated AW/ACD was 304, which suggests little benefit for any higher recall rate.

interpretive).⁵¹ During the reading process, some readers may have the ability to more easily recognise masses, where others may be better at detecting architectural distortion.⁵² Many breast cancers may also have subtle malignant appearances,⁵³ and if the reader only reads a small volume of cases, there is some evidence to show that a low prevalence to real-life breast cancer cases can reduce reader efficacy.¹³

Table 3
International comparison result of recall rates and cancer detection rates for initial/first and subsequent screening mammograms in women.

Country	Recall (%)		Cancer detection per 1000 mammograms	
	Initial	Subsequent	Initial	Subsequent
Netherlands	1.4 ^a		5.3	^a
Switzerland	3.4	2.4	3.9	4.0
England	8.0 ^a		6.8	^a
Canada	9.2	5.1	6.2	4.1
Australia	10.8	3.8	8.2	4.3
USA	15.1	9.0	7.5	3.4

Figures collated from Refs. 58,79.

^a Data is not available.

Reader work practices (logistics)

Differences in the method of screening programmes can also influence recall rate. When compared with single-reading, double interpretation of screening mammograms has been shown to improve CDR,^{54,55} especially for less experienced readers.⁵⁶ To date, many international screening programmes, such as those in Australia and Europe,⁴⁴ apply a double reader strategy when interpreting mammograms. For example, in Australia, each case is read by two readers with any difference in decision (recall versus no recall) to be decided by a third arbitrating reader.⁵⁷ With double reading, mammograms need to be interpreted by consensus, in which recall occurs only with agreement of the readers involved; or a decision on reader disagreement may be obtained by panel arbitration. In the USA, single reading is the usual practice whereby only one reader will interpret the mammographic case. The probability a woman may be recalled has been found to be higher if only one reader considers the mammogram abnormal and this can contribute to the higher percentage of recall rates among screened women in USA population.⁵⁸

In a large UK study, researchers postulated that arbitration and consensus between readers using the double reader strategy can lower recall rates, especially in detecting high difficulty cancers.⁵⁹ Furthermore, they suggested that the double reading of screening mammograms can increase the cancer detection rates when

compared to single reading without an overall increase in the recall rates across the screened population.⁵⁹

The characteristics of the screened population

The wide variation in recall rates observed in screening mammography can also be explained by different cohorts of women screened. The variety of the women's physical and social characteristics (such as age, physical breast density, use of hormone replacement therapy (HRT), a familial history of breast cancer) and "radiological" characteristics/history (that is optical/mammographic density, time between screenings, previous mammograms for comparison) influence the likelihood of false positive results.^{36,60–62}

Younger women and use of HRT have an increased risk of false positive results due to a direct link with increased breast density.⁶⁰ The appearance of the breast changes with a patient's hormonal status which in turn alters the mammographic appearance, including the optical density of the breast. These associated mammographic appearances in breast density may challenge the reader when interpreting the mammogram and lead to false results or high rate of recall. In addition, some characteristics of individual patients may influence the readers' decision to report a case abnormal. These include previous breast surgery/biopsies and a family history of breast cancer if recorded.^{63–65} On the other hand, women who had previous mammograms (priors) that are available for comparison have a lower percentage of false positive results.⁶⁶

The Breast Imaging Reporting and Data System (BIRADS) is system of classification of abnormalities and a subjective measure of breast density which is commonly used internationally to record findings from screening mammography. This qualitative measure of breast density, from 1 to 4, depends on the breast reader's perception of the percentage of dense breast tissue overall as displayed by the mammograms.⁶⁷ The differences in breast tissue composition (fat and fibroglandular) will affect the mammographic appearance and this naturally varies between women. Women with extremely high breast density are more likely to have higher chances of having false positive results and also being recalled.^{61,62,68,69} A recent study by Al Mousa et al. investigated cancer detection in highly dense breasts and found that FFDM could overcome the lower sensitivity in high breast density previous seen with SFM.⁷⁰ It was suggested that the use of image processing tools available in FFDM can enhance the visibility of cancers that overlay dense fibroglandular regions of the breast, hence improving cancer detection and this study has important future research implications for reducing recall rates in highly dense breasts.⁷¹

In 2014, there was a change in the BIRADS classification of abnormalities, particularly in relation to the management recommendation for BIRADS 3.⁷² Previously, a woman who has had her mammogram classified as a BIRADS 3 result will be recalled for follow up assessment to observe any changes on what is most likely a benign lesion. The fifth edition of the BIRADS classification now allows readers to add more information in the BIRADS 3 report including specific details such as the need for biopsy such as when a benign lesion increases in size.⁷² This allows for the expansion of the recall concept into recognition that recalling does not necessarily equate to suspicious regions or cancer but may relate to the management of benign findings.

Recall rates and cancer detection rates in screening mammography

When estimating accuracy in mammography screening, it is important to understand the probability of detecting breast cancer. If breast cancers are detected at an early stage, a range of effective management or treatment options can be offered to the women. Exactly how recall rates alter in relation to cancer detection rates is not well understood and the relationship between sensitivity and

recall rate is not necessary linear.³ Some studies have argued that an increased increment in recall rate could improve the number of cancers detected during screening (cancer detection rate) due to the earlier detection of interval cancers.⁴ However others have suggested that increasing recall rate simply increases the number of false-positive decisions.⁵

Table 2 collates the empirical based studies investigating the link between recall rates and sensitivity. Some studies have shown recall rates having a positive correlation with sensitivity and a negative correlation with PPV.^{3,73} One study that attempted to map the linear correlation between both recall rates and cancer detection was published by Gur et al.⁵⁴ This limited study, which included a group of 10 highly experienced readers, found significant inter-reader variability for both recall and cancer detection rates ranging from 7.7% to 17.2% and from 2.6 to 5.4 per 1000 mammograms. From this study, the authors concluded that readers may find more cancers if they increase their recall rates. However, studies with larger numbers of readers involved, such as Yankaskas et al. (2001), found that the cancer detection rate was only affected at the lower range of recall rates and should not affect CDR at higher recall rates more than 7%.³ It must be acknowledged, however, that the effect of changes in recall rates from Gur et al. and Yankaskas et al. did not quantify false positive findings. False positive cases were found to increase in a study by Otten et al. (2005), who discovered that lower recall rates between 1% and 4%, more cancers can be detected earlier, however, if the recall rate is higher than 4%, the number of false positive cases decisions is higher.⁴

The effect of an increased recall rate can be seen at certain target recall rates, particularly at the lower range between 1 and 5.5%.^{3,4} In attempting to quantify an "optimum" recall rate, a retrospective study by Schell et al. found target recall rates at 10% for first screening and 6.7% for subsequent screening yielded the best trade-off with sensitivity.⁷³

International data and interventions to reduce the recall rates

Recent literature has investigated a variety of performance measures, such as sensitivity, specificity and PPV for screening mammography. An international study by Yankaskas et al. (2004) has documented the variation in recall rates per country, providing evidence of the large difference between recall decisions that can be made for women attending screening in Europe, North America and Australia.⁵⁸ Table 3 shows a summary of available data detailing an international comparison of results of recall rates and cancer detection rates for first and subsequent screening among women. It can be seen that the variation ranges from 1.4% in Nordic countries to about 15% in western countries such as the United States, with recall rates for a first screen being higher than those for subsequent screens.^{3,18} However, no significant increase in the cancer detection rate was observed with a substantial increase of the recall rate.⁵⁸

Viewing these international variations, it is uncertain why some countries have screening programmes that allow for very limited recalls but with a possible trade-off for cancer detection.⁴ The United States operates with much higher recall rates and this may reflect a medical and screening environment where litigation is common and feared.⁷⁴ However what is clear is that there appears to be little evidence for the establishment of recall rates from similar type countries that may have similar breast cancer statistics let alone when there are obvious differences in the quality of health between countries, such as those from a developing to developed country. Health professionals should be aware of their own countries recall rates in relation to others as a catalyst for discussing optimised recall rates for the population screened.

Control over the variability of mammographic performance quality and the effectiveness of mammography programmes have

been addressed through the regulation of mammography facilities such as the Mammography Quality Standards Act (MQSA) 1992 in the USA^{43,75} and BreastScreen Australia National Accreditation Standards (NAS).⁶ These health policies provide a benchmark for the population-based screening programmes. Different organisation groups have recommended different recall policies as guidelines to evaluate the performance in the screening population in their respective nations or states.^{6,76}

The recommended target recall rate across Australian and European programmes varies significantly in their breast screening classification. Australia and Europe have separated recall rates into first screening and subsequent screening, with the first screen of a woman showing a higher recall rate than subsequent screens. The recommended recall rate in European countries (as averaged across nations with published data) is lower than in Australia for first screening with 7% and 10% respectively but with a similar recall rate for subsequent screening (5%). However, some European nations have recall rates that are lower than the standard recommendations in the European Guidelines,^{44,76} with 5% for the first screening and 3% for subsequent screenings across their population-based screening programme, such as the Netherlands and Switzerland.⁵⁸

Despite screening programmes having desired target recall rates, there is a lack of scientific evidence regarding the effect of enforcing target recall rates on the readers' performance (particularly in relation to cancer detection rates). The strategy of limiting a reader's recall rates using "target recall rates" has not been explored in the known literature and such a strategy may alter the readers' decision making process and can introduce biases in their recall decisions.^{77,78} However recall rates and false positive decisions that result in additional imaging and over diagnosis continue to be a strong concern for women participating in screening mammography programmes and hence research in this field is timely.

Conclusion

Recall rates and false positives decisions are a feature of clinical practice and breast readers' decision making strategies. Together, these two performance metrics play a significant role in the effectiveness of a screening programme extending from participation and imaging through to diagnosis and assessment. A better understanding of how multiple factors across multi-disciplines influence the recall rate in screening mammography is essential for optimising clinical practice, especially for practitioners involved in the medical imaging sphere. A recommendation for the future is to better determine optimal or target recall rates that may be applicable internationally or that may need to be set for distinct populations that are prone to overcalling and false positive decisions.

Conflict of interest

None.

References

- Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, et al. Recall and detection rates in screening mammography. *Cancer* 2004;100(8):1590–4.
- Halladay JR, Yankaskas BC, Bowling JM, Alexander C. Positive predictive value of mammography: comparison of interpretations of screening and diagnostic images by the same radiologist and by different radiologists. *Am J Roentgenol* 2010;195(3):782–5.
- Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *Am J Roentgenol* 2001;177(3):543–9.
- Otten JDM, Karsssemeijer N, Hendriks JHCL, Groenewoud JH, Fracheboud J, Verbeek ALM, et al. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. *J Natl Cancer Inst* 2005;97(10):748–54.
- Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, et al. Comparison of screening mammography in the United States and the United Kingdom. *J Am Med Assoc* 2003;290(16):2129–37.
- BreastScreen Australia. National accreditation standards: BreastScreen Australia quality 2008. Available from: [http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/A03653118215815BCA257B41000409E9/\\$File/standards.pdf](http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/A03653118215815BCA257B41000409E9/$File/standards.pdf) [cited 2014 May 21].
- Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *J Med Screen* 2012;19(Suppl. 1):57–66.
- Brodersen J, Siersma VD. Long-term psychosocial consequences of false-positive screening mammography. *Ann Fam Med* 2013;11(2):106–15.
- Lafata JE, Simpkins J, Lamerato L, Poisson L, Divine G, Johnson CC. The economic impact of false-positive cancer screens. *Cancer Epidemiol Biomarkers Prev Publ Am Assoc Cancer Res Cosponsored by Am Soc Prev Oncol* 2004;13(12):2126–32.
- Australian Institute of Health and Welfare. *Breast cancer in Australia: an overview*. Canberra: AIHW; 2012. p. 1039–3307. Contract No.
- Brodersen J, Jørgensen KJ, Gøtzsche PC. The benefits and harms of screening for cancer with a focus on breast screening. *Pol Arch Med Wewnętrznej* 2010;120(3):89–94.
- Kopans DB, Smith RA, Duffy SW. Mammographic screening and "over-diagnosis". *Radiology* 2011;260(3):616–20.
- Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS One* 2013;8(5):e64366.
- Berns EA, Hendrick RE, Cutter GR. Performance comparison of full-field digital mammography to screen-film mammography in clinical practice. *Med Phys* 2002;29(5):830–4.
- Lewin JM, Edward Hendrick R, D'Orsi C, Isaacs P, Moss L, Karellas A, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. *Radiology* 2001;218(3):873–80.
- Hofvind S, Møller B, Thoresen S, Ursin G. Use of hormone therapy and risk of breast cancer detected at screening and between mammographic screens. *Int J Cancer* 2006;118(12):3112–7.
- Fracheboud J, de Koning HJ, Boer R, Groenewoud JH, Verbeek ALM, Broeders MJM, et al. Nationwide breast cancer screening programme fully implemented in the Netherlands. *Breast* 2001;10(1):6–11.
- Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. *J Natl Cancer Inst* 2003;95(18):1384–93.
- Feig SA, Yaffe MJ. Current status of digital mammography. *Semin Ultrasound CT MRI* 1996;17(5):424–43.
- Feig SA, Yaffe MJ. Digital mammography. *Radiographics* 1998;18(4):893–901.
- Juel JM, Skaane P, Hoff SR, Johannessen G, Hofvind S. Screen-film mammography versus full-field digital mammography in a population-based screening program: the Sogn and Fjordane study. *Acta Radiol (Stockholm, Sweden 1987)* 2010;51(9):962–8.
- Heddsen B, Ronnow K, Olsson M, Miller D. Digital versus screen-film mammography: a retrospective comparison in a population-based screening program. *Eur J Radiol* 2007;64(3):419–25.
- Ng K, Muttarak M. Advances in mammography have improved early detection of breast cancer. *J Hong Kong Coll Radiol* 2003;6:126–31.
- Skaane P, Young K, Skjennald A. Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading—Oslo I study. *Radiology* 2003;229(3):877–84.
- Skaane P, Skjennald A, Young K, Egge E, Jøbsen I, Sager EM, et al. Follow-up and final results of the Oslo I study comparing screen-film mammography and full-field digital mammography with soft-copy reading. *Acta Radiol (Stockholm, Sweden 1987)* 2005;46(7):679–89.
- Del Turco MR, Mantellini P, Ciatto S, Bonardi R, Martinelli F, Lazzari B, et al. Full-field digital versus screen-film mammography: comparative accuracy in concurrent screening cohorts. *AJR Am J Roentgenol* 2007;189(4):860–6.
- Vigeland E, Klaasen H, Klingens TA, Hofvind S, Skaane P. Full-field digital mammography compared to screen film mammography in the prevalent round of a population-based screening programme: the Vestfold County study. *Eur Radiol* 2008;18(1):183–91.
- Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005;353(17):1773–83.
- Rawashdeh MA, Lewis SJ, Lee W, Mello-Thoms C, Reed WM, McEntee M, et al., editors. *Experience in reading digital images may decrease observer accuracy in mammography*. SPIE; 2015.
- Bennett RL, Sellars SJ, Blanks RG, Moss SM. An observational study to evaluate the performance of units using two radiographers to read screening mammograms. *Clin Radiol* 2012;67(2):114–21.
- Debono JC, Poulos AE, Houssami N, Turner RM, Boyages J. Evaluation of radiographers' mammography screen-reading accuracy in Australia. *J Med Radiat Sci* 2015;62(1):15–22.
- Miglioretti D, CC G, Carney P, Omega T, Buist D, Sickles E, et al. When radiologists perform: best the learning curve in screening mammogram interpretation. *Radiology* 2009;253(3).

33. Gale AG. Performs: a self-assessment scheme for radiologists in breast screening. *Semin Breast Dis* 2003;6(3):148–52.
34. BREAST. BreastScreen reader assessment strategy. Available from: <http://www.breastaustralia.com> [cited 2015 24.03.2015].
35. Alcantara SB, Reed W, Willis K, Lee W, Brennan P, Lewis S. Radiologist participation in multi-disciplinary teams in breast cancer improves reflective practice, decision making and isolation. *Eur J Cancer Care* 2014;23(5):616–23.
36. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst* 2002;94(18):1373–80.
37. Alberdi RZ, Llanes AB, Ortega RA, Exposito RR, Collado JM, Verdes TQ, et al. Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. *Eur Radiol* 2011;21(10):2083–90.
38. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, et al. How experience and training influence mammography expertise. *Acad Radiol* 1999;6(10):575–85.
39. Mello-Thoms C, Hardesty L, Sumkin J, Ganott M, Hakim C, Britton C, et al. Effects of lesion conspicuity on visual search in mammogram reading. *Acad Radiol* 2005;12(7):830–40.
40. Nakashima R, Kobayashi K, Maeda E, Yoshikawa T, Yokosawa K. Visual search of experts in medical image reading: the effect of training, target prevalence, and expert knowledge. *Front Psychol* 2013;4.
41. Haneuse S, Buist DSM, Miglioretti DL, Anderson ML, Carney PA, Onega T, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. *Radiology* 2012;262(1):69–79.
42. Esserman L, Cowley H, Eberle C, Kirkpatrick A, Chang S, Berbaum K, et al. Improving the accuracy of mammography: volume and outcome relationships. *J Natl Cancer Inst* 2002;94(5):369–75.
43. U.S. Department Of Health and Human Services. *An overview of the final regulations implementing the Mammography Quality Standards Act of 1992*. Rockville, Md: U.S. Department of Health and Human Services; 1997. p. 16–9.
44. National Health Service Breast Screening Radiologist Quality Assurance Committee. *Quality assurance guidelines for radiologists*. England: NHSBSP Publications; 1997. National Health Service Breast Screening Programme publication no 15 Sheffield.
45. Kan L, Olivetto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. *Radiology* 2000;215(2):563–7.
46. Given-Wilson R, Blanks R. Does quantity of film reading affect quality? *Clin Radiol* 2011;66(2):97–8.
47. Miglioretti DL, Smith-Bindman R, Abraham L, Brenner RJ, Carney PA, Bowles EJ, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. *J Natl Cancer Inst* 2007;99(24):1854–63.
48. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst* 2004;96(24):1840–50.
49. Reed WM, Lee WB, Cawson JN, Brennan PC. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. *Acad Radiol* 2010;17(11):1409–13.
50. Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings. *Radiology* 2013;269(1):61–7.
51. Berlin L. Accuracy of diagnostic procedures: has it improved over the past five decades? *Am J Roentgenol* 2007;188(5):1173–8.
52. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretation of mammograms. *N Engl J Med* 1993;331(22):1493–9.
53. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics* 2003;23(4):881–95.
54. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, et al. Recall and detection rates in screening mammography: a review of clinical experience - implications for practice guidelines. *Cancer* 2004;100(8):1590–4.
55. Ciatto S, Ambrogetti D, Bonardi R, Catarzi S, Risso G, Rosselli Del Turco M, et al. Second reading of screening mammograms increases cancer detection and recall rates. Results in the florence screening programme. *J Med Screen* 2005;12(2):103–6.
56. Tabar L, Vitak B, Chen TH, Yen AM, Cohen A, Tot T, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology* 2011;260(3):658–63.
57. Australian Institute of Health and Welfare. *BreastScreen Australia data dictionary: version 1.1*. 2015.
58. Yankaskas BC, Klabunde CN, Ancelle-Park R, Rennert G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? *J Med Screen* 2004;11(4):187–93.
59. Blanks RG, Wallis MG, Moss SM. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. *J Med Screen* 1998;5(4):195–201.
60. Christiansen CL, Wang F, Barton MB, Kreuter W, Elmore JG, Gelfand AE, et al. Predicting the cumulative risk of false-positive mammograms. *J Natl Cancer Inst* 2000;92(20):1657–66.
61. Lehman CD, White E, Peacock S, Drucker MJ, Urban N. Effect of age and breast density on screening mammograms with false-positive findings. *Am J Roentgenol* 1999;173(6):1651–5.
62. Cook AJ, Elmore JG, Miglioretti DL, Sickles EA, Aiello Bowles EJ, Cutter GR, et al. Decreased accuracy in interpretation of community-based screening mammography for women with multiple clinical risk factors. *J Clin Epidemiol* 2010;63(4):441–51.
63. Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. *J Am Med Assoc* 1997;277(1):49–52.
64. Castells X, Molins E, Macia F. Cumulative false positive recall rate and association with participant related factors in a population based breast cancer screening programme. *J Epidemiol Community Health* 2006;60(4):316–21.
65. Sala M, Salas D, Asuncion N, Zubizarreta R, Castells X. Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening. *Ann Oncol* 2012;23(1):104–11.
66. Burnside ES, Sickles EA, Sohlich RE, Dee KE. Differential value of comparison with previous examinations in diagnostic versus screening mammography. *AJR Am J Roentgenol* 2002;179(5):1173–7.
67. American College of Radiology. *American College of Radiology Breast Imaging Reporting and Data System (BIRADS)*. 4th ed. Reston, VA: American College of Radiology.
68. Boyd NFMDD, Guo HM, Martin LJP, Sun LM, Stone JM, Fishell EMDF, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med* 2007;356(3):227–36.
69. Mandelson MT, Oestreich N, Porter PL, White D, FINDER CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. *J Natl Cancer Inst* 2000;92(13):1081–7.
70. Al Mousa DS, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? *Clin Radiol* 2014;69(4):333–41.
71. Al Mousa DS, Mello-Thoms C, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Mammographic density and cancer detection: does digital imaging challenge our current understanding? *Acad Radiol* 2014;21(11):1377–85.
72. American College of Radiology. *Reporting system*. In: *BI-RADS-mammography 2013 (Internet)*; 2013. p. 121–40. Available from: <http://www.acr.org/Quality-Safety/Resources/BIRADS/Mammography>.
73. Schell MJ, Yankaskas BC, Ballard-Barbash R, Qaish BF, Barlow WE, Rosenberg RD, et al. Evidence-based target recall rates for screening mammography. *Radiology* 2007;243(3):681–9.
74. Berlin L. Radiologic errors and malpractice: a blurry distinction. *Am J Roentgenol* 2007;189(3):517–22.
75. Feig SA, D'Orsi CJ, Hendrick RE, Jackson VP, Kopans DB, Monsees B, et al. American College of Radiology guidelines for breast cancer screening. *Am J Roentgenol* 1998;171(1):29–33.
76. Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Ann Oncol Off J Eur Soc Med Oncol/ESMO* 2008;19(4):614–22.
77. Gunderman RB. Biases in radiologic reasoning. *Am J Roentgenol* 2009;192(3):561–4.
78. Fleck MS, Samei E, Mitroff SR. Generalized "satisfaction of search": adverse influences on dual-target search accuracy. *J Exp Psychol Appl* 2010;16(1):60–71.
79. Australian Institute of Health and Welfare. *BreastScreen Australia monitoring report 2010–2011*. 2013.

CHAPTER 3

EXTENDED METHODS

This chapter is an extended version of method that performed in the experiment work for the studies presented in this thesis. The work presented in this chapter provides a detailed explanation of the process involved pertaining to the research work for this thesis. The methodological approach of this work aided in minimizing the potential cofounders and biases that existed in actual screening reading. The work in this chapter was incorporated into four main phases; test set development, workstation set-up, reading sessions and data analysis.

Ethical Approval

Ethics approval was obtained from University of Sydney Human Ethics Research Committee for this study (Project number 2014/484). Documents related to the ethics of this study can be found in the appendix (Appendices A, B, C).

Participants

In acknowledging there are other screen readers who provide breast screening interpretation including radiologists, breast physicians, mammographers and breast practitioners, the participants in this thesis are Australian and the study is contextualized to be in the Australian healthcare setting. Almost 95% of the mammographic interpretation in Australia is provided by radiologists and a very small amount of the mammograms are read by breast physicians. Radiographers do not report on mammograms in Australia and there is no legislative pathway for them to do so. They are not recognized readers of medical images by the Commonwealth Department of Health. Therefore, the term used to describe readers of screening mammograms in this thesis is breast radiologists.

All participants in this study are board-certified breast radiologists with 9 to 26 years of experience in interpreting screening mammograms and a median of 8,000 mammographic readings per year. **Table 1** details the demographics of the participating breast radiologists. This study was conducted at the Medical Imaging Optimisation and Perception Group (MIOPeG) laboratory at the Brain and Mind Centre (BMC) of the University of Sydney and all participants took part voluntarily.

All participants were recruited from BreastScreen New South Wales (BSNSW). An invitation to participate in this study was sent to all radiologists who reported mammograms for BSNSW. The potential participants were contacted and provided with a Participant Information Sheet (Appendix B) in order to obtain more information about the study. A mutually suitable time for them to attend the BMC to interpret the cases at three separate times was determined. Prior to each reading session, a brief explanation and demonstration was given to the participants on how to read images using the customised recording software. A consent form (Appendix C) was signed before data collection took place. As an expression of gratitude for participation in this study, all participating breast radiologists were given a small gift voucher worth 100 AUD.

Table 1 Demographic details of participating breast radiologists

Reader number	Number of years of experience	Number of mammography cases read per year	Number of hours per week reading mammograms
1	15	30 000	10
2	26	10 000	3
3	15	10 000	10
4	9	6 000	24
5	20	3 500	6
Median	15	8 000	10

Research design

This study was a laboratory-based experiment. The initial stage of this study involved preparing the laboratory environment for the participants which incorporated three steps/phases;

- i. Workstation set-up
- ii. Monitor calibration
- iii. Test set development

Workstation set-up

The laboratory reading environment and reading procedure were designed to be as authentic as possible to the clinical environment in BreastScreen NSW. The images in the test set were displayed on a pair of EIZO Radioforce GS510 medical-grade monitors (Ishikawa, Japan) driven by a Sectra (Linköping, Sweden) workstation which is a web-based picture archiving and communication system (PACS). Two general monitors (Dell Inc, United States) were used; **A**) one for displaying the work list of mammographic images in the Sectra system and **B**) the other displaying the customised recording software interface (**Figure 1**). The workstation was placed in a room with no natural light and the surrounding wall painted with a light grey matte colour, to minimize specular reflection.

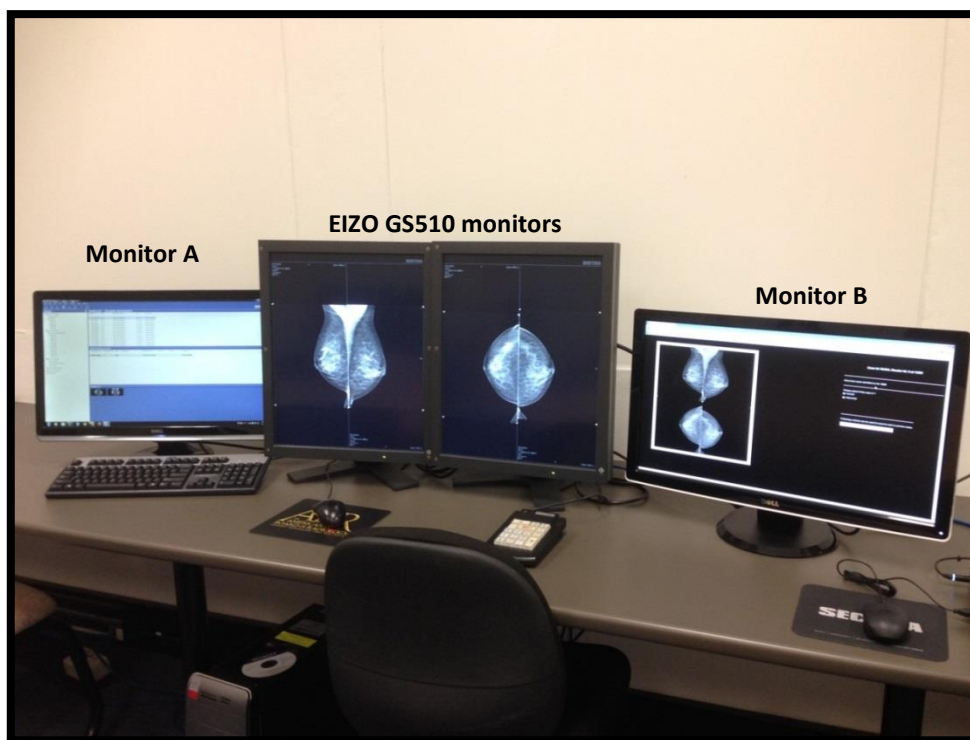


Figure 1 Workstation setup for participants in MIOPeG laboratory.

Monitor calibration

Prior to each reading session, the monitors were calibrated to conform to the Digital Imaging and Communication in Medicine (DICOM) Gray-scale Display Function (GSDF) part 14 standard (1). The purpose of the calibration was to ensure the medical image display used in this study was operating at consistent acceptable levels.

Firstly, the reading monitors were positioned away from direct light sources to minimize reflection. Next, each monitor was warmed-up for at least 30 minutes to stabilize the monitor output before any measurements were taken. Both monitors were calibrated using Verilum software (IMAGE Smiths Inc., Germantown, Maryland, US), whilst

monitor performance was assessed according to the American Association of Physics in Medicine (AAPM) Task Group 18 Quality Control (TG18-QC) guidelines (2).

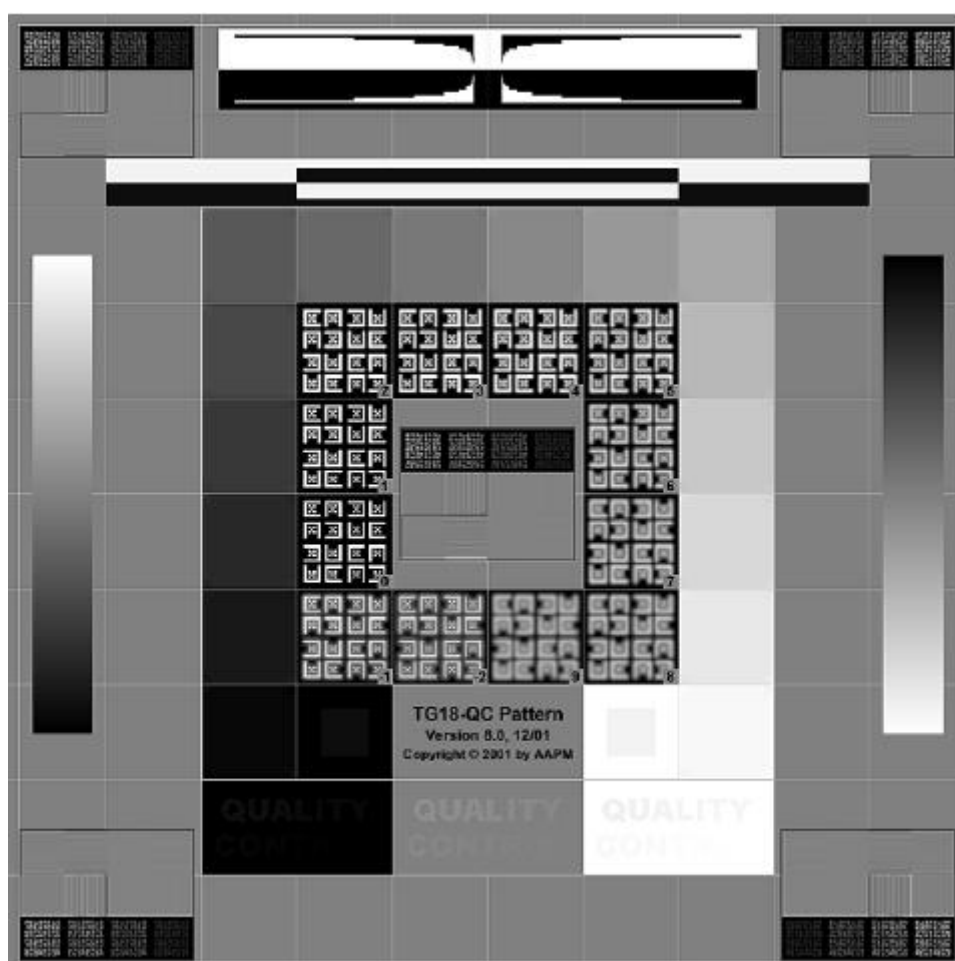


Figure 2 The TG18-QC comprehensive test pattern used in the study.

The calibration of the monitors was performed using VeriLUM® 5.2 software (IMAGE Smiths Inc., USA). It incorporates a sensor that measures the minimum luminance and maximum luminance of the monitors at the centre of the display screen for 30-50 seconds as in **Figure 3**.

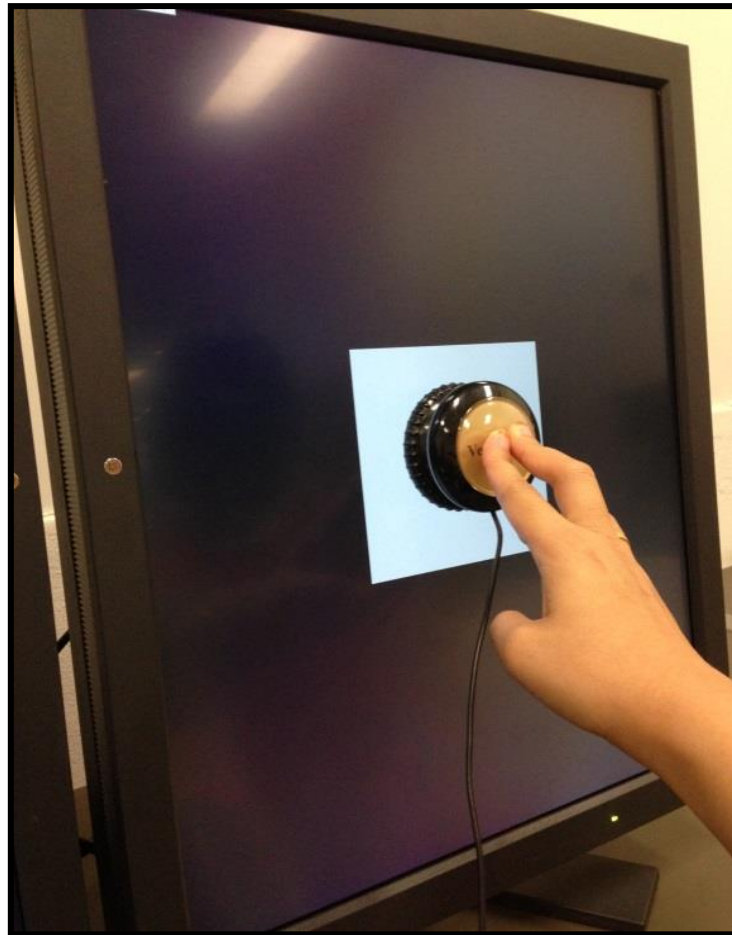


Figure 3 Measuring minimum luminance and maximum luminance of the monitors at the centre of the display screen.

The calibrated results of both reading monitors are presented in **Table 2**. Monitor 1 and monitor 2 demonstrated the maximum luminance of 421cd/m^2 and 433cd/m^2 respectively and the same minimum luminance of 0.88 cd/m^2 . The contrast ratio for monitor 1 was 488:1 slightly lower than monitor 2 with a contrast ratio of 492:1. These results were in the tolerance levels recommended by the AAPM (2).

Table 2 Calibrated monitor results

Luminance parameter	Monitor 1	Monitor 2
Maximum luminance (cd/m ²)	421	433
Minimum luminance (cd/m ²)	0.88	0.88
Contrast ratio	488: 1	492: 1

The luminance of monitor was measured using a spectrometer CS-2000 (Konica Minolta, Japan) (**Figure 4**). This telescopic-type luminance meter was placed 50cm from a display screen. Measurements were made with a viewing angle of 0.2° in the absence and presence (10-18 lux) of ambient lighting. This enabled calculation of the contrast ratio and luminance ratio.



Figure 4 Konica Minolta Spectroradiometer CS-2000 (Japan)

The spectrometer was positioned towards the display faceplate at a distance of 50 cm. For each of the reporting monitors, all measurements were taken three times. Mean values for all parameters were calculated. The monitors demonstrated the output shown in **Table 3**.

- i. Maximum luminance (L_{\max}): This refers to the maximum amount of light emitted from the monitor. Luminance was measured at the center of the highest luminance patch of the TG18- LN test patterns. The SI unit is candela per meter square (cd/m^2) or foot lambert (FL).

- ii. Minimum luminance (L_{\min}): This refers to the minimum amount of light emitted from the display. It was measured at the center of TG18-LN01. The SI unit is candela per meter square (cd/m²) or foot lambert (FL).
- iii. Contrast ratio: This is the ratio of maximum to minimum luminance in the absence of ambient lighting (L_{\max}/L_{\min}).
- iv. Just Noticeable Difference (JND): JND refers to change in luminance outputted by the system that can be perceived by the human visual system as change in shades of grey using an inbuilt look-up-table (2).
- v. Maximum luminance difference between two primary monitors: the percentages differences between the maximum luminance of the two primary reporting monitors at each workstation.
- vi. Luminance non-uniformity: The maximum deviation of luminance across a monitor displaying a uniform pattern (2). This was measured using a TG18-UNL10 test pattern. Measurement of luminance was performed at the four corners and center of the test pattern, and the percentage deviation was calculated as shown below:

$$\text{Percentage difference, } L_{dev} = \frac{L_{\max}(2) - L_{\max}(1)}{\text{The higher } L_{\max}(2 \text{ or } 1)} \times 100\%$$

Table 3 Output results of reporting monitors

Luminance parameter	AAPM Guidelines for primary displays	Monitor 1	Monitor 2
Maximum luminance, L_{\max} (cd/m ²)	≥ 170 ≥ 450	413	416
Minimum luminance, L_{\min} (cd/m ²)	≤ 1.5	0.92	0.93
Contrast ratio (CR)	≥ 250 ≥ 300	449: 1	447:1
Just Noticeable Difference (JND)	NA	609	609
Percentage difference	$\leq 5\%$ $\leq 10\%$	0.7%	0.7%
Luminance non-uniformity (Percentage uniformity)	$\leq 30\%$	8.5%	8.2%

Ambient lighting measurement

The ambient lighting of workstation was measured using a chroma meter (Konica Minolta, Japan) at a distance of 70 cm from the display screen. The measurement was performed under two different monitor conditions; either turned off or on. Only ambient light reflections were measured when the monitor displays were turned off, whilst another measurement was taken when the monitor was displaying the mammographic images. Ambient lighting levels were kept between 10-20 lux throughout the study.

Customised recording software

A customised software program was developed for this study. The software allows the reader to identify and mark the location of any cancers found in the test set without writing responses on paper, or dictating a decision, as is generally done in clinical practice. The software also provides direct feedback to the reader about the number of cases recalled in each round of the study and allows them to alter the number according to any prescribed recall rate.

This customised recording software was developed in the Java programming environment (Java version 1.7) by a computer programmer. All images used in this software were extracted from the test set stored on a university SECTRA workstation and were resized to the desirable monitor size (53cm x 30cm). Each mammographic case consisted of 2 views of both breasts (craniocaudal (CC) and mediolateral oblique (MLO)). These images were then stitched together to form a combined image as shown in **Figure 5**. This software was designed to run on any computer or system with 256 MB RAM and Java virtual machine support capabilities. This dedicated customised recording software only supported standard image formats such as JPEG, PNG, GIF, JPG, TIF but not DICOM. The database files were continuously stored on a H2 database from where the software was started (the current working directory). As well as facilitating the lesion location selected by the reader, the software also captured the data of lesion coordinates allowing statistical analysis using a Jackknife free response operating characteristic (JAFROC) method (3).



Figure 5 Digital mammography display for user interface in customised recording software

The response to each case was registered via a mouse event. When the reader identified a suspicious lesion on the image, he/she must first decide whether the case is warranted for recall or not, as defined by BreastScreen Australia’s classification system (**Table 6**). Once the recall decision was made, the reader was required to mark the location and give a confidence score using the “mark lesion” button. A square lesion box with drop-down list of confidence score ranging from 1 to 5 popped up adjacent to the images (**Figure 6**). The lesion box was able to be moved and resized to the respective lesion location and size. If the case did not warrant recall for further assessment and was deemed normal, the reader could move to the next case by clicking the “next” button and a score of 1 was automatically assigned to the case. A score of 2 was given if a benign lesion was thought to be present. Readers could also review and change their decisions before the reading session ended.

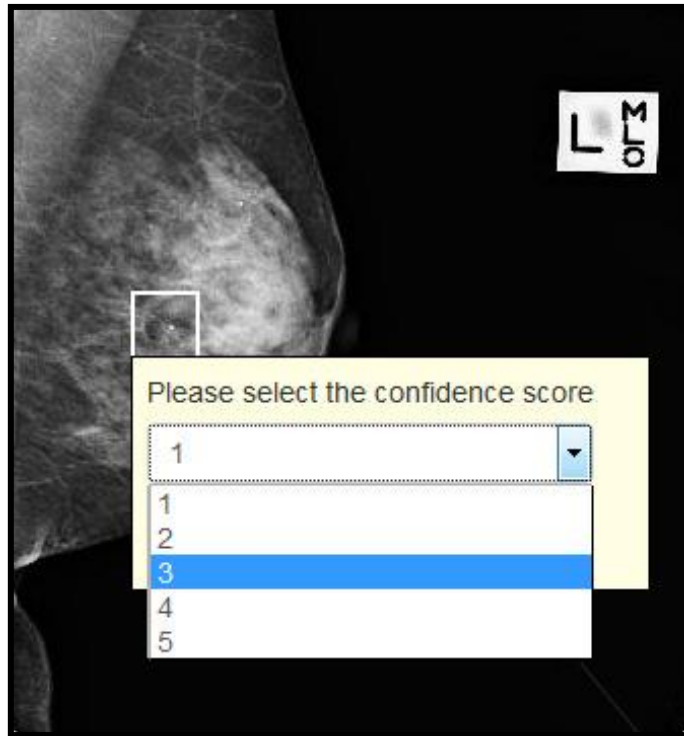


Figure 6 Snapshot of the lesion box and confidence score displays on left medio-lateral oblique (MLO) view.

The customised software was not integrated with the Sectra workstation but rather the reader recorded their interpretation and located any lesions on a laptop that displayed identical images in the same order as those displayed on the reporting monitors.

Application set-up

The following steps guide how to run the software and start the application server:

1. Find file known as *launch.bat* in SOFTWARE folder.
2. Double click the *launch.bat* file.
3. Wait until a black window known as shell (**Figure 7**) is loaded. This shell must be run before any reading session takes place.

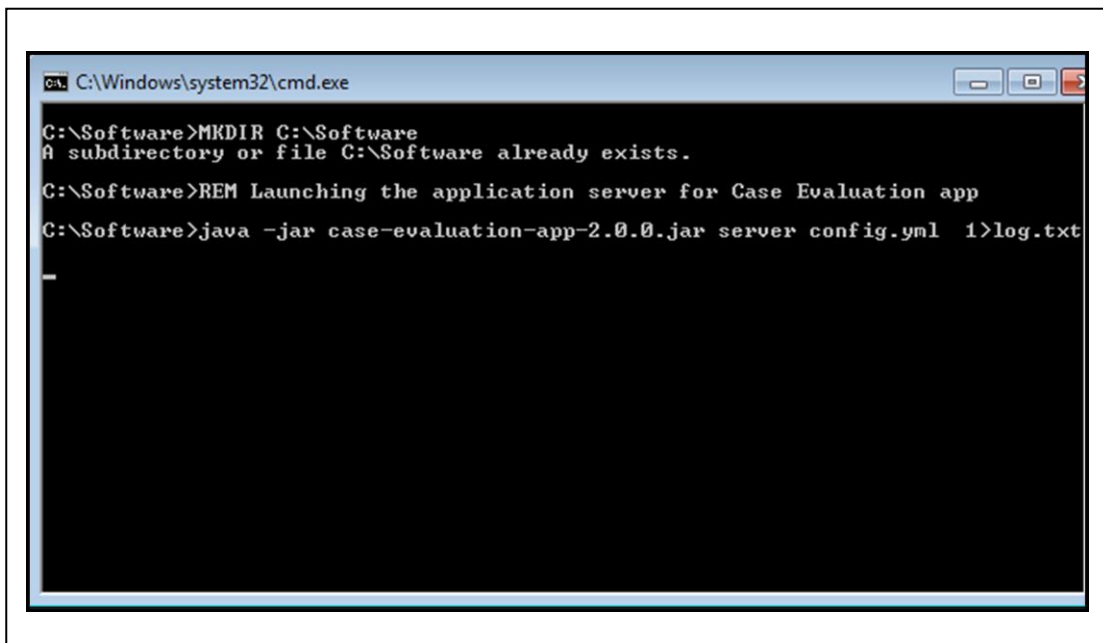


Figure 7 Snapshot of a black window (shell)

4. Open the internet browser.
5. Type the following address or Uniform Resource Locator (URL) into the address bar: <http://localhost:8080/assess/<ReaderID>ReadingSession> (e.g <http://localhost:8080/assess/1/1>)

The above URL was linked to the test set prepared for reader number 1 for his/her first reading session. A different reader was assigned with specific reader number for each reading session.

Test set development

One identical test set of images was used throughout this study and comprised of two hundred mammographic cases which were obtained from the BreastScreen NSW digital imaging library. Permission to use the images was granted from BreastScreen NSW. All patients' identification details were removed from the images. The selection of the mammographic cases used in test set was gathered by a radiographer with more than three years of experience in mammography and validated for inclusion by the NSW State breast radiologists. The number of abnormal cases used in the set was designed to simulate Australian recall rate for the first screen (10%). All mammographic cases used in the test set were generated between January 2006 and October 2011, and previously read and reported by two or three radiologists during routine screening.

The test set contained 20 abnormal and 180 normal cases, with the abnormal cases having a single biopsy-proved malignancy. All normal cases in the test set were determined by a report finding of normal/return to screen produced at the 2 years follow up assessment. That is, both mammograms were taken 2 years apart and both reported as normal by two independent blinded breast readers. Each mammographic examination consisted of two-view digital mammograms (cranio-caudal view (CC) and medio-lateral oblique (MLO) of both breasts) with a range of lesion difficulties, from subtle to obvious cancer presentations which were determined by the percentage of readers detecting individual lesions. The 20 abnormal cases had various types of abnormal mammographic appearances (stellate mass, non-specific density (NSD), architectural distortion (AD) and

mixed appearance of calcification and AD, and stellate and NSD), with 16 cases the cancer was visible on both mammographic views (CC and MLO) of a given breast, while four cases had a visible lesion on either one of the mammographic views, which resulted in a total of 36 malignant lesions available for localization. The distribution of lesion characteristics was chosen by clinical audit and it represents the distribution of lesions that are commonly missed in clinical practice in Australia consecutive audit within BreastScreen and are therefore representative of the normal, benign and malignant cancer cases within the screened community. The variation in lesion types was representative of the natural variation and there was no intent to alter the prevalence of a specific cancer type to have it included in the study.

All images passed the BreastScreen Australia (BSA) quality assessment, and 5% of the 200 cases were digitized film-screen and 95% full field digital mammography (FFDM). As the test set was designed to represent actual clinical prevalence, only images that fulfilled image criteria (positive and negative cases) were consecutively chosen to be included in the test set until the predetermined number of abnormal and normal was reached. No other specific criteria were used in the selection of cases. For the purpose of simulating the first screening read, no prior images were provided to the readers.

Justification for 200 mammographic cases

The 20 breast cancer cases in the test set were randomly mixed with 180 normal cases in a 1:10 ratio. These 200 mammographic cases formed an appropriate sample size to create the test set for this study determined and confirmed by sample size tables for receiver

operating characteristics studies modelled by Obuchowski, NA (4). To evaluate the adequacy of sample size used for this study, the following measurements were considered.

- i. The average of ROC AUC of 0.80 (high accuracy)
- ii. Large observer variability (interobserver =0.05, interobserver=0.01)
- iii. Large relative frequency of normal and abnormal cases in the study, R (value of R=4/1)

For the sensitivity at a false-positive rate less than or equal to 0.10, the estimated number of cases needed for four observers the test set was 50. However, for the purpose of resembling the clinical prevalence of breast cancer, the test set was enriched with an elevated number of mammographic cases. The case number of 200 was also chosen to allow a test set that did not overload the readers to finish the task in one sitting.

Truth (Gold standard)

The specific location of the cancer site and malignancies in the 20 abnormal cases were then identified by one expert breast radiologists who is involved in training assessment, quality, clinical policies of BreastScreen NSW and also is responsible for the clinical management of a screening centre. To assist the expert to determine the location and the particular mammographic appearance of the depicted abnormalities, this expert had access to the biopsy reports and the prior images if available. These locations were defined as the truth and were used in analysing the cases marked by the readers.

Mammographic appearances of cancers in the test set: detailed classification

The cancer cases and lesions examined in this work presented with varying mammographic appearances. All malignancies present in the 20 cancer cases were characterized by the expert breast radiologists according to the lesion type, breast density and location of lesion with respect to the fibroglandular tissue and these details are provided in **Table 5**. These characteristics were based on Australian Synoptic Breast Imaging Report of the National Breast Cancer Centre (NBCC) (5) as described in **Table 4**, to assist with understanding the terms of lesion abnormalities used in this thesis.

Table 4 Definition of lesion abnormalities according to Australian Synoptic Breast Imaging Report

Lesion abnormalities	Definition
Calcification	Deposition or collections of calcium compounds in breast tissue of sufficient size to be seen on mammogram and malignancy are characterised by size (0.05-0.5mm), calcification distribution (cluster, multiple cluster, or sometimes scattered), pleomorphism and variation of density.
Stellate lesion	Spiculations of variable length radiating from a central point or mass. If a central mass is present, it may be small or large, and of low, mixed or high density compared to surrounding breast parenchyma.
Architectural distortion (AD)	Abnormal configuration of the ductal and ligamentous structures of breast parenchyma compared with the remainder of the breast tissue markings and often appears with spiculation, focal retraction, distortion of the parenchymal edge, and disorganisation of markings.
Non-specific density (NSD)	Asymmetry of breast tissue seen in one of the breast, on either one or two mammographic views with poorly defined characteristics of breast density.

There were four different mammographic appearance types of cancer presented in this test set; 8 cases had stellate lesion; 6, non-specific density (NSD); 2, architectural distortion (AD); 3, mixed appearance of calcification and AD; and 1, mixed appearance of stellate and NSD. For breast density, 2 cases had entirely fat (<25% glandular tissue); 7, scattered fibroglandular densities (25-50% glandular tissue); 8, heterogeneously dense (51-75% glandular tissue); and 3, extremely dense appearances (>75% glandular tissue). **Figure 8** shows the example of mammographic cases with four specific breast density. In the test set in this study, 55% of the cancer cases had heterogeneously and extremely dense breast, 10% had entirely fat and 30% had scattered fibroglandular densities. This distribution of breast cancer in this study is similar to the breast density population that commonly found in women aged 40 to 75 worldwide (6-8).

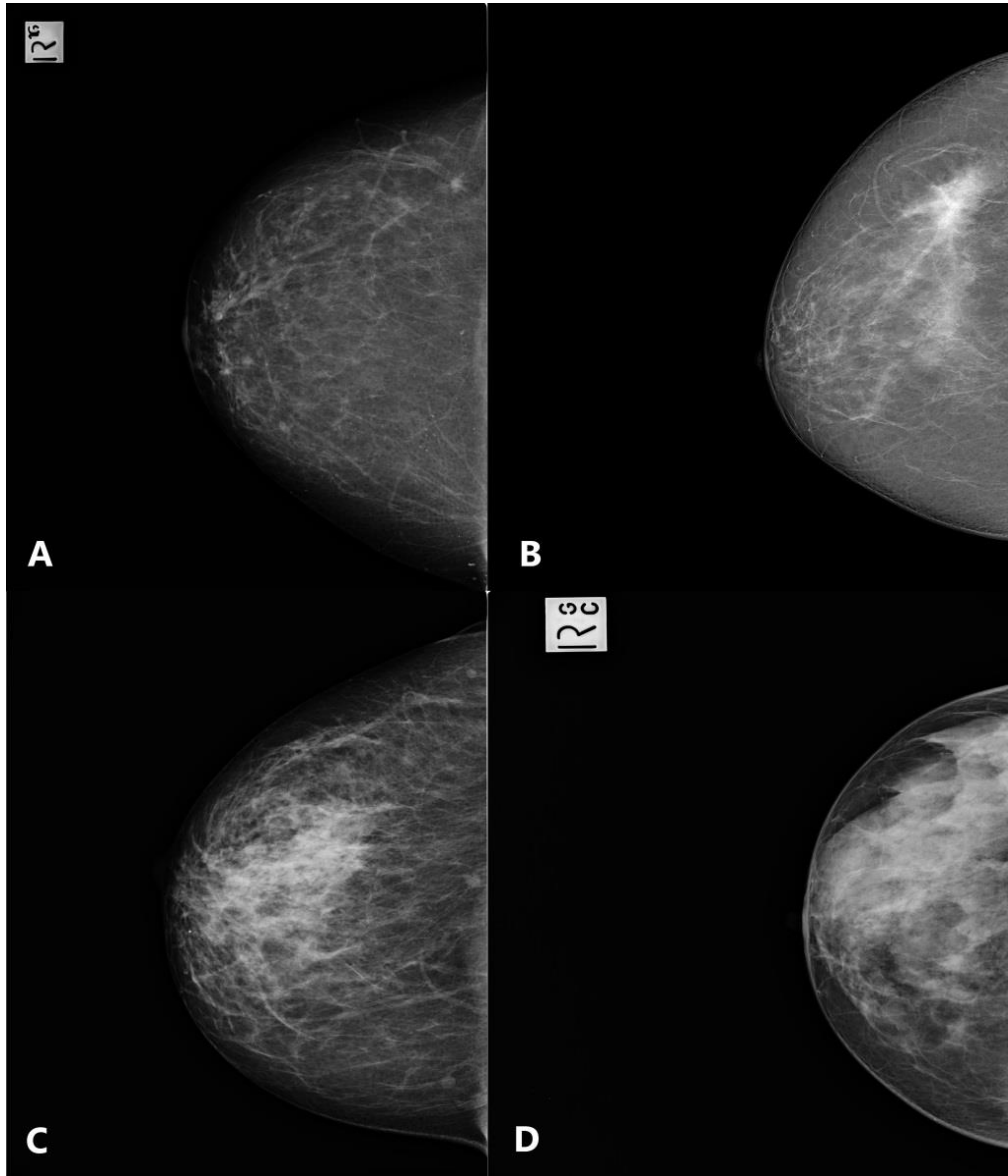


Figure 8 Example of mammographic images with different breast density A) < 25% glandular tissue, B) 25-50% glandular tissue, C) 51-75% glandular tissue, D) >75% glandular tissue

In term of lesion location, lesions in 11 cases were overlapped the fibroglandular tissue; lesions in 6 cases being outside the fibroglandular tissue and lesions in 3 cases at the edge of the fibroglandular tissue (**Figure 9**).

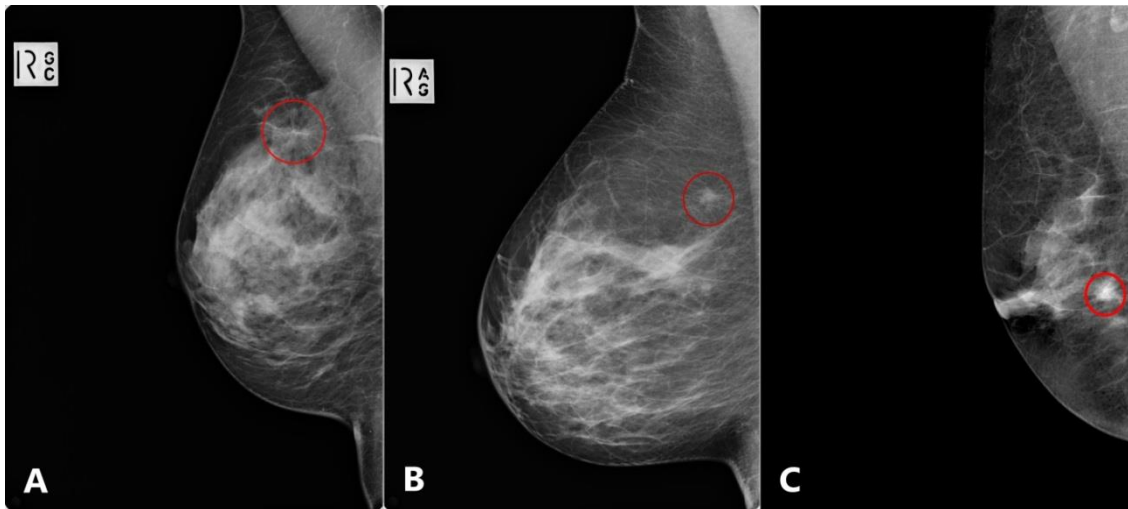


Figure 9 Mammographic images above showed lesion (circled) location to fibroglandular tissue A) overlapping fibroglandular tissue, B) outside fibroglandular tissue and C) at the edge of fibroglandular tissue

Table 5 Mammographic appearances of 20 cancer cases present in the test set

Case ID	Type of lesion	Breast density	Lesion location
MJBJ	AD	25-50% glandular tissue	Overlay fibroglandular tissue
MJAS	Calcification + AD	>75% glandular tissue	Overlay fibroglandular tissue
MJBG	Calcification + AD	51-75% glandular tissue	Overlay fibroglandular tissue
MJBK	NSD	51-75% glandular tissue	Overlay fibroglandular tissue
MJBL	Stellate	>75% glandular tissue	Overlay fibroglandular tissue
MJCF	NSD	>75% glandular tissue	Overlay fibroglandular tissue
MJCQ	NSD	51-75% glandular tissue	The edge of fibroglandular tissue
MJCR	Stellate	25-50% glandular tissue	Outside fibroglandular dense tissue
MJCX	Stellate	< 25% glandular tissue	Outside fibroglandular dense tissue
MJDA	Stellate	51-75% glandular tissue	Outside fibroglandular dense tissue
MJDH	Stellate	51-75% glandular tissue	Outside fibroglandular dense tissue
MJDU	Stellate	51-75% glandular tissue	Overlay fibroglandular tissue
MJEA	Stellate	25-50% glandular tissue	Overlay fibroglandular tissue
MJEB	NSD	25-50% glandular tissue	Outside fibroglandular dense tissue
MJEG	Calcification + AD	25-50% glandular tissue	Overlay fibroglandular tissue
MJGR	Stellate	25-50% glandular tissue	Overlay fibroglandular tissue
MJHD	AD	< 25% glandular tissue	Outside fibroglandular dense tissue
MJHH	NSD	51-75% glandular tissue	Overlay fibroglandular tissue
MJHJ	NSD	25-50% glandular tissue	The edge of fibroglandular tissue
MJHK	Stellate	51-75% glandular tissue	The edge of fibroglandular tissue

*AD, architectural distortion

†NSD, non-specific density

Mammograms randomization

In order to eliminate any possible reading order biases that may arise in this experiment, the order of cases in the test set was randomized in a different order for each reader and for each reading session. The Sectra and customised recoding lists were synchronised post randomisation.

Prior to the randomization, a free online random number generator, RANDOM.ORG that is available at <https://www.random.org> was used to randomise the sequence of cases in the test set and a pre-randomized list was prepared for this study process. This list was used as a reference in randomizing mammographic cases in the SECTRA workstation and in the customised recording software.

Randomizing cases in Sectra system

The 200 mammographic cases in the Sectra system were then randomized with reference to the pre-randomised list prepared earlier. A test set with the new randomized sequence was de-identified from the patients' information and provided with a new unique ID. This new unique ID was created in the Sectra by adding a number in front of the existing case ID, for example, if the existing case ID: MJAA, new case ID is 003 MJAA.

Randomizing cases in customised recording software

The customised recording software designed for this study was independent software which was not integrated with the Sectra system. A separate case worklist was created to ensure the displayed cases on the customised recording software matched with the cases displayed on the reporting monitors of the Sectra system. For each reader, the sequence in the customised recording software was the same as the one created in the Sectra system.

Image presentation (display protocol)

Once these 200 mammographic images were randomly sorted, images were displayed in the following order as it would be viewed in BreastScreen clinic (**Figure 10**) and readers could alter the image order backwards and forwards.

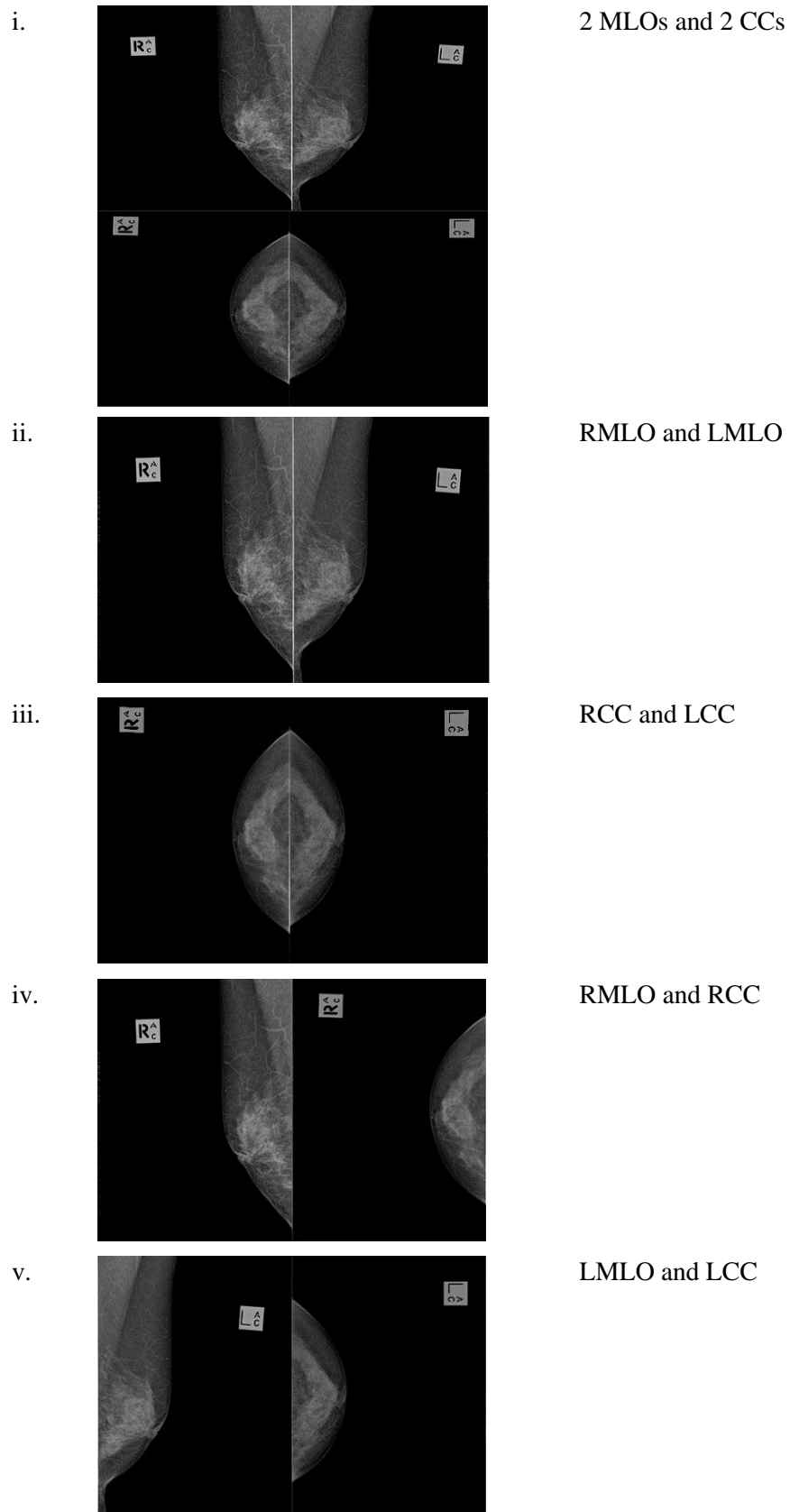


Figure 10 Image display sequence for reading process

Reading task

All participants read the same 200 cases at three reading sessions. At first reading session, the readers were allowed to recall as many cases as they believed necessary. For the second and third reading sessions, readers were given a prescribed recall rate at the beginning of the session and required to interpret each mammographic case. The readers were briefed how to review and rate the test set, followed by a demonstration on how to use the customized recording and Sectra software. No information was provided on the number of cases with abnormal findings or the number of lesions on the images.

During the reading sessions, the participants were required to identify each mammographic case that they considered recalling in keeping with their free or specified target recall rate. The readers were asked to mark the location of the suspicious lesions within the recalled cases using customised recording software and provide a confidence rating ranging from 1 to 5 for the location of detectable lesions of each recalled cases (1: definitely normal, 2: probably benign, 3: uncertain, 4: probably malignant, 5: definitely malignant). No time limit was set for the reporting session.

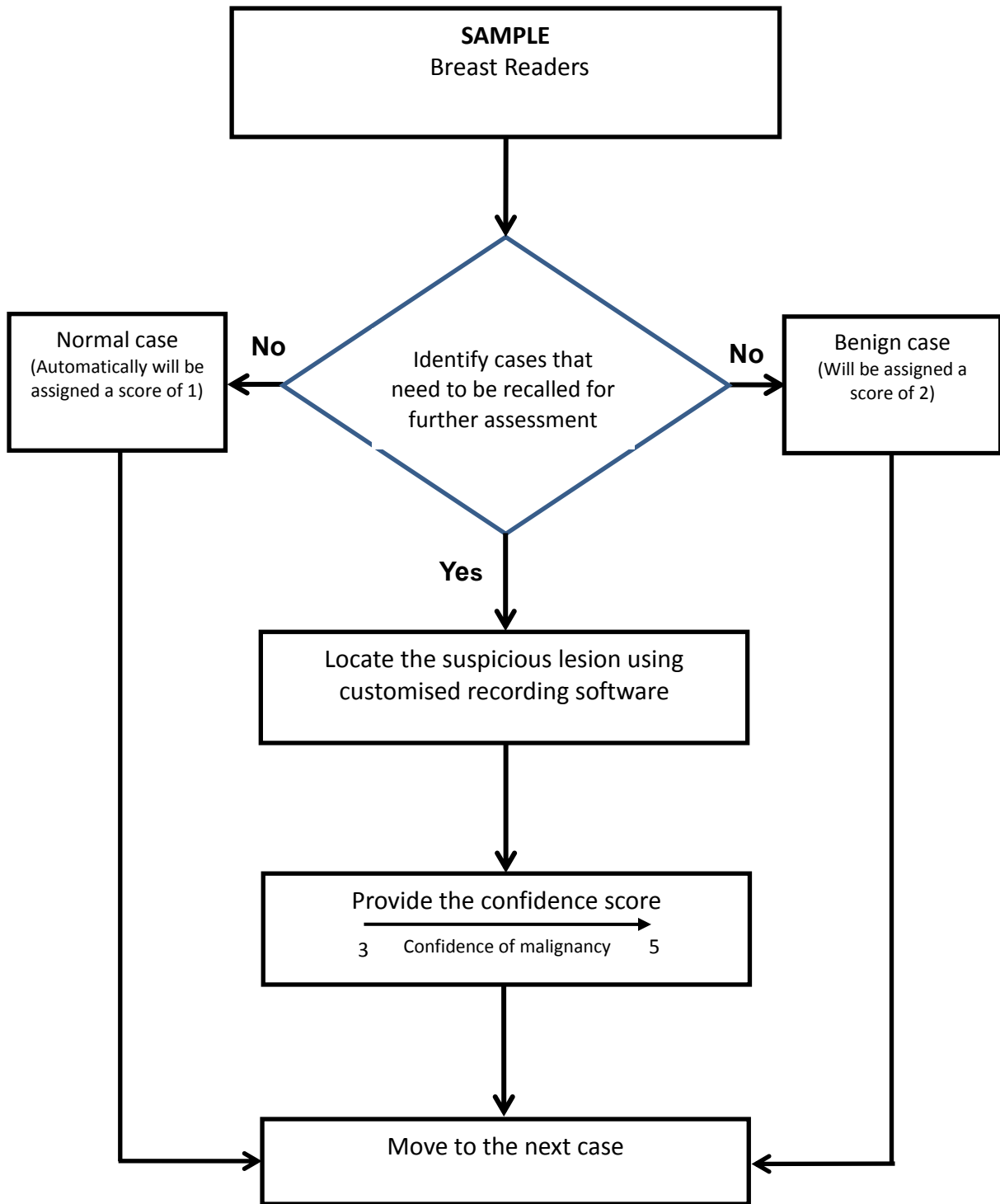


Figure 11 Reading process flow for readers when performing the reading task

Reading conditions

This longitudinal study was divided into three reading sessions. Each session had a different recall rate condition and was separated by a minimum of two months to reduce any memory effects and learning bias. The total study time was six months for the three reads for each reader.

At the first reading session, no numerical percentage recall rate was imposed and termed “free recall” when interpreting the cases. At this condition, the reader could recall as many cases as necessary. For the subsequent readings (second and third session), the number of mammographic cases that readers could recall was then constrained to a certain number/percentage. Readers were restricted to recall a maximum of 30 cases (15%) in their second reading and 20 cases (10%) in the third reading. The determination of the percentage of the recalled cases for second reading session of 15% was based on international results of recall rates, in which the highest recall rate was 15.1%, as applied in the USA (9). For the third session, the percentage of 10% was determined to align with the first screening recall of BreastScreen Australia and proportion of abnormal cases in the test set was made to comply with the Australian standard (10).

Breast Screen Australian classification system

This confidence scoring system used in this work was aligned to BreastScreen Australia (4) practice for classifying mammographic lesions. **Table 6** shows the details of the numerical score for respective recall decisions made by each reader during the reading

session. A score of 1 and 2 would be considered to a normal and benign decision respectively, and a score of 3, 4 or 5 would be equated to a recommendation of a suspicious case which is recalled for further assessment. The readers were not required to classify the types of lesions for each recalled case.

Table 6 Recall recommendation based on Australian Classification System

Category	Description based on the Australian Classification System	Recall Recommendation
1	No significant imaging abnormality	Normal : No recall to assessment
2	Benign findings and no further imaging is required	
Threshold of recall		
3	Intermediate/equivocal findings and further investigation is required	Suspicious malignant - recall to assessment
4	Suspicious findings of malignancy and further investigation is required	
5	Malignant findings and further investigation is required	

Data Analysis

Data from this experimental work were analysed in two phases as shown in **Figure 12**; the first phase of analysis began by evaluating the overall performance of the readers when reading the screening mammograms at three varying levels of recall rates.

The initial part of this work in assessing reader performance was investigating three performance metrics; sensitivity, specificity and receiver operating characteristic (ROC) area under the curve (AUC). Further analysis was then performed in later work to explore the effect of reduced recall rates on readers' performance through a methodology that incorporates the lesion location information. Through this methodology, the readers' ability to correctly locate lesions and give a confidence rating based on these decisions was assessed. With the addition of the number of participating readers, the performance of readers to correctly locate lesion was analysed through case location sensitivity, lesion location sensitivity, and Jack-knife free-response ROC (JAFROC) figure of merit (FOM).

The second phase of the analysis focused on exploring mammographic lesion appearances. The analysis was performed by identifying the cancer appearances that were more likely to be missed when readers performed at lower recall rates.

Chapter 5 in this thesis explains and discusses thoroughly the first part of this analysis work, whilst Chapter 6 reports the results of subsequent analysis.

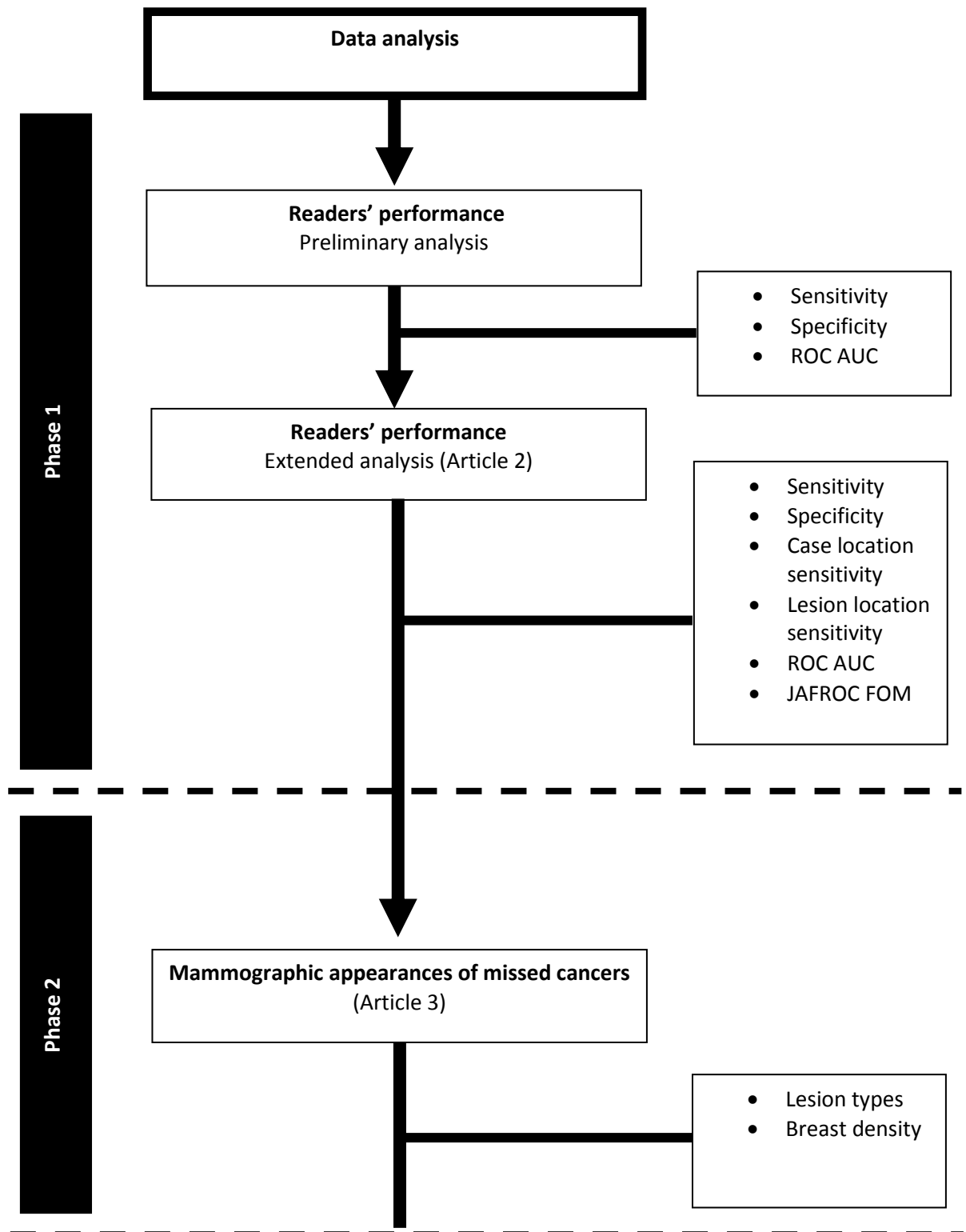


Figure 12 Schematic overview of the steps involved in the analysis process

Recall rate to assessment

The rate of recall to assessment at the first reading session (free recall condition) for each reader was defined as the proportion of recalled cases that reader considered needed further assessment. It was calculated by the follow equation:

$$\text{Recall rate} : \frac{\text{Total number of recalled cases}}{\text{Total mammographic cases in the test}} \times 100\%$$

Performance metrics

Overall, for all reading sessions, reader performance was assessed using sensitivity, specificity, case location sensitivity, lesion location sensitivity, receiver operating characteristic (ROC) area under the curve (AUC) and Jack-knife free-response ROC (JAFROC) figure of merit (FOM). Details of each performance metric used for this study are further explained in Chapter 5.

ROC analysis

Receiver operating characteristic (ROC) analysis is a binary paradigm focused on a single rating per case that is commonly used in assessing the accuracy of a radiologist's interpretation of mammography images. The evaluation of a reader's accuracy is based on the reciprocal relationship of sensitivity (*true positive fraction* (TPF)) and specificity (*false positive fraction* (FPF)) at various test positive thresholds which are graphically presented

as an ROC curve. The area under the ROC curve (AUC) was used as a figure-of-merit to represent the overall performance index between 0.0 to 1.0, with 0.0 indicating the worst test and 1.0 being the perfect accuracy test. An ROC AUC of 0.5 corresponds to a random guess, giving a completely uninformative test.

In this study, the reader was given a task to interpret a set of mammography images/cases where the radiologist needed to rate their confidence in the presence or absence of cancers for each case where the reader was allowed to give only one rating for each case (positive or negative). Analysis using this binary approach disregarded the number of cancer lesions present or their precise location.

JAFROC analysis

JAFROC analysis is an advanced quantitative analysis of reader performance when interpreting mammography images. This analysis incorporates a free-response paradigm that allows lesion location information to be included when analysing reader performance. In this study, a TP score was given to a lesion when a reader successfully marked and localized the lesion correctly within the acceptance radius. Furthermore, with the additional information of lesion location, this method demonstrates higher statistical power as compared to the ROC analysis (11). Through this method of analysis, the readers were allowed to locate multiple suspicious lesion locations during the interpretation process. The non-parametric area under the alternative free-response receiver operating characteristic (AFROC) curve was used as the figure of merit for JAFROC and the graph was plotted as

the lesion localisation fraction (LLF) versus false positive fraction (FPF). An acceptance radius surrounding each lesion was used as a proximity criterion in determining whether the lesion was correctly marked or not.

For this study, a 60-pixel acceptance radius surrounding each lesion (within 20 millimeters) was defined to classify each mark. This acceptance radius was chosen as it encompassed the largest malignant lesion present in the test set. The marked lesions were identified either as positive or negative by comparing selections with the truth table compiled by the expert breast radiologist. A true positive (TP) was scored if a lesion was marked within the acceptance radius and received a confidence score between 3 and 5. A false positive (FP) was defined for any incorrect localization on normal cases, or if the localisation was outside the 60 pixels range of a lesion in abnormal cases. A lesion that was correctly localized but received a confidence score of 2 was considered as a false positive lesion. A true negative (TN) outcome was recorded if the case was correctly identified as normal or lesion-free. A false negative was scored when cancer lesions were not marked.

Data extraction

All data stored in the customised recording software were exported in Microsoft Excel (*xls*) file format by accessing the following URLs. This browser will return a file, an excel file, that can be saved on the computer for further analysis. The step-to-step how to extract the data from the data from the software is described as below:

- i. **To access all case readings:** insert the following address or URL into the address bar: <http://localhost:8080/export/cases>. A specific file named as *Assessments.xls* will be appeared and can be saved on the computer.
- ii. **To access the lesion coordinate details:** insert the following address or URL into the address bar: <http://localhost:8080/export/tags>. A specific file named as *Tags.xls* will be appeared and can be saved on the computer.
- iii. **To access details of user activity** (as details of readings that were modified such as changed or removed later: insert the following address or URL into the address bar: <http://localhost:8080/export/audits>. A specific file named as *Audits.xls* will be appeared and can be saved on the computer.
- iv. **To export the rating score of specific case and reader:** insert the following address or URL into the address bar: <http://localhost:8080/assessment/{userId}/{sessionId}/{caseId}>. For example: <http://localhost:8080/assessment/1/1/MJAB>; this URL will link to specific case identified as MJAB for reader number 1 for his/her first reading session.

- v. **To export the lesion marking of a specific case:** insert the following URL into the address bar/ browser: <http://localhost:8080/tag/{userId}/{caseId}>. For example: <http://localhost:8080/tag/1/1/MJAB>.

Lesion localization

The extracted data from the customised recording software were compiled in an Excel datasheet. The figures in the dataset provide the location of four corners of the lesion box in coordinates of X and Y. The coordinates at the centre of the lesion were calculated from the available information extracted from the customised recording software. From figures given in the dataset, X_1 and Y_1 were the coordinates of left most top corner of the lesion box (A), while X_2 and Y_2 were the coordinates of the right most bottom of the lesion box (C).



The coordinates (X_3 and Y_3) of the centre of the lesion were therefore calculated using the formula below:

$$\text{Centre of the lesion } (X_3, Y_3) = \{(X_2 - X_1) / 2 + X_1\}, \{(Y_2 - Y_1) / 2 + Y_1\}$$

For example; if $(X_1, Y_1) = (518, 250)$, $(X_2, Y_2) = (548, 280)$;

$$\begin{aligned} \text{The coordinate of } (X_3, Y_3) &= \{(548-518)/2 + 518\}, \{(280-250)/2 + 250\} \\ &= \underline{(533, 265)} \end{aligned}$$

The lesion location of 20 abnormal cases from the truth table were calculated and only known by the researchers (**Table 7**).

Table 7 The truth location of lesion coordinates of abnormal cases

No.	Case ID	Lesion coordinates (X, Y)
1.	MJAS	(215, 253) (542, 258)
2.	MJBG	(122, 370)
3.	MJBJ	(123, 244) (451, 297)
4.	MJBK	(142, 418) (492, 354)
5.	MJBL	(87, 257)
6.	MJCF	(486, 380)
7.	MJCQ	(208, 316) (539, 303)
8.	MJCR	(140, 285) (421, 315)
9.	MJCX	(86, 229) (479, 311)
10.	MJDA	(118, 289) (549, 323)
11.	MJDH	(119, 228) (519, 290)
12.	MJDU	(159, 428) (478, 403)
13.	MJEA	(216, 175) (508, 208)
14.	MJEB	(242, 275) (654, 260)
15.	MJEG	(188, 261) (488, 301)
16.	MJGR	(183, 252) (629, 242)
17.	MJHD	(223, 446) (586, 459)
18.	MJHH	(225, 269)
19.	MJHJ	(200, 542) (513, 483)
20.	MJHK	(238, 375) (551, 411)

Comparison recording software output to truth (gold standard)

The locations of marked lesions were compared with the truth table. The decision made during the reading task by each reader was categorised into TP, FP, TN or FN based on the detection criteria previously defined. The data were then analysed using JAFROC Version 4.2 software (3, 11) developed by Dev P. Chakraborty and available at www.devchakraborty.com.

Statistical analysis

Reader Performance

To compare readers' performance across three recall conditions, individual performance results were pooled across the readers and a non-parametric statistical analysis was performed using SPSS Software version 22.0. Data in each reading condition was treated independently and the statistical analysis was performed in two steps;

- i. Kruskal-Wallis test: this test was performed across the three reading sessions with statistical significance set at $P < 0.05$ at 95% confidence intervals (CIs). If the results from the Kruskal-Wallis test shown statistically significance, post-hoc analysis was needed to identify which groups were significantly different from each other.

- ii. Mann-Whitney U test: Each recall condition was grouped into three pairs (free recall and 15%, 15% and 10%, 10% and free recall) and comparison was made using Mann-Whitney U test. Bonferroni adjustment was applied to the alpha values by dividing the alpha level by the number of comparisons made ($n=3$) to control for Type 1 error. Results therefore with the revised alpha level, a $P < 0.017$ were deemed to represent significant differences.

Mammographic appearances of missed cancers

Contrary to analysis of Study 1, analysis of study 2 focused on determining whether the detection of any specific cancer types altered when the recall rates were reduced (Condition 15% and 10%), which narrowed the analysis to the 20 abnormal cases only. The mammographic appearances of the cancer lesions used in this study were classified into three categories:

- i. lesion type;
- ii. breast density and
- iii. location of the lesions

Localization and identification of mammographic appearances of 20 abnormalities were made by the same expert breast radiologist who determined the ground truth of the test set. The mammographic appearances were chosen according to the synoptic and standardised breast reporting report practice in Australia. This synoptic breast imaging is similar to the Breast Imaging Reporting and Data System, BI-RADS criteria, 4th edition, endorsed by the Royal Australian and New Zealand College of Radiologists (RANZCR). Breast density was categorised into 4 groups:

1. The breast is almost entirely fat (<25% glandular)
2. There are scattered fibroglandular densities (25–50% glandular)
3. The breast is heterogeneously dense (51-75% glandular)
4. The breast tissue is extremely dense (>75% glandular)

The following chapter, Chapter 4, gives results of performance at the individual level, before presenting the results of the group performance via journal manuscripts. Chapter 5 has recently been submitted to The Breast Journal (TBJ), and is currently under review. Chapter 6 has been accepted for publication (in Press) by the British Journal of Radiology (BJR) (<http://www.birpublications.org/toc/bjr/0/0>).

References

1. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine (DICOM) Part 14: Grayscale Standard Display Function 2004. Available from: http://dicom.nema.org/dicom/2004/04_14pu.pdf.
2. Samei E, Badano A, Chakraborty D, Compton K, Cornelius C, Corrigan K, et al. Assessment of display performance for medical imaging systems: Executive summary of AAPM TG18 report. *Medical physics*. 2005;32(4):1205-25.
3. Chakraborty DP. Analysis of Location Specific Observer Performance Data: Validated Extensions of the Jackknife Free-Response (JAFROC) Method. *Academic Radiology*. 2006;13(10):1187-93.
4. Obuchowski NA. Sample Size Tables For Receiver Operating Characteristic Studies. *American Journal of Roentgenology*. 2000;175(3):603-8.
5. National Breast Cancer Centre. Breast imaging: a guide for practice. The Royal Australian and New Zealand College of Radiologists; 2014.
6. Salem C, Atallah D, Safi J, Chahine G, Haddad A, El Kassis N, et al. Breast Density and Breast Cancer Incidence in the Lebanese Population: Results from a Retrospective Multicenter Study. *Biomed Res Int*. 2017;2017:7594953.
7. Sprague BL, Gangnon RE, Burt V, Trentham-Dietz A, Hampton JM, Wellman RD, et al. Prevalence of mammographically dense breasts in the United States. *Journal of the National Cancer Institute*. 2014;106(10).
8. Checka CM, Chun JE, Schnabel FR, Lee J, Toth H. The relationship of mammographic density and age: implications for breast cancer screening. *AJR American journal of roentgenology*. 2012;198(3):W292-5.
9. Yankaskas BC, Klabunde CN, Ancelle-Park R, Rennert G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? *Journal of medical screening*. 2004;11(4):187-93.
10. BreastScreen Australia. National Accreditation Standards: BreastScreen Australia Quality 2008 [cited 2014 May 21]. Available from: <http://www.cancerscreening.gov.au>
11. Chakraborty DP. Recent advances in observer performance methodology: jackknife free-response ROC (JAFROC). *Radiation Protection Dosimetry*. 2005;114(1-3):26-31.

CHAPTER 4

EXTENDED RESULTS

This chapter serves as bridging section for chapter 5 and 6 and provides some complimentary results related to articles 2 and 3. The results presented in this chapter assess the readers' performance at an individual level.

Individual performance within the free recall

Results of individual recall rates for the breast radiologists ranged from 18.5% to 34.0% when reading at the free recall condition as shown in **Figure 13**. Reader 3 recalled the highest number of lesions among the readers, with 34.0%, while Reader 4 had the lowest rate (18.5%). With an average of 25.6%, this recall rate yielded a considerably higher rate than recommended by BreastScreen Australian for a woman's first screen, which is currently set at 10% (1).

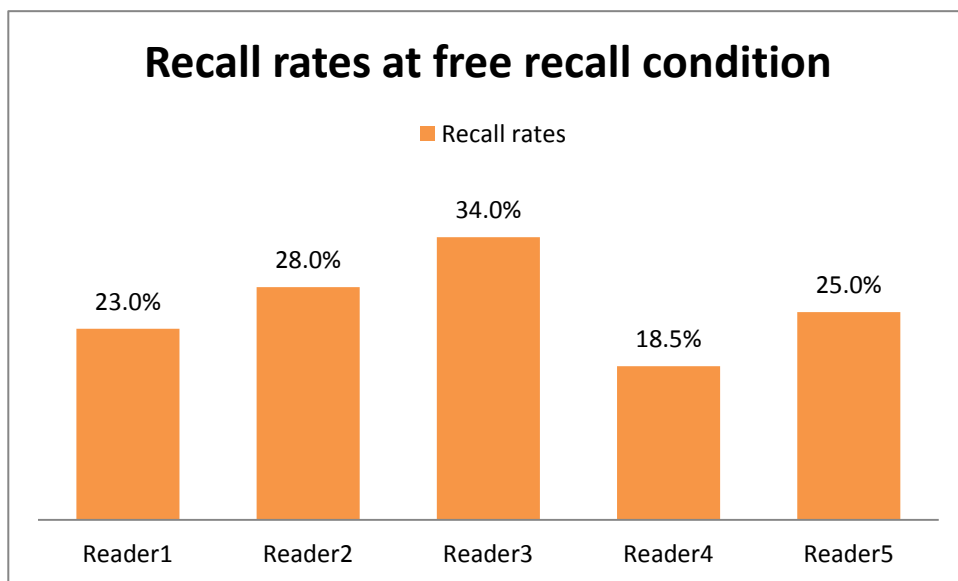


Figure 13 Histogram of individual recall rate of each reader at the free recall condition

ROC curve

The individual relationship between sensitivity and the false positive fraction (FPF) is illustrated through ROC curves. **Figure 14** illustrates the corresponding ROC curves for all three recall conditions for each reader. Variability can be seen between readers and also within reads. The overall performance of readers in this study was reported by the ROC area under the curve (AUC) as presented in Chapter 5. ROC AUC is able to summarize the whole ROC graph into a single performance index, whereby it combines the measures of sensitivity and specificity.

AFROC curve

The Jack-knife free-response ROC (JAFROC) figure of merit (FOM) of individual readers, when performing at three different recall conditions is illustrated by the non-parametric area under the alternative free-response receiver operating characteristic (AFROC) curve as shown in **Figure 15**. The relationship shown in the curve was defined as the lesion localisation fraction (LLF) versus FPF and the overall performance with regards to JAFROC FOM is reported in Chapter 5.

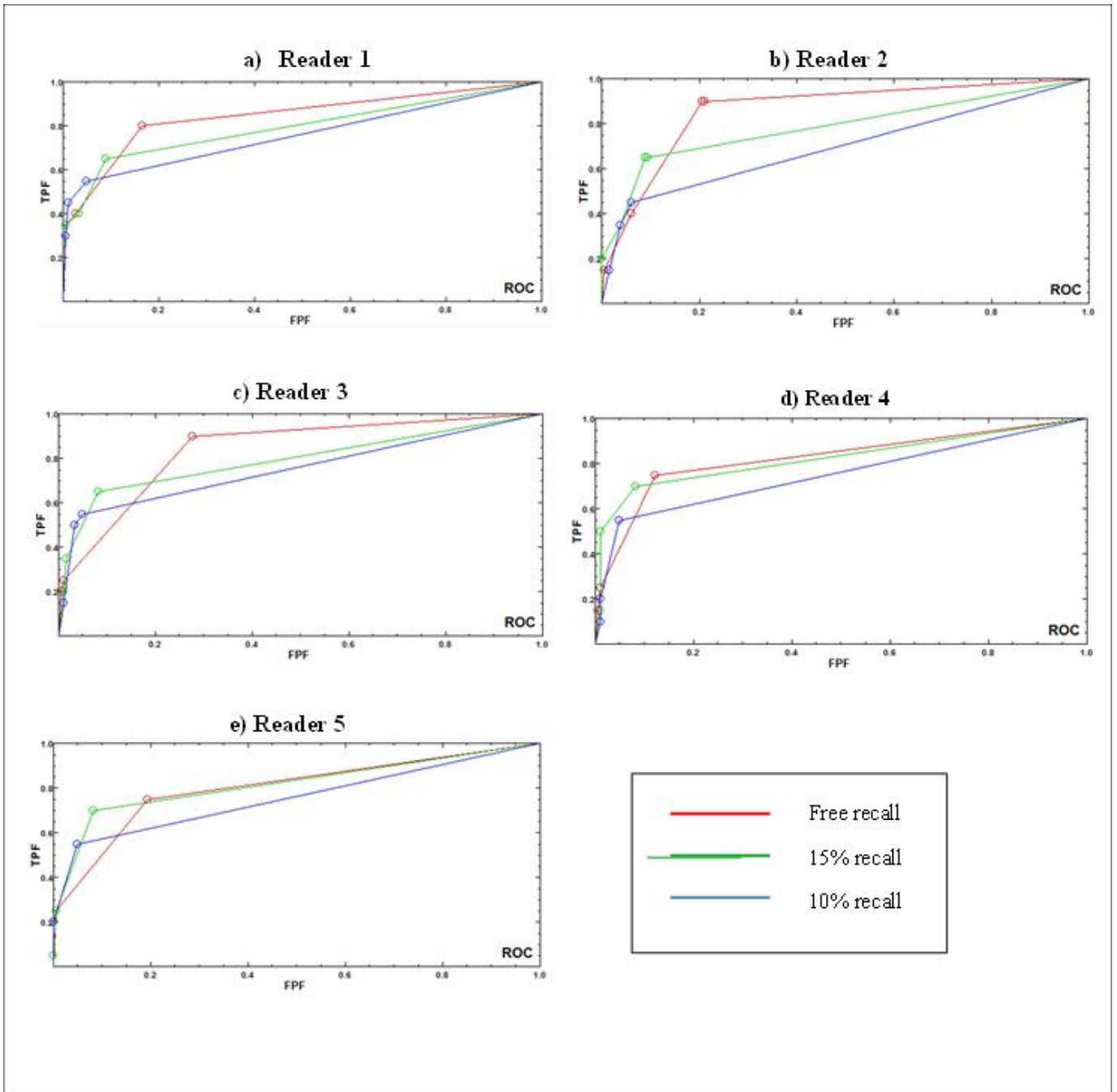


Figure 14 The empirical ROC curves for each reader when performing at three different recall conditions (i. Free recall, ii. 15% recall and iii. 10% recall).

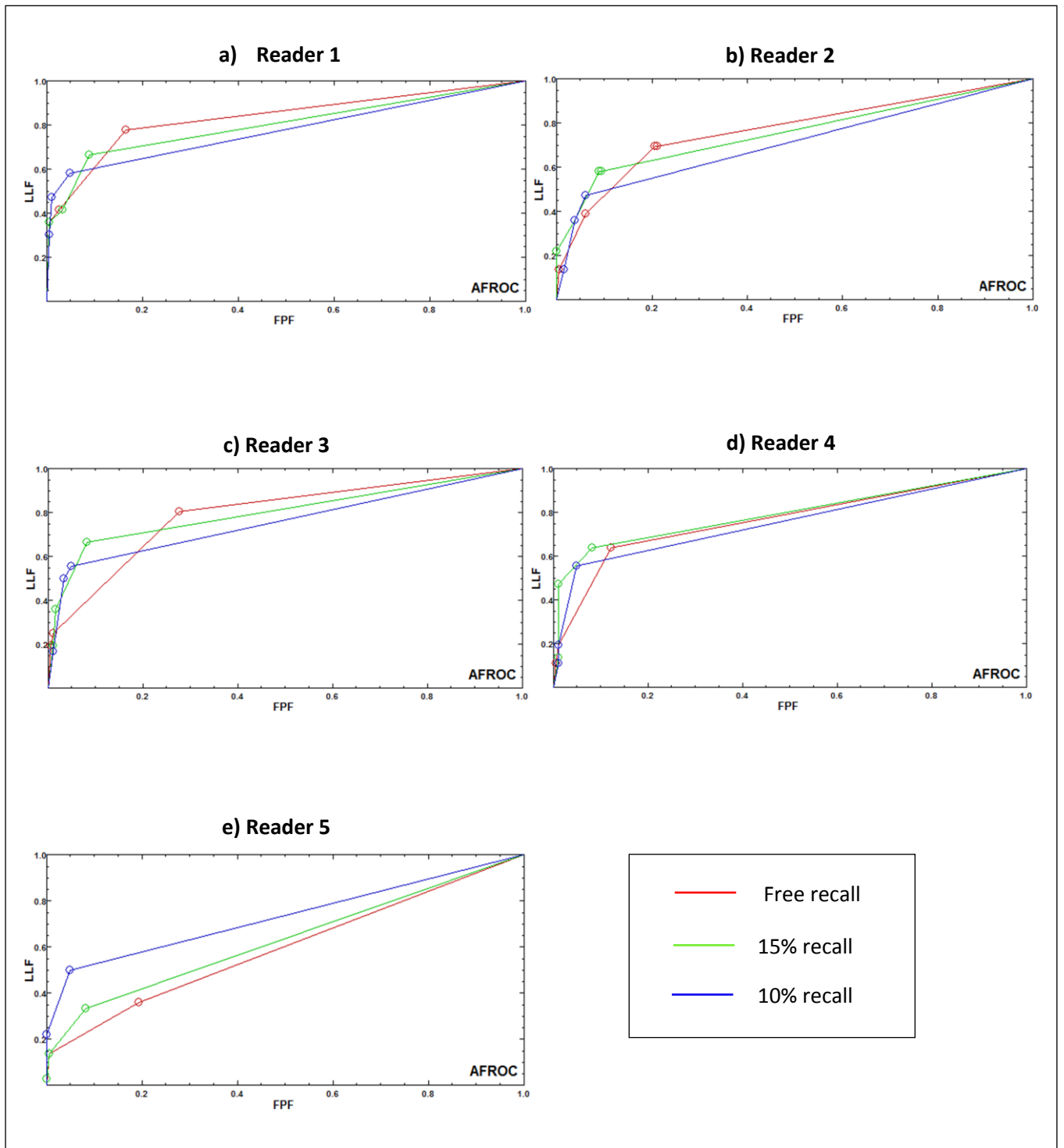


Figure 15 The empirical AFROC curves for each reader when performing at three different recall conditions (i. Free recall, ii. 15% recall and iii. 10% recall).

Lesion detectability at free recall, 15% and 10% recall rates

Table 8 shows lesion detectability for each of the cancer cases across all reading sessions (S1: free recall, S2: 15% recall and S3: 10% recall) at individual reader level. Cancer cases that were correctly identified by readers were assigned as true positive (TP) and were given a green colour. For every missed cancer, the case was marked with red colour and assigned a false negative (FN). Readers demonstrated wide differences in the detectability of lesions within and between readers. Across the 20 cancer cases, lesion detectability was consistent for 6 cancer cases (MJBL, MJCX, MJDA, MJDH, MJE A and MJGR) regardless any recall conditions imposed on the readers. The marking of one cancer was found to be very difficult by majority of the readers (n=4) at all recall conditions, and only reader 5 was able to detect the cancer lesion MJCF at free and 15% recall rates, however this reader then missed the lesion at the recall condition of 10%.

Inconsistency in lesion detectability at the individual level for certain cancer lesions was observed when reading at different recall conditions. For example, for Reader 1, the ability to recognise two cancer cases (MJAS, MJB J) and correctly recall these as malignant was inconsistent. Those two cases were identified as malignant at the free recall condition, however at the 15% recall condition, the cases were reported as normal and hence did not warrant any recall for further assessment. Furthermore, later in the experiment, these cases were identified as malignant at the 10% condition and hence recalled. This decision pattern was also observed in Reader 3 for case MJAS and Reader 5 for case MJB J. It is interesting to note that there were cancer cases that were not reported at the free call and 15% conditions but were recalled at the last reading condition (10%), as demonstrated by Reader 3 (MJBK). However, Readers 1, 4 and 5 changed their decision on cases from

being malignant (15% recall) to normal case at the third read (10% recall) when reading case MJEB (Reader 1), MJHK (Reader 4), MJBG and MJHD (Reader 5). The extended results from this work are reported in Chapter 6.

Following the classification of marked or missed lesions across all 3 conditions by the 5 readers, the individual reader's results were then grouped into three difficulty levels (lower, medium and higher difficulty) and the variability in readers' recall decision was discussed according to the mammographic appearance of the cancer cases in Chapter 6.

Table 8 Distribution of individual recall decisions for each cancer case at three recall conditions (S1: Free recall; S2: 15% recall; S3: 10% recall)

Case ID	Reader 1			Reader 2			Reader 3			Reader 4			Reader 5		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
MJAS	TP	FN	TP	TP	FN	FN	TP	FN	TP	TP	TP	TP	TP	FN	FN
MJBG	TP	FN	FN	TP	FN	FN	TP	FN	FN	FN	FN	FN	FN	TP	FN
MJBJ	TP	FN	TP	TP	TP	TP	TP	TP	TP	TP	TP	FN	TP	FN	TP
MJBK	TP	TP	FN	TP	TP	FN	FN	FN	TP	TP	TP	TP	FN	FN	FN
MJBL	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJCF	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN	FN	TP	TP
MJCQ	TP	TP	TP	TP	TP	FN	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJCR	TP	TP	FN	TP	FN	FN	TP	TP	FN	FN	FN	FN	TP	TP	TP
MJCX	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJDA	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJDH	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJDU	FN	FN	FN	TP	TP	FN	TP	TP	FN	TP	TP	FN	FN	FN	FN
MJEA	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJEB	FN	TP	FN	TP	FN	FN	TP	FN	FN	TP	FN	FN	TP	FN	FN
MJEG	TP	TP	TP	TP	TP	TP	TP	TP	FN	TP	FN	FN	TP	TP	TP
MJGR	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJHD	TP	TP	TP	FN	TP	TP	TP	TP	TP	TP	TP	TP	FN	TP	FN
MJHH	TP	FN	FN	TP	TP	FN	TP	FN	FN	FN	FN	FN	TP	TP	FN
MJHJ	FN	FN	FN	TP	FN	FN	TP	TP	TP	TP	TP	TP	TP	TP	TP
MJHK	TP	TP	FN	TP	TP	FN	TP	FN	FN	FN	TP	FN	FN	FN	FN

*True positive, TP

†False positive, FN

References

1. BreastScreen Australia. National Accreditation Standards: BreastScreen Australia Quality 2008 [cited 2014 May 21]. Available from: <http://www.cancerscreening.gov.au>

CHAPTER 5

RADIOLOGISTS' PERFORMANCE AT REDUCED RECALL RATES IN MAMMOGRAPHY: A LABORATORY STUDY

Chapter 5 is submitted as:

N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, B.P.Soh, S. Lewis. *Radiologists' performance at reduced recall rates in mammography: A laboratory study*. *The Breast Journal*. (2017). (TBJ-00165-2017.R1)

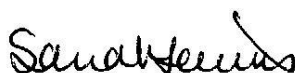
STATEMENT FROM AUTHOR

Statement from authors confirming authorship contribution of the PhD candidate

As co-authors of the paper “Radiologists’ performance at reduced recall rates in mammography: A laboratory study”, we confirm that Norhashimah Mohd Norsuddin has made the following contributions:

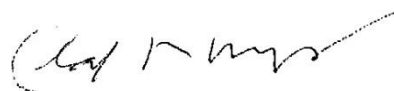
- Conception and design of the research
- Data collection
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Asc. Professor Sarah Lewis



Date: 21/03/2017

Asc. Professor Claudia
Mello-Thoms



Date: 21/03/2017

Dr Warren Reed



Date: 21/03/2017

Dr Baolin Pauline Soh



Date: 21/03/2017

RADIOLOGISTS' PERFORMANCE AT REDUCED RECALL RATES IN MAMMOGRAPHY: A LABORATORY STUDY

Abstract

Rationale and objectives: Target recall rates are often used as a performance indicator in mammography screening programs with the intention of reducing false positive decisions, over diagnosis and anxiety for participants. However, the relationship between target recall rates and cancer detection is unclear, especially when readers are directed to adhere to a predetermined rate. The purpose of this study was to explore the effect of setting different recall rates on radiologist's performance.

Materials and Methods: Institutional ethics approval was granted and informed consent was obtained from each participating radiologist. Five experienced breast imaging radiologists read a single test set of 200 mammographic cases (20 abnormal and 180 normal). The radiologists were asked to identify each case that they required to be recalled in three different recall conditions; free recall, 15% and 10% and mark the location of any suspicious lesions.

Results: Wide variability in recall rates was observed when reading at free recall, ranging from 18.5% to 34.0%. Readers demonstrated significantly reduced performance when reading at prescribed recall rates, with lower sensitivity ($H=12.891$, $P=0.002$), case location sensitivity ($H=12.512$, $P=0.002$) and ROC AUC ($H=11.601$, $P=0.003$) albeit with an increased specificity ($H=12.704$, $P=0.002$). However, no significant changes were

evident in lesion location sensitivity ($H=1.982$, $P=0.371$) and JAFROC FOM ($H=1.820$, $P=0.403$).

Conclusion: In this laboratory study, reducing the number of recalled cases to 10% significantly reduced radiologists' performance with lower detection sensitivity, although a significant improvement in specificity was observed.

Keywords: Recall rates, sensitivity, specificity, reader's performance, screening mammography

Introduction

A comparison of international screening programs has shown a wide range of recall rates in clinical practice across different countries, from 1.4% in the Netherland to 15% in the United States (US) for the first mammography screening examination (1). Recalling a large number of women is considered to improve the number of cancers detected, however previous comparative studies have demonstrated that high recall rates do not significantly improve the cancer detection rate (2, 3). Additionally, a higher recall rate may only contribute to a substantial increase in false positive findings which may result in unnecessary assessments, patient anxiety and additional financial costs hampering the success of breast screening programs (4-6).

The positive correlation between false positive results and recall rates (7) has prompted many screening programs to impose specific recall targets in order to optimize the trade-off between recall rates and cancer detection. These recall policies are also used to evaluate the performance of breast readers in the respective programs and provide guidelines for best practice. Variation also exists within screening mammography programs, with higher target recall rates for the first or initial screening as compared to subsequent screening. For example, BreastScreen Australia (BSA) suggests that the clinical recall rate should be at 10% for initial screens with a recall rate for subsequent screening at 5% (8, 9).

Extensive studies have shown varying results regarding an appropriate recall rate that will give the best trade-off between recall rates and cancer detection rates (10-12). A prospective study by Yankaskas et al. (2001) suggested screening practices at recall rates

between 4.9% and 5.5% will yield efficiency in cancer detection versus false positive results, whereas subsequent work by Schell et al. (2007) demonstrated maximum sensitivity and minimal false positives occurred at a 10% recall rate for the first screening and at 6.5% for subsequent screening (10, 12). Another study from the Netherlands has indicated that recall rates of more than 4% only contribute to a higher number of false positive decisions, not the number of cancers detected (11).

The purpose of this study was to explore the effect of setting varying recall rates on the performance of radiologists viewing the same test set of images over three separate reading sessions. We have explored this through a methodology that assesses the radiologists' ability to correctly locate lesions and give a confidence rating based on the decisions.

Materials and Methods

Institutional ethical approval was granted and informed consent was obtained from all participants involved in this study. It was performed at the Medical Imaging Optimisation and Perception Group (MIOPeG) laboratory at the Brain Mind Centre (BMC) of the University of Sydney between February 2015 and January 2016.

Participants

Five experienced radiologists who specialized in breast imaging with a median of 15 years (range, 9 to 26 years) of experience of interpreting mammograms, reading between 3500 and 30000 (median, 8000) mammograms each year and spent a median of 10 hours a week

reading mammography cases volunteered to participate in this study (**Table 1**). The participating radiologists were given a small gift voucher as an expression of gratitude for their participation on completion of the study.

Table 1 Demographic details of participating radiologists

Reader number	Number of years of experience	Number of mammography cases read per year	Number of hours per week reading mammograms
1	15	30 000	10
2	26	10 000	3
3	15	10 000	10
4	9	6 000	24
5	20	3 500	6
Median	15	8 000	10

Experimental protocol

Cases

A test set containing 200 de-identified digital mammographic examinations obtained from the BreastScreen NSW (BSNSW) digital imaging library was presented in a randomized order to each reader for three separate reading sessions. Each case comprised two

mammographic views, the cranio-caudal (CC) and the medio-lateral oblique (MLO) respectively of each breast. There were 180 cases with normal findings and 20 cases with abnormal findings in the test set; all the abnormal cases contained a single biopsy-proved malignancy. An expert breast radiologist who is involved in training assessment, quality, clinical policies of BreastScreen NSW and also has responsibility for clinical management of a screening center then identified the 'truth' locations of all malignant cases. The expert did not participate as an observer in this experiment and had access to prior images with biopsy confirmed malignancy results. Normal cases were validated after 2 years normal screening follow up.

Reading environment

This longitudinal study was divided into three separate reading sessions. Each session had a different recall rate condition and was separated from the previous reading by a minimum of two months to reduce any memory effects. The total study time was six months for the three reads for each reader. At the first reading session, no numerical percentage recall rate was imposed and readers were tasked with a "free recall" when interpreting the cases; that is, they could recall as many cases as they believed necessary. In the second session, the number of mammographic cases that readers could recall was set at 30 cases (15%), and reduced to 20 cases (10%) for the third reading.

The laboratory reading environment was designed to be as authentic to the clinical environment as possible, using an identical clinical workstation as used by BreastScreen New South Wales (BSNSW), Australia. All images in the test set were displayed on a pair

of five-megapixel (5MP) EIZO Radioforce GS510 medical-grade monitors (Ishikawa, Japan) with ambient lighting maintained at 20 to 40 lux throughout the reading sessions. Prior to study commencement, calibration was performed on the monitor displays to adhere to the Digital Imaging and Communications in Medicine (DICOM) Part 14 Standard (13).

Reading Task

Figure 1 shows the flowchart process of the study methodology. During the reading sessions, readers identified each mammographic case that they considered malignant in keeping with their free or specified target recall rate. Following that, readers used customized recording software to mark the location of any suspicious lesions from the recalled cases. This software was designed to record all the coordinates of marked lesions for each of the recalled cases on a laptop adjacent to the two 5MP monitors. Readers were not permitted to exceed their target recall rates and the software provided a continuously updated count on the number of cases marked as “recall”. All lesions marked on each recalled case were given a confidence score ranging from 1 to 5, with a greater number indicating a higher confidence of malignancy. A score of 1 or 2 indicated a normal and benign lesion respectively. A 60-pixel acceptance radius surrounding each lesion was considered the acceptable radius as it encompassed the largest lesion present in the test set. A lesion was considered correctly detected when the location was within 60 pixels from the center of the true location of the cancer and it was given a confidence score between 3 to 5.

Readers were not provided with any clinical information associated with the cases including the prevalence of abnormal cases and no prior images were available. The readers had unlimited time and were able to scroll back through the cases if they wished to alter their decision or needed to reduce the number of cases recalled to align with the specific target recall rate condition. Readers were also able to digitally manipulate the images including windowing, zooming and panning as in the actual clinical setting.

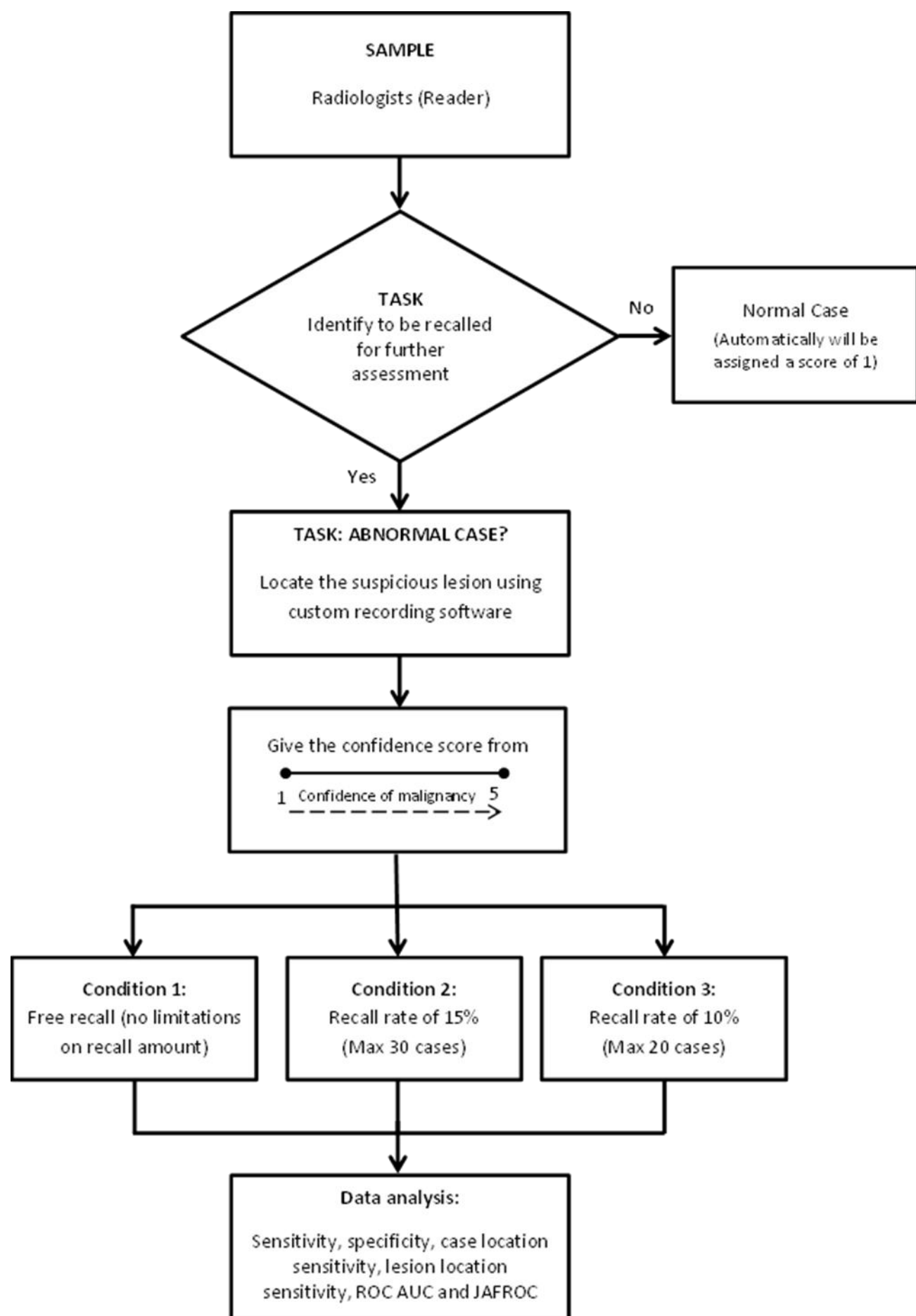


Figure 1 Flowchart process of reading task was conducted in the study

Data Analysis

For all reading sessions, reader performance was assessed using sensitivity, specificity, case location sensitivity, lesion location sensitivity, receiver operating characteristic (ROC) area under the curve (AUC) and Jack-knife free-response ROC (JAFROC) figure of merit (FOM). The marked lesions were identified either as positive or negative by comparing selections with the truth table compiled by the expert breast radiologist. All performance data were analyzed using JAFROC Version 4.2 software and statistical analysis was performed using SPSS software version 22.0.

A true positive was scored if a lesion was marked within the acceptance radius and received a confidence score between 3 to 5. A false positive was defined for any incorrect localization on normal or benign cases, or if it was outside the 60 pixels range of a lesion in abnormal cases. A true negative outcome was recorded if the case was correctly identified as normal or lesion-free. A false negative was scored when cancer lesions were not marked. The performance metrics used for this study are explained as follows:

- Sensitivity was defined as the proportion of abnormal cases correctly identified by the reader.
- Specificity was defined as the proportion of normal cases correctly identified by the reader.
- Case location sensitivity measures the proportion of positive cases correctly marked, where at least one lesion was correctly identified on the correct location in the case.

- Lesion location sensitivity is the proportion of positive lesions correctly marked; it was calculated by dividing the number of lesions correctly detected by the total number of positive/abnormal lesions, where the positive lesions detected were on the correct locations for each lesion.
- ROC analysis is a binary paradigm focused on a single rating per case. In this study a TP score was given to a case when the reader correctly identified the correct side of the breast containing cancer, without the need to show the specific location of the lesion.
- JAFROC analysis is a free-response paradigm that allows lesion location information to be included when analysing reader performance.

The analysis was done in two steps. Firstly, a Kruskal-Willis test was performed across the three reading sessions with statistical significance set at $P < 0.05$. Secondly, post-hoc analysis using the Mann-Whitney U test was performed to identify which groups were significantly different from each other. For this purpose, Bonferroni adjustment was applied to the alpha values by dividing the alpha level by the number of comparisons made. Results with the revised alpha level, $P < 0.017$, were deemed to represent significant differences.

Results

Table 2 demonstrates the readers' scores for all performance metrics; sensitivity, specificity, case location sensitivity, lesion location sensitivity, ROC AUC and JAFROC FOM at the conditions of free recall, 15% and 10% recall rate. The median recall rate for

all readers when reading at free recall was 25.0% and ranged from 18.5% to 34.0%, which was higher than that recommended by BreastScreen Australia for initial screening (10%). By limiting the number of recalled cases (from free recall to 10% recall rate), readers demonstrated reduced performance, with a decrease in sensitivity (from median 0.80 to 0.55), case location sensitivity (from median 0.80 to 0.55), lesion location sensitivity (from median 0.64 to 0.56), and ROC AUC (from median 0.84 to 0.75). However, there was a median increase in specificity from 0.81 to 0.95.

Significant changes were observed in reduced sensitivity ($H=12.891$, $P=0.002$), case location sensitivity ($H=12.512$, $P=0.002$) and ROC AUC ($H=11.601$, $P=0.003$) along with an increased specificity ($H=12.704$, $P=0.002$) across all reading conditions (**Table 3**). No significant differences were noted for lesion location sensitivity ($H=1.982$, $P=0.371$) and JAFROC FOM ($H=1.820$, $P=0.403$). Although a significant difference was found in ROC AUC, post-hoc Mann-Whitney U test showed no significant differences in ROC AUC when reading at free recall and 15% ($z=-2.200$, $P=0.028$).

Table 2 Results for sensitivity, lesion location sensitivity, case location sensitivity, specificity, ROC AUC and JAFROC FOM at free call, 15% and 10% conditions

Reader number	Sensitivity			Specificity			Case location sensitivity			Lesion location sensitivity			ROC AUC			JAFROC FOM		
	Free recall	15%	10%	Free recal l	15%	10%	Free recal l	15%	10%	Free recall	15%	10%	Free recal l	15%	10%	Free recal l	15%	10%
1	0.80	0.65	0.55	0.83	0.91	0.95	0.80	0.60	0.55	0.78	0.67	0.58	0.84	0.79	0.76	0.83	0.80	0.78
2	0.90	0.65	0.45	0.79	0.91	0.93	0.80	0.60	0.45	0.56	0.56	0.47	0.86	0.79	0.70	0.76	0.76	0.71
3	0.90	0.65	0.55	0.72	0.92	0.95	0.80	0.65	0.55	0.81	0.67	0.56	0.84	0.79	0.75	0.80	0.80	0.76
4	0.75	0.70	0.55	0.88	0.92	0.94	0.70	0.65	0.55	0.64	0.64	0.56	0.83	0.82	0.75	0.76	0.79	0.75
5	0.75	0.70	0.55	0.81	0.92	0.95	0.65	0.60	0.55	0.36	0.33	0.50	0.80	0.82	0.76	0.59	0.63	0.73
Median	0.80	0.65	0.55	0.81	0.92	0.95	0.80	0.60	0.55	0.64	0.64	0.56	0.84	0.79	0.75	0.76	0.79	0.75

†JAFROC FOM , Jack-knife free-response figure of merit

‡ROC AUC , Receiver operating characteristic area under the curve

Table 3 Kruskal-Wallis analysis and post hoc Mann-Whitney U test of sensitivity, lesion location sensitivity, case location sensitivity, specificity, ROC AUC and JAFROC FOM

	Kruskal-Wallis Test	Post-hoc test (Mann-Whitney U test)		
		Free recall VS 15%	15% VS 10%	Free recall VS 10%
		P value (P < 0.017)		
	P value (P < 0.05)			
Sensitivity	0.002*	0.008*	0.006*	0.007*
Specificity	0.002*	0.008*	0.007*	0.008*
Case Location Sensitivity	0.002*	0.013*	0.006*	0.006*
Lesion Location Sensitivity	0.371	0.598	0.243	0.245
ROC AUC	0.003*	0.028	0.009*	0.009*
JAFROC FOM	0.403	0.917	0.251	0.251

**Values shown in bold represent statistically significant difference*

†JAFROC FOM , Jack-knife free-response figure of merit

‡ROC AUC, Receiver operating characteristic area under the curve

Discussion

Varied specified recall rates have been introduced by various national and international organizations as a guideline for optimizing the performance of their breast screening programs. However, there is a lack of evidence supporting the actual effect of differing target recall rates upon radiologists' performance. Our results show that each of the five readers' sensitivity was highest when operating in the free recall condition as compared to a reduced specified recall rate, with a higher median ROC AUC (0.84) and a JAFROM FOM (0.73). When the readers were tasked with reducing their recalled cases from free recall to specific recall rates (15% and 10%), their performance declined noticeably, with a significant reduction in sensitivity, case location sensitivity and ROC AUC.

The higher sensitivity at higher specified recall rates observed in our study is in agreement with Gur et al and Schell et al, who suggest that recalling more cases may result in a higher cancer detection rate (7, 12). However, the effectiveness of a breast screening program is not merely dependent on the number of cancers detected but also in reducing the number of unnecessary recall among screened women. It is well documented that false-positive recalls are associated with psychological consequences and economic burden (14-19).

In the current study, readers demonstrated a significant improvement in their specificity as recall rates reduced ($P=0.002$). By lowering the number of cases allowed to be recalled, readers may have needed to sacrifice some cases that they considered to be abnormal at a higher recall rate, which in turn resulted in fewer false positive decisions.

The increased specificity observed here concurs with previous findings by Otten et al and Elmore et al, where higher recall rates correlated with an increase in false positive findings and lower specificity (1, 11).

Comparing our results directly with previous studies in the literature is complex as we began our reading sessions with the highest recall rate (free recall) and stepped it down to 15% and then 10% which complied with the Australian standard for the first screening recall rate. In the Dutch study by Otten and colleagues (2005), they reported the effect of increasing recall rates up to 10% on cancer detection, with the best efficiency between cancer detection and specificity seen at recall rates below 5%. Otten et al found that between recall rates of 0.9 and 4.0%, the cancer detection rate increased approximately 17.0% respectively. However, when they modeled a further increase in the recall rate above 5%, the results suggested a relatively small increase in the cancer detection rate (approximately 0.6%), with a higher number of false positive results (11).

Similarly, Yankaskas et al (2001) also reported a positive effect between recall rate and sensitivity at recall rates of 5%. As the recall rates increased to 13.4%, a non-significant increase in sensitivity was observed in this prospective study (10). In our current work, we were unable to test the effect of lowering the recall rate further to 5% due statistically to the number of cancer cases in the test set of 200. However, the readers in our study demonstrated their best cancer detection at free recall, with a significant decrease in sensitivity at 15% and 10% respectively. In acknowledging a difference in study outcomes, the best explanation may lie in the variation between reading in the clinical environment and our experimental design. Also in contrast to our study in the research by

Otten et al (11) and Yankaskas et al (10), the previous clinical information and prior images were available potentially altering reader performance (20, 21).

Although small reductions were seen in lesion location sensitivity, there was no statistically significant difference for this performance metric at lower recall rates. Conversely, case location sensitivity demonstrated significant differences across all reading conditions. The discrepancy in these results may be explained by the fact that lesion location sensitivity has a stricter criterion in decision making, where readers must correctly identify the location of the lesion for every lesion in cancer cases. In contrast, the criterion applied for case location sensitivity was for readers to correctly identify where at least one lesion was marked in each case. It also interesting to note that, at an individual level, reader 5 demonstrated a substantial difference in lesion location sensitivity values (below 0.40) as compared to the other four readers, regardless of the recall rate conditions, without a significant change in the case location sensitivity. In this example, we see the clinical reporting behaviour of readers, whereby this reader noted they are more likely to mark only one mammographic view for a case that required further assessment, rather than lesion-specific marking as per our instructions. The reader may have considered the task done if located on just one image as this is still a recall in clinical practice. To our knowledge, no commentary on the effect of recall rates on readers' performance has been reported using JAFROC analysis from the prior studies. With greater statistical power over traditional methodologies reporting observer's performance, such as ROC (22, 23), our data yielded no significant differences in JAFROC FOM as observed across all reading sessions. Quite simply, in the true positive cases that the radiologists did recall, they were able to identify the lesion location with good precision.

Although a 10% recall rate in our final specified target recall rate is the same as Australian clinical practice for the first mammographic screening, often readers are not held to this in a strict sense. All five readers in our study demonstrated a higher individual recall rate (free recall) ranging from between 18.5% and 34.0% when no limitation was imposed. This result may be because the readers were aware that the study was conducted in laboratory conditions and their decisions would not influence patient care and no patients would actually be recalled. The readers may also have been expecting a higher number of abnormalities, or an enriched test set, in a laboratory study (20, 24). Additionally, the removal of clinical history and prior images in the test set may have affected their behavior when interpreting the mammograms as they usually have access to this additional information in clinical practice and more likely to recall in these circumstances to be on the safe side.

Our study was designed to resemble clinical mammography practice, hence there was a requirement to use a reasonably high number of cases ($n=200$) and images ($n=800$). Therefore fatigue may be present for some readers and this may have contributed to some loss of attention, particularly for difficult to detect lesions (25). Our sample consisted of highly experienced radiologists (median of 15 years) with more than 3000 mammographic cases reading per year, which may have contributed to the similarity in their performance. A larger study with a diversity of reader experience and case load would allow clarification of the importance of experience when adhering to specific recall rates. It would be interesting to explore how readers with a range of experience adhere to different recall rates and also how varying recall rates would affect the readers' performance, and may yield some important relationships between lesion location sensitivity and experience.

Nevertheless, current data from our preliminary work has demonstrated important empirical evidence on how limiting the number of recalled cases in screening can impact on the behavior of readers. A unique aspect of this study is that recall rates were controlled by treating predetermined recall rates as the primary indicator; a confounder known to increase variability in past studies (10-12).

Conclusion

In summary, our data suggests that specific target recall rates caused a significant reduction in readers' sensitivity with an associated increase in specificity. However, no differences in readers' JAFROC scores were demonstrated, indicating a level of consistency in readers identifying and locating lesions within this test set. Reducing recall rates by enforcing a specific target recall rate may result in a corresponding reduction in cancer detection and may impact on the behaviour of readers in their recall-decision strategies for abnormal cases. Further work on how specific lesion features or characteristics were reflected in the readers' recall decision is required to understand further the nature of improved the balance between true and false positive decisions.

References

1. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International Variation in Screening Mammography Interpretations in Community-Based Programs. *Journal of the National Cancer Institute*. 2003;95(18):1384-93.
2. Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, et al. Comparison of screening mammography in the united states and the united kingdom. *JAMA*. 2003;290(16):2129-37.
3. Kemp Jacobsen K, O'Meara ES, Key D, S.M. Buist D, Kerlikowske K, Vejborg I, et al. Comparing sensitivity and specificity of screening mammography in the United States and Denmark. *International Journal of Cancer*. 2015;137(9):2198-207.
4. Alamo-Junquera D, Murta-Nascimento C, Macia F, Bare M, Galceran J, Ascunce N, et al. Effect of false-positive results on reattendance at breast cancer screening programmes in Spain. *Eur J Public Health*. 2012;22(3):404-8.
5. Sim MJ, Siva SP, Ramli IS, Fritschi L, Tresham J, Wylie EJ. Effect of false-positive screening mammograms on rescreening in Western Australia. *The Medical journal of Australia*. 2012;196(11):693-5.
6. Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, et al. Systematic review of the psychological consequences of false-positive screening mammograms. *Health technology assessment (Winchester, England)*. 2013;17(13):1-170, v-vi.
7. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, et al. Recall and detection rates in screening mammography. *Cancer*. 2004;100(8):1590-4.
8. BreastScreen Australia. National Accreditation Standards: BreastScreen Australia Quality 2008 [cited 2014 May 21]. Available from: [http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/A03653118215815BCA257B41000409E9/\\$File/standards.pdf](http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/A03653118215815BCA257B41000409E9/$File/standards.pdf).
9. Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*. 2008;19(4):614-22.
10. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of Recall Rates with sensitivity and positive predictive values of screening mammography. *American journal of roentgenology*. 2001;177(3):543-9.
11. Otten JDM, Karssemeijer N, Hendriks JHCL, Groenewoud JH, Fracheboud J, Verbeek ALM, et al. Effect of Recall Rate on Earlier Screen Detection of Breast Cancers

Based on the Dutch Performance Indicators. Journal of the National Cancer Institute. 2005;97(10):748-54.

12. Schell MJ, Yankaskas BC, Ballard-Barbash R, Qaqish BF, Barlow WE, Rosenberg RD, et al. Evidence-based Target Recall Rates for Screening Mammography. Radiology. 2007;243(3):681-9.

13. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine (DICOM) Part 14: Grayscale Standard Display Function 2004. Available from: http://dicom.nema.org/dicom/2004/04_14pu.pdf.

14. Bond M, Pavey T, Welch K, Cooper C, Garside R. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technology Assessment. 2013;17(13):170.

15. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. Ann Intern Med. 2007;146(7):502-10.

16. Brodersen J, Siersma VD. Long-term psychosocial consequences of false-positive screening mammography. Annals of family medicine. 2013;11(2):106-15.

17. Castells X, Molins E, Macia F. Cumulative false positive recall rate and association with participant related factors in a population based breast cancer screening programme. Journal of epidemiology and community health. 2006;60(4):316-21.

18. Lafata JE, Simpkins J, Lamerato L, Poisson L, Divine G, Johnson CC. The economic impact of false-positive cancer screens. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2004;13(12):2126-32.

19. Maxwell AJ, Beattie C, Lavelle J, Lyburn I, Sinnatamby R, Garnett S, et al. The effect of false positive breast screening examinations on subsequent attendance: retrospective cohort study. Journal of medical screening. 2013;20(2):91-8.

20. Soh BP, Lee W, McEntee MF, Kench PL, Reed WM, Heard R, et al. Screening Mammography: Test Set Data Can Reasonably Describe Actual Clinical Reporting. Radiology. 2013;268(1):46-53.

21. Carney PA, Cook AJ, Miglioretti DL, Feig SA, Bowles EA, Geller BM, et al. Use of clinical history affects accuracy of interpretive performance of screening mammography. Journal of Clinical Epidemiology. 2012;65(2):219-30.

22. Chakraborty DP. Recent advances in observer performance methodology: jackknife free-response ROC (JAFROC). Radiation Protection Dosimetry. 2005;114(1-3):26-31.

23. Chakraborty DP. Analysis of Location Specific Observer Performance Data: Validated Extensions of the Jackknife Free-Response (JAFROC) Method. Academic Radiology. 2006;13(10):1187-93.

24. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The Prevalence Effect in a Laboratory Environment: Changing the Confidence Ratings. *Academic Radiology*. 2007;14(1):49-53.
25. Ciatto S, Ambrogetti D, Bonardi R, Catarzi S, Risso G, Rosselli Del Turco M, et al. Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. *Journal of medical screening*. 2005;12(2):103-6.

CHAPTER 6

AN INVESTIGATION INTO THE MAMMOGRAPHIC APPEARANCES OF MISSED CANCERS WHEN RECALL RATES ARE REDUCED.

Chapter 6 is accepted for publication as:

N. Mohd Norsuddin, C. Mello-Thoms, W. Reed, M.Rickard, S. Lewis. *An investigation into the mammographic appearances of missed cancers when recall rates are reduced.* British Journal of Radiology (BJR). (In Press) (2017).

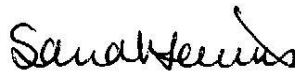
STATEMENT FROM AUTHOR

Statement from authors confirming authorship contribution of the PhD candidate

As co-authors of the paper “An investigation into the mammographic appearances of missed breast cancer when recall rates are reduced”, we confirm that Norhashimah Mohd Norsuddin has made the following contributions:

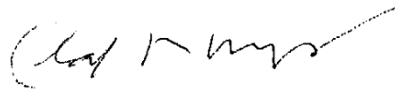
- Conception and design of the research
- Data collection
- Analysis and interpretation of the findings
- Writing the paper and critical appraisal of content

Asc. Professor Sarah Lewis



Date: 25/03/2017

Asc. Professor Claudia Mello-Thoms



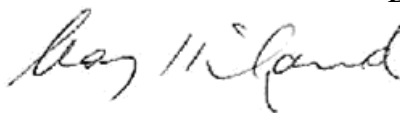
Date: 24/03/2017

Dr Warren Reed



Date: 24/03/2017

Dr Mary Rickard



Date:

26/03/17

**AN INVESTIGATION INTO THE MAMMOGRAPHIC
APPEARANCES OF MISSED BREAST CANCER WHEN RECALL
RATES ARE REDUCED**

Abstract

Objectives: This study investigated whether certain mammographic appearances of breast cancer are missed when radiologists read at lower recall rates.

Methods: Five radiologists read one identical test set of 200 mammographic (180 normal cases and 20 abnormal cases) three times and were requested to adhere to three different recall rate conditions: free recall, 15% and 10%. The radiologists were asked to mark the locations of suspicious lesions and provide a confidence rating for each decision. An independent expert radiologist identified the various types of cancers in the test set, including the presence of calcifications and the lesion location including specific mammographic density.

Results: Radiologists demonstrated lower sensitivity and receiver operating characteristic (ROC) area under the curve (AUC) for non-specific density (NSD)/asymmetric density (H=6.27, P=0.04 and H=7.35, P=0.03 respectively) and mixed features (H=9.97, P=0.01 and H=6.50, P=0.04 respectively) when reading at 15% and 10% recall rates. No significant change was observed on cancer characterized with stellate masses (H=3.43, P=0.18 and H=1.23, P=0.54 respectively) and architectural distortion (AD) (H=0.00, P=1.00 and H=2.00, P=0.37 respectively). Across all recall conditions, stellate masses were likely to be recalled (90.0%) while NSDs were likely to be missed (45.6%).

Conclusion: Cancers with a stellate mass were more easily detected and are likely to continue to be recalled, even at lower recall rates. Cancers with non-specific density and mixed features were most likely to be missed at reduced recall rates.

Advances in knowledge:

Internationally, recall rates vary within screening mammography programs considerably, with a range between 1% to 15% and very little is known about the type of breast cancer appearances found when radiologists interpret screening mammograms at these various recall rates. Therefore, understanding the lesion types and the mammographic appearances of breast cancers that affected by readers' recall decisions should be investigated.

Introduction

Several studies have demonstrated substantial variability in recall rates among radiologists reporting in breast screening programs, with large international variations ranging from 1% to 15.1%.¹⁻³ Although many countries use the Breast Imaging Reporting and Data System (BIRADS) as a standardized method of reporting mammograms, a considerable variability in assessment is still seen, even when reporting the same mammographic images by different readers.^{1, 4-6}

Many factors influence the difficulty of reaching a correct diagnosis for normal and abnormal cases. Specific mammographic lesion features have been found to significantly contribute to cancer detection, especially lesion conspicuity.⁷ Ikeda et al found that 22% of missed cancers in the Malmö Screening Trial showed subtle mammographic signs of malignancy, with lesions that present with architectural distortion (AD) being the most challenging malignancy feature for readers to detect.⁸ Furthermore, dense breast tissue has been found to be a strong confounder for lesion detection and cases with high mammographic breast density were more likely to be recalled.⁹⁻¹¹

With a large variation in the recommended target recall rates within screening mammography programs internationally¹²⁻¹⁴, very little is known about the mammographic features or breast cancer appearances which are affected by recall decisions at reduced rates. A recent study by Onega et al¹⁵ with 119 radiologists reading 109 screening mammograms found low recall agreement for lesions with architectural distortions (AD) and asymmetric densities features. Identifying cancer appearances that are more likely to be missed when recall rates are reduced is clinically important because it can inform readers' decisions in recalling women for further assessment. Therefore, the

purpose of this study was to investigate which types of mammographic appearances of breast cancer are most likely to be missed when radiologists read at lower recall rates.

Materials and methods

Sample

Institutional ethics approval was granted for this study (Project number: 2014/484). Five breast radiologists who reported for BreastScreen New South Wales (BSNSW) with 15 to 26 years of experience participated in this study. The radiologists read between 2,000 and 30,000 (median 8,000) mammograms each year and spent a median of 10 hours a week reading mammography cases.

Cases

A test set of screening mammograms, comprising of 200 cases, was obtained from the BSNSW digital imaging library. An enriched test set containing 180 normal cases and 20 abnormal cases, with each abnormal case containing a single biopsy-proved malignancy, was obtained. Each mammographic examination consisted of a two-view digital mammogram, a cranio-caudal view (CC) and a medio-lateral oblique (MLO) of both breasts with a range of lesion conspicuity, from subtle to obvious cancer presentations and a variety of normal mammographic appearances. In 16 cases the cancer was visible on both mammographic views (CC and MLO) of a given breast, while four cases had a visible lesion on either one of the mammographic views, which resulted in a total of 36 malignant lesions available for localization. The “truth” locations and mammographic appearances of

all malignant cases were identified by an expert radiologist (M.R), who is involved in training, quality assessment, clinical policies of BreastScreen NSW and also is responsible for the clinical management of a screening centre. This expert radiologist had access to the biopsy report and prior images, which assisted in determining the location and the mammographic appearances of the abnormalities based on the Australian Synoptic Breast Imaging Report of the National Breast Cancer Centre (NBCC) that is endorsed by the Royal Australian and New Zealand College of radiologists (RANZCR).^{16, 17} Normal cases were validated after 2 years normal screening follow up. Lesion descriptors used in this study have been defined in **Table 1**.

Table 1 Definition of lesion terms used for classification ¹⁷

Lesion abnormalities	Definition
Calcification	Deposition or collections of calcium compounds in breast tissue of sufficient size to be seen on mammogram and malignancy is characterised by size (0.05-0.5mm), distribution (cluster, multiple cluster, or sometimes scattered), pleomorphism and variation of density.
Stellate lesion	Spiculations of variable length radiating from a central point or mass characterized with small or large, and of low, mixed or high density compared to surrounding breast parenchyma.
Architectural distortion	Abnormal configuration of the ductal and ligamentous structures of breast parenchyma compared with the remainder of the breast tissue markings and often appears with spiculation, focal retraction, distortion of the parenchymal edge, and disorganisation of markings.
Non-specific density	Asymmetry of breast tissue seen in one of the breasts, on either one or two mammographic views with poorly defined characteristics of breast density.

The categories of mammographic cancers' appearances were: stellate, AD, non-specific density (NSD), mixed appearance of calcification and AD, and stellate and NSD. The mammographic appearances of the cancer lesions were classified into three categories: lesion type, breast density and location of the lesions in the images. Breast density was graded according to the Breast Imaging Reporting and Data System, BI-RADS criteria, 4th edition: 1–the breast is almost entirely fat (<25% glandular), 2–there are scattered fibroglandular densities (approximately 25–50% glandular), 3–the breast is heterogeneously dense (approximately 51–75% glandular) or 4–the breast tissue is extremely dense (>75% glandular).

Reading sessions

This study was conducted in a laboratory reading environment designed to closely resemble the clinical environment, with all images displayed on the same calibrated, high-specification workstation with five-megapixel EIZO Radioforce GS510 medical-grade monitors (Ishikawa, Japan). Readers were required to read all the 200 mammographic cases in three separate reading sessions using different recall rates. At the first reading condition, no numerical percentage recall rate was imposed and readers were tasked with a “free recall” when interpreting the cases; that is, they could recall as many cases as they believed necessary. In the second condition, the number of mammographic cases that readers could recall was restricted to 30 cases (15%) based on international recall rates, to reflect 15.1% in the United States.³ For the third session, readers were restricted to recall only 20 cases (10%) to align with the first screening recall required by BreastScreen Australia. To reduce any memory effect, each reading session was separated by a

minimum of two months and the reading order of images was randomized for each reading session and each reader.

Reading task

During the three reading sessions, readers indicated whether they would recall or not recall the mammographic cases as per their usual clinical practice for BreastScreen Australia. For each recalled case, readers were required to mark the location of the detected lesions on both mammographic views and score them on the scale of one to five (with five being the highest confidence of malignancy) using a custom made recording software. This scoring system is aligned to BreastScreen Australia practice for classifying mammographic lesions; a score of 1 and 2 indicated a normal and benign decision respectively, and a score of 3, 4 or 5 would be considered as a recall for assessment. Readers were not permitted to exceed their target recall rates at the end of each recall condition, however, if they exceeded during the reading session, the readers were able to scroll back through the cases and alter their decision to ensure the number of cases recalled aligned with the prescribed target recall rate condition. Readers were also able to digitally manipulate the images including windowing, zooming and panning as in the actual clinical setting. For the purpose of simulating the first screening read, no prior images or clinical history were provided during the three reading sessions. The readers were not aware of the prevalence of abnormal cases in the test set.

Data analysis

Reader performance was assessed by sensitivity and receiver operating characteristic (ROC) area under the curve (AUC). Sensitivity was defined by the proportion of cancers correctly marked by readers. Even though the readers were encouraged to mark a perceived lesion on both views, for analysis purposes, a true positive (TP) was assigned when the reader correctly marked it in one -view of the positive breast only. All marked lesions were compared with those contained in the truth table. A significant difference of both metrics was compared across the three recall conditions using a non-parametric Kruskal-Wallis test. The false positive rate was calculated by the number of false positive decisions made on normal cases divided by the total number of normal cases.

Further analysis for this study focused on determining whether the detection of any specific cancer types was altered when the recall rates were reduced (Condition 15% and 10%), which narrowed the analysis to the 20 abnormal cases only and the analysis was performed at a case-based level. Cancer difficulty was scored out of 15, which was the total number of readers ($n = 5$) multiplied by the number of reading conditions ($n = 3$). Cancer difficulty was then classified as the sum of lesions that were correctly marked throughout all recall conditions resulting in three difficulty levels as follows;

1. **Lower difficulty:** Lesion in the case was correctly marked by readers at least 12 times across the 3 reading conditions
2. **Medium difficulty:** Lesion in the case was correctly marked by readers between 5 to 11 times across the 3 reading conditions
3. **Higher difficulty:** Lesion in the case that was marked by readers less than 5 times across the 3 conditions

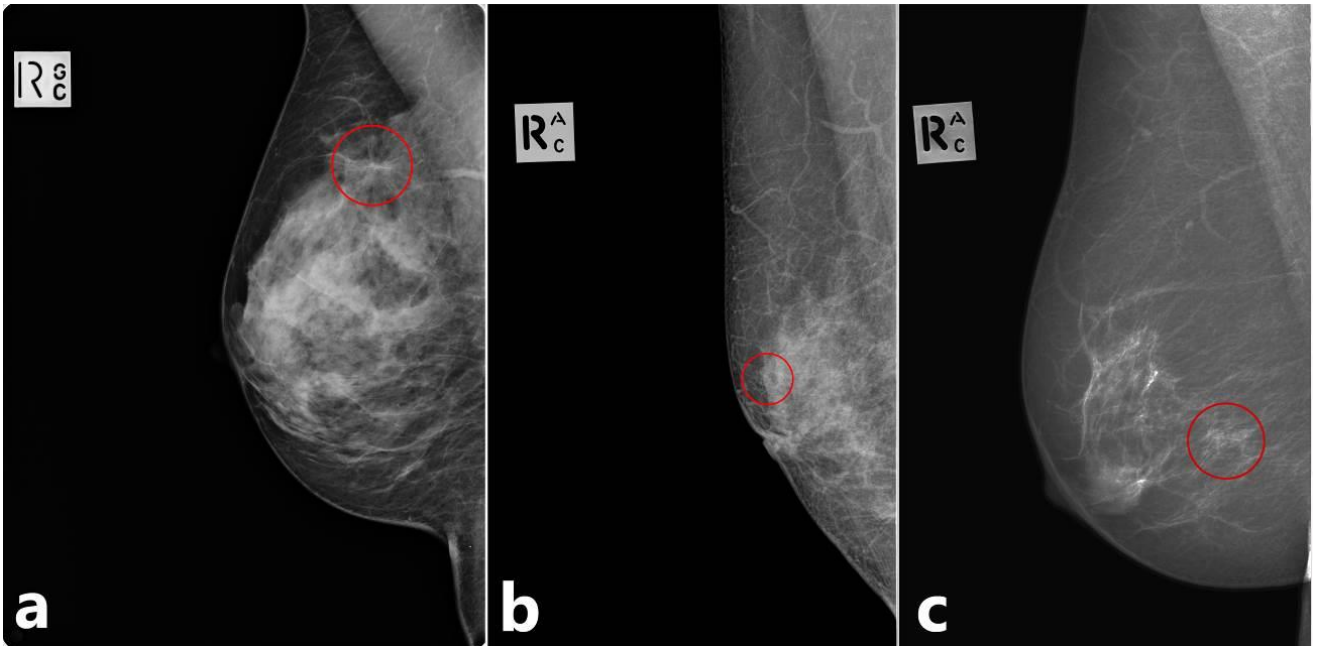


Figure 1 Examples of lesion features present in this study a) stellate mass; b) mixed features of calcification and architectural distortion (AD); c) non-specific density (NSD)

Results

For each cancer type, our results demonstrated that readers have higher sensitivity (0.80) and ROC AUC (0.84) when reading at the free recall (mean recall rate of 25.6%) condition as compared to 15% (0.65 and 0.79 respectively) and 10% (0.55 and 0.75 respectively).

Changes in sensitivity at 15% and 10% recall rates were compared against the baseline free recall using ROC AUC. There was a significant decrease in sensitivity at 15% and 10% for NSD (H=6.27, P=0.04 and H=7.35, P=0.03 respectively), and for mixed features (H=9.97, P=0.01 and H=6.50, P=0.04 respectively). There was no significant difference in sensitivity at 15% and 10% for stellate lesions (H=3.43, P=0.18 and H=1.23, P=0.54 respectively), and AD (H=0.00, P=1.00 and H=1.23, P=0.37 respectively) (Table 2). An average false positive rate of 0.17 (range 0.12 to 0.21) was observed for free recall, 0.08 (range 0.08 to 0.09) for 15% and 0.05 (range 0.05 to 0.06) for 10%.

Table 2 Mean values of sensitivity and receiver operating characteristic (ROC) area under the curve (AUC) of each mammographic feature at free recall, 15% and 10% recall rates

Lesion type	Mean sensitivity			P value
	Free recall	15% recall	10% recall	
Stellate	0.95	0.90	0.83	0.180
AD	0.80	0.80	0.80	1.000
NSD	0.67	0.47	0.20	0.043*
Mixed features	0.80	0.45	0.30	0.007*

	Mean ROC AUC			P value
	Free recall	15% recall	10% recall	
Stellate	0.93	0.93	0.89	0.541
AD	0.81	0.86	0.88	0.368
NSD	0.75	0.70	0.58	0.025*
Mixed features	0.79	0.68	0.62	0.039*

*Significant differences (P<0.05)

Table 3 shows an analysis of mammographic appearances of cancer cases in relation to case difficulty at free recall, 15% and 10% recall rates. Ten cancer cases were grouped as ‘lower difficulty’ cases with six of the cancers characterized with stellate masses. Cancers with NSD were the most common cancer features found among eight cases in the ‘medium difficulty’ group. The ‘higher difficulty’ cases were characterized with NSD and mixed features of calcification+AD.

In this study, cancers characterized with stellate mass features were most likely to be recalled by all five readers regardless of any recall conditions. At the 15% recall condition, cancer with mixed features of calcification+AD (for example, MJAS) showed the highest reduction in recall decisions (from 5 to 1), followed by cancers with NSD (from 4 to 1). When the recall rate was further reduced to 10%, six cancers were less likely to be recalled, with all of these cases having been recalled at ‘free recall’ and 15% recall, but missed by all readers at the 10% recall condition. NSD was found to be the most common feature that was missed by all readers at the 10% recall, followed by cancers with mixed features of calcifications+AD and stellate+NSD. It is noted that two cancers with AD features showed an increase in detection at 15% (from 3 to 5) and 10% recalls (from 3 to 4).

When considering the lesion type and mammographic breast density together, the analysis revealed most of the cancers in the lower difficulty group were in cases with low mammographic density ($\leq 50\%$ glandular, BIRADS 1 or 2) (see **Table 3**). Conversely, a greater number of cancers in the medium difficulty and higher difficulty group were located in cases with high mammographic density ($\geq 51\%$ glandular, BIRADS 3-4).

Table 3 Distribution of detection and cancer appearances (lesion type, breast density and lesion location) for each cancer in relation to case difficulty at free recall, 15% and 10% recall rates.

Case ID	Number of readers detected cancer for each reading session			Total	Lesion type	Breast density (BIRADS)
	Free recall	15%	10%			
<i>Lower difficulty</i>						
MJBL	5	5	5	15	Stellate	>75%
MJCX	5	5	5	15	Stellate	< 25%
MJDA	5	5	5	15	Stellate	51-75%
MJEA	5	5	5	15	Stellate	25-50%
MJGR	5	5	5	15	Stellate	25-50%
MJDH	5	5	5	15	Stellate	51-75%
MJCQ	5	5	4	14	NSD	51-75%
MJEG	5	4	3	12	Calcifications + AD	25-50%
MJBJ	5	3	4	12	AD	25-50%
MJHD	3	5	4	12	AD	< 25%
<i>Medium difficulty</i>						
MJHJ	4	3	3	10	Stellate	51-75%
MJAS	5	1	3	9	Calcifications + AD	>75%
MJBK	3	3	2	8	NSD	51-75%
MJCR	4	3	1	8	Stellate	25-50%
MJDU	3	3	0	6	Stellate + NSD	51-75%
MJHH	4	2	0	6	NSD	25-50%
MJHK	3	3	0	6	NSD	25-50%
MJEB	4	1	0	5	NSD	51-75%
<i>High difficulty</i>						
MJBG	3	1	0	4	Calcifications + AD	51-75%
MJCF	1	1	0	2	NSD	>75%

*AD, architectural distortion

†NSD, non-specific density

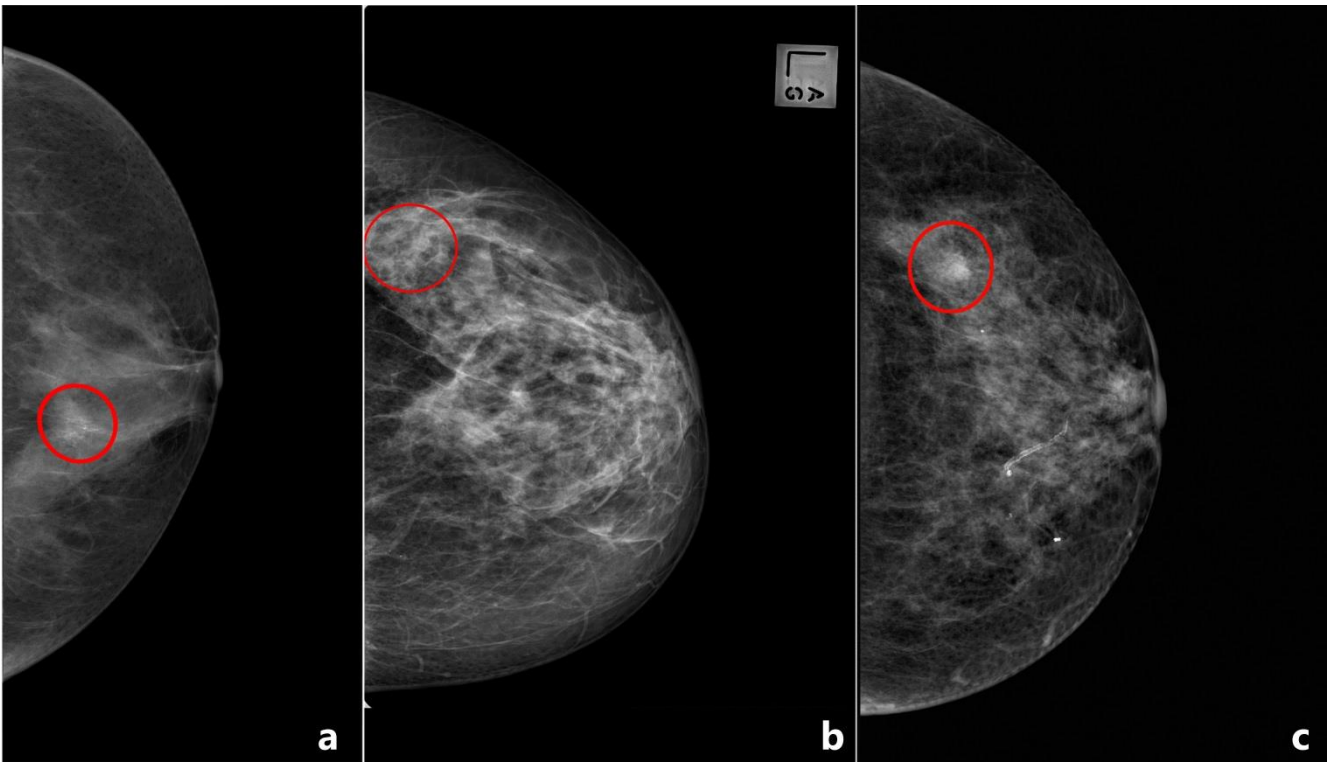


Figure 2 All three cancers above characterized with non-specific density (NSD) but with variability in lesion detectability and level of difficulty at reduced recall rates; **a)** MJCQ: lower difficulty; **b)** MJBK: medium difficulty and **c)** MJCF: higher difficulty.

Discussion

This study provides a unique perspective on the variability in mammographic interpretation by evaluating the mammographic appearances/features that were more likely to be recalled when recall rates are reduced. In agreement with our findings, a recent study by Onega et al ¹⁵ has shown that asymmetric densities contributed to low agreement for recall cases.

In this study, readers were able to recall cancers with stellate masses regardless of any recall condition. Retrospective analysis of the 20 cancer cases has demonstrated that cancers characterized with NSD were less likely to be recalled as soon as the readers were asked to reduce their recall rate to 15% and 10%. NSD was then followed by cancers with mixed features, which in this study were calcifications+AD and stellate+NSD. This is likely due to their subtle and indirect signs of malignancy. These subtle features of malignancy have been recognized in previous studies as being frequently missed by readers.^{8, 18-20} A study by Duncan et al ²⁰ when reviewing the mammographic features of 112 incidental screen-detected cancers, found greater asymmetric density and parenchymal deformity in the missed cancers than in those detected. Readers may have also interpreted irregular opacities in NSD as benign in breasts composed of tissue with irregular densities which do not warrant recall.^{21, 22}

Considering the features of cancer lesions are varied, past research has shown that cancers that present with mixed appearances (more than one mammographic feature) are more likely to be missed and readers are more susceptible to omission error when cancers display mixed mammographic appearances.²³ Our study supports this, whereby the mammographic features of stellate lesions were associated with spiked linear extensions

radiating outwards and ill-defined spicules from the central lesion which indicate the desmoplastic reaction of breast cancer into surrounding tissue.²⁴ The distinctive features of stellate lesions have a positive predictive value (PPV) of 84%-91% and are most common mammographic features of invasive breast cancer.²⁵ These features were easily recognized by our readers and were recalled for further assessment. However, when stellate features, were associated with other mammographic features, such as NSD with ill-defined borders or AD, these lesions became less suspicious and hence were not recalled at lower rates. Similarly, cancers characterized by combination features of calcifications and AD are less likely to be recalled than cancers with AD alone, although in a study by Craft et al²⁶ found up to 48% of cancers with calcifications were associated with malignancy. Calcification has also been found to have uncertain malignant potential when associated with atypical breast lesions.²⁷ It is interesting to note that we also found inconsistency in readers' decisions when recalling cancers characterized with AD. Unlike other cancer features, the fine linear structures of AD normally seen on mammograms can resemble superimposed normal breast tissue, but it also can appear as a stellate shape and an accompanying feature of other abnormalities, which turn out to be a breast cancer. A recent study on the ability of readers to detect AD by Suleiman et al²⁸ has shown that readers had greater difficulty detecting cancers with AD than other cancer features, with significantly lower sensitivity and receiver operating characteristic (ROC) area under the curve (AUC) results. With the limited number of cases that were allowed to be recalled in the conditions of 15%, and 10% recall rates, several features of AD such as trabecular thickening, which disrupt the normal breast tissue pattern, may also lead to uncertainty in readers' decision-making, resulting in recalling for further assessment.^{15, 28}

Other perceptual factors such as mammographic density may have an attractor and distractor effect that also influenced recall decisions.^{10, 29-32} Our high difficulty cases occurred in conjunction with high mammographic density. In addition, higher mammographic density may also increase the likelihood of cancers being missed in mammography.^{10, 29-32} In this study, cancers within cases of low mammographic density (BIRADS 1 and 2) were more likely to be recalled and cancers present in cases with higher mammographic density (BIRADS 3 and 4) were likely to be missed. With some limitations specific to two dimensional (2D) mammography, it is possible that lesions with subtle malignancy signs such as NSD might be obscured by dense parenchyma. This finding is concurrent with earlier findings by Bird et al¹⁹ who reported that 24% of missed cancer in screening mammography were due to higher mammographic density. Additional views such as coned compression or magnifications, which give better contrast and spatial detail on the targeted area, were not available to our readers and may have improved recognition of the lesions; however, these views are not part of a standard screening protocol. Further research in this area employing eye-tracking analysis may aid in understanding the visual search patterns of readers making their decisions under strict recall conditions. An additional area for further research may include the role of training to improve the identification of more difficult lesions, including the effect of experience upon consistent recalling of certain lesion types.

Reflecting on the clinical significance of the results of this study, the fact that the readers continued to recall stellate lesions even at reduced recall rates may be due to the biological significance of these findings. This is because stellate lesions are often recognised as highly likely to be malignant³³ but not often associated with high histologic grade.³⁴ The correlation between mammographic features and histologic grade was

evident in a study by De Nunzio et al when investigating 212 patients with invasive cancer which found that lesions presenting as stellate had significant correlation with low histologic grade, which suggests stellate as a good prognostic features and associated with reduced breast cancer mortality.³⁴ This was supported later by findings from Alexander et al, where patients with this type of lesion had better survival rate (more than 95%) as compared to other mammographic features.³³ Unlike stellate, cancer characterized with NSD was associated with high histologic grade and larger size (up to 90milimeter).³⁴ As high grade cancer is faster growing compared to low grade cancer, this may give a shorter window for this type of cancer to be detected. Such factors may have been taken into account by the radiologists in their decisions at strict recall conditions, as they may improve the survival rate of the screened women. Conversely, other lesions such as calcifications, mixed features and NSD have high likelihood of being benign lesions,³⁵ thus perhaps justifying the reduced need to recall these lesions.

This study was conducted in a laboratory environment rather than a clinical setting which may have affected the readers' reporting pattern. A previous comparison study by Gur et al³⁶ with nine experienced radiologists demonstrated higher recall rates when the reading took place in the laboratory as compared to in-clinic reading. On the other hand, in our study radiologists were "forced" to recall the most significant cases that required further assessments and were not allowed to exceed a prescribed recall rate. In some cases, although the radiologists found more cases might need to be recalled, they had to in effect 'let go' some of the cases due to the strict recall rule. In addition, it may be argued that a relatively low number of each cancer type was presented in this study. However, we believe that it has given important insights of the type of breast cancer that affect upon

reader's recall decisions. A study with a greater number of cancer types and readers will minimize biasing in results.

Conclusion

This study provides important insights into the types of cancer cases that contribute to the greatest uncertainty or are missed at low recall rates. Cancers with a stellate mass were more easily detected and are likely to continue to be recalled, even at lower recall rates. Lesions with non-specific density and mixed features were most likely to be recalled at reduced recall rates. By understanding which cancer features are likely to be missed, a dedicated training intervention can be developed to improve readers' performance when considering an optimal recall rate.

References

1. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in Interpretive Performance at Screening Mammography and Radiologists' Characteristics Associated with Accuracy. *Radiology*. 2009;253(3):641-51.
2. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International Variation in Screening Mammography Interpretations in Community-Based Programs. *Journal of the National Cancer Institute*. 2003;95(18):1384-93.
3. Yankaskas BC, Klabunde CN, Ancelle-Park R, Rennert G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? *Journal of medical screening*. 2004;11(4):187-93.
4. Ciatto S, Ambrogetti D, Bonardi R, Catarzi S, Risso G, Rosselli Del Turco M, et al. Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. *Journal of medical screening*. 2005;12(2):103-6.
5. Redondo A, Comas M, Macia F, Ferrer F, Murta-Nascimento C, Maristany MT, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *The British journal of radiology*. 2012;85(1019):1465-70.
6. Duijm LE, Louwman MW, Groenewoud JH, van de Poll-Franse LV, Fracheboud J, Coebergh JW. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. *British journal of cancer*. 2009;100(6):901-7.
7. Rawashdeh MA, Bourne RM, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses. *Academic Radiology*. 2013;20(5):576-80.
8. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Brenner RJ, Sickles EA. Analysis of 172 Subtle Findings on Prior Normal Mammograms in Women with Breast Cancer Detected at Follow-up Screening. *Radiology*. 2003;226(2):494-503.
9. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology*. 2001;219(1):192-202.
10. Boyd NFMDD, Guo HM, Martin LJP, Sun LM, Stone JM, Fishell EMDF, et al. Mammographic Density and the Risk and Detection of Breast Cancer. *The New England Journal of Medicine*. 2007;356(3):227-36.
11. Boyd NF, Martin LJ, Bronskill M, Yaffe MJ, Duric N, Minkin S. Breast tissue composition and susceptibility to breast cancer. *Journal of the National Cancer Institute*. 2010;102(16):1224-37.

12. BreastScreen Australia. National Accreditation Standards: BreastScreen Australia Quality 2008 [cited 2014 May 21]. Available from: <http://www.cancerscreening.gov.au>
13. U.S. Department Of Health and Human Services. An overview of the final regulations implementating the Mammography Quality Standards Act of 1992. Rockville, Md: U.S. Department of Health and Human Services. 1997:16-9.
14. National Health Service Breast Screening Radiologist Quality Assurance Committee. Quality assurance guidelines for radiologists. National Health Service Breast Screening Programme publication no 15 Sheffield, England: NHSBSP Publications. 1997.
15. Onega T, Smith M, Miglioretti DL, Carney PA, Geller BA, Kerlikowske K, et al. Radiologist Agreement for Mammographic Recall by Case Difficulty and Finding Type. *Journal of the American College of Radiology*. 2012;9(11):788-94.
16. National Breast Cancer Centre. Breast imaging: a guide for practice. The Royal Australian and New Zealand College of Radiologists; 2014.
17. Australian Institute of Health and Welfare. BreastScreen Australia data dictionary : version 1.1. 2015.
18. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics*. 2003;23(4):881-95.
19. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology*. 1992;184(3):613-7.
20. Duncan KA, Needham G, Gilbert FJ, Deans HE. Incident round cancers: what lessons can we learn? *Clinical Radiology*. 1998;53(1):29-32.
21. Samardar P, Paredes ESd, Grimes MM, Wilson JD. Focal Asymmetric Densities Seen at Mammography: US and Pathologic Correlation. *RadioGraphics*. 2002;22(1):19-33.
22. Kopans DB, Swann CA, White G, McCarthy KA, Hall DA, Belmonte SJ, et al. Asymmetric breast tissue. *Radiology*. 1989;171(3):639-43.
23. Peters G, Jones CM, Daniels K. Why is microcalcification missed on mammography? *Journal of medical imaging and radiation oncology*. 2013;57(1):32-7.
24. Cherel P, Becette V, Hagay C. Stellate images: anatomic and radiologic correlations. *European Journal of Radiology*. 2005;54(1):37-54.
25. Maja Podkrajšek JŽ, Marko Hočevár. What is the most common mammographic appearance of T1a and T1b invasive breast cancer? *Radiology and Oncology*. 2008;42(4):173-80.
26. Craft M, Bicknell AM, Hazan GJ, Flegg KM. Microcalcifications Detected as an Abnormality on Screening Mammography: Outcomes and Followup over a Five-Year Period. *International Journal of Breast Cancer*. 2013;2013:7.

27. Houssami N, Ciatto S, Bilous M, Vezzosi V, Bianchi S. Borderline breast core needle histology: predictive values for malignancy in lesions of uncertain malignant potential (B3). *British journal of cancer*. 2007;96(8):1253-7.
28. Suleiman WI, McEntee MF, Lewis SJ, Rawashdeh MA, Georgian-Smith D, Heard R, et al. In the digital era, architectural distortion remains a challenging radiological task. *Clinical Radiology*. 2016;71(1):e35-40.
29. Al Mousa DS, Brennan PC, Ryan EA, Lee WB, Tan J, Mello-Thoms C. How mammographic breast density affects radiologists' visual search patterns. *Academic Radiology*. 2014;21(11):1386-93.
30. Al Mousa DS, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? *Clinical Radiology*. 2014;69(4):333-41.
31. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast Density as a Predictor of Mammographic Detection: Comparison of Interval- and Screen-Detected Cancers. *Journal of the National Cancer Institute*. 2000;92(13):1081-7.
32. Lehman CD, White E, Peacock S, Drucker MJ, Urban N. Effect of age and breast density on screening mammograms with false-positive findings. *American Journal of Roentgenology*. 1999;173(6):1651-5.
33. Alexander MC, Yankaskas BC, Biesemier KW. Association of Stellate Mammographic Pattern with Survival in Small Invasive Breast Tumors. *American Journal of Roentgenology*. 2006;187(1):29-37.
34. De Nunzio MC, Evans AJ, Pinder SE, Davidson I, Wilson ARM, Yeoman LJ, et al. Correlations between the mammographic features of screen detected invasive breast cancer and pathological prognostic factors. *The Breast*. 1997;6(3):146-9.
35. Domingo L, Romero A, Blanch J, Salas D, Sánchez M, Rodríguez-Arana A, et al. Clinical and radiological features of breast tumors according to history of false-positive results in mammography screening. *Cancer Epidemiology*. 2013;37(5):660-5.
36. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*. 2008;249(1):47-53.

CHAPTER 7

DISCUSSION AND CONCLUSION

Background

Mammography reporting is challenging due to the heterogeneity of the breast parenchyma and subtlety of breast lesions. Some lesions may be difficult to categorize and patients having such lesions are often recalled for additional examination(s). However, the majority of the recalled cases tend to be benign or negative (1). To reduce unnecessary recalls, costs and patient stress associated with additional imaging, different countries and screening programs have guidelines for recalling clients and have set recall rates recommendations (2-6).

The literature review (Chapter 2) explored the multiple factors that affect the recall rates results of screening mammography. Large variations in the recall policies are evident within screening mammography programs across different countries (1, 7). Differing methods of screening (imaging technologies), the characteristics of the screened population, practices among radiologists or breast readers and health policies in the respective countries are all suggested as confounding factors responsible for the variability in the reported recall rates in prior studies (8). Furthermore, there is a paucity of information on how varying recall rates affect the performance of radiologists and their decision to recall abnormal cases in screening mammography. This paucity of evidence warrants further investigation. Therefore, this thesis explored the impact of setting different target recall rates on the performance of breast radiologists (Chapter 5). It also assessed the mammographic characteristics of breast lesions that influenced breast radiologists' decision to recall (in Chapter 6). This chapter (Chapter 7) integrates and analyses the outcomes of the two studies conducted.

The outcome of this work is also compared to published literature on this subject and the implications of recall rate recommendations upon the performance of breast radiologists in breast cancer detection are discussed. Secondly, the findings of the types and mammographic characteristics of cancers that are more or less likely to be recalled are summarized and discussed. Finally, the clinical implications of these findings to clinicians and policy-makers are also considered. The limitations of the thesis and recommendations for future studies are suggested.

Discussion of Study 1: To investigate the effect of reduced recall rates on breast radiologists' performance using receiver operating characteristic (ROC) and Jackknife free response operating characteristic (JAFROC) analysis.

False positive findings have psychological consequences for clients, such as anxiety, and increase the cost of breast screening programs (9-13). It is expected that by limiting the number of recalled cases in screening programs, the number of false positive outcomes are reduced. Data produced from this thesis shows that the false positive rate of breast radiologists was significantly reduced by 3-9% when recall rates were lowered. On the other hand, specificity increased at lower recall rates (0.95 at 10% and 0.92 at 15%). This was achieved at the expense of lower sensitivity. The sensitivity decreased substantially from an average of 0.80 at free recall to 0.65 at 15% recall and 0.55 at 10%. Sensitivity was higher at free recall compared to 15% (0.80 vs 0.65). In other words, breast radiologists were able to detect more cancer lesions when they were allowed to work at free recall (or without a set recall rate target).

Target recall rates have the potential to impact upon breast radiologists' behaviour. In order not to exceed the target recall rate stipulated in the guidelines, breast radiologists may need to adopt a stricter reporting criterion. Breast radiologists may choose to overlook lesions they consider less suspicious whilst recalling those they perceive to be more indicative of cancer. As reported by Carney et al in their study when analysing the impact of a web-based educational program upon the performance of community US radiologists with higher recall rates, 72.3% of participating radiologists would change their routine clinical behaviour in order to meet their recall rates and they did this by re-reviewing cases that had a smaller likelihood of being cancer (14).

Some breast radiologists take into account a woman's risk of breast cancer such as breast density, age, presentation of symptoms, the use of HRT prior to the time of screening, family history or previous biopsy history in their decision making process. However, these details were not provided to the participants in this laboratory study, so it is difficult to assess the impact of such information on their recall decision and we therefore need to conclude that the decisions made by our breast radiologists were made exclusively based on the mammographic images (appearances or characteristics of breast lesion).

Mammographic characteristics such as dense tissue and the subtlety of breast lesions have been demonstrated to impact on breast radiologists' perception and contribute to missed cancers (15, 16). Even when lesions are visible, the analysis of the lesion depends on the impression created in the mind of the radiologist (17) which is very subjective. Another challenge in the interpretation task is if the lesion is rarely perceived by the radiologists. A visual search study by Wolfe et al found that lesions that occur less frequently in a screening scenario affected lesion detection and led to low performance (18). Thus lesion identification and detection by radiologists may be influenced by such internal variables (visual perception, subjectivity of human readers) which may be responsible for the differences in their decision at prescribed target recall rates.

It should be noted that decision making for further assessment in mammography incorporates the interaction between visual perception and clinical judgement, both of which can be influenced by such factors as training and experience. Each breast radiologist employs different approaches in gathering and using information when making a decision to recall (19). It has been shown in a previous study that breast radiologists relied on their previous knowledge and experience when they perceived the presence of a lesion and

decided upon its cancer characteristics (20). In a study by Nodine et al, experienced breast radiologists generally demonstrated a higher performance than inexperienced residents when reading mammograms (20). This is because the decision-making process by experienced breast radiologists is based on their prior information from previous training and experiences, which is known as the memory-cueing hypothesis. Thus experienced breast radiologists search their memory for cues to recognize abnormalities that assist them in making a decision (21). In contrast, resident doctors and junior radiologists with limited experience tend to engage in heuristic decision-making (21).

Due to having a very homogenous group of readers in this study (with mean experience of 17 years and mean of 11,900 mammogram readings per year), the impact of these characteristics on performance at various recall rates in this thesis has not been evaluated statistically. Nevertheless, the readers did work in different practice settings and may have learned their skills under different mentorship conditions. Rothschild et al (2013) found practice site/location can significantly affect a radiologist's recall rate in screening mammography in the US, as recall rates were higher for radiologists practising in hospitals than community-based screening (22). Cumulative clinical knowledge acquired from various training and mentorships does make a difference in clinical decision making and thus may lead to variability in breast radiologists' decisions. This can be confirmed by the findings from this thesis that showed individual recall rates that ranged from 18.5% to 34.0% when reading at the free recall condition. Therefore, it is logical that this variation in recall rates in the current study may be due to the breast radiologists acting according to their own individual practices and expertise.

It is also worth noting that some breast radiologists are more risk averse than others. The differences in risk aversion may have been responsible for inter-reader

differences in decisions to recall in this work. All readers were aware of the laboratory nature of the study however real clinical decision making is more challenging when there is potential for litigation and medical diagnostic errors such as missed cancer. This may explain why the recall rates in the United States (US) are higher (currently around 15% of women are recalled) than breast screening in Australia and other national breast screening programs (23), as there is a documented history of litigation (24). For this reason, a breast radiologist may feel subtle pressure to recall a mammographic case instead of making a return to screening decision alongside the knowledge that a missed cancer due to error is likely to have a significant implication on patient's survival. Previous studies demonstrated that such errors have direct impact on physicians emotionally (25, 26). Therefore, heightened concerns about missing a cancer may be a key reason why the radiologists recall more cases for further assessment.

Another interesting aspect found in this study is the individual recall rates demonstrated were higher than recommended by Australian clinical practice (10%), ranging from 18.5% to 34.0%. However, when the recall rates were reduced to 15% and 10%, a substantial decrease in sensitivity was observed. A possible explanation for this was that the breast radiologists were aware the reading task was being performed in the laboratory setting where there were no effects on patient care or cost associated with further assessments. A previous study by Gur and colleagues when assessing the performance of nine board-certified American radiologists has found that radiologists have lower sensitivity when reading mammograms in the laboratory as opposed to in the clinic (0.89 vs 0.92 respectively) and recall more cases for further assessment (27). Furthermore, the radiologists may also presume the test set was enriched with abnormal cases as well as having the notion of being "tested" as individuals, perhaps resulting in biased responses

(28). Therefore, the radiologists may have altered their routine clinical behaviour and acted in a different manner from the way they would normally do in a clinic setting.

The experimental set up for the reading environment in this study was designed to be as authentic as possible to the clinical setting in BreastScreen NSW. Although it is argued that the radiologist's behaviour in a laboratory setting might be different in the real clinical setting, a study by Soh et al (29) demonstrated significant levels of reader agreement when reading the same mammographic cases in these two reading conditions: clinical and laboratory environment, with Kendall's coefficient of concordance (Kendall's W) varying from an acceptable to a good level (0.69 to 0.72), which supports the applicability of the work in this thesis.

When reading in the free recall condition, where the breast radiologists had an opportunity to recall more mammographic cases that they perceived as suspicious, a higher cancer detection rate was demonstrated. Hence the work from this thesis concurs with findings of a number of previous studies (23, 30, 31) demonstrating that higher recall rates improve the sensitivity of breast radiologists. However, free recall was associated with high false positive rates and this would likely hamper the effectiveness of a breast screening program. To ensure that a breast screening program is effective and successful, screening is not merely dependent on the number of cancers detected but also in reducing the number of unnecessary recall decisions among screened women, as false-positive recalls are associated with psychological consequences and an economic burden (9, 10, 32-35). One must weigh the advantages of early detection against the disadvantages of false positive errors, and just as detection of more cancers may improve the survival rate and lower the mortality from breast cancer, a high false positive rate will reduce confidence, may lessen participation and likely be unsustainable financially to maintain.

Another interesting point to discuss is the possible effect on radiologists' reporting behaviours by the practice of double reading in BreastScreen Australia (BSA), which provides some context for the study. Although this is not part of this study, anecdotally, it was noted that this strategy may influence breast radiologists' behaviour in mammography screen reporting. Giving verbal feedback after the reading sessions in the current study, the participating breast radiologists spoke of their experience of being part of this double reading strategy context. In acknowledging the reading protocol in this study was set as a single reading, all our participating breast radiologists have worked in some capacity for BSA, where the double-reading strategy is applied. Of the five breast radiologists, the minimum number of years of working in breast screening was nine, therefore, it is possible that they may have been engrained into a double reading strategy/clinical work practice, and this may be difficult to diffuse. Knowing that in the clinical practice the mammographic images will be read by an additional blinded reader before or after them, the breast radiologists may have regarded the system of double reading as a safety net or buffer against error when they work clinically. Previous studies have demonstrated that breast radiologists with an established practice pattern have difficulty in changing their reporting behaviour (36, 37).

Discussion of Study 2: To assess which types of mammographic appearances of breast cancer are more likely to be missed when breast radiologists read at lower recall rates.

The accuracy of radiological image reporting requires efficient search, perception and decision-making skills. Even when lesions are fixated and perceived, the categorization of such lesions as benign or malignant depends on their mammographic appearance (38-40). Whilst the features of some lesions can be more easily recognized, others present with features that are very subtle or non-specific (such as nonspecific density, NSD) for malignancy. Subtle mammographic features of breast lesions and NSD may create decision-making challenges for the breast radiologists and may be overlooked or dismissed at lower recall rates (41). The work in chapter 6 examined which types of mammographic appearances of breast cancer are most likely to be missed when breast radiologists read at lower recall rates. Findings of the work showed a significant reduction in sensitivity for cancer lesions presenting as non-specific density ($P=0.04$) and mixed features ($P=0.01$) at 15% recall rates but no difference in the detection of stellate lesions ($P=0.18$) and architectural distortion (AD) ($P=1.00$). Stellate lesions were also the most commonly recalled lesions by the breast radiologists, whilst the recall rates of lesions with mixed features (calcification+AD) and non-specific densities (NSD) were significantly reduced at 10% recall condition.

According to Mello-Thoms et al (2006) the perception of a lesion depends on its visibility and background parenchymal changes (40). Stellate lesions have very typical features and create parenchymal changes that distinguish them from their background. They also have spiked linear extensions radiating outwards and spicules that can be easily recognized by radiologists (42). Therefore, the high sensitivity of the radiologist cohort for

this type of lesions in this work may be due to their conspicuity, typical mammographic appearance and associated parenchymal perturbations. Furthermore, according to the Breast Imaging Reporting and Data Systems (BI-RADS) classification scheme, stellate lesions are almost always malignant (43) and commonly present less than 1.5 cm in size (44). This type of lesion is often not associated with a high histologic grade in breast cancer (45) and is usually curable with early treatment.

When Tabar and colleagues (2004) investigated the correlation of mammographic features of small invasive breast carcinomas (measuring 1-14mm) and long-term prognosis, they found that patients with stellate lesions had a survival rate more than 90% (46). A later study by Alexander et al (2006) investigating the favourable outcomes of 201 patients diagnosed with invasive breast cancer also demonstrated comparable findings (44), which suggests stellate lesions are a reliable indicator of good prognostic features (low histologic grade, high survival rate) in clinical decision-making. Therefore, our radiologists may have considered the clinical significance of the findings in their decisions to recall. This may explain the higher recall rates of the stellate lesion type of the current work.

Conversely, lesions presenting with NSD demonstrated low sensitivity at the low recall rate (15% and 10%) conditions. Radiologists also demonstrated a low agreement in their decision to recall when lesions were associated with NSD (also called asymmetric density) (30). Although this lesion appearance sometimes precedes malignancy with a high histologic grade (31), an NSD is often regarded as benign rather than malignant (47). This may be due to the mammographic appearance of NSD that often mimics normal parenchymal heterogeneity with poorly defined characteristics of density (32). These NSDs are usually composed mostly of densities that are difficult to distinguish from

fibroglandular tissue in the breast. Unlike stellate lesions, the mammographic features of NSD often present as indirect and subtle signs of breast cancer. The indirect and subtle signs are those related to tissue reaction in the area of breast cancer (more than signs from the cancer itself) and often associated with other benign signs (48). Furthermore, the proportion of this type of lesion found in invasive breast cancer cases reported was very small as opposed to other mammographic features (44, 46, 49). This suggests the radiologists have relatively little exposure to cancers with this mammographic appearance.

In a previous visual search study by Wolfe et al it was found that readers would change their decision criteria at low prevalence which led to miss error and lower performance (50). Thus when a target recall rate is set, features that mimic benign conditions and lack evidence of malignancy may have a high likelihood to be overlooked and dismissed. This can be either due to an uncommon cancer presentation or a variability in the knowledge of the radiologists (24, 33), even though the lesion is visible on a mammogram and perceived by some radiologists.

The same reason may also explain the inconsistency in radiologists' decision making when detecting lesions characterized with AD as observed in this study. AD is often associated with many benign conditions such as radial scars, sclerosing adenosis, fat necrosis, previous surgery, trauma and infection, thus perhaps justifying the reduced need to recall these lesions (51-53). A 10% reduction in sensitivity was reported by van Breest and colleagues (2016) for women with a previous history of surgery mainly due to the scarring of breast parenchyma which mimics AD in mammograms (51). Even if AD is perceived, it is a difficult task for a radiologist to make a decision whether it is benign or malignant. A study by Suleiman et al demonstrated a significantly lower sensitivity and lower ROC AUC for AD lesions compared with the detection of other lesions types (52).

Due to this array of confounding conditions, AD may be easy to overlook and misinterpreted as being benign, with AD demonstrated to constitute to approximately 18% of missed cancers in a study reported by Burrell et al (53).

It is also noteworthy that another perceptual factor such as mammographic breast density affects cancer detection in 2-dimensional mammography. Data from this thesis demonstrates that cancer lesions present in mammograms with higher mammographic breast density (BIRADS 3 and 4) were likely to be missed when reading at limited recall rates. Due to the similarity of mammographic appearance between fibroglandular tissues and breast lesions, high mammographic density can obscure the breast lesions with subtle malignancy signs and increase the risk of interval cancer four times higher than breasts with low mammographic density (54).

Findings from this thesis suggest the role of mammographic features of breast lesions are a predictor of breast radiologists' decision making in recalling for further assessment. Cancer with indirect and subtle signs of malignancy may challenge breast radiologists when making their decision to recall at low rates. In this study, NSD was the type of breast lesion that had the highest likelihood of being unrecalled and missed.

Implications of the findings this thesis

Quality guidelines for breast cancer screening programs are intended to utilize data to improve quality outcomes and to help clinicians understand the gap between their own performance and national targets (2, 3, 55, 56). In this study, a significant reduction in the sensitivity of breast radiologists was evident when the target recall rates were reduced from free recall to 15% and 10%. It could be suggested as an implication from this thesis that reducing recall rates is associated with decreased performance in cancer detection. However, clinical decisions must be a balance of risk and benefit. Although a high recall rate may result in an increase in sensitivity as demonstrated at the free recall condition in this thesis, it also increases the number of FP results. Thus, this will raise great concern not only among the women attending screening (5, 13) but also with policy makers as this may in turn increase the cost associated with the additional radiological examinations. It has been reported that the high cost related to false positive decisions causes an enormous burden to services and individuals and it is estimated that more than USD \$100 million per year is paid for false positive outcomes in the most populous state of the United States, California (57), however, no comparable figure was found for the Australian population. As a comparative figure, it was estimated that 24% of the health expenditure by the Australian Government was allocated to breast cancer, with AUD \$118 million spent on screening mammography services through the BreastScreen Australia Program (58). Interestingly, an increment of 32% in breast cancer expenditure for women was reported in Australia from \$252 in 2000-2001 to \$331 for the 2004-2005 financial year (58) and largely this growth contributed to out-of-hospital medical expenses such as visits to general practitioners and specialists, as well as pathology and imaging services (58).

Lowering recall rates in this study decreased the ability of breast radiologists to correctly detect cancer. The current work confirmed that benign-appearing soft tissue findings and indirect signs of malignancy such as non-specific density (NSD) had a higher probability not to be recalled at low recall rates as compared to other mammographic features. The detection of NSD lesions in mammograms is particularly important as this lesion is associated with high histologic grade cancer. Delaying in detecting such high malignancy cancer can decrease the survival rate of screened women.

A strategy that has been established to assist in this regard is through educational feedback mechanism such as the BreastScreen Reader Assessment Strategy (BREAST) (59). BREAST provides opportunities for breast radiologists to assess a test bank of mammographic cases and feedback is provided at the end of the reading session that provides a learning/training platform for breast radiologists through self-directed educational activities (59). Through this mechanism, researchers and clinicians should now be looking at an improvement upon educational strategies that can use the current findings to help clinicians understand the nature and type of positive lesions that are likely for them to give away at a low recall rate. Thus, training interventions can be tailored and personalised to improve the diagnosis of breast cancer types that pose detection challenges. Conversely, a major challenge in breast screening is obtaining feedback on truth and the lesion type of a recalled disease, which is often unknown prior to biopsy and pathology correlation. Therefore, an educational feedback mechanism is essential for breast radiologists, particularly in Australia, where many people (both breast radiologists and clients) live remotely.

Another possible mechanism to improve breast radiologists' performance is by optimal pairing of breast radiologists in a double-reading environment. Although double reading in mammography is known to improve the efficacy of breast screening, there is little knowledge about how to optimise the process of pairing breast radiologists who complement each other in terms of skill mix. Currently in Australia, breast radiologists are randomly paired without considering the breast radiologists performance characteristics at an individual level. Both readers may or may not have a difficulty in recognizing and detecting some particular type of breast lesions. By pairing them randomly, there may be a possibility the breast radiologists will miss some breast lesions and this may hamper the efficacy of double reading strategy. By studying the characteristics and type of breast lesions that have low detectability and low recall rate for each reader based on evidence, optimal pairing may result in a greater number of malignant being detected and recalled for further assessment.

Nevertheless, the experience of readers reading the mammographic images does change over time and this may challenge and limit the implementation of this strategy in the practice. This is because people may change their work behaviour, acquire knowledge or have changed knowledge as the results of their work practices. For example, if they have time away from reading mammograms for 6 months or more, this may limit their resources of working memory and affect their performance in reading mammographic images. Conversely, readers who are working through a large number of mammographic cases form a database of knowledge in their working memory that aids in decision-making. Another potential challenge that needs considering is the possibility all the radiologists who were employed in one service may not be able to be paired together due to some particular/logistical reasons and although the ideal is for radiologists to pair with someone

who complements their skill, this may not always be possible and the pairing strategy may be difficult to monitor consistently.

Strengths of the thesis

In this laboratory-based reader performance study, an experiment was conducted in a controlled environment which allowed for minimization of the confounding factors such as viewing conditions and establishment of the effects of recall rates. The recall rate was treated as a primary indicator (independent variable) for assessing the performance of breast radiologists in screening mammography.

Additionally, the work in this thesis has explored performance through a methodology that assesses the breast radiologists' ability to correctly locate lesions. The thesis answers an important question regarding whether setting a target recall rate affects behaviour of breast radiologists when interpreting a mammogram with regards to sensitivity, case sensitivity, lesion location sensitivity, specificity, receiver operating characteristic (ROC) area under the curve (AUC) and Jackknife free-response receiver operating characteristic (JAFROC) figure of merit (FOM).

Data provided from JAFROC FOM analysis and lesion location from the work of this thesis is beneficial to clinicians as it does reflect actual behaviour of breast radiologists interpreting mammographic images. With greater statistical power over traditional methodologies reporting observer's performance, such as ROC (60, 61), the accuracy of breast radiologists to correctly identify each lesion on mammograms can be identified.

Limitations of the studies

Some caution must be in place when interpreting data produced from this thesis due to the following limitations.

Firstly, a relatively small number of readers (n=5) were used for the study. The reasonably large number of cases used in the test set (n=200) and the number of reads required for three reading sessions by each reader made it difficult to recruit a larger reader cohort. Although some breast radiologists initially showed their interest in participating, they could not commit to performing the repetitive reading task at the designed laboratory setting due to their tight clinical schedules. This limited the number of readers and this may have been a potential hindrance to explore more significant changes.

Secondly, there were relatively small numbers of each lesion type in the sample. This may have minimized our ability to demonstrate the radiologists' performance at reduced recall rates for each lesion type, thus limiting generalizability of the results. Furthermore, the fact that the number of cases was chosen from a real population screening program audit makes it impossible to control and may lead to bias. However, the results provided from this thesis should inform further research activities.

Thirdly, information on positive recall lesion type (true and false) cannot be provided at any stage of this study as it was not part of the study protocol; where the readers did not require to report the lesion type. This was done in order to keep as close with clinical practice where breast radiologists do not have to describe the lesion type they

wish to recall. Collection of this information in a future research project would be of great interest.

Fourthly, the test set was compromised of non-digital images (digitised images) and digital images. This may have potentially contributed to the variations in individual case image quality and hence may have affected radiologists' accuracy when reading the mammograms. However, this was done in order to keep as close with clinical practice where the radiologists cannot choose what type of mammographic images that they have to read. Some prior cases may be original film-screen images that have been digitised, however all cases passed the BreastScreen NSW image quality analysis.

Recommendation for future work

Closely related to my project, these are some extensions that would help clarify my results:

- 1. Larger sample size:** A larger study with a diversity of reader experience and case load would allow clarification of the importance of experience when adhering to specific recall rates. It would be interesting to explore how readers with a range of experience adhere to different recall rates and also how varying recall rates would affect the readers' performance, and may yield some important relationships between lesion location sensitivity and experience.
- 2. Availability of prior images:** It would be of great interest to provide prior images (where available) on the mammographic cases within the test set and explore how this may affect breast radiologists' decision-making and performance in future work. This was a direct suggestion from the participating breast radiologists in the current work. Although prior images were deliberately omitted from the test set in order to simulate a first screening read, previous studies demonstrated that availability of prior images had an impact on breast radiologists' performance and assisted in their clinical decision making (62, 63). Thus, exploring such an effect will establish how the availability of prior images can impact breast radiologists' decision to recall, although this would result in an increase in reading time for each session, and may limit the interest of breast radiologists to participate in lengthy recall studies.

3. **Visual search study:** A study with eye-tracking analysis should also be employed in future work as this can offer an explanation for the findings and a more comprehensive understanding about how a lesion is missed and how the decisions are made. This is particularly important as the work in Chapter 6 of the thesis has demonstrated that detection rate for certain type of breast lesions (such as non-specific density) are unlikely to be recalled when the recall rates decreases, however without eye tracking analysis, it cannot be known for certain if the breast radiologists fixated on the lesions or not.

4. **False positive recall decisions analysis:** Further analysis should be done on data gathered for false positive decisions to identify the nature of the features that were marked for recall but were not cancer. The work in this thesis did not record the mammographic characteristics of false positive recalls. Information on the false positive recall rates, the false positive lesion type and whether the same false positive lesions were recalled by multiple readers may benefit clinicians when considering an optimal recall rate in screening mammography. Study of these false positive recall decisions may contribute to the body of knowledge as to why some normal mammographic images are difficult to interpret and hard to recognize. The outcome of this analysis may suggest ways to reduce the false positive recall rate and improve specificity in breast screening.

Conclusion

Findings provided in this thesis suggest that reducing recall rates through recommended target recall rates may reduce the false positive (FP) rate in screening as the specificity increases; however, this could also have direct effect on the early detection of breast cancer as reduced recall rates may result in a corresponding reduction in cancer detection. Detection of some types of breast cancer may be delayed and survival outcomes decrease with a reduced recall rate. Further work from this thesis also demonstrates the appearances of subtle and indirect signs of malignancy such as non-specific density (NSD) appear to be strongly related to reducing detection of cancer at low recall rates. While cancer with obvious and direct signs of malignancy such as stellate masses has a higher detectability regardless any recall rates. The evidence produced in this thesis provides insights about how to improve the performance of radiologists in breast screening mammography through appreciating the multifactorial nature of screening. Hence, this may inform the future work in designing key educational strategies towards optimizing diagnostic efficacy.

Reference

1. Australian Institute of Health and Welfare. BreastScreen Australia monitoring report 2012–2013. Canberra: AIHW, 2015 Cancer series no 95 Contract No.: CAN 93.
2. BreastScreen Australia. National Accreditation Standards: BreastScreen Australia Quality 2008 [cited 2014 May 21]. Available from: <http://www.cancerscreening.gov.au>
3. National Health & Medical Research Council. Clinical practice guidelines: management early breast cancer. Second ed. Canberra: NHMRC; 2001.
4. Feig SA, D'Orsi CJ, Hendrick RE, Jackson VP, Kopans DB, Monsees B, et al. American College of Radiology guidelines for breast cancer screening. American Journal of Roentgenology. 1998;171(1):29-33.
5. Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2008;19(4):614-22.
6. National Health Service Breast Screening Radiologist Quality Assurance Committee. Quality assurance guidelines for radiologists. National Health Service Breast Screening Programme publication no 15 Sheffield, England: NHSBSP Publications. 1997.
7. Australian Institute of Health and Welfare. BreastScreen Australia monitoring report 2010-2011. Canberra: AIHW, 2013 Cancer series no.77 Contract No.: CAN 74.
8. Mohd Norsuddin N, Reed W, Mello-Thoms C, Lewis SJ. Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography. 2015;21(4):334-41.
9. Lafata JE, Simpkins J, Lamerato L, Poisson L, Divine G, Johnson CC. The economic impact of false-positive cancer screens. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2004;13(12):2126-32.
10. Maxwell AJ, Beattie C, Lavelle J, Lyburn I, Sinnatamby R, Garnett S, et al. The effect of false positive breast screening examinations on subsequent attendance: retrospective cohort study. Journal of medical screening. 2013;20(2):91-8.
11. Burman ML, Taplin SH, Herta DF, Elmore JG. Effect of False-Positive Mammograms on Interval Breast Cancer Screening in a Health Maintenance Organization. Annals of Internal Medicine. 1999;131(1):1-6.
12. Sim MJ, Siva SP, Ramli IS, Fritschi L, Tresham J, Wylie EJ. Effect of false-positive screening mammograms on rescreening in Western Australia. The Medical journal of Australia. 2012;196(11):693-5.

13. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *Journal of medical screening*. 2012;19(suppl 1):57-66.
14. Carney PA, Aiello Bowles EJ, Sickles EA, Geller BM, Feig SA, Jackson S, et al. Using a Tailored Web-based Intervention to Set Goals to Reduce Unnecessary Recall. *Academic Radiology*. 2011;18(4):495-503.
15. Rawashdeh MA, Lewis SJ, Lee W, Mello-Thoms C, Reed WM, McEntee M, et al., editors. Experience in reading digital images may decrease observer accuracy in mammography. SPIE; 2015.
16. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology*. 1992;184(3):613-7.
17. Tavakoli Taba S, Hossain L, Heard R, Brennan P, Lee W, Lewis S. Personal and Network Dynamics in Performance of Knowledge Workers: A Study of Australian Breast Radiologists. *PLOS ONE*. 2016;11(2):e0150186.
18. Wolfe JM, Horowitz TS, Kenner NM. Cognitive psychology: rare items often missed in visual searches. *Nature*. 2005;435(7041):439-40.
19. Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*. 2001;7(2):97-107.
20. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, et al. How experience and training influence mammography expertise. *Academic Radiology*. 1999;6(10):575-85.
21. Cox JR, Griggs RA. The effects of experience on performance in Wason's selection task. *Memory & Cognition*. 1982;10(5):496-502.
22. Rothschild J, Lourenco AP, Mainiero MB. Screening Mammography Recall Rate: Does Practice Site Matter? *Radiology*. 2013;269(2):348-53.
23. Yankaskas BC, Klabunde CN, Ancelle-Park R, Rennert G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? *Journal of medical screening*. 2004;11(4):187-93.
24. Berlin L. Radiologic Errors and Malpractice: A Blurry Distinction. *American Journal of Roentgenology*. 2007;189(3):517-22.
25. West CP, Huschka MM, Novotny PJ, et al. Association of perceived medical errors with resident distress and empathy: A prospective longitudinal study. *JAMA*. 2006;296(9):1071-8.
26. Elmore JG, Taplin SH, Barlow WE, Cutter GR, D'Orsi CJ, Hendrick RE, et al. Does Litigation Influence Medical Practice? The Influence of Community Radiologists' Medical Malpractice Perceptions and Experience on Screening Mammography. *Radiology*. 2005;236(1):37-46.

27. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*. 2008;249(1):47-53.
28. Soh BP, Lee W, Kench PL, Reed WM, McEntee MF, Poulos A, et al. Assessing reader performance in radiology, an imperfect science: Lessons from breast screening. *Clinical Radiology*. 2012;67(7):623-8.
29. Soh BP, Lee WB, McEntee MF, Kench PL, Reed WM, Heard R, et al. Mammography test sets: reading location and prior images do not affect group performance. *Clin Radiol*. 2014;69(4):397-402.
30. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, et al. Recall and detection rates in screening mammography. *Cancer*. 2004;100(8):1590-4.
31. Schell MJ, Yankaskas BC, Ballard-Barbash R, Qaqish BF, Barlow WE, Rosenberg RD, et al. Evidence-based Target Recall Rates for Screening Mammography. *Radiology*. 2007;243(3):681-9.
32. Bond M, Pavey T, Welch K, Cooper C, Garside R. Systematic review of the psychological consequences of false-positive screening mammograms. *Health Technology Assessment*. 2013;17(13):170.
33. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. *Ann Intern Med*. 2007;146(7):502-10.
34. Brodersen J, Siersma VD. Long-term psychosocial consequences of false-positive screening mammography. *Annals of family medicine*. 2013;11(2):106-15.
35. Castells X, Molins E, Macia F. Cumulative false positive recall rate and association with participant related factors in a population based breast cancer screening programme. *Journal of epidemiology and community health*. 2006;60(4):316-21.
36. Carney PA, Abraham L, Cook A, Feig SA, Sickles EA, Miglioretti DL, et al. Impact of an Educational Intervention Designed to Reduce Unnecessary Recall during Screening Mammography. *Academic Radiology*. 2012;19(9):1114-20.
37. Flocke SA, Litaker D. Physician Practice Patterns and Variation in the Delivery of Preventive Services. *Journal of General Internal Medicine*. 2007;22(2):191-6.
38. Mello-Thoms C, Hardesty L, Sumkin J, Ganott M, Hakim C, Britton C, et al. Effects of Lesion Conspicuity on Visual Search in Mammogram Reading¹. *Academic Radiology*. 2005;12(7):830-40.
39. Al Mousa DS, Brennan PC, Ryan EA, Lee WB, Tan J, Mello-Thoms C. How mammographic breast density affects radiologists' visual search patterns. *Academic Radiology*. 2014;21(11):1386-93.
40. Mello-Thoms C. How Does the Perception of a Lesion Influence Visual Search Strategy in Mammogram Reading? *Academic Radiology*. 2006;13(3):275-88.

41. Otten JDM, Karssemeijer N, Hendriks JHCL, Groenewoud JH, Fracheboud J, Verbeek ALM, et al. Effect of Recall Rate on Earlier Screen Detection of Breast Cancers Based on the Dutch Performance Indicators. *Journal of the National Cancer Institute*. 2005;97(10):748-54.
42. Cherel P, Becette V, Hagay C. Stellate images: anatomic and radiologic correlations. *European Journal of Radiology*. 2005;54(1):37-54.
43. American College of Radiology. ACR BI-RADS ATLAS Reporting system. . 2013. In: In, BI-RADS- Mammography 2013 [Internet]. [121-40]. Available from: <http://www.acr.org/Quality-Safety/Resources/BIRADS/Mammography>.
44. Alexander MC, Yankaskas BC, Biesemier KW. Association of Stellate Mammographic Pattern with Survival in Small Invasive Breast Tumors. *American Journal of Roentgenology*. 2006;187(1):29-37.
45. De Nunzio MC, Evans AJ, Pinder SE, Davidson I, Wilson ARM, Yeoman LJ, et al. Correlations between the mammographic features of screen detected invasive breast cancer and pathological prognostic factors. *The Breast*. 1997;6(3):146-9.
46. Tabar L, Tony Chen HH, Amy Yen MF, Tot T, Tung TH, Chen LS, et al. Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma. *Cancer*. 2004;101(8):1745-59.
47. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Brenner RJ, Sickles EA. Analysis of 172 Subtle Findings on Prior Normal Mammograms in Women with Breast Cancer Detected at Follow-up Screening. *Radiology*. 2003;226(2):494-503.
48. Martin JE, Moskowitz M, Milbrath. Breast cancer missed by mammography. *American Journal of Roentgenology*. 1979;132(5):737-9.
49. Maja Podkrajšek JŽ, Marko Hočevar. What is the most common mammographic appearance of T1a and T1b invasive breast cancer? *Radiology and Oncology*. 2008;42(4):173-80.
50. Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen*. 2007;136(4):623-38.
51. van Breest Smallenburg V, Duijm LE, Voogd AC, Jansen FH, Louwman MW. Mammographic changes resulting from benign breast surgery impair breast cancer detection at screening mammography. *Eur J Cancer*. 2012;48(14):2097-103.
52. Suleiman WI, McEntee MF, Lewis SJ, Rawashdeh MA, Georgian-Smith D, Heard R, et al. In the digital era, architectural distortion remains a challenging radiological task. *Clinical Radiology*. 2016;71(1):e35-40.
53. Burrell HC, Evans AJ, Wilson AR, Pinder SE. False-negative breast screening assessment: what lessons can we learn? *Clin Radiol*. 2001;56(5):385-8.

54. Kavanagh AM, Byrnes GB, Nickson C, Cawson JN, Giles GG, Hopper JL, et al. Using Mammographic Density to Improve Breast Cancer Screening Outcomes. *Cancer Epidemiology Biomarkers & Prevention*. 2008;17(10):2818-24.
55. Quality Assurance Guidelines For Breast Cancer Screening Radiology. NHS Cancer Screening Programmes, 2011 Publication No 59.
56. U.S. Department Of Health and Human Services. An overview of the final regulations implementating the Mammography Quality Standards Act of 1992. Rockville, Md: U.S. Department of Health and Human Services. 1997:16-9.
57. Cyrlak D. Induced costs of low-cost screening mammography. *Radiology*. 1988;168(3):661-3.
58. Australian Institute of Health and Welfare. Breast cancer in Australia: An overview. Canberra: AIHW, 2012 Cancer series No 71 Contract No.: CAN 67.
59. BREAST. BreastScreen Reader Assessment Strategy [cited 2015 24.03.2015]. Available from: <http://www.breastaustralia.com>.
60. Chakraborty DP. Recent advances in observer performance methodology: jackknife free-response ROC (JAFROC). *Radiation Protection Dosimetry*. 2005;114(1-3):26-31.
61. Chakraborty DP. Analysis of Location Specific Observer Performance Data: Validated Extensions of the Jackknife Free-Response (JAFROC) Method. *Academic Radiology*. 2006;13(10):1187-93.
62. Varela C, Karssemeijer N, Hendriks JHCL, Holland R. Use of prior mammograms in the classification of benign and malignant masses. *European Journal of Radiology*. 2005;56(2):248-55.
63. Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. *JAMA*. 1997;277(1):49-52.

APPENDICES

Appendix A: Ethics Approval



Research Integrity
Human Research Ethics Committee

Monday, 4 August 2014

Dr Sarah Lewis
Medical Imaging and Radiation Sciences; Faculty of Health Sciences
Email: sarah.lewis@sydney.edu.au

Dear Sarah

I am pleased to inform you that the Health Low Risk Subcommittee has approved your project entitled "Forced recall rates in screening mammography".

Details of the approval are as follows:

Project No.: 2014/484
Approval Date: 1 August 2014
First Annual Report Due: 1 August 2015
Authorised Personnel: Lewis Sarah; Mello-Thoms Claudia; Reed Warren; Mohd Norsuddin Norhashimah; Brennan Patrick;

Documents Approved:

Date Uploaded	Type	Document Name
27/06/2014	Participant Info Statement	PIS
13/05/2014	Participant Consent Form	Consent Form
27/06/2014	Advertisements/Flyer	Recruitment flyer
27/06/2014	Questionnaires/Surveys	Participant questionnaire of demographics

HREC approval is valid for four (4) years from the approval date stated in this letter and is granted pending the following conditions being met:

Condition/s of Approval

- Continuing compliance with the National Statement on Ethical Conduct in Research Involving Humans.
- Provision of an annual report on this research to the Human Research Ethics Committee from the approval date and at the completion of the study. Failure to submit reports will result in withdrawal of ethics approval for the project.
- All serious and unexpected adverse events should be reported to the HREC within 72 hours.
- All unforeseen events that might affect continued ethical acceptability of the project should be reported to the HREC as soon as possible.
- Any changes to the project including changes to research personnel must be approved by the HREC before the research project can proceed.

Research Integrity
Research Portfolio
Level 6, Jane Foss Russell
The University of Sydney
NSW 2006 Australia

T +61 2 8627 8111
F +61 2 8627 8177
E ro.humanethics@sydney.edu.au
sydney.edu.au

ABN 15 211 513 464
CRICOS00026A

- Note that for student research projects, a copy of this letter must be included in the candidate's thesis.

Chief Investigator / Supervisor's responsibilities:

1. You must retain copies of all signed Consent Forms (if applicable) and provide these to the HREC on request.
2. It is your responsibility to provide a copy of this letter to any internal/external granting agencies if requested.

Please do not hesitate to contact Research Integrity (Human Ethics) should you require further information or clarification.

Yours sincerely



Dr Rachel Skinner
Chair
Health Low Risk Subcommittee

This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (2007), NHMRC and Universities Australia Australian Code for the Responsible Conduct of Research (2007) and the CPMP/ICH Note for Guidance on Good Clinical Practice.

Appendix B: Participant information sheet



Discipline of Medical Radiation Sciences
Faculty of Health Sciences

ABN 15 211 513 464

DR. SARAH LEWIS
Senior Lecturer in Diagnostic Radiography

Room M217
Building M C42
Cumberland Campus
The University of Sydney
PO Box 170 Lidcombe
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 9149
Facsimile: +61 2 9351 9146
Email: sarah.lewis@sydney.edu.au
Web: <http://www.sydney.edu.au/>

Forced Recall Rate in Screening Mammography

PARTICIPANT INFORMATION STATEMENT

(1) What is this study about?

You are invited to take part in a research study about the effect of specific recall rate on breast readers' performance when interpreting mammographic images. The study compares the breast readers' observer performance in the laboratory setting with different levels of permissible recall rates over 3 reads. This proposed research is expected to provide significant results for increasing effectiveness in reader performance and diagnostic imaging technologies in laboratory setting.

You have been invited to participate in this study because you are a radiologist. This Participant Information Statement tells you about the research study. Knowing what is involved will help you decide if you want to take part in the research. Please read this sheet carefully and ask questions about anything that you don't understand or want to know more about.

Participation in this research study is voluntary. So it's up to you whether you wish to take part or not.

By giving your consent to take part in this study you are telling us that you:

- ✓ Understand what you have read
- ✓ Agree to take part in the research study as outlined below
- ✓ Agree to the use of your personal information as described.

You will be given a copy of this Participant Information Statement to keep.

(2) Who is running the study?

The study is being carried out by the following researchers:

- Dr Sarah Lewis (Senior Lecturer in Diagnostic Radiography)
- Dr Warren Reed (Senior Lecturer in Diagnostic Radiography)
- Dr Claudia Mello-Thoms (Associate Professor in Diagnostic Radiography)

- Ms Norhashimah Mohd Norsuddin (Postgraduate Research Student in Medical Radiation Sciences)

Ms Norhashimah Mohd Norsuddin is conducting this study as the basis for the degree of Doctor of Philosophy at The University of Sydney. This will take place under the supervision of Dr Sarah Lewis (Senior Lecturer in Diagnostic Radiography).

(3) What will the study involve for me?

In this study, you will be asked to interpret 200 mammographic images at three reading sessions with different recall rates in a laboratory environment. Before the start of each reading session, you will be asked to complete a general demographic questionnaire. During the reading session, you will identify and localize each digital mammogram that you consider needs to be recalled for a particular recall rate. The characteristics of each recalled image and the coordinates of all lesions found on the images will be recorded on screen. The reading sessions will be separated by a minimum of 4 months and will be conducted in the Medical Imaging Optimisation and Perception Group (MIOPeG) laboratory at the Brain and Mind Research Institute (BMRI), The University of Sydney.

(4) How much of my time will the study take?

Each reading session will take approximately 3 hours to complete the test set. Altogether the reading sessions will take 9 hours for 3 different recall rates.

(5) Who can take part in the study?

This study is open to all breast readers (radiologists or breast physicians) who are involved in mammography reporting. Participation in this study is entirely voluntary.

(6) Do I have to be in the study? Can I withdraw from the study once I've started?

Being in this study is completely voluntary and you do not have to take part. Your decision whether to participate will not affect your current or future relationship with the researchers or anyone else at the University of Sydney and BREASTSCREEN New South Wales.

If you decide to take part in the study and then change your mind later, you are free to withdraw at any time. You can do this by email to sarah.lewis@sydney.edu.au. Any collated data will also be withdrawn should you choose to withdraw from the study at a later stage.

(7) Are there any risks or costs associated with being in the study?

Aside from giving up your time, we do not expect that there will be any risks or costs associated with taking part in this study.

(8) Are there any benefits associated with being in the study?

We cannot guarantee or promise that you will receive any direct benefits from being in the study.

(9) What will happen to information about me that is collected during the study?

By providing your consent, you are agreeing to us collecting personal information about you for the purposes of this research study. Your information will only be used for the purposes outlined in this Participant Information Statement, unless you consent otherwise.

Your information will be stored securely and your identity/information will be kept strictly confidential, except as required by law. Study findings may be published, but you will not be individually identifiable in these publications.

We will keep the information we collect for this study, and we may use it in future projects. We don't know at this stage what these other projects will involve. We will seek ethical approval before using the information in these future projects.

(10) Can I tell other people about the study?

Yes, you are welcome to tell other people about the study.

(11) What if I would like further information about the study?

When you have read this information, Ms Norhashimah Mohd Norsuddin or Dr Sarah Lewis will be available to discuss it with you further and answer any questions you may have. If you would like to know more at any stage during the study, please feel free to contact Ms Norhashimah Mohd Norsuddin (Postgraduate Research Student in Medical Radiation Sciences) at nmoh5894@unisyd.edu.au or Dr Sarah Lewis (Senior Lecturer in Diagnostic Radiography) at sarah.lewis@unisyde or +61 2 9351 9149.

(12) Will I be told the results of the study?

You have a right to receive feedback about the overall results of this study. You can tell us that you wish to receive feedback by emailing us at sarah.lewis@sydney.edu.au. This feedback will be in the form of a word document summary. You will receive this feedback after the study is finished.

(13) What if I have a complaint or any concerns about the study?

Research involving humans in Australia is reviewed by an independent group of people called a Human Research Ethics Committee (HREC). The ethical aspects of this study have been approved by the HREC of the University of Sydney (Project number: 2014/484). As part of this process, we have agreed to carry out the study according to the *National Statement on Ethical Conduct in Human Research (2007)*. This statement has been developed to protect people who agree to take part in research studies.

If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the university using the details outlined below. Please quote the study title and protocol number.

The Manager, Ethics Administration, University of Sydney:

- **Telephone:** +61 2 8627 8176
- **Email:** ro.humanethics@sydney.edu.au
- **Fax:** +61 2 8627 8177 (Facsimile)

This information sheet is for you to keep

Appendix C: Participant consent form



Discipline of Medical Radiation Sciences
Faculty of Health Sciences

ABN 15 211 513 464

DR. SARAH LEWIS
Senior Lecturer in Diagnostic Radiography

Room M218
Building M C42
Cumberland Campus
The University of Sydney
PO Box 170 Lidcombe, NSW
AUSTRALIA
Telephone: +61 2 9351 9149
Facsimile: +61 2 9351 9146
Email: sarah.lewis@sydney.edu.au
Web: <http://www.sydney.edu.au/>

FORCED RECALL RATE IN SCREENING MAMMOGRAPHY

PARTICIPANT CONSENT FORM

I, [PRINT NAME], agree to take part in this research study.

In giving my consent I state that:

- ✓ I understand the purpose of the study, what I will be asked to do, and any risks/benefits involved.
- ✓ I have read the Participant Information Statement and have been able to discuss my involvement in the study with the researchers if I wished to do so.
- ✓ The researchers have answered any questions that I had about the study and I am happy with the answers.
- ✓ I understand that being in this study is completely voluntary and I do not have to take part. My decision whether to be in the study will not affect my relationship with the researchers or anyone else at the University of Sydney and BREASTSCREEN New South Wales now or in the future.
- ✓ I understand that I can withdraw from the study at any time.
- ✓ I understand that personal information about me that is collected over the course of this project will be stored securely and will only be used for purposes that I have agreed to. I understand that information about me will only be told to others with my permission, except as required by law.
- ✓ I understand that the results of this study may be published, and that publications will not contain my name or any identifiable information about me.

I consent to:

- **Being contacted about future studies** YES NO
- **Receiving feedback about my personal results** YES NO

Would you like to receive feedback about the overall results of this study?

YES NO

If you answered **YES**, please indicate your preferred form of feedback and address:

Postal: _____

Email: _____

.....
Signature

.....
PRINT name

.....
Date

Appendix D: Participant questionnaire



Discipline of Medical Radiation Sciences Faculty of Health Sciences

CHIEF INVESTIGATORS

Dr Sarah Lewis, Dr Warren Reed, A/Prof Claudia Mello-Thoms, Prof Patrick Brennan

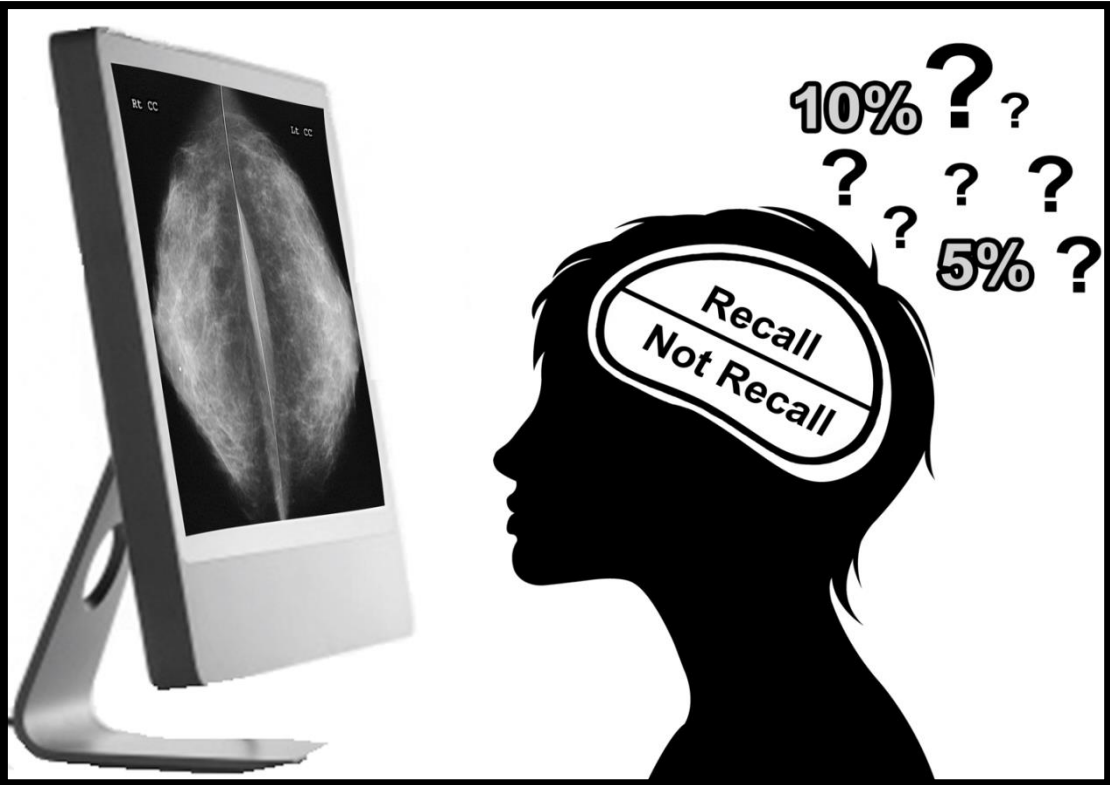
PARTICIPANT QUESTIONNAIRE

Participant Code: _____

1. How long have you specialized in breast radiology? years.
2. Do you report in other modalities of radiology? If so, what are they:
3. Number of years qualified from RANZCR years
4. On average, how many mammography cases would you read per year?
5. On average how much time (in hours) do you spend reporting mammograms per week?
..... hours per week
6. On average, how much time would you spend on a single reporting session?
..... hours
7. On average, how many cases would you report on in a single session? cases
8. If applicable, how much time do you spend reading other modalities per week (on average)? hours per week
9. What mammographic system(s) do you read from?
 Screening
 Digital
 Both
10. Do you wear glasses or contact lenses to read mammograms?
 YES: Glasses/Lenses
 No
11. Are you right- or left-handed person? Please make a selection below.
 LEFT handed
 RIGHT handed
 Ambidextrous

Comments:
.....
.....

Appendix E: 3 Minute thesis competition



Appendix F: HDR poster presentation, 2014

OPTIMISING RECALL RATES THROUGH OBSERVER PERFORMANCE IN SCREENING MAMMOGRAPHY



Norhashimah Mohd Norsuddin, Warren Reed, Claudia Mello-Thoms, Sarah Lewis
Faculty of Health Sciences, The University of Sydney, East Street, P.O. Box 170, Lidcombe, NSW 2141, Australia

Introduction & aim

False positive recall rates have been associated with significant psychological and economical costs for screened women [1]. The recall rate as linked to the cancer detection rate is not well understood [2]. International recall rates vary dramatically, from below 1% in Nordic countries to over 15% in the United States (Table 1). This variation also exist across states and territories in Australia with an average recall rate of 10.7% in 2011 for first screening mammograms [3]. The aim of this study is to investigate how altering recall rates via incremental changes in a laboratory setting will affect the performance by Australian radiologists in reading screening mammograms. The research question is: Can recall rates in mammographic screening programs be optimised?

Table 1: International comparison of recall rates, PPV and cancer detection rates for first screening mammograms in women aged 50–64 years [2]

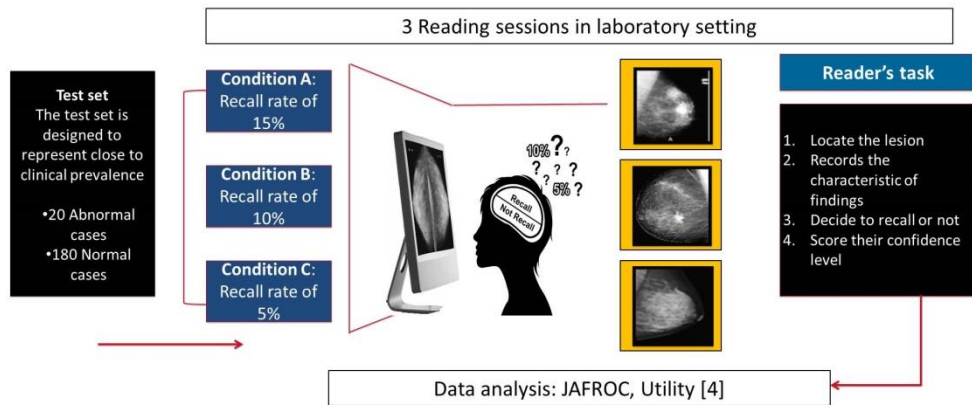
Country	Recall (%)	95% CI	Cancer detection per 1000 mammograms
NETHERLAND	1.4	1.3-1.5	5.3
SWITZERLAND	3.4	2.1-4.7	3.9
ENGLAND	8.0	7.9-8.1	6.8
CANADA	9.2	9.0-9.4	6.2
JAPAN	11.3	10.4-12.2	7.8
USA	15.1	14.5-15.7	7.5

Table 2: Recall rates for first screening round in women aged 50–69 from 1999 to 2011 in Australia [3]

	National Standard	1999	2003	2004	2005	2006	2007	2008	2009	2010	2011
Women aged 50-69 years	<10%	7.7	9.4	9.9	9.8	9.9	9.9	9.9	10.7	11.1	10.7

Methods

This study will compare the breast readers' observer performance in a laboratory setting with different levels of permissible recall rates over 3 reads at approximately 3 months apart to negate memory effect. Data collection will commence in November 2014 and will track 5-7 readers across the conditions.



Significance of study

Determination of an optimum recall rate has the potential to make early breast cancer detection programs more effective, reduce anxiety and increase confidence in mammography services and contribute a better understanding of radiology decision making in BreastScreen Australia.

Acknowledgement

National Breast Cancer Foundation (NBCF) and BREAST for the provision of digital images and display equipment.

References

[1] Hersch, J., Jansen, J., Barratt, A., Irwig, L., Houssami, N., Howard, K., ... McCaffery, K. (2013). Women's views on overdiagnosis in breast cancer screening: a qualitative study. *BMJ*, 346, f158.
 [2] Yankaskas, B. C., Klabunde, C. N., Ancelle-Park, R., Rennett, G., Wang, H., Fracheboud, J., ... Bulliard, J.-L. (2004). International comparison of performance measures for screening mammography: can it be done? *Med Screen*, 11(4), 187-193.
 [3] Australian Institute of Health and Welfare. (2013). *BreastScreen Australia monitoring report 2010-2011*. (Cancer series no.77). Retrieved from <http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129544880>.
 [4] Abbey, C. K., Gallas, B. D., Boone, J. M., Niklason, L. T., Hadjiiski, L. M., Sahiner, B., & Samuelson, F. W. (2014). Comparative Statistical Properties of Expected Utility and Area Under the ROC Curve for Laboratory Studies of Observer Performance in Screening Mammography. *Academic Radiology*, 21(4), 481-490.

Appendix G: Physics and Perception Meeting, 2014

Forced Recall Rates in Screening Mammography

Norhashimah Mohd Norsuddin

Supervisor: Sarah Lewis

Co-Supervisor: Claudia Mello-Thoms, Warren Reed



Appendix H: Presentation for Professor Craig Abbey, 2014

Forced Recall Rates in Screening Mammography :

Laboratory Based Experiment

Norhashimah Mohd Norsuddin

DISCIPLINE OF MEDICAL RADIATION SCIENCES



Appendix I: Presentation for Professor Steve Hillis, 2015

Forced Recall Rates in Screening Mammography

Norhashimah Mohd Norsuddin

Supervisor: Sarah Lewis

Co-Supervisor: Claudia Mello-Thoms, Warren Reed



Appendix J: Presentation for Jeremy Wolfe, 2015

How specific recall rates affect observer performance in screening mammography?


Norhashimah Mohd Norsuddin

Supervisor: Sarah Lewis

Co-Supervisor: Claudia Mello-Thoms, Warren Reed



Appendix K: IWDM 2016 13th International Workshop on Breast Imaging, Malmö, Sweden



THE UNIVERSITY OF SYDNEY

Lower Recall Rates Reduced Readers' Sensitivity in Screening Mammography

Norhashimah Mohd Norsuddin, Claudia Mello-Thoms, Warren Reed, Patrick C. Brennan, and Sarah Lewis

Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences, Faculty of Health Sciences, The University of Sydney, Lidcombe NSW, Australia

Introduction

Higher recall rates have been related to increased false positive decisions, causing significant psychological stress for the women screened and economical costs for the mammography screening service [1, 2].

Considering that false positive results positively correlate with high recall rates [3], some organizations have recommended specific recall policies as guidelines to evaluate the performance of readers in the screening program. Recall rates vary dramatically from below 1% in the Netherlands to over 15% in the United States [4].

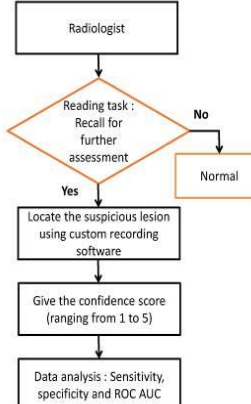
In Australia, the National Accreditation Standards (NAS) has recommended a target recall rate of 10% for the first screening and 5% for subsequent screening across the national mammography screening program known as BreastScreen. The intention of introducing recommended recall target rates is to optimize the trade-off between sensitivity and recall rates. However, extensive research surrounds the question of what is an optimum recall rate and how reader behaviour can be altered to comply with such target rates [3-6].

Methods

Test set

A single enriched test set data of 200 mammographic cases (180 normal; 20 abnormal) was used in this study, with randomisation of cases between readers and conditions.

Reading Task



Results

Table 1. Reader sensitivity, specificity, ROC AUC and median values at free call, 15% and 10% conditions

Reader	Sensitivity			Specificity			ROC AUC		
	Control (Free recall)	15%	10%	Control (Free recall)	15%	10%	Control (Free recall)	15%	10%
1	0.80	0.65	0.55	0.83	0.91	0.95	0.84	0.79	0.76
2	0.90	0.65	0.45	0.79	0.91	0.93	0.86	0.79	0.70
3	0.90	0.65	0.55	0.72	0.92	0.95	0.84	0.79	0.75
4	0.75	0.70	0.55	0.88	0.92	0.95	0.83	0.82	0.75
Median	0.85	0.65	0.55	0.81	0.92	0.95	0.84	0.79	0.75

Table 2. Kruskal-Wallis analysis and post hoc Mann-Whitney U test of sensitivity, specificity and ROC AUC

Test	Post-hoc test (Mann-Whitney U test)		
	Control VS 15%	15% VS 10%	Control VS 10%
Sensitivity	0.006	0.017	0.015
Specificity	0.007	0.019	0.018
ROC AUC	0.007	0.017	0.019

Discussion

The results from this study demonstrate that readers operating in a free recall condition (mean recall rate of 25.6%) detected more cancers compared to a reduced specified recall rate, with a higher median ROC of 0.84 and sensitivity of 0.85. When the readers were tasked with reducing their recalled cases from free recall to specific recall rates (15% and 10%), their performance declined noticeably, with a significant reduction in sensitivity (P=0.006) and ROC AUC (P=0.007).

Higher sensitivity at higher permissible recall rates as observed in our study is in agreement with research by Gur et al and Schell et al, suggesting that recalling more cases may result in a higher cancer detection rate [3, 6]. Readers demonstrated a significant improvement in their specificity at lower recall rates (P=0.007). By lowering the number of cases allowed to be recalled, readers may have needed to sacrifice some cases that they considered to be abnormal at a higher recall rate, resulting in fewer false positive decisions.

Conclusion

Setting lower target recall rates significantly reduced the readers' performance in correctly identifying breast cancers, with a significant improvement in specificity

References

1. Hørdahl, S., et al., False positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *Journal of Medical Screening*, 2012, 19, p. 31-40.
2. Hamaoka, H., et al., International comparison of performance measures for screening mammography: can it be done? *Journal of Medical Screening*, 2006, 13(6), p. 369-376.
3. Gur, D., et al., Recall rate detection rates in screening mammography. *Cancer*, 2004, 100(8), p. 1595-1596.
4. Hamaoka, H., et al., Association of Recall Rates with Sensitivity and Positive Predictive Value of Screening Mammography. *American journal of roentgenology*, 2001, 177(2), p. 243-247.
5. O'Brien, J.D., et al., Effect of Recall Rate on Patient Screen Detection of Breast Cancer based on the Dutch Performance Indicators. *Journal of the National Cancer Institute*, 2010, 102(10), p. 748-754.
6. Schell, M., et al., Evidence-based target recall rates for screening mammography. *Radiology*, 2007, 243(2), p. 681-687.

Aim of study

This study compares breast readers' performance in a laboratory setting at three levels of recall rates

Reading Conditions



1. Control (Free recall)
2. Recall rate of 15%
3. Recall rate of 10%




Figure 2: Screenshot of custom recording software interface when the radiologist marking and scoring a lesion



Figure 1: Workstation setup for participants in the MIOPeG laboratory

Appendix L: Sydney Cancer Conference 2016, Australian Technology Park, Sydney, Australia



THE UNIVERSITY OF SYDNEY

Detection of non-specific density lesions is lowered when recall rates are reduced

Norhashimah Mohd Norsuddin^{a,b}, Claudia Mello-Thoms^a, Warren Reed^a and Sarah Lewis^a

^aMedical Image Optimisation and Perception Group (MIOPeG),
Discipline of Medical Radiation Sciences, Faculty of Health Sciences, The University of Sydney, 33 Lidcombe NSW 2141, Australia
^bDiagnostic Imaging & Radiotherapy Programme, Faculty of Health Sciences, The National University of Malaysia, Kuala Lumpur, Malaysia

Introduction

Internationally, there is a large variation in the target recall rates for screening mammography programs [1-3]. However, how such variations affect radiologists' decision making in detecting breast cancers are not well understood, nor are the types of cancers that are likely to be missed when radiologists perform at low recall rates.

Aim of study

The aim of this study was to investigate which mammographic appearances of breast cancer are likely to be missed when radiologists read at reduced recall rates.

Methods

Test set

A single enriched test set data of 200 mammographic cases (180 normal; 20 abnormal) was used in this study, with randomisation of cases between readers and conditions.

Reading Task

Recall for further assessment
YES or NO

Locate the suspicious lesion

Give a confidence score (1-5)

Reading Session

Feb 2016

1st

24 months interval

2nd

24 months interval

3rd

Jan 2016

Free recall

15% recall rate

10% recall rate

Analysis of 20 abnormal cases

Level of difficulty	Total number of detection across 3 recall conditions
Lower	12-15
Medium	5-11
Higher	< 5

Results

Table 1 Percentage of total recall and non-recall decisions on cancer cases throughout all reading sessions

Mammographic appearances	Recall (%)	Not recall (%)
Stellate	77.3	22.7
AD	70.0	30.0
NSD	51.3	48.7
Calcification + AD	58.7	41.3
Stellate + NSD	36.6	63.3

Table 2 Distribution of detection and cancer appearances for each cancer in relation to case difficulty at free recall, 15% and 10% recall rates.

Case ID	Number of readers detected cancer for each reading session			Total	Lesion type	
	Free recall	15% recall	10% recall			
Lower difficulty	MJBI	5	3	4	12	AD
	MJBL	5	5	5	15	Stellate
	MJCL	5	5	4	14	NSD
	MJCC	5	5	5	15	Stellate
	MJDA	5	5	5	15	Stellate
	MJDH	5	5	5	15	Stellate
	MJEA	5	5	5	15	Stellate
Medium difficulty	MJEG	5	4	3	12	Calcifications+AD
	MJGR	5	5	5	15	Stellate
	MJHD	3	5	4	12	AD
	MJAS	5	1	3	9	Calcifications+AD
Higher difficulty	MJHK	3	3	2	8	NSD
	MJCR	4	3	1	8	Stellate
	MJDU	3	3	0	6	Stellate+NSD
	MJEB	4	1	0	5	NSD
	MJHH	4	2	0	6	NSD
Higher difficulty	MJHJ	4	3	3	10	Stellate
	MJIK	3	3	0	6	NSD
Higher difficulty	MJBG	3	1	0	4	Calcifications+AD
	MJCF	1	1	0	2	NSD

Discussion

This study demonstrated that cancer characterised with non-specific density (NSD) was less likely to be recalled at lower rates, followed by calcifications and architectural distortions (AD), due to their subtle malignancy signs.

Stellate mammographic features with spiked linear extensions were easily recognized by our readers and were recalled for further assessment [4]. However, when stellate features were associated with other mammographic features, such as NSD with ill-defined borders, these lesions became less suspicious, and hence were not recalled at lower rates.

Cancer characterized with mixed features of calcifications and AD was the type of cancer that most of the readers could detect at free recall condition. However, when the recall rates was reduced to 15%, this type of cancer was less likely to be recalled with the highest drop in readers' detection was found in case MJAS. At 10% recall, cancer with NSD was the type of cancer that mostly gave up by all readers under this strict recall.

This finding is possibly due to a decision making error [5]. The readers may have identified the lesion earlier at free recall but decided that area was only attributable to the normal variability of breast tissue at reduced recall rates.

Conclusion

Overall, mammographic appearances of cancer lesion affects readers' recall decisions when recall rates were reduced, where cancer characterized with non-specific density were most likely to be missed.

References

1. Emrose, J.G., et al., Variability in Interpretive Performance of Screening Mammography and Radiologist Characteristics Associated with Accuracy. *Radiology*, 2009, 233(3): p. 841-851.
2. Emrose, J.G., et al., International Variation in Screening Mammography Interpretations in Community-Based Programs. *Journal of the National Cancer Institute*, 2003, 95(18): p. 1384-1393.
3. Yarikakis, B.C., et al., International comparison of performance measures for screening mammography: can it be done? *Journal of Medical Screening*, 2004, 11(4): p. 187-193.
4. Chene, F., V. Becette, and C. Hozy, Stellate images: anatomic and radiologic correlations. *European Journal of Radiology*, 2005, 54(1): p. 37-54.
5. Nordin, C.F., et al., Nature of Suspicious in Screening Mammograms for breast masses. *Academic Radiology*, 1996, 3: p. 1000-1006.

Corresponding author

Norhashimah Mohd Norsuddin,
Email: nmoh5894@uni.sydney.edu.au

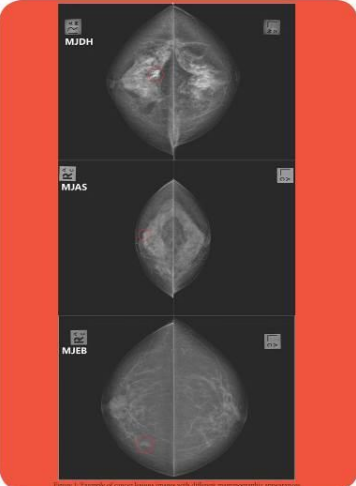


Figure 1. Four pairs of cancer lesion images with different mammographic appearances

Appendix M: IWDM 2016 Proceeding paper

Lower Recall Rates Reduced Readers' sensitivity in screening mammography

Appendix M is published as

Norhashimah Mohd Norsuddin, Claudia Mello-Thoms, Warren Reed, Patrick C. Brennan and Sarah Lewis, "*Lower Recall Rates Reduced Readers' sensitivity in screening mammography*", IWDM 2016, LNCS 9699, pp. 116-121, 2016, DOI:10.1007/978-3-319-41546-8_15

Lower Recall Rates Reduced Readers' Sensitivity in Screening Mammography

Norhashimah Mohd Norsuddin^{1,2}(✉), Claudia Mello-Thoms¹, Warren Reed¹,
Patrick C. Brennan¹, and Sarah Lewis¹

¹ Medical Imaging Optimisation and Perception Group (MIOPeG),
Discipline of Medical Radiation Sciences, Faculty of Health Science, The University of Sydney,
Cumberland Campus, East Street, Lidcombe, NSW 2141, Australia
nmoh5894@uni.sydney.edu.au, {claudia.mello-thoms,
warren.reed,patrick.brennan,sarah.lewis}@sydney.edu.au

² Diagnostic Imaging and Radiotherapy Programme, Faculty of Health Sciences,
The National University of Malaysia (UKM), 50300 Kuala Lumpur, Malaysia

Abstract. Higher recall rates have been related to increased false positive decisions, causing significant psychological and economical costs for both screened women and the mammography screening service respectively. This study compares breast readers' performance in a laboratory setting under varying levels of recall rates. Four experienced radiologists volunteered to read a single test set of 200 mammographic cases over three separate conditions. The test set contained of 180 normal and 20 abnormal cases and the participants were asked to identify each case that required to be recalled in line with three different target recall rates: control (unspecified or free recall (first read)), 15 % (second read) and 10 % (third read). Readers were required to mark the location of any malignancies using custom made detection software. The recall rates for the control condition ranged between 18.5 % and 34 %. Statistically significant differences were observed in sensitivity for control (median = 0.85) vs 15 % (median = 0.65, $z = -2.381$, $P = 0.017$), 15 % vs 10 % (median = 0.55, $z = -2.428$, $P = 0.015$) and control vs 10 % ($z = -2.381$, $P = 0.017$). ROC AUC was significantly different for control (median = 0.84) vs 15 % (median = 0.79, $z = -2.381$, $P = 0.017$) and 15 % vs 10 % (median = 0.75, $z = -2.381$, $P = 0.017$). Specificity significantly improved at lower recall rate of 10 % (median = 0.95) vs 15 % (median = 0.92, $z = -2.428$, $P = 0.017$). Setting specific target recall rates for readers significantly limited their performance in correctly identifying cancers. In this study, decreasing the number of recalled cases down to 10 %, significantly reduced cancer detection, with a significant improvement in specificity ($P \leq 0.05$).

Keywords: Recall rate · Sensitivity · Specificity · Screening mammography · Breast cancer

1 Introduction

Mammography is an effective imaging tool for detecting breast cancer at an early stage, however a relatively large number of screened women undergo unnecessary imaging

and invasive tests due to false positive findings. Such error may hamper the success of breast screening programs and reduce public confidence in medical screening programs [1, 2]. Some attention has been paid to variation in reader performance, particularly around recall rates using international comparisons of women attending breast screening programs. Overall, a large range of recall rates exists in screening practices, with recall rates quoted in the literature from below 1.4 % in the Netherlands up to 15 % in the United States [3] for initial screening. This variation was determined to some extent by factors including the mammographic technologies available at the point of screening, such as screen-film mammography (SFM) and full field digital mammography (FFDM) [4, 5]. Additionally, a woman's presentation can influence recall rate comparisons, especially when considering age, screening history, use of hormone therapy, breast density, possible previous invasive procedures and familial breast cancer history [6–9].

Considering that false positive results positively correlate with recall rates [10], some organizations have recommended specific recall policies as guidelines to evaluate the performance of readers in the screening population in their respective nations or states. In Australia, the National Accreditation Standards (NAS) has recommended a target recall rate of 10 % for the first screening and 5 % for subsequent screening across the population-based screening program. However it is known that some readers operate at considerably higher rates [11]. The intention of introducing recommended recall target rates is to optimize the trade-off between sensitivity and recall rate. However, extensive research has shown varying results around the question of an optimum recall rate [12–14]: a retrospective study by Schell et al. (2007) suggested that the best trade-off with sensitivity was at a 10 % recall rate for the first screen [14]; earlier work by Yankaskas et al. (2001) suggested that the effect of an increased recall rate on false positives can only be seen at the lower, limited range of recall rates between 4.9 % and 5.5 % [12, 13]; a Dutch study reported higher recall rates of more than 4 % are associated with an increase in false-positive decisions without an increase in the cancer detection rate [13].

Currently, there is a lack of evidence supporting the real effect of changing target recall rates upon readers' performance, particularly in relation to cancer detection rates. Other factors include the difficulty of altering embedded decision making practices and the advent of digital technologies. The purpose of this study is to further explore the relationship between recall rates and breast readers' performance in screening mammography to potentially improve the efficacy of breast screening programs.

2 Methods

Four experienced breast imaging radiologists who regularly report on screening mammograms for BreastScreen New South Wales (BSNSW) participated in this study. The mean number of years of experience was 16 and the mean case load read per year was 11,900 screening mammograms. Institutional ethical approval for this study was granted, informed consent was obtained from all participants and permission to use the images of patient materials was waived.

3 Results

Table 1 shows readers' scores for sensitivity, specificity and ROC AUC for the control, 15 % and 10 % recall rates respectively. The Kruskal-Wallis test showed significant differences in sensitivity ($P = 0.006$), specificity ($P = 0.007$) and ROC AUC ($P = 0.007$) across the three reading sessions (Table 1). The post-hoc Mann-Whitney U test shows significant changes in sensitivity, for control (free recall) versus (vs) 15 % ($z = -2.381$, $P = 0.017$) 15 % vs 10 % ($z = -2.428$, $P = 0.015$), and control vs 10 % ($z = -2.381$, $P = 0.017$) respectively. Specificity was significantly different for 15 % v 10 % ($z = -2.397$, $P = 0.017$) and the ROC AUC was significantly different for control vs 15 % ($z = -2.381$, $P = 0.017$) and 15 % vs 10 % ($z = -2.381$, $P = 0.017$) (Table 2).

Table 1 Reader sensitivity, specificity, ROC AUC and median values at free call, 15% and 10% conditions

Reader	Sensitivity			Specificity			ROC AUC		
	Control (Free recall)	15%	10%	Control (Free recall)	15%	10%	Control (Free recall)	15%	10%
1	0.80	0.65	0.55	0.83	0.91	0.95	0.84	0.79	0.76
2	0.90	0.65	0.45	0.79	0.91	0.93	0.86	0.79	0.70
3	0.90	0.65	0.55	0.72	0.92	0.95	0.84	0.79	0.75
4	0.75	0.70	0.55	0.88	0.92	0.95	0.83	0.82	0.75
Median	0.85	0.65	0.55	0.81	0.92	0.95	0.84	0.79	0.75

Table 2 Kruskal-Wallis analysis and post hoc Mann-Whitney U test of sensitivity, specificity and ROC AUC

	Kruskal-Wallis Test	Post-hoc test (Mann-Whitney U test)		
		Control VS 15%	15% VS 10%	Control VS 10%
	P value ($P \leq 0.05$)	P value ($P \leq 0.017$)		
Sensitivity	0.006	0.017	0.015	0.017
Specificity	0.007	0.019	0.017	0.018
ROC AUC	0.007	0.017	0.017	0.019

4 Discussion

The results from this study demonstrate that readers operating in a free recall condition (mean recall rate of 25.6 %) detected more cancers when compared to a reduced specified recall rate (mean ROC of 0.84 and 0.75 respectively, $P = 0.006$). The largest reduction occurred when readers were tasked with reducing their recalled cases from free recall to 15 %. The sensitivity change observed in our study is in agreement with those reported by Gur et al. and Schell et al., suggesting that recalling more cases may result in a higher cancer detection rate [10, 14]. Our data does not concur with the Otten et al. study [13],

4. Berns, E.A., Hendrick, R.E., Cutter, G.R.: Performance comparison of full-field digital mammography to screen-film mammography in clinical practice. *Med. Phys.* **29**(5), 830–834 (2002)
5. Lewin, J., et al.: Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. *Radiology* **218**(3), 873–880 (2001)
6. Castells, X., Molins, E., Macia, F.: Cumulative false positive recall rate and association with participant related factors in a population based breast cancer screening programme. *J. Epidemiol. Community Health* **60**(4), 316–321 (2006)
7. Carney, P.A., et al.: Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann. Intern. Med.* **138**(3), 168–175 (2003)
8. Lehman, C.D., et al.: Effect of age and breast density on screening mammograms with false-positive findings. *Am. J. Roentgenol.* **173**(6), 1651–1655 (1999)
9. Boyd, N.F., et al.: Mammographic density and the risk and detection of breast cancer. *New Engl. J. Med.* **356**(3), 227–236 (2007)
10. Gur, D., et al.: Recall and detection rates in screening mammography. *Cancer* **100**(8), 1590–1594 (2004)
11. BreastScreen Australia. National Accreditation Standards: BreastScreen Australia Quality (2008). [http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/A03653118215815BCA257B41000409E9/\\$File/standards.pdf](http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/A03653118215815BCA257B41000409E9/$File/standards.pdf). Accessed 21 May 2014
12. Yankaskas, B.C., et al.: Association of recall rates with sensitivity and positive predictive values of screening mammography. *Am. J. Roentgenol.* **177**(3), 543–549 (2001)
13. Otten, J.D.M., et al.: Effect of recall rate on earlier screen detection of breast cancers based on the dutch performance indicators. *J. Nat. Cancer Inst.* **97**(10), 748–754 (2005)
14. Schell, M.J., et al.: Evidence-based target recall rates for screening mammography. *Radiology* **243**(3), 681–689 (2007)
15. Elmore, J.G., et al.: International variation in screening mammography interpretations in community-based programs. *J. Nat. Cancer Inst.* **95**(18), 1384–1393 (2003)
16. Soh, B.P., et al.: Mammography test sets: reading location and prior images do not affect group performance. *Clin. Radiol.* **69**(4), 397–402 (2014)
17. Soh, B.P., et al.: Screening mammography: test set data can reasonably describe actual clinical reporting. *Radiology* **268**(1), 46–53 (2013)

Appendix N: Performance metrics

Performance metrics	Description
Sensitivity	<p>Measures the percentage or fraction of actual positive cancer cases that are correctly identified by the reader. Often described as a decimal. Mathematically can be expressed as</p> $\text{Sensitivity} = \frac{\text{number of true positive (TP)}}{\text{number of true positive (TP)} + \text{true negative (TN)}}$ $\text{Sensitivity} = \frac{TP}{TP + TN}$
Specificity	<p>Measures the percentage of fraction of cancer free cases that are correctly identified by the reader. Often described as a decimal. Mathematically can be expressed as</p> $\text{Specificity} = \frac{\text{number of true negative (TN)}}{\text{number of true negative (TN)} + \text{number of false positive (FP)}}$ $\text{Specificity} = \frac{TN}{TN + FP}$
Receiver operating characteristic (ROC) analysis	<p>ROC analysis is a binary paradigm focused on a single rating per case. the patient either does or does not have disease, i.e., truth is binary. The radiologist's task is to state whether the patient does or does not have disease, i.e., the response is binary. The resulting 2 x 2 truth-response table defines good decisions (true positives (TP), true negatives (TN)) and bad decisions (false positives (FP) and false negatives (FN)). The performance measure rewards good decisions and penalizes bad decisions.</p> <p>In this study, a TP score was given to a case when the reader correctly identified the correct side of the breast containing cancer, without the need to show the specific location of the lesion. This analysis is based on the ROC equivalent ratings inferred</p>

from the free-response data; where the rating of the highest-rated mark is included for comparison as it is the current gold standard. The ROC figure of merit is the area under the ROC curve (AUC). Radiologists will be asked to rate each case for probability of disease. The truth is known to the person running the study but not the radiologists. DBM-MRMC software is available on two websites that analyze MRMC-ROC data, and if the p-value is less than 5% one concludes that there is a statistically significant performance difference between at least two modalities.

$$\text{True positive fraction (TPF)} = \frac{\text{number of TP cases}}{\text{total number of abnormal cases}}$$

$$\text{True negative fraction (TNF)} = \frac{\text{number of TN cases}}{\text{total number of abnormal cases}}$$

$$\text{False positive fraction (FPF)} = \frac{\text{number of FP cases}}{\text{total number of normal cases}}$$

$$\text{False negative fraction (FNF)} = \frac{\text{number of FN}}{\text{Total number of abnormal cases}}$$

Therefore;

$$TNF = 1 - FPF \text{ and } FNF = 1 - TPF$$

The ROC curve is the plot of TPF along the y-axis vs. FPF as the confidence level is varied. The ROC curve is contained within the unit square. AUC = area under the ROC curve and $0 \leq AUC \leq 1$. AUC is the probability that an abnormal image is rated higher than a normal image.

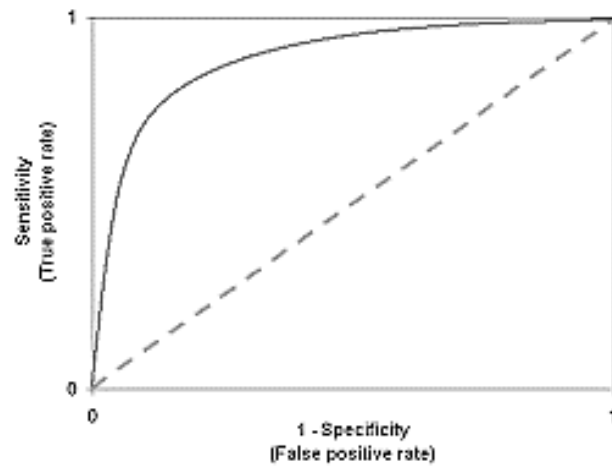


Figure 1 ROC graph

Jack-knife
free-response
ROC
(JAFROC)
analysis

JAFROC analysis is an advanced quantitative analysis of reader performance when interpreting mammography images. This analysis incorporates a free-response paradigm that allows lesion location information to be included when analysing reader performance. The JAFROC figure-of-merit (FOM) is the non-parametric (Mann-Whitney-Wilcoxon U-Statistic) area θ^{JAFROC} under the AFROC curve and it is defined by

$$\theta = \frac{1}{N_N \cdot N_L} \sum_{i=1}^{N_N} \sum_{j=1}^{N_L} \psi(X_i, Y_j)$$

$$\psi(X_i, Y_j) = \begin{cases} 1.0 & \text{if } Y_j > X_i \\ 0.5 & \text{if } Y_j = X_i \\ 0.0 & \text{if } Y_j < X_i \end{cases}$$

Where;

N_N = the number of normal cases

N_L = total number of lesions

X_i = the rating of the highest rated mark on the i^{th} normal case

Y^j = the rating of them j^{th} lesion.

**unmarked normal cases and unmarked lesions are assigned the -2000 rating.

Whereas the non-lesion localization marks on the abnormal cases are not be counted.

In this study, a TP score was given to a lesion when a reader successfully

marked and localized the lesion correctly within the acceptance radius. Furthermore, with the additional information of lesion location, this method demonstrates higher statistical power as compared to the ROC analysis. Through this method of analysis, the readers were allowed to locate multiple suspicious lesion locations during the interpretation process. The non-parametric area under the alternative free-response receiver operating characteristic (AFROC) curve was used as the figure of merit for JAFROC and the graph was plotted as the lesion localisation fraction (LLF) versus false positive fraction (FPF) or non-lesion localization (NLL). The software implementing JAFROC is available on www.devchakraborty.com.

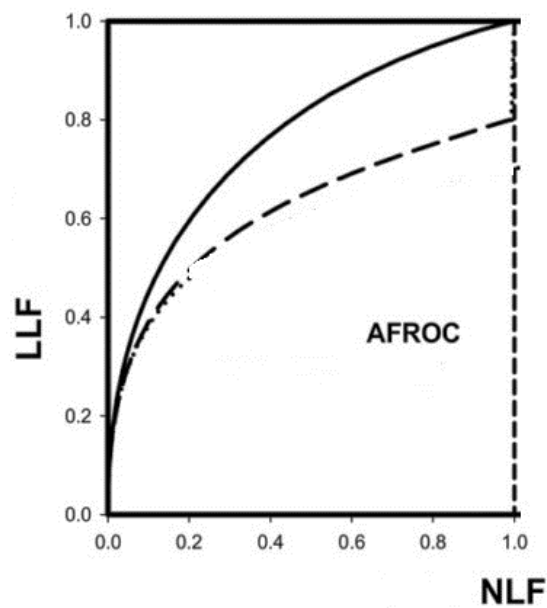


Figure 2 AFROC graph