Accuracy Evaluation of Radiographers Screen Reading Mammograms

Josephine Carmel Debono

This thesis is submitted in fulfillment of the requirements of the
Degree of Master of Applied Science

University of Sydney

2012
Declaration of Originality

I hereby declare that this thesis is my original work.

To the best of my knowledge it contains no other persons work, or previously published material unless otherwise acknowledged.

This thesis has not been accepted for an award at any other institute of higher learning.

Signed: .......................... Date: ..............................

Josephine Carmel Debono

SID: 307077462
For my family

“Consult not your fears but your hopes and your dreams.  
Think not about your frustrations, but about your unfulfilled potential.  
Concern yourself not with what you tried and failed in,  
but with what it is still possible for you to do”.

Pope John XXIII
Acknowledgements

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.

My gratitude goes to Adjunct Professor Ann Poulos who has provided academic and supportive guidance along the journey. Your reviews and counsel have been invaluable.

Further appreciation goes to Associate Professor Nehmat Houssami, especially during the research design. Your wealth of experience, advice and encouragement has been welcomed.

Thanks goes to Professor John Boyages for listening to my suggestion of carrying out an evaluation of radiographer screen reading accuracy. Your foresight has provided me with the unique opportunity to undertake this research to potentially impact on the efficiency of the BreastScreen Australia program and the future roles of the radiographer profession.

The advanced analyses in this thesis were performed by a collaborating statistician, Robin Turner, and I am grateful for her most valuable contribution to this research. Your insight and intellect have been valued.

Last but most certainly not least, I wish to acknowledge 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 recognise my incredibly affirming husband Ray, my extraordinary daughter Rachel for her immeasurable support, my talented son James, my encouraging daughter Justine and son-in-law Luke, and my gorgeous little granddaughter Amalia, who has been a wonderful distraction, and who will hopefully learn that she too has the potential to achieve anything that she sets in her sights.
Table of Contents

Declaration of Originality.............................................................................................................. ii
Acknowledgements......................................................................................................................... iv
Table of Contents.............................................................................................................................. v
Abstract........................................................................................................................................ viii
Publication arising from this study................................................................................................ x
  Evaluating radiographers’ diagnostic accuracy in screen reading: what constitutes a quality study? x
Presentations arising from this study.............................................................................................. xxviii
List of Tables ................................................................................................................................ xxix
List of Figures ............................................................................................................................... xxx
Chapter 1: Introduction .................................................................................................................. 1
  1.1. Aim............................................................................................................................... 4
  1.2. Structure of thesis......................................................................................................... 4
Chapter 2: Background .................................................................................................................. 6
  2.1. Workforce shortages ................................................................................................. 7
     2.1.1. UK experience ...................................................................................................... 8
  2.2. Context of screen reading ........................................................................................... 9
     2.2.1. Images produced ................................................................................................. 9
     2.2.2. Screen reading process ...................................................................................... 12
  2.3. Diagnostic accuracy ..................................................................................................... 14
     2.3.1. Screen readers ................................................................................................... 14
     2.3.2. Abnormality detection ...................................................................................... 15
  2.4. Current radiographer roles in the screening process .................................................. 16
Chapter 3: Literature Review ........................................................................................................ 17
  Literature search ...................................................................................................................... 18
  Data extraction ....................................................................................................................... 19
  Evaluation of the literature ..................................................................................................... 20
  3.1. Samples in the reviewed studies .................................................................................. 20
     3.1.1. Images ............................................................................................................... 20
     3.1.2. Screen readers .................................................................................................. 25
  3.2. Methods in the reviewed studies .................................................................................. 26
     3.2.1. Gold standard .................................................................................................... 26
     3.2.2. Screen reading process .................................................................................... 27
  3.3. Results of the reviewed studies .................................................................................... 29
     3.3.1. Screen reader performance ............................................................................... 29
  3.4. Summary....................................................................................................................... 35
Chapter 4: Method ................................................................................................................. 37
  4.1. Aim ................................................................................................................................ 38
    4.1.1. Ethics approval ....................................................................................................... 38
  4.2. Samples .......................................................................................................................... 38
    4.2.1. Screen reader recruitment ....................................................................................... 38
    4.2.2. Image test set .......................................................................................................... 38
  4.3. Procedure .................................................................................................................... 42
    4.3.1. Image presentation .................................................................................................. 42
    4.3.2. Reporting documents .............................................................................................. 44
    4.3.3. File location communication .................................................................................. 44
    4.3.4. Participant orientation ............................................................................................. 45
    4.3.5. Screen reading process ............................................................................................ 45
  4.4. Data entry ................................................................................................................... 46
  4.5. Comparison to gold standard ......................................................................................... 47
  4.6. Data analysis .................................................................................................................. 48
Chapter 5: Results ..................................................................................................................... 51
  5.1. Samples .......................................................................................................................... 52
    5.1.1. Screen reader sample .............................................................................................. 52
    5.1.2. Image test set sample .............................................................................................. 54
  5.2. Individual screen reader accuracy .................................................................................. 54
    5.2.1. Sensitivity and specificity ....................................................................................... 54
    5.2.2. Recall rates .............................................................................................................. 58
    5.2.3. ROC ........................................................................................................................ 58
  5.3. Pooled screen reader accuracy ....................................................................................... 59
    5.3.1. Sensitivity and specificity ....................................................................................... 59
    5.3.2. ROC ........................................................................................................................ 60
    5.3.3. Accuracy according to lesion type .......................................................................... 61
    5.3.4. Learning effect ........................................................................................................ 62
  5.4. Inter-observer accuracy .................................................................................................. 62
Chapter 6: Discussion ............................................................................................................... 63
  6.1. Evaluation of sample ..................................................................................................... 64
    6.1.1. Screen reader sample .............................................................................................. 64
    6.1.2. Test sets versus consecutive populations ................................................................. 65
  6.2. Individual screen reader accuracy .................................................................................. 66
    6.2.1. Sensitivity and specificity ....................................................................................... 66
    6.2.2. Recall rates .............................................................................................................. 70
    6.2.3. ROC and AUC ........................................................................................................ 71
  6.3. Pooled screen reader accuracy ....................................................................................... 72
    6.3.1. Pooled sensitivity and specificity ........................................................................... 72
    6.3.2. Pooled ROC ............................................................................................................ 73
Abstract

The aim of this thesis is to evaluate the accuracy of radiographers screen reading mammograms. Currently, the goal of the BreastScreen Australia program of early detection of breast cancer to minimise morbidity and mortality is being compromised due to radiologist workforce shortages. These shortages contribute to delays in women receiving their screening mammogram results. Further future delays are anticipated due to the ageing population and the recommendation to increase the target age range of eligible women. A potential solution to this problem is for radiographers to take on the role as one of the two screen readers.

Prior to consideration of such a strategy, it is necessary to evaluate the accuracy of radiographers screen reading mammograms. Previous international radiographer screen reading accuracy studies have reported acceptable accuracy levels, and two Australian pilot studies report comparable accuracy levels to these international studies. These studies are encouraging; however, there is a noticeable absence of a large well-designed study undertaken in an Australian setting in this area of research. This Australian study aims to improve on the design of previous studies by applying a rigorous gold standard and including further essential study design characteristics, such as a large number of radiographers.

A literature review of previous radiographer screen reading accuracy studies identified the components of a well-designed study and provided the rationale for the method used for the study comprising this thesis. To evaluate the accuracy of radiographers screen reading mammograms, 10 radiographers employed by the Westmead Breast Cancer Institute were recruited to blindly and independently screen read an image test set of 500 mammograms, including normal and a representative spectrum of abnormal pathology. The radiographers indicated whether or not they considered an abnormality to be present, using a standardised reporting form, and accuracy was determined through comparison to the gold standard of known outcomes consisting of pathology results and 6-year follow-up.
Accuracy evaluation of these radiographer screen readers found that individual sensitivity levels ranged between 76.0% and 92.0%, and individual specificity levels ranged between 74.8% and 96.2%. Recall rates ranged between 14.2% and 57.0%. Pooled screen reader accuracy across the screen readers estimated sensitivity as 82.2% and specificity as 89.5%. Areas Under the Reading Operating Characteristic Curve (AUC) ranged between 0.842 and 0.923. Radiographer screen readers were more accurate at detecting calcifications and discrete masses, in comparison to stellate lesions, architectural distortion and non-specific densities. As they read more images there was also a moderate improvement in sensitivity, and significant improvement in specificity. There was minimal inter-observer variability between the screen readers.

This study has provided strong evidence that, even without formal screen reading training, this sample of radiographers in an Australian setting have good accuracy levels when screen reading mammograms, however, these were associated with high recall rates. It is expected that with formal screen reading training, these accuracy levels will improve further and recall rates will likely reduce, such that radiographers have the potential to be one of the two screen readers in the BreastScreen Australian program, contributing to timeliness and improved program outcomes.

Further research is recommended following the employment of radiographers as screen readers in Australian population settings. Prior to employment as screen readers, formal screen reading training, aimed at improving lesion detection in areas of weakness, while building upon areas of known strength, is essential. Further research aimed at investigating both breast density and reader fatigue as detractors of accuracy would be beneficial. Evaluation following further screen reading volume, under digital technology conditions and of radiographer/radiologist pairs would also be advantageous.
Publication arising from this study

The manuscript that follows has been submitted for peer review for publication in the European Journal of Radiology, 2012.

Evaluating radiographers’ diagnostic accuracy in screen reading: what constitutes a quality study?

Josephine Carmel Debono\textsuperscript{a} and Dr Ann Poulos\textsuperscript{b}

\textsuperscript{ab} Faculty of Health Sciences, University of Sydney, PO Box 170, Lidcombe, NSW 1825, Australia

Telephone Number: +61 2 9351 9514

Fax Number: +61 2 9351 9146

Email: \textsuperscript{a}jdeb9656@uni.sydney.edu.au

\textsuperscript{a}jodebono61@gmail.com

\textsuperscript{b}ann.poulos@sydney.edu.au

\textsuperscript{a} Corresponding Author

Both authors declare no actual or potential conflict of interest.
Keywords: Quality evaluation tools; diagnostic accuracy; screen readers; radiographers

Abstract

Accurate screen reading of mammogram images is critical for the early detection of breast cancer, the goal of population screening programs. Radiology workforce issues have been addressed by the training and employment of radiographers as screen readers. Studies have been undertaken to evaluate the accuracy of radiographers in this role. Since diagnostic accuracy in screen reading underpins the goal of breast screening to detect breast cancer early and reduce mortality, the quality of these studies is paramount. van den Biggelaar et al. in their 2008 systematic review found only a small number of these could be described as well-designed. This finding raises the question of what constitutes a well-designed quality study in this area of research.

A number of tools for evaluating the quality of diagnostic accuracy studies have been identified in the literature; however observer characteristics have sometimes been underconsidered in spite of their ability to affect study outcomes.

A literature search identified eleven studies investigating the diagnostic accuracy of radiographers as screen readers. This reported study has identified quality issues which compromise the validity and reliability of outcomes of studies focusing on radiographers’ accuracy in screen reading. During the process of evaluating these studies a quality assessment tool specifically for evaluating the quality and reporting of studies investigating diagnostic accuracy of screen readers has been derived (DASQUART). This tool, with further refinement and validation will make a contribution to promoting well-designed studies in this important area of research and practice.
Introduction

Diagnostic accuracy in medical imaging is essential for appropriate patient management and treatment. Accurate screen reading of mammogram images is critical for the early detection of breast cancer, the goal of population screening programs. Screen readers of mammogram images are predominantly, but not exclusively radiologists. Currently there are workforce issues in radiology which impact on their availability for screen reading. In the UK this shortage has been addressed by the training and employment of radiographers as screen readers. A range of studies have investigated the diagnostic accuracy of radiographers in this role and provide evidence that radiographers have comparable accuracy to radiologists. More recent studies provide evidence of the ability of radiographers to contribute to improvements in the efficiency of the screening process and most importantly that combining radiologist and radiographer screen reading has been found to improve cancer detection rates.

Since diagnostic accuracy in screen reading underpins the goal of breast screening to detect breast cancer early and reduce mortality, the quality of these studies is paramount. A systematic review published in 2008 by van den Biggelaar et al. excluded articles without evidence of sensitivity and specificity and an appropriate gold standard, resulting in a total of six. This systematic review raised questions of what constitutes a well-designed study and how quality is defined in studies investigating screen reading accuracy by radiographers. More specifically, the authors emphasised the necessity of determining the key components of a well-designed study in this area of research to increase the rigour and applicability of the outcomes to the clinical environment.
Quality evaluation tools for studies of diagnostic accuracy

A number of tools for evaluating the quality of diagnostic accuracy studies have been identified in the literature. The STAndards for the Reporting of Diagnostic accuracy studies (STARD), was developed from an initiative to improve the accuracy and completeness of reporting studies of diagnostic accuracy. Subsequently the Quality of Diagnostic Accuracy Studies (QUADAS) tool was developed and validated by Whiting, Rutjes, Reitsma et al. (2003) to determine the quality of primary studies in systematic reviews of diagnostic accuracy.

Subsequently Whiting, Rutjes, Dinnes et al. (2005) conducted a systematic review of existing quality assessment tools to examine both the extent and type of quality assessment being incorporated in diagnostic accuracy systematic reviews. Aspects of quality considered in their review were classified as: potential for bias, conduct of the study, applicability of the results and quality of reporting. Following data extraction the data was synthesized according to purpose and summarized as items.

This classification is useful since it is all-inclusive and includes items of quality drawn from an extensive review of systematic reviews. As well as determining the individual items related to quality in diagnostic accuracy studies, the classification also synthesizes these items into aspects of quality. Importantly this classification includes quality items relating to the reporting of studies. The comprehensive nature of this classification facilitates the appropriation of the quality items or criteria to a specific area of diagnostic accuracy research.
Importance of observer characteristics and variability

The importance of observer characteristics and variability on diagnostic accuracy in medical imaging have been emphasized by Brealey and Westwood \(^{21}\), who claim that observers are frequently ignored in diagnostic accuracy studies in medical imaging in spite of their ability to affect the study outcomes. The number of observers, for example, influences the internal and external validity of research studies, while the profession and experience of observers affect estimates of accuracy. Brealey and Westwood \(^{21}\) strongly recommend the inclusion of observer assessment criteria in a quality assessment tool evaluating diagnostic accuracy in medical imaging.

The aim of this study was firstly to evaluate the quality of the studies investigating the diagnostic accuracy of radiographers as screen readers using a quality evaluation tool constructed by combining the criteria for quality of Whiting, Rutjes, Dinnes et al. (2005) \(^{17}\), and Brealey and Westwood (2007) \(^{21}\). Secondly the applicability and appropriateness of the criteria were determined and an adapted quality evaluation tool was developed specifically for use in evaluating diagnostic accuracy in screen reading studies.
Method

A literature search was undertaken within the Medline, PubMed and Cinahl databases, using combinations of the terms: mammogram, radiographer, technologist, screen reading, accuracy and interpretation. There were no limits applied for publication dates. An initial review of titles and abstracts enabled the exclusion of papers that were clearly not relevant to the subject of interest. Studies investigating the diagnostic accuracy of radiographers reading mammograms were selected. Further studies were located using the reference lists. As only a small number ($n = 11$) of papers were located, no further inclusion/exclusion criteria were applied.

Quality evaluation

For this evaluation the classifications and items of Whiting et al. (2005)\textsuperscript{17} were combined with the observer characteristics recommended by Brealey and Westwood (2007)\textsuperscript{21} to develop a comprehensive all-inclusive quality assessment tool for diagnostic accuracy studies using imaging, see table 1. Items were redefined as criteria and then applied to the studies to evaluate their quality by two researchers experienced in mammography and the diagnostic process of screen reading.
Table 1 - Classification of items included in quality assessment tools (Source: Whiting et al., 2005 p.3) plus observer characteristics (Brealey and Westwood, 2007 p.676)

<table>
<thead>
<tr>
<th>ID</th>
<th>Item</th>
<th>Description of item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Potential for bias</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>Reference Standard</td>
<td>Was an appropriate reference standard used to determine the presence or absence of the target condition?</td>
</tr>
<tr>
<td>A2</td>
<td>Disease Progression bias</td>
<td>Could a change in disease state have occurred between application of the index test and reference standard?</td>
</tr>
<tr>
<td>A3</td>
<td>Verification bias</td>
<td>Did all subjects receive verification of the target condition using the same reference standard?</td>
</tr>
<tr>
<td>A4</td>
<td>Incorporation bias</td>
<td>Did the index test form part of the reference test?</td>
</tr>
<tr>
<td>A5</td>
<td>Treatment paradox</td>
<td>Was treatment started based on the result of the index test before the reference standard was applied?</td>
</tr>
<tr>
<td>A6</td>
<td>Review bias</td>
<td>Were index test results interpreted without knowledge of the results of the reference standard, and vice versa?</td>
</tr>
<tr>
<td>A7</td>
<td>Clinical review bias</td>
<td>Was clinical information available when test results were interpreted?</td>
</tr>
<tr>
<td>A8</td>
<td>Observer/instrument variation</td>
<td>Was observer/instrument variation likely to have affected estimates of test performance?</td>
</tr>
<tr>
<td>A9</td>
<td>Handling of uninterpretable results</td>
<td>Were uninterpretable results included in the analysis?</td>
</tr>
<tr>
<td>A10</td>
<td>Arbitrary choice of threshold value</td>
<td>Was the threshold value chosen independently of the results of the study? i.e., it should not have been chosen to optimize estimates of test performance</td>
</tr>
<tr>
<td><strong>B. Applicability</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>Spectrum composition</td>
<td>Was the population studied similar to the one in which you are interested?</td>
</tr>
<tr>
<td>B2</td>
<td>Population recruitment</td>
<td>Was the method of population recruitment adequate to include an appropriate spectrum of patients?</td>
</tr>
<tr>
<td>B3</td>
<td>Disease prevalence/severity</td>
<td>Was the spectrum of disease prevalence and severity similar to the one in which you are interested?</td>
</tr>
<tr>
<td>B4</td>
<td>Change in technology of index test</td>
<td>Is it likely that the technology of the test has changed since the study was conducted?</td>
</tr>
<tr>
<td><strong>C. Conduct of the study</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>Subgroup analysis</td>
<td>Were subgroup analyses appropriate and specified?</td>
</tr>
<tr>
<td>C2</td>
<td>Sample size</td>
<td>Were an appropriate number of participants included in the study?</td>
</tr>
<tr>
<td>C3</td>
<td>Objectives</td>
<td>Were study objectives relevant to the study question?</td>
</tr>
<tr>
<td>C4</td>
<td>Protocol</td>
<td>Was a study protocol developed before the study started and did the investigators adhere to it?</td>
</tr>
<tr>
<td><strong>D. Reporting of the study</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>Inclusion criteria</td>
<td>Were inclusion criteria clearly reported?</td>
</tr>
<tr>
<td>D2</td>
<td>Test execution</td>
<td>Were sufficient details provided on how the index test was performed to permit its replication?</td>
</tr>
<tr>
<td>D3</td>
<td>Reference execution</td>
<td>Were sufficient details provided on how the reference standard was performed to permit its replication?</td>
</tr>
<tr>
<td>D4</td>
<td>Normal defined</td>
<td>Did the authors clearly report what they considered to be a normal test result?</td>
</tr>
<tr>
<td>D5</td>
<td>Appropriate results</td>
<td>Were appropriate results presented? e.g., sensitivity, specificity, likelihood ratios</td>
</tr>
<tr>
<td>D6</td>
<td>Precision of results</td>
<td>Was some estimate of the precision of the results presented? e.g., confidence interval</td>
</tr>
<tr>
<td>D7</td>
<td>Drop-outs</td>
<td>Were all patients that entered the study accounted for?</td>
</tr>
<tr>
<td>D8</td>
<td>Data table</td>
<td>Was an n x n table of test performance reported?</td>
</tr>
<tr>
<td>D9</td>
<td>Utility of test</td>
<td>Was some indication of how useful the test might be in practice?</td>
</tr>
<tr>
<td><strong>E. Observer Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>Image allocation to observers</td>
<td>How were images allocated to be read by the observers?</td>
</tr>
<tr>
<td>E2</td>
<td>Number of observers</td>
<td>Was the number of observers presented?</td>
</tr>
<tr>
<td>E3</td>
<td>Observer experience</td>
<td>Was the experience of the observers described?</td>
</tr>
<tr>
<td>E4</td>
<td>Observer training</td>
<td>Was the training of the observers described?</td>
</tr>
<tr>
<td>E5</td>
<td>Observer profession</td>
<td>Was the profession of the observers presented?</td>
</tr>
<tr>
<td>E6</td>
<td>Observer variability</td>
<td>Was there an assessment of observer variability?</td>
</tr>
<tr>
<td>E7</td>
<td>Analysis of observer variability</td>
<td>Was observer variability considered in the analyses of test accuracy?</td>
</tr>
</tbody>
</table>
Development of quality assessment tool for screen reading

During the process of evaluation the criteria were appropriated to the specific quality aspects of studies reporting on screen reading. Adaptations to the criteria were identified that increased the relevance and applicability of the tool for the specific purpose of the evaluation of the diagnostic accuracy of screen readers interpreting mammograms in breast screening facilities.

Results

Quality evaluation

Eleven studies were identified in the literature relating to the diagnostic accuracy of radiographers reading screening mammograms and are presented in table 2.

Table 2 – Screen reading studies

<table>
<thead>
<tr>
<th>Authors and year of publication in chronological order.</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Haiart &amp; Henderson., 1991</td>
<td>A comparison of interpretation of screening mammograms by a radiographer, a doctor and a radiologist.</td>
</tr>
<tr>
<td>2 Bassett et al., 1995</td>
<td>Effects of a program to train radiologic technologists to identify abnormalities on mammograms.</td>
</tr>
<tr>
<td>3 Pauli et al.,(1), 1996</td>
<td>Comparison of radiographer / radiologist double film reading with single reading in breast cancer screening.</td>
</tr>
<tr>
<td>5 Tonita et al., 1999</td>
<td>Medical radiologic technologist review: effects on a population-based breast cancer screening program.</td>
</tr>
<tr>
<td>6 Wivell et al., 2003</td>
<td>Can radiographers read screening mammograms?</td>
</tr>
<tr>
<td>8 Holt., 2006</td>
<td>Evaluating radiological technologists’ ability to detect abnormalities in film-screen mammographic images: A decision analysis pilot project.</td>
</tr>
<tr>
<td>9 Duijm et al., 2007</td>
<td>Additional double reading of screening mammograms by radiologic technologists: impact on screening performance parameters.</td>
</tr>
<tr>
<td>1 Duijm et al., 2008</td>
<td>Introduction of additional double reading of mammograms by radiographers: effects on a biennial screening programme outcome.</td>
</tr>
<tr>
<td>1 Duijm et al., 2009</td>
<td>Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome.</td>
</tr>
</tbody>
</table>

No studies were excluded from the review. Study quality of each of the 11 studies was evaluated using the developed tool; the results of these evaluations are presented in table 3. The ‘Total’ row under each category A-E indicate the numbers of negative responses to the criteria for each study while the numbers in the final ‘Total’ column indicate the numbers of
negative responses to each of the 34 criteria. If a partial negative response was indicated then 0.5 was allocated.

Table 3 - Evaluation of reviewed studies using the constructed quality tool (table 1)

<table>
<thead>
<tr>
<th>Study</th>
<th>Haiart et al., 1991</th>
<th>Bassett et al., 1995</th>
<th>Pauli et al., (1) 1996</th>
<th>Pauli et al., (2) 1996</th>
<th>Tonita et al., 1999</th>
<th>Wivell et al., 2003</th>
<th>Sumkin et al., 2003</th>
<th>Holt 2006</th>
<th>Duijm et al., 2007</th>
<th>Duijm et al., 2008</th>
<th>Duijm et al., 2009</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>1</td>
</tr>
<tr>
<td>A2</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>1</td>
</tr>
<tr>
<td>A3</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0</td>
</tr>
<tr>
<td>A4</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>Partial</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7.5</td>
</tr>
<tr>
<td>A5</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>A6</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>Partial</td>
<td>√</td>
<td>0</td>
</tr>
<tr>
<td>A7</td>
<td>N/S</td>
<td>-</td>
<td>√</td>
<td>N/S</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td>A8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11</td>
</tr>
<tr>
<td>A9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11</td>
</tr>
<tr>
<td>A10</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2.5</td>
<td>3.5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3.5</td>
<td>33.5</td>
</tr>
</tbody>
</table>

B. Applicability of results

<table>
<thead>
<tr>
<th>Study</th>
<th>Haiart et al., 1991</th>
<th>Bassett et al., 1995</th>
<th>Pauli et al., (1) 1996</th>
<th>Pauli et al., (2) 1996</th>
<th>Tonita et al., 1999</th>
<th>Wivell et al., 2003</th>
<th>Sumkin et al., 2003</th>
<th>Holt 2006</th>
<th>Duijm et al., 2007</th>
<th>Duijm et al., 2008</th>
<th>Duijm et al., 2009</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>√</td>
<td>N/S</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>1</td>
</tr>
<tr>
<td>B2</td>
<td>√</td>
<td>N/S</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td>B3</td>
<td>√</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>B4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>

C. Conduct of the study

<table>
<thead>
<tr>
<th>Study</th>
<th>Haiart et al., 1991</th>
<th>Bassett et al., 1995</th>
<th>Pauli et al., (1) 1996</th>
<th>Pauli et al., (2) 1996</th>
<th>Tonita et al., 1999</th>
<th>Wivell et al., 2003</th>
<th>Sumkin et al., 2003</th>
<th>Holt 2006</th>
<th>Duijm et al., 2007</th>
<th>Duijm et al., 2008</th>
<th>Duijm et al., 2009</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>4</td>
</tr>
<tr>
<td>C3</td>
<td>√</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0</td>
</tr>
<tr>
<td>C4</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

D. Reporting of the study

<table>
<thead>
<tr>
<th>Study</th>
<th>Haiart et al., 1991</th>
<th>Bassett et al., 1995</th>
<th>Pauli et al., (1) 1996</th>
<th>Pauli et al., (2) 1996</th>
<th>Tonita et al., 1999</th>
<th>Wivell et al., 2003</th>
<th>Sumkin et al., 2003</th>
<th>Holt 2006</th>
<th>Duijm et al., 2007</th>
<th>Duijm et al., 2008</th>
<th>Duijm et al., 2009</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>D2</td>
<td>Partial</td>
<td>√</td>
<td>√</td>
<td>Partial</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>1</td>
</tr>
<tr>
<td>D3</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>1</td>
</tr>
<tr>
<td>D4</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 4 presents the adapted quality tool specifically developed for diagnostic accuracy in screen reading and derived from the quality evaluation process. This tool, for ease of identification, is named the DASQUART (Diagnostic Accuracy Study Quality And Reporting Tool).

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D5</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>√</td>
</tr>
<tr>
<td><strong>D6</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>√</td>
</tr>
<tr>
<td><strong>D7</strong></td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>D8</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>D9</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>2.5</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2.5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E1</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>E2</strong></td>
<td>√</td>
<td>√</td>
<td><strong>Partial</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>E3</strong></td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>E4</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>E5</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>E6</strong></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>√</td>
</tr>
<tr>
<td><strong>E7</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>2</td>
<td>2</td>
<td>2.5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

N/S Not stated
N/A Not Applicable

E. Observer characteristics
Table 4: Developed tool named *DASQUART* for determining quality in studies investigating diagnostic accuracy in screen reading

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description of Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Reference Standard An appropriate reference standard of pathology and at least 1 year follow-up used to determine the presence or absence of breast cancer</td>
</tr>
<tr>
<td>A2</td>
<td>Disease Progression bias An interval cancer could not occur between the initial mammogram and the reference standard</td>
</tr>
<tr>
<td>A3</td>
<td>Verification bias Same reference standard applied across the study</td>
</tr>
<tr>
<td>A4</td>
<td>Incorporation bias The reading of the screening mammogram does not form part of the reference standard</td>
</tr>
<tr>
<td>A6</td>
<td>Review bias Mammograms read blinded to knowledge of reference standard and interpretation by other readers</td>
</tr>
<tr>
<td>A7</td>
<td>Clinical review bias Previous image rounds available for comparison</td>
</tr>
<tr>
<td>A8</td>
<td>instrument variation No reporting instrument variation which will affect estimates of test performance e.g. Use of BIRADS® lexicon</td>
</tr>
<tr>
<td>A9</td>
<td>Handling of uninterpretable results Uninterpretable results included in the analysis</td>
</tr>
<tr>
<td>A10</td>
<td>Arbitrary choice of threshold value Threshold value of normal chosen independently of results</td>
</tr>
<tr>
<td>B1</td>
<td>Spectrum composition Image sample similar to one of interest (test sets e.g. PERFORMS, BREAST and consecutive screening)</td>
</tr>
<tr>
<td>B2</td>
<td>Population recruitment Image sample selected adequate to include appropriate spectrum (test sets e.g. PERFORMS, BREAST and consecutive screening)</td>
</tr>
<tr>
<td>B3</td>
<td>Disease prevalence/ severity Spectrum of breast cancer prevalence similar to one of interest (test sets e.g. PERFORMS, BREAST and consecutive screening)</td>
</tr>
<tr>
<td>B4</td>
<td>Change in technology of index test No change in mammography technology which will affect applicability of results</td>
</tr>
<tr>
<td>C1</td>
<td>Subgroup analysis Sub-group analyses were appropriate and specified</td>
</tr>
<tr>
<td>C2</td>
<td>Sample size Appropriate number of images included in study</td>
</tr>
<tr>
<td>C3</td>
<td>Objectives Study objectives relevant to study question</td>
</tr>
<tr>
<td>C4</td>
<td>Study design The purpose, method, results and conclusions demonstrate logical coherence and consistency</td>
</tr>
<tr>
<td>D1</td>
<td>Inclusion criteria Included in systematic reviews</td>
</tr>
<tr>
<td>D2</td>
<td>Test execution (a) images Sufficient details of mammogram reading reported to permit its replication. Details include number of images read in total and at one sitting, how images were selected (test sets), degree of difficulty (test sets), types of breast cancers included (test sets). Time taken to read, background lighting and type of monitors</td>
</tr>
<tr>
<td>D3</td>
<td>Reference execution Sufficient details provided of reference standard used to permit its replication</td>
</tr>
<tr>
<td>D4</td>
<td>Normal defined Authors clearly reported what was considered a normal reading result</td>
</tr>
<tr>
<td>D5</td>
<td>Appropriate results Appropriate results of accuracy presented, e.g. sensitivity and specificity and ROC analysis</td>
</tr>
<tr>
<td>D6</td>
<td>Precision of results Estimate of precision of results presented as appropriate</td>
</tr>
<tr>
<td>D7</td>
<td>Drop-outs All images and observers accounted for</td>
</tr>
<tr>
<td>D8</td>
<td>Data table Test performance reported in a data table</td>
</tr>
<tr>
<td>D9</td>
<td>Utility of test Clinical relevance of the test emphasised</td>
</tr>
<tr>
<td>E1</td>
<td>Image allocation to observers Image allocation to observers described</td>
</tr>
<tr>
<td>E2</td>
<td>Number of observers Number of observers presented</td>
</tr>
<tr>
<td>E3</td>
<td>Observer experience Experience of observers described</td>
</tr>
<tr>
<td>E4</td>
<td>Observer training Training of observers described</td>
</tr>
<tr>
<td>E5</td>
<td>Observer profession Profession of observers presented</td>
</tr>
<tr>
<td>E6</td>
<td>Analysis of observer variability Observer variability in analysis e.g. Kappa statistic</td>
</tr>
</tbody>
</table>
Discussion

The outcome of the quality evaluation of the studies is presented in table 3. Whiting, Harbord and Kleijen (2005) 22 emphasised the need to investigate individual quality items and their association with estimates of diagnostic accuracy rather than produce scores. So while identification of negative responses to criteria may be a simplistic method of scoring quality, this reported quality evaluation has identified a ranking of studies as well as the category demonstrating the largest number of negative responses, that of section A: Potential for bias. Potential bias can severely compromise outcomes and must be minimised wherever possible.

Bias can be minimised by ensuring the research design is similar to the screen reading process in practice using criteria A1-A8. The highest number of negative responses for bias potential were A4 (7.5), A8 (11) and A9 (11). Incorporation bias (A4) did occur in the studies since it is an immutable aspect of the screen reading process. Potential confounders which affect test performance and relate to the varying classification systems used (A8) can be reduced by using a validated reporting instrument such as the BIRADS® classification lexicon 23. Uninterpretable results (A9) were not included in the reviewed studies since the results were known prior to the test.

Since potential bias is the predominant detractor of quality in the reviewed studies it is suggested that further work needs to identify the association of the criteria within category A with the estimates of diagnostic accuracy produced in the studies, and determine a hierarchy of the impact of negative responses to the criteria for outcome estimates of accuracy.

Negative responses in section B: Applicability of results, were highest in B4. This is explained by the introduction of digital technology since the reviewed articles were published. It is possible that this new technology may provide increased diagnostic accuracy in screen reading, and so results of the reviewed studies may not be generalisable to facilities using digital equipment. This change however did not influence the applicability of the results at the time of publication.
Section C: Conduct of the study demonstrated low but significant negative responses to C2, sample size. Appropriate sample size is a critical component of a research study and in the field of research covered by the reviewed studies, sample refers to both number of images read and number of observers reading the images. This criterion therefore requires clarification.

Section D: Reporting of the study criteria D6 (8) and D7 (10) demonstrated large numbers of negative responses. Precision of results and accounting for all the images (rather than patients) were lacking in some studies. The way in which these criteria are expressed does not readily apply to screen reading.

Section E: Observer characteristics demonstrated high numbers of negative responses in criteria E6 (7) and E7 (10) which are fundamentally the same. Observer variability should be analysed statistically through the use of the Kappa statistic or similar, as appropriate.

In summary, these evaluation results emphasise the need for a specific evaluation tool for diagnostic accuracy in screen reading. The specific screen reading processes which minimize bias can be clearly enunciated, appropriate sample sizes of images and observers identified and criteria relating to study reporting increased in relevance.

**Quality evaluation tool for studies in diagnostic accuracy in screen reading**

The quality tool used to evaluate the reviewed studies (table 1) was appropriated to provide a specific tool for diagnostic accuracy studies in screen reading named the DASQUART and is presented in table 4. The quality criteria of Whiting, Rutjes, Dinnes et al. (2005) 17, and additional criteria related to medical imaging of Brealey and Westwood (2007) 21 have been adapted to enhance relevance, clarity and precision and to contribute to the development of a user-friendly quality assessment tool.
Changes to existing criteria

To maintain consistency in the structure of the tool definitive statements rather than questions are presented throughout as descriptions of criteria. A positive response to these statements indicates an aspect of quality. Criteria for which a negative response indicates quality have been changed (A4, A8, and B4). One criterion not relevant to this area of study has been removed (A5) since treatment does not typically begin until verification has been made through pathology results. Criterion A8 of observer variation is similar to criteria E1-E7 of observer characteristics and has been removed. Only instrument variation, specifically the reporting form used to interpret the images, now comprises A8. For criterion C2, participants are changed to images while number of observers (screen readers) comprises E2. The inclusion in D5 of ROC analysis effectively combines sensitivity and specificity values, and in D7 patients are replaced by images and observers.

Additional criteria

Criterion D2 now provides further detail to allow replicability as well as identify variables which influence diagnostic accuracy to further appropriate the tool to this area of study. Details included are related to the screen reading process and include: number of images read at one sitting, how images were selected, degree of difficulty of interpretation, details of types of breast cancer, time taken to read and environmental conditions such as lighting and type of monitors.

Evidence for criteria

This appropriation has been carried out using evidence from the literature: van den Biggelaar et al. 16 (A1, D5), Brennan et al. 24, Reed et al. 25 (D2), Brealey and Westwood (E1-E7), in addition to using details of the breast screening process contained within the BreastScreen Australia National Accreditation Standards (NAS) 1. The NAS is not only based on rigorous international evidence relating to best practice 1, but also encourages the research design in these studies to mimic the real-life environment of screen reading and consequently provide
the most clinically useful outcomes. One aspect of the screen reading process which is typically impractical for research purposes is screen reading consecutive populations. This has led to the use of test sets in research studies. However for these studies to be clinically useful a correlation between test set results and real-life clinical results is essential.

**Test sets and clinical practice**

Much debate surrounds the testing of diagnostic accuracy using test sets which have artificially inflated breast cancer prevalence versus consecutive screening images which mimic the real-life clinical situation. Minimal or no correlation between test set outcomes and clinical outcomes has been identified by Scott Evan, Gale et al. 26, Rutter and Taplin 27. Gur, Bandos, Cohen et al. reported a significant difference between performance in the clinic than completing test sets 28. A study by Pauli et al. found a strong correlation between test set outcomes and consecutive screening outcomes when used together in the same research design 12. These studies however used varying numbers of breast cancers, images and types of breast cancer to comprise the test set.

This variation can be overcome by the use of a validated test set such as PERFORMS (Scott and Gale) 29 and BREAST (Brennan, Lee and Tapia) 30 which increases the rigour of the study and provides consistency in the important aspects of study as spectrum composition, spectrum of images and spectrum of disease (B1-B3). The degree of difficulty in terms of types of cancers, proportions of breast density and numbers of images read would be consistent, and so comparisons between study outcomes could be more readily applied.

Incorporating validated test sets into the quality evaluation tool specifically developed to evaluate screen reading accuracy, may well lead to an identification and understanding of the specific causal agents for any lack of correlation between clinical audits and screen reading test sets, which as Soh et al. state is needed to facilitate the process of evaluating the diagnostic accuracy of screen readers in practice 31.
Conclusion

This reported study has identified the quality issues which compromise the validity and reliability of outcomes of studies which focus on radiographers’ accuracy in screen reading. During the process of evaluating these studies a quality assessment tool specifically for evaluating the quality of studies investigating the diagnostic accuracy of screen readers has been derived. This tool, with further refinement and validation will make a contribution to promoting well-designed studies in this important area of research and practice.
References:


Presentations arising from this study

Local Presentation. 9th August, 2010: *Westmead Breast Cancer Institute Tumour Board Meeting*, Westmead. NSW.

International Conference Presentation. 12th September, 2010. *16th International Society of Radiographers and Radiologic Technologists World Congress*. Gold Coast. QLD.


Local Presentation. 18th November, 2010. *Westmead Breast Cancer Institute Staff Training Day*. Westmead. NSW.

National Conference Presentation. 16th April, 2011. *8th Annual Scientific Meeting of Medical Imaging and Radiation Therapy*. Adelaide. SA.


National Conference Presentation. 21st April, 2012. *9th Annual Scientific Meeting of Medical Imaging and Radiation Therapy*. Sydney. NSW.

International Conference Presentation. 9th June, 2012. *17th International Society of Radiographers and Radiologic Technologists World Congress*. Toronto, Canada.

List of Tables

Table 1- BSA recommendation requirement and BI-RADS® classification .......................13
Table 2- Studies reviewed ....................................................................................................19
Table 3- Study setting and number of images ......................................................................22
Table 4- Sample composition, cancer prevalence and lesion proportion ..............................22
Table 5- Reported lesion proportion of recalled images in Duijm et al. (2007, 2008, 2009) studies ........................................................................................................................................25
Table 6- Number and mammographic experience of screen readers ..................................26
Table 7- Gold Standard ........................................................................................................27
Table 8- Screen reading process ..........................................................................................28
Table 9- Data analysis ...........................................................................................................31
Table 10- Screen reading performance- sensitivity and specificity (where applicable post-training results are reported) ........................................................................................................32
Table 11- Screen Reading Performance of varying screen reading combinations of Duijm et al. studies ..................................................................................................................33
Table 12- ROC classification guide .....................................................................................34
Table 13- Distribution of benign and malignant lesion proportions .....................................41
Table 14- Data abbreviations ...............................................................................................47
Table 15- Number of mammograms read by each screen reader ........................................48
Table 16- Varying thresholds of positive and negative .........................................................50
Table 17- Schedule of radiographer screen reading .............................................................53
Table 18- Individual screen reader accuracy-with BI-RADS® classification of '3, 4 or 5' considered positive .....................................................................................................................55
Table 19- Individual screen reader accuracy-with BI-RADS® classification of '4 or 5' considered positive ..........................................................57
Table 20- Pooled sensitivity and specificity (including 95% CI) at each threshold ...............60
Table 21- Analysis of accuracy by radiographic appearance of lesion type ..........................61
Table 22- Learning effect- accuracy at two time intervals as more mammograms screen read .................................................................................................................................62
Table 23- Comparison of screen reader accuracy using Areas Under the Curve (AUC) (including 95% CI) ....................................................................................................................62
List of Figures

Figure 1- Normal mammogram- Cranio-Caudal view on left and Medio-Lateral Oblique view on right ..............................................................10

Figure 2- Magnified malignant presentation- calcifications .................................................10

Figure 3- Magnified malignant presentation- stellate lesion ................................................11

Figure 4 - Magnified malignant presentation- architectural distortion ..............................11

Figure 5- Data extraction .....................................................................................................20

Figure 6- Method of image test set compilation ..................................................................40

Figure 7- Radiographer screen reader experience .............................................................52

Figure 8- Screen reader sensitivity .....................................................................................55

Figure 9- Screen reader specificity ....................................................................................56

Figure 10- ROC scatter plot of screen readers 1-10 ............................................................56

Figure 11- Recall rates .........................................................................................................58

Figure 12- ROC for each screen reader ..............................................................................59

Figure 13- Pooled estimates of sensitivity and specificity for threshold of screen reading of BI-RADS® classification of '3, 4 and 5' versus '1 and 2' with estimated ROC ..................60
Chapter 1: Introduction
This thesis aims to evaluate the accuracy of radiographers screen reading mammograms. The goal of the BreastScreen Australia (BSA) program of reduction of breast cancer mortality and morbidity through early detection is being compromised due to radiologist workforce shortages. These workforce shortages contribute to delays in women receiving their screening mammogram results (ABC News, 2004, 2005; The Australian, 2011). This problem of delays in screening mammogram results is anticipated to increase in the future as a result of the ageing population, leading to further health and skilled labour shortages (Australian Government Productivity Commission, 2005; Australian Government, 2010; Commonwealth of Australia, 2010). Simultaneously, the ageing population will continue to increase the volume of eligible women participating in the program, further compounded by the recommendation to increase the breast screening target age from the current 50 to 69 years to a target of 45 to 74 years of age (Commonwealth Department of Health and Ageing Australia, 2009; Moran & Warren-Forward, 2012).

A strategy was applied in the United Kingdom (UK) as a solution to similar breast screening program delays due to radiologist shortages and high screening volume. Mammogram volume increased as a result of the introduction of multiple factors, including double screen reading to increase accuracy in the mid 1990’s, two-view mammography in 2003, and a review in maximum target age range from 65 to 70 years of age in 2004 (Pauli, Hammond, Cooke, & Ansell, 1996a; Price, Miller, & Mellor, 2002; Bennett, Sellars, Blanks, & Moss, 2011). In the UK radiographers have undertaken training and been employed alongside radiologists as second screen-readers (Price et al., 2002; National Health Scheme Cancer Screening Programmes, 2011). This has been successful in ensuring the volume of mammograms were accurately and efficiently double screen read in the UK breast screening program (Wivell, Denton, Eve, Inglis, & Harvey, 2003; Bennett et al., 2011). This strategy was supported by the 2000 National Health Scheme Cancer Plan that recommended working practices were based upon skills and ability rather than profession (United Kingdom. Department of Health, 2000; National Health Scheme Cancer Screening Programmes, 2011). This has been further
developed into double radiographer screen reading in some units due to its’ success (Bennett et al., 2011; National Health Scheme Cancer Screening Programmes, 2011). This same strategy can potentially be applied within the Australian context to maintain the goal of early detection of breast cancer in the BSA program. However, prior to radiographers taking on the role as one of two screen readers in an Australian setting, it is critical to evaluate the accuracy of these radiographers as screen readers.

Acceptable radiographer screen reading accuracy levels, when compared to an appropriate gold standard of pathology results and a minimum 1-year follow-up, have been reported in previous, predominantly international, screen reading studies (Haiart & Henderson, 1991; Bassett et al., 1995; Pauli et al., 1996a; Tonita, Hillis, & Lim, 1999). Furthermore, there is evidence that the addition of radiographers as screen readers increased cancer detection rates (Duijm, Groenewoud, Fracheboud, & de Koning, 2007; Duijm et al., 2008; Duijm et al., 2009).

In their systematic review, however, van den Biggelaar, Nelemans, & Flobbe (2008) identified very few well-designed studies evaluating the accuracy of radiographers screen reading. These authors stated the need for a well-designed study using an appropriate gold standard, consisting of known pathology and follow-up of a minimum of 1-year rather than the radiologist reports (2008, p. 92). The question remaining, therefore, is what does constitute a well-designed study? There are many characteristics to consider and clearly one of the most important in this type of study is the evaluation of accuracy through the comparison to an appropriate gold standard. A limited number of previous international studies did use appropriate gold standards (Haiart & Henderson, 1991; Bassett et al., 1995; Pauli et al., 1996a; Pauli, Hammond, Cooke, & Ansell, 1996b; Tonita et al., 1999), but were lacking in other important areas.

There were no Australian studies conducted prior to 2008; since the review of van den Biggelaar et al. (2008) was published, however, two Australian pilot studies have reported comparable radiographer accuracy levels to these previous international studies (Holt &
Pollard, 2010; Moran & Warren-Forward, 2010). These results are encouraging and clearly indicate the need for a large well-designed study to evaluate the accuracy of radiographers screen reading in the Australian context. Both Australian pilot studies failed to use appropriate gold standards (Holt & Pollard, 2010; Moran & Warren-Forward, 2010). The design of this Australian study aims to improve on previous studies of the evaluation of accuracy through comparison to a rigorous gold standard, while applying rigorous study design characteristics.

1.1. Aim

The aim of this thesis is to undertake a well-designed Australian study to evaluate the accuracy of radiographers screen reading mammograms.

The objectives to address this aim are:

1. To identify the characteristics of a well-designed study in diagnostic accuracy in screen reading of breast images.
2. To evaluate the accuracy of radiographers as screen readers in an Australian setting.
3. To determine accuracy according to lesion types.
4. To identify the effect of learning on accuracy.

1.2. Structure of thesis

This thesis evaluates the accuracy of radiographers as screen readers. Prior to the actual thesis, an article for publication and a list of presentations arising from this research have been included. The article submitted for peer review to the European Journal of Radiology reports on the identification of the specific characteristics of a well-designed study in diagnostic accuracy of screen readers and the associated contribution of these characteristics to the development of a quality evaluation tool. Chapter 2 provides the descriptive material comprising the background context to this thesis and includes details of the workforce issues forming the rationale for the thesis, as well as specific details of the screening process. Chapter 3 presents a critical evaluation of the literature which provides justification for the
methods employed in the research study described in Chapter 4. Chapter 5 reports the results of the study evaluating accuracy of radiographers as screen readers; discussions of these findings are located in Chapter 6, while a summary and evaluation of outcomes, together with conclusions drawn, are summarized in Chapter 7.

It needs to be noted that in this thesis I refer to the radiographer readers of the screening mammograms in this reported study as *screen readers*, and to the process of reading screening mammograms as *screen reading*. 
Chapter 2: Background
The goal of international population-based mammography screening is to reduce morbidity and mortality of breast cancer through early detection (Australia. Department of Health and Ageing; Forrest, 1986; World Health Organization (WHO), 2002, 2006). This goal of early detection of breast cancer is dependent upon high quality images and accurate, timely screen reading. This chapter firstly describes the problem of workforce shortages which underpins the research focus of this thesis; shortages of radiologists within screening programs are undermining the timeliness of screen reading. Secondly, the context within which screen reading takes place is described. This includes the images produced and the screen reading process, which can be further subdivided into the reporting form and the screen reading environment. Thirdly, the individual variables of screen readers which influence diagnostic accuracy are presented, together with evidence of degrees of difficulty when identifying specific types of abnormalities. Finally, the current roles of radiographers in the screening process are identified.

2.1. Workforce shortages

Radiologist workforce shortages are evident in Australia and overseas (Jones, 2000; Bhargavan, Sunshine, & Schepps, 2002; Royal Australian New Zealand College of Radiologists (RANZCR), 2006; European Society of Radiology, 2008; The Royal College of Radiologists; Australian Diagnostic Imaging Association (ADIA) n.d.). In Australia, the current and anticipated demand for radiologists increasingly exceeds the supply. In 2010 there were 74 radiologists providing services per million of the Australian population, while international levels average 100 radiologists per million of the population. Furthermore, the projected need of 1000 additional radiologists by 2021 to reach international average levels is considered unachievable (Australian Medical Association (AMA), 2007; Royal Australian New Zealand College of Radiologists (RANZCR), 2010).

The current shortage has contributed to delays in women receiving their screening mammogram results, and undergoing diagnostic work-up in assessment clinics, as appropriate. One of the performance objectives of the BSA National Accreditation Standards
(NAS) states that women should receive their results from screening in a timely manner (BreastScreen Australia, 2008). This means a minimum of 90% of screened women should have a letter sent to them within 14 days, and that all women should be notified of their results within 28 days (BreastScreen Australia, 2008). These standards, however, are being compromised, and there have been reported delays in women receiving their results due to radiologist shortages in Australia. These workforce shortages can be attributed to a number of causes including the retiring and ageing workforce; work-lifestyle balance; female, often child-rearing participation in the workforce; heavy workloads; and interventional radiology demands (Jones, 2000; Royal Australian New Zealand College of Radiologists (RANZCR), 2010).

To address these radiology shortages, strategies have been suggested for solving the problem; for example, Smith and Baird (2007) suggested that radiographers could alleviate some of these workforce pressures by taking on some reporting roles, a successful solution to similar shortages overseas (Smith & Baird, 2007). They state that this strategy has reduced patient waiting times for reporting in the UK, and that accuracy is comparable to that of radiologists.

2.1.1. UK experience

Radiographer screen reading was implemented in 1989 in the UK National Health Service Breast Screening Programme (NHSBSP) in response to pressures of high screening volume that co-existed with radiology workforce shortages (Price et al., 2002; Wivell et al., 2003). This strategy of radiographers screen reading alongside radiologists as one of two screen readers continues to successfully maintain accuracy and timeliness of the NHSBSP. In some units, a further development of double radiographer screen reading has been introduced successfully (United Kingdom. Department of Health, 2000; Bennett et al., 2011; National Health Scheme Cancer Screening Programmes, 2011). A similar strategy of radiographer screen reading could be utilised in Australia to maintain timeliness of the BSA program. The pressures that the UK screening program is facing are partially due to workforce shortages and target age range increases (The Royal College of Radiologists, 2010; Bennett et
In the Australian breast screening program, current pressures include workforce shortages and an ageing population (Moran & Warren-Forward, 2012). Further pressures are anticipated, not only due to the ageing population, but also due to the recommendation for increases in mammogram target age range. If radiographers take on some screen reading roles, there is the potential to optimise the timeliness of the BSA program.

2.2. Context of screen reading

The context of screen reading involves the normal and abnormal breast images that are produced by the radiographers for screen reading. The next step of the process of screen reading requires the recording of the screen readers’ observations on a reporting form. The screen reading environment where this recording takes place makes a contribution to the overall accuracy of the process (W Reed, Poulos, Rickard, & Brennan, 2009).

2.2.1. Images produced

Mammogram images are radiographs of breast tissue presenting as dense, glandular (white) tissue and less dense, fatty (shades of grey to black) tissue. Mammogram images demonstrating malignant breast lesions may be identified through the presence of calcifications (white specks); discrete (small, vague) masses; stellate (star-shaped) lesions; architectural distortion (irregular tissue pattern); or non-specific (asymmetric) densities (Bassett, 2000; Tabar, Tot, & Dean, 2005). The standard mammogram images in Australia are comprised of the Cranio-Caudal (CC) view and the Medio-Lateral Oblique (MLO) view. Figure 1 presents a normal mammogram, comprised of two images of each breast, specifically the cranio-caudal and the medio-lateral oblique. Figures 2, 3 and 4, present magnified views of examples of malignant breast lesions, namely calcifications, stellate lesions and architectural distortion.
Figure 1- Normal mammogram- Cranio-Caudal view on left and Medio-Lateral Oblique view on right
Source: (University of Washington Radiology, 2007-2008)

Figure 2- Magnified malignant presentation- calcifications
Source: (Philpotts, 2009)
Figure 3 - Magnified malignant presentation- stellate lesion

Source: (Alexander, Yankaskas, & Biesemier, 2006)

Figure 4 - Magnified malignant presentation- architectural distortion

Source: (Bassett, 2000)
These images have inherent features contributing to challenges in screen reading accuracy. Small differences in contrast between normal and abnormal tissue create difficulties in perception of important lesion details, such as margins and lucency (Tabar et al., 2005; Kopans, 2006). Individual variations in proportion of breast density have been found to influence accuracy in interpretation (Carney et al., 2003; Wang, Xw, Li, Huang, & Tang, 2012). In addition to the perception of breast tissue creating interpretation difficulty, sub-optimal positioning techniques have been identified as a major contributor to accurate visualisation of abnormalities on mammogram images (Taplin et al., 2002). It has therefore been necessary to create a screen reading process that addresses these challenges.

2.2.2. Screen reading process

During the screen reading process, once the images are deemed acceptable for reading, each of two readers, currently predominantly radiologists in Australia, typically blindly and independently views the images to increase accuracy (BreastScreen Australia, 2008). If no consensus is reached, a third independent reader typically makes the final decision (BreastScreen Australia, 2008). High screen reading sensitivity (true positive rate) is imperative for early breast cancer detection, and it has been established that double independent screen reading detects 10-15% more cancers than a single reader (Anderson, Muir, Walsh, & Kirkpatrick, 1994; Thurfjell, Lernevall, & Taube, 1994; Beam, Sullivan, & Layde, 1996; Blanks, Wallis, & Moss, 1998; Dinnes et al., 2001; Georgian-Smith et al., 2007; Gromet, 2008; Caumo et al., 2011). Each screen reading is recorded on the screen reading reporting form.

Screen reading reporting form

The BSA program requires both independent reads of each mammogram to be combined into a single recommendation as either Normal (1), or Suspicious (2) (NSW Program for Mammographic Screening, 1995; BreastScreen Australia, 2008), as presented in Table 1. Many health services within the BSA program employ the practice of the reader indicating their screen read on a standardised reporting form using a modified Breast Imaging Reporting
and Data System (BI-RADS®) classification lexicon of ‘1’ (no lesion); ‘2’ (benign lesion); ‘3’ (probably benign); ‘4’ (probably malignant); or ‘5’ (malignant) (American College of Radiology, 2003; Obenauer, Hermann, & Grabbe, 2005) as presented in Table 1. (See Appendix A for an example of a reporting form).

Table 1- BSA recommendation requirement and BI-RADS® classification

<table>
<thead>
<tr>
<th>BSA recommendation requirement</th>
<th>Modified BI-RADS® lexicon of mammogram classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Normal</td>
<td>‘1’ (no lesion)</td>
</tr>
<tr>
<td></td>
<td>‘2’ (benign lesion)</td>
</tr>
<tr>
<td>2-Suspicious</td>
<td>‘3’ (probably benign)</td>
</tr>
<tr>
<td></td>
<td>‘4’ (probably malignant)</td>
</tr>
<tr>
<td></td>
<td>‘5’ (malignant)</td>
</tr>
</tbody>
</table>

In the BSA program, therefore, a screen read of ‘1’ (no lesion) or ‘2’ (benign lesion) equates to a recommendation of Normal (1). Mammogram screen reads of ‘3’ (probably benign), ‘4’ (probably malignant) and ‘5’ (malignant) equate to a recommendation of Suspicious (2), as presented in Table 1.

Following screening, when a mammogram has a recommendation of Suspicious (2), the woman is recalled for follow-up at an assessment clinic (recall to assessment) where the nature of the lesion is then determined. The significance of this screen reading and subsequent assessment process, is that the presence of an abnormality is determined at the time of screen reading, while its’ nature is then determined following further investigation at an assessment clinic (BreastScreen Australia, 2008).

Minimising unnecessary recall of women to assessment is essential in reducing potential anxiety levels. This is achieved by keeping false positive rates as low as possible and simultaneously maximising screen reading accuracy (BreastScreen Australia, 2008). When a Normal (1) recommendation is made the woman is returned to the routine 2- year screening interval of the program.

The ability of the screen reader to accurately distinguish between abnormal and normal mammograms is the essential requirement of screen reading and has been found to be related
to the screen reading environment and the individual variables of the screen readers themselves (W Reed et al., 2009).

**Screen reading environment**

The screen reading environment plays an important role in influencing decision making in the diagnostic process (Krupinski & Jiang, 2008; W Reed et al., 2009). The images in screen reading populations comprise a small proportion of abnormal images among hundreds of normal images, contributing to screen reading accuracy difficulties (Andolina & Lille, 2011). Screen reader concentration on the task at hand is therefore critical for accuracy levels. Concentration can be influenced by the busyness of images; noise; lighting levels; and ergonomics (DeLong, DeLong, & Clarke-Pearson, 1988; Prabhu, Gandhi, & Goddard, 2005; Brennan et al., 2008; Pollard et al., 2009; W Reed et al., 2009). Busy images may contain a range of density variations due to fatty and dense breast tissue combinations that potentially need more time and concentration to read (W Reed et al., 2009). Noise levels need to be kept to a minimum to maximise reader concentration. BSA NAS state that lighting levels need to be \( \leq 50 \) lux to optimise viewing conditions (BreastScreen Australia, 2008; W Reed et al., 2009). Ergonomic factors to consider to minimise lapse in reader concentration include computer, chair and desktop heights in relation to the readers, and eye strain potential (Prabhu et al., 2005). These environmental conditions potentially influence the abilities of the screen readers when detecting abnormalities.

**2.3. Diagnostic accuracy**

Diagnostic accuracy is dependent upon the screen readers and their ability to accurately detect abnormalities.

**2.3.1. Screen readers**

Individual variables of screen readers have been found to influence accuracy. Screen reader experience, combined with ability to detect abnormalities and report them correctly, contribute to the accuracy of screening programs (Nodine et al., 1999; Nodine, Mello-Thoms,
Kundel, & Weinstein, 2002; Joy, Penhoet, & Petitti, 2005; International Cancer Screening Network, n.d.). Screen reading experience and cancer detection ability is linked to the volume of mammograms read over a given period of time (Schmidt, Hartwagner, Spork, & Groel, 1998; Kan, Olivotto, Warren Burhenne, Sickles, & Coldman, 2000; BreastScreen Australia, 2008). This is the reason the BSA NAS state that screen readers must read a minimum of 2000 mammograms per year (BreastScreen Australia, 2008). Variability of individual reader accuracy may also be influenced by management of fatigue levels, dependent upon factors such as screen reading volume, reporting time of day, or whether or not reporting is following a full day of reporting elsewhere (Scott & Gale, 2007; W Reed et al., 2009). Screen readers also experience degrees of difficulty in the detection of specific abnormalities.

2.3.2. Abnormality detection

Specific abnormality characteristics in mammograms have varying degrees of detection difficulty (Yankaskas, Schell, Bird, & Desrochers, 2001; Beam, Conant, & Sickles, 2002). Calcifications have been associated with high false negative rates and may be overlooked due to the difficulty in determining normal or suspicious appearance (Burrell, Evans, Wilson, & Pinder, 2001; Muttarak, Kongmebhol, & Sukhamwang, 2009). Stellate lesions may be difficult to detect due to the surrounding tissue obscuring the spicules that radiate outwards, often from a central mass (Gibbons, 1998; Kirwan, Denton, Nash, Humphreys, & Michell, 2000; Cherel, Becette, & Hagay, 2005). Architectural distortion has a subtle appearance and may present similarly to normal overlapping breast tissue, causing it to be a commonly missed cancer and false negative presentation (Knutzen & Gisvold, 1993; Burrell et al., 2001; Yankaskas et al., 2001; Banik, Rangayyan, & Desautels, 2011; Shaheen, Schimmelpenninck, Stoddart, Raymond, & Slanetz, 2011; Banik, Rangayyan, & Desautels, 2012). These inherent lesion detection difficulties contribute to the overall accuracy of breast screening programs. Another factor to consider during the screening process is the current role of the radiographer.
2.4. Current radiographer roles in the screening process

Currently in Australia, radiographers are responsible for evaluating their images according to set criteria using the PGMI (Perfect, Good Moderate, Inadequate) method of evaluation of clinical image quality (BreastScreen Australia, 2008). Initially, during the screen reading process, radiographers evaluate the daily mammogram images they produce to ensure optimal screen reading quality. Current image quality criteria include correct exposure; adequate breast tissue visualisation; and absence of skin folds (National Health Scheme Breast Screening Programme, 1998; BreastScreen Australia, 2008). While reviewing images for screen reading quality acceptance, radiographers are typically viewing a large volume of mammogram images, especially since digital imaging has expanded in Australia. Furthermore, while working in assessment clinics, where additional investigative work-up views are undertaken, radiographers are informally gaining mammogram screen reading ability daily. Radiographers may also be encouraged to sit-in during reporting sessions in some workplace settings, potentially impacting on individual accuracy levels.

In the UK, the role of radiographers in screening programs has expanded to include screen reading. The majority of research investigating accuracy levels of radiographers as screen readers has therefore emanated from the UK and Europe (Haiart & Henderson, 1991; Pauli et al., 1996a, 1996b; Wivell et al., 2003; Duijm et al., 2007; Duijm et al., 2008; Duijm et al., 2009). It is essential to undertake a detailed literature review incorporating an evaluation of these studies to determine the optimal design for the research study reported in this thesis. This literature review is presented in Chapter 3.
Chapter 3: Literature Review
The purpose of this literature review was to undertake a critical analysis of previous studies investigating the accuracy of radiographers as screen readers. A literature search and data extraction was carried out, followed by an evaluation of study components. This evaluation provides justification for the method reported in Chapter 4, as well as the rationale for the research study reported in Chapter 5.

**Literature search**

A literature search was undertaken using the Medline, Pubmed and Cinahl databases, with various combinations of the keywords mammogram, radiographer, technologist, accuracy, screen reading and interpretation. There were no limits applied for publication dates. An initial review of titles and abstracts enabled the exclusion of papers that were clearly not relevant to the subject of interest. The inclusion criteria used to determine accepted papers were that studies must have investigated the accuracy of radiographers screen reading mammograms. Further papers were selected using manual searches of the reference lists within the accepted papers. A total of 13 studies published between 1991 and 2010 were identified and are presented in Table 2.
Table 2- Studies reviewed

<table>
<thead>
<tr>
<th>Authors</th>
<th>Country of Study</th>
<th>Study Title - arranged chronologically.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bassett et al., 1995</td>
<td>United States</td>
<td>Effects of a program to train radiologic technologists to identify abnormalities on mammograms.</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>United Kingdom</td>
<td>Comparison of radiographer / radiologist double film reading with single reading in breast cancer screening.</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>United Kingdom</td>
<td>Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions.</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>Canada</td>
<td>Medical radiologic technologist review: effects on a population-based breast cancer screening program.</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>United Kingdom</td>
<td>Can radiographers read screening mammograms?</td>
</tr>
<tr>
<td>Holt., 2006</td>
<td>Canada</td>
<td>Evaluating radiological technologists ability to detect abnormalities in film-screen mammographic images: a decision analysis pilot project.</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>Netherlands</td>
<td>Additional double reading of screening mammograms by radiologic technologists: impact on screening performance parameters.</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>Netherlands</td>
<td>Introduction of additional double reading of mammograms by radiographers: effects on a biennial screening programme outcome.</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>Netherlands</td>
<td>Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome.</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>Australia</td>
<td>Radiographers‘ ability to perceive and classify abnormalities on mammographic images-results of a pilot project.</td>
</tr>
</tbody>
</table>

**Data extraction**

Data extraction involved identification of the components of the samples, methods, and results of the reviewed studies and are presented in Figure 5.
Evaluation of the literature

The samples, methods and results of the reviewed studies are presented below.

3.1. Samples in the reviewed studies

In diagnostic accuracy studies of screen readers, the sample comprises both the breast images read and the screen readers themselves.

3.1.1. Images

The images selected for reading in these reviewed studies were either actual consecutive screening populations or constructed image test sets compiled of screening or diagnostic mammogram images. Screening images are typically those of asymptomatic women, whereas
diagnostic images are typically of women who present with a symptom. True consecutive screening population image sets, read prospectively, are the ideal representative sample, however compiled image test sets that have been previously read can also provide clinically useful outcomes. The setting and number of images used in the study, together with the sample composition, cancer prevalence and lesion proportion, influence how representative the constructed image test set sample is to the true screening population under consideration.

**Study setting and number of images**

The study setting and number of images read in the reviewed studies are presented in Table 3. Studies were predominantly undertaken in the screening setting, with the exception of Holt’s study undertaken in the diagnostic setting (Holt, 2006). Holt (2006) selected diagnostic images to minimise bias potential due to the obvious difference of diagnostic and screening image labeling in their facility. There was considerable sample size variation in number of images read, ranging from two pilot test set studies comprising 50 images (Holt, 2006; Moran & Warren-Forward, 2010) through to true consecutive population studies comprising up to 106 093 mammograms (Duijm et al., 2009). There were five constructed test set studies and the small sample size of three pilot studies of less than 60 images is expected, however, one test set study with a small sample size of 79 images has limited generalisability. The true consecutive population studies, and the constructed test set of Bassett et al. (1995), were comprised of larger sample sizes, maximising the generalisability of these studies.
Table 3- Study setting and number of images

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting</th>
<th>Number of images read</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>Screening</td>
<td>3 362 by radiographer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 080 by radiologist</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>Screening</td>
<td>627 post training</td>
</tr>
<tr>
<td></td>
<td></td>
<td>611 pre training</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>Screening</td>
<td>17 202</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>Screening</td>
<td>79 test set</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17 202 consecutive</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>Screening</td>
<td>27 863</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>Screening</td>
<td>1 000 test set</td>
</tr>
<tr>
<td></td>
<td></td>
<td>54 000 consecutive</td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td>Screening</td>
<td>3 019</td>
</tr>
<tr>
<td>Holt, 2006</td>
<td>Diagnostic</td>
<td>50</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>Screening</td>
<td>61 251</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>Screening</td>
<td>66 225 by radiologists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>78 325 by radiologists and radiographers</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>Screening</td>
<td>106 093</td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td>Screening</td>
<td>50</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>Screening</td>
<td>60</td>
</tr>
</tbody>
</table>

Test sets versus consecutive populations

Sample composition, relating to consecutive population screening or tests sets, is presented in Table 4. Screen readers participating in consecutive screening enables true accuracy evaluation, however, the screen reading of thousands of images needed to produce meaningful results is clearly an onerous task due to the average of one cancer detected per 200 women screened (BreastScreen Australia, 2008). On the other hand, constructed test sets may alleviate the need for high reading volume; however, producing results that are applicable to the screening population is highly dependent upon the test set being constructed with an appropriate range of representative images.

Table 4- Sample composition, cancer prevalence and lesion proportion

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Composition</th>
<th>Cancer Prevalence</th>
<th>Lesion Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>Consecutive</td>
<td>0.45% / 0.38%</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>Test Set</td>
<td>7.7% / 6.9%</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>Consecutive</td>
<td>0.8%</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>Test Set / Consecutive</td>
<td>20.3% / Not Reported</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>Consecutive</td>
<td>Not Reported</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>Consecutive</td>
<td>Not Reported</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td>Reader Selection</td>
<td>Not Reported</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Holt, 2006</td>
<td>Pilot Test Set</td>
<td>14%</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>Consecutive</td>
<td>0.76%</td>
<td>Reported</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>Consecutive</td>
<td>Not Reported</td>
<td>Reported</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>Consecutive</td>
<td>0.73%</td>
<td>Reported</td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td>Pilot Test Set</td>
<td>36%</td>
<td>Not Reported</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>Pilot Test Set</td>
<td>Not Reported</td>
<td>Not Reported</td>
</tr>
</tbody>
</table>
Cancer prevalence and lesion proportion

There has been some debate and ambiguity surrounding the degree of accuracy correlation between test sets and consecutive population screening. Pauli et al. undertook an accuracy study utilising a test set to assess performance at regular intervals in a consecutive population study and demonstrated that good results from test set screen reading are transferable to reliable consecutive population screen reading (Pauli et al., 1996b). Scott et al. undertook a study to examine the relationship between test sets and true screen reading practice and reported some correlation (Scott, Evans, Gale, Murphy, & Reed, 2009), while Gur et al. (2008) reported that ‘The “Laboratory” Effect’ (p. 47) during retrospective screen reading led to consecutive screen reading accuracy in the clinical environment in fact being better than test set results. Soh et al. (2012) indicated that there is a need for more understanding and identification of the causes of any correlation ambiguity between test sets and consecutive population screening. It is possible that this ambiguity may be due to cancer prevalence and lesion proportion in the constructed test set.

The sample composition, cancer prevalence and lesion proportion of the image sets used in the reviewed studies are presented in Table 4.

Cancer prevalence

The study image sets were predominantly composed of consecutive populations or cancer-enriched image test sets. Four test set studies reported inflated cancer prevalence ranging between 6.9% and 36% (Bassett et al., 1995; Pauli et al., 1996b; Holt, 2006; Moran & Warren-Forward, 2010), while eight studies were comprised of consecutive populations with true cancer prevalence ranging between 0.38% to 0.8% (Haiart & Henderson, 1991; Pauli et al., 1996a, 1996b; Tonita et al., 1999; Wivell et al., 2003; Duijm et al., 2007; Duijm et al., 2008; Duijm et al., 2009). One study allowed the readers to select mammograms they wished to interpret; potentially creating bias through selection of less difficult images they felt sufficiently confident to read (Sumkin et al., 2003).
Cancer prevalence was not reported in five studies, reducing replicability of these studies (Tonita et al., 1999; Sumkin et al., 2003; Wivell et al., 2003; Duijm et al., 2008; Holt & Pollard, 2010). Reader awareness of cancer prevalence was identified in Wivell et al.’s (2003) study and could increase expectation bias, potentially inflating the false positive rate. No other studies reported reader awareness of cancer prevalence, potentially improving study rigour. Consecutive screening populations exhibit low cancer prevalence, although readers remain unaware of exact levels at the time of screen reading. Prevalence expectation has not been found to effect accuracy levels in experienced radiologist readers, thus supporting the use of cancer-enriched image test sets in accuracy studies (W Reed, Ryan, McEntee, Evanoff, & Brennan, 2011). It is possible however, that with inexperienced readers, prevalence expectation may have a greater influence. Furthermore, many radiographers have limited screen reading experience. It would therefore be prudent to minimise potential expectation bias in future accuracy studies using cancer-enriched image test sets by ensuring screen readers are kept unaware of test set cancer prevalence.

**Lesion proportion**

Lesion proportions were reported in the Duijm et al. (2007, 2008, 2009) studies, increasing replicability, and these are presented in Table 5. The reporting of lesion proportion is important to determine the generalisability of image test sets to the consecutive screening population. Ideal image sets include a representation of all lesion types, ranging from subtle to more obvious presentations. It is important to know the lesion types being read to enable analysis of specific lesion type accuracy and to determine the possible causes of inaccuracy. Reporting of both cancer prevalence and lesion proportion within the sample composition are strengths of the Duijm et al. (2007, 2008, 2009) studies; however, further detailed analysis of reader accuracy in relation to lesion type detection was not undertaken.
Table 5- Reported lesion proportion of recalled images in Duijm et al. (2007, 2008, 2009) studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Densities</th>
<th>Calcifications</th>
<th>Densities with Calcifications</th>
<th>Assymmetric densities</th>
<th>Architectural distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duijm et al., 2007</td>
<td>67.7%</td>
<td>22.1%</td>
<td>6.2%</td>
<td>1.3%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>66.4%</td>
<td>22%</td>
<td>6.9%</td>
<td>1.5%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>67%</td>
<td>21.5%</td>
<td>6.7%</td>
<td>1.9%</td>
<td>3%</td>
</tr>
</tbody>
</table>

In Australia, radiographers have no experience of formal screen reading because radiologists have traditionally dominated in this role (Smith & Baird, 2007). It may therefore be appropriate and practically useful in research studies to use a carefully constructed representative image test set with inflated cancer prevalence incorporating adequate lesion proportion with varying levels of detection difficulty. This is likely to maximise applicability and generalisability of the results.

3.1.2. Screen readers

The profession, number of screen readers, and their mammographic experience, potentially influence accuracy levels and generalisability of results. The number of screen readers in these studies varied from 1 to 33 radiographers, and from 1 to 9 experienced radiologists, as presented in Table 6. Most studies used both radiographers and radiologists as screen readers, with three studies omitting to report radiologist numbers (Pauli et al., 1996a; Holt & Pollard, 2010; Moran & Warren-Forward, 2010), limiting replicability. Radiographer mammographic experience varied from 1 month to 16 years. There is no known evidence to suggest that mammographic experience influences screen reading accuracy levels through varying levels of exposure to viewing mammographic images. It is possible, however, that because of their role in the screen reading process, radiographers may become accustomed to detecting abnormalities. This is discussed in section 2.4. Table 6 presents the mammographic experience of the radiographer screen readers. Four studies omitted to report radiographer mammographic experience, limiting generalisability. Screen reader numbers and mammographic experience affect the generalisability of results to the radiographer/radiologist
population; however, their level of agreement or inter-observer variability is of considerable importance in determining the validity of the results and is discussed in section 3.3.

Table 6- Number and mammographic experience of screen readers

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of radiographers</th>
<th>Radiographer mammographic experience</th>
<th>Number of radiologists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>1</td>
<td>Not reported</td>
<td>1</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>8</td>
<td>1 year minimum</td>
<td>7</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>7</td>
<td>6-121 months</td>
<td>Number not reported</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>7</td>
<td>6-121 months</td>
<td>9</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>3</td>
<td>Not reported</td>
<td>5</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>3</td>
<td>Not reported</td>
<td>4</td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td>33</td>
<td>2-26 years</td>
<td>9</td>
</tr>
<tr>
<td>Holt, 2006</td>
<td>5</td>
<td>5-20 months</td>
<td>2</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>21</td>
<td>1-124 months</td>
<td>8</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>21</td>
<td>1-124 months</td>
<td>8</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>21</td>
<td>1-124 months</td>
<td>8</td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td>11</td>
<td>Not reported</td>
<td>Number not reported</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>12</td>
<td>3.5-16 years</td>
<td>Number not reported</td>
</tr>
</tbody>
</table>

3.2. Methods in the reviewed studies

The methods used in accuracy studies, in particular, the applied gold standard and the screen reading process, influence potential bias levels, study rigour and generalisability.

3.2.1. Gold standard

An appropriate gold standard is essential in studies determining screen reader accuracy. Gold standards used in the studies are presented in Table 7. An appropriate gold standard is defined by van den Biggelaar et al. (2008) in their systematic literature review. They state that a minimum of 1-year follow-up and pathology results or the use of a validated test set are appropriate gold standards (van den Biggelaar, Nelemans, & Flobbe, 2008). A validated test set ensures known outcomes. Follow-up confirms that a woman found to have a normal screening mammogram result has not since been diagnosed with breast cancer. This is recommended in the case of either a true missed cancer in the screening process or in the case of a cancer presenting within the screening interval. These interval cancers are known to develop within the normal screening interval (van Dijck, Verbeek, Hendriks, & Holland, 1993; Reitsma, Rutjes, Scholten, Bossuyt, & Zwinerman, 2005). In Australia, this screening
interval is recommended to be 2 years (BreastScreen Australia, 2008), so the author of this thesis suggests it would be more prudent to apply a gold standard of pathology results combined with a minimum 2-year follow-up, or a validated test set. In Australia, interval matching involves the matching of all cancers recorded by the New South Wales (NSW) Cancer Registry with the BreastScreen NSW database to ensure all breast cancers are recorded in the BreastScreen database (BreastScreen Australia, 2008). This process determines those mammograms assessed as normal in the screening process that are subsequently found to be malignant within the 2-year screening interval due to the reasons mentioned above (van Dijck et al., 1993).

Using this robust definition, five studies used an appropriate gold standard as presented in Table 7 (Pauli et al., 1996b; Tonita et al., 1999; Duijm et al., 2007; Duijm et al., 2008; Duijm et al., 2009), however many studies fell short of this ideal. One study (Sumkin et al., 2003) used the radiologists’ report as their gold standard despite the known variation in radiologist accuracy (Elmore, Wells, Lee, Howard, & Feinstein, 1994; Thurfjell et al., 1994; Gur et al., 2004; Duijm et al., 2009; W Reed et al., 2009). It is crucial that in future studies, this robust gold standard is applied when determining screen reader accuracy to maximise study rigour.

### Table 7- Gold Standard

<table>
<thead>
<tr>
<th>Study</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>Pathology + 1-year follow-up *</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>Biopsy + 1-year follow-up *</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>Pathology + 1.5-year follow-up *</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>Validated In-House Test Set</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>Pathology + 2-year follow-up *</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>3-year follow-up *</td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td>Radiologists report</td>
</tr>
<tr>
<td>Holt, 2006</td>
<td>1-year follow-up *</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>Pathology + 2-year follow-up *</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>Pathology + 2-year follow-up *</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>Pathology + 2-year follow-up *</td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td>Radiologist consensus + pathology</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>6-year follow-up *</td>
</tr>
</tbody>
</table>

* follow-up of woman to confirm that normal screening results remain unchanged

### 3.2.2. Screen reading process

The screen reading process used influences the rigour of the studies. Important aspects of the method design include the screen readers being blinded to the reports of other screen readers;
the availability, if appropriate, of previous comparison images; and the reading classification system used. These variables are presented in Table 8.

Table 8- Screen reading process

<table>
<thead>
<tr>
<th>Study</th>
<th>Blinded</th>
<th>Previous Images</th>
<th>Reading Classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>Yes</td>
<td>Not Reported</td>
<td>Normal/</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>Yes</td>
<td>No</td>
<td>Benign/Suspicious</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>Yes</td>
<td>Yes</td>
<td>Normal/Abnormal</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>Unknown</td>
<td>Yes</td>
<td>Recall/Return to Screening</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>Yes</td>
<td>Not Reported</td>
<td>Normal/Abnormal</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>Yes</td>
<td>Yes</td>
<td>Recall/Return to Screening</td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td>Yes</td>
<td>Yes</td>
<td>Follow-up or not</td>
</tr>
<tr>
<td>Holt, 2006</td>
<td>Yes</td>
<td>No</td>
<td>Scale of 1 → 4</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>No</td>
<td>Yes</td>
<td>Positive/Negative</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>No</td>
<td>Yes</td>
<td>Recall/No Recall</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>No</td>
<td>Yes</td>
<td>Positive/Negative</td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td>Yes</td>
<td>Yes</td>
<td>Scale of 1 → 5</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>Yes</td>
<td>Not Reported</td>
<td>Scale of 1 → 5</td>
</tr>
</tbody>
</table>

**Blinded screen reading**

The established practice of double independent reading in screening programs is undertaken to increase the accuracy of breast cancer detection when compared to single reading (Caumo et al., 2011), as indicated in section 2.2.2. Most studies (9 of 13) employed readers to independently and blindly screen read, minimising bias by eliminating the influence of the decisions of other readers. In the three studies of Duijm et al. (2007, 2008, 2009), however, the screen readers were not blinded. Another study referred cases with discrepancies between readers for a third opinion to a consultant radiologist (Tonita et al., 1999). This reader was blinded to the other opinions; however, it was not ensured that this consultant radiologist was not one of the original readers, potentially increasing bias. Independent, blinded screen reading is normal practice in the Australian setting (BreastScreen Australia, 2008) and therefore should be replicated in future studies.

**Previous images**

In consecutive population screening, previous images, when available, are typically viewed alongside current images for comparison to maximise accuracy, particularly minimising false positives. Clearly initial screening episodes do not allow for previous comparison images.
Methods of the reviewed studies vary as presented in Table 8, with two previous studies not providing any previous comparison images, potentially increasing screen reading difficulty and lowering performance (Bassett et al., 1995; Holt, 2006). A further three studies did not report previous image availability, limiting reproducibility (Haiart & Henderson, 1991; Tonita et al., 1999; Moran & Warren-Forward, 2010). Interestingly, most studies provided comparison images; however, there was no evidence this affected screen reading performance as accuracy levels remained comparable to those studies that failed to provide previous images. Nevertheless, previous images should be available to replicate the screening setting and maximise the potential for diagnostic accuracy.

Reading classifications

Reading classifications varied between studies and are presented in Table 8. Screen readers in 10 of the studies were instructed to read mammograms relatively non-specifically. The readers stated whether the area of interest was benign or suspicious, needed follow-up, or was normal or abnormal. A positive read was determined when a mammogram was classified as being suspicious, needing follow-up, or was abnormal. Three studies employed a specific classification scale similar to the BI-RADS® classification lexicon (Holt, 2006; Holt & Pollard, 2010; Moran & Warren-Forward, 2010) enabling Reader Operating Characteristic (ROC) analysis, unfortunately not undertaken. The use of the BI-RADS® classification lexicon described in section 2.2.2., maximises applicability to current work practice in many screening programs (American College of Radiology, 2003) and simultaneously enables more robust ROC accuracy evaluation as discussed below.

3.3. Results of the reviewed studies

3.3.1. Screen reader performance

Screen reader performance was determined in the reviewed studies using a range of statistical analyses, shown in Table 9. Rigorous reading performance analysis includes the evaluation of individual screen reader accuracy; combining or pooling screen reader accuracy to increase
the ability to generalise the study sample results to the population (van Houwelingen, Zwinderman, & Stijnen, 1993; Reitsma et al., 2005); and inter-observer accuracy. Individual and pooled accuracy encompasses sensitivity, specificity, ROC analysis, accuracy according to lesion type, and accuracy at varying levels of test positive thresholds.

Limited data analysis was identified in all of the reviewed studies. Seven studies reported individual sensitivity and specificity levels, while pooled data analysis was only undertaken by five studies. One study alone analysed the Area Under the ROC Curve (AUC) (Pauli et al., 1996b), however the remaining studies did not undertake ROC analysis. One study analysed lesion accuracy (Wivell et al., 2003), while another analysed threshold accuracy (Holt, 2006). No studies reported an analysis of inter-observer accuracy. Limited data analysis provides sub-optimal evaluation and understanding of reader accuracy, often with a potential associated limit in ability to generalise results to the population from the study sample. More rigorous data analysis is needed to remedy these deficiencies in future studies. This analysis needs to include individual sensitivity and specificity; pooled sensitivity and specificity across the readers; individual and pooled ROC and AUC; lesion type accuracy; accuracy at differing test positive thresholds; and inter-observer accuracy.
Table 9- Data analysis

<table>
<thead>
<tr>
<th>Study</th>
<th>Individual Sensitivity</th>
<th>Individual Specificity</th>
<th>Pooled Sensitivity</th>
<th>Pooled Specificity</th>
<th>AUC</th>
<th>Lesion Accuracy</th>
<th>Threshold Accuracy</th>
<th>Inter-Observer Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>✓</td>
<td>✓</td>
<td>✓ (Mean Only)</td>
<td>✓ (Mean Only)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Holt, 2006</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>✓</td>
<td>No</td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td>No</td>
<td>No</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Holt &amp; Pollard, 2010</td>
<td>No</td>
<td>No</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Sensitivity

Comparable sensitivity (true positive rate) ranges, of 61% to 91.42% for radiographers and 63.9% to 100% for radiologists, were identified in the reviewed studies, and are presented in Table 10. The single diagnostic setting study (Holt, 2006) reported comparable results to the consecutive screening studies despite an inflated positive rate. Two studies report equal sensitivity of radiologists and radiographers (Pauli et al., 1996a, 1996b) and it is pertinent to note that institution one of the Bassett et al. (1995) study reported higher radiographer sensitivity than radiologists. Notably, one study suggested that review by a radiologist and a radiographer could replace double-reading by two radiologists (Tonita et al., 1999). There did not appear to be any clear explanation for the range in accuracy levels of both radiographers.
and radiologists in the reviewed studies; however, differences in rigour of individual studies may have contributed. Variations of the size, setting, composition, cancer prevalence and lesion proportion of the image set, combined with screen reader experience, may potentially have influenced accuracy levels. In addition, the gold standard applied and the screen reading process itself, as well as analysis of accuracy, could account for accuracy differences. The need for consistency in study rigour to facilitate comparison between studies has been clearly demonstrated.

Table 10- Screen reading performance- sensitivity and specificity (where applicable post-training results are reported)

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Radiographers</td>
<td>Radiologists</td>
</tr>
<tr>
<td>Haiart &amp; Henderson, 1991</td>
<td>80%</td>
<td>83%</td>
</tr>
<tr>
<td>Bassett et al., 1995</td>
<td>Inst. 1</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>Inst. 2</td>
<td>84%</td>
</tr>
<tr>
<td>Pauli et al., 1996a</td>
<td></td>
<td>73%</td>
</tr>
<tr>
<td>Pauli et al., 1996b</td>
<td></td>
<td>83%</td>
</tr>
<tr>
<td>Tonita et al., 1999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wivell et al., 2003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sumkin et al., 2003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holt, 2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duijm et al., 2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duijm et al., 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moran &amp; Warren-Forward, 2010</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Specificity**

Lower radiographer specificity (true negative rate) of 45% to 99.1% compared to 81% to 99.2% for radiologists was demonstrated in six studies as presented in Table 10. Conversely, one study reported a slightly higher radiographer specificity of 99.1% compared to radiologists of 99% (Duijm et al., 2007). The generally lower specificity rates of radiographers indicate a superior ability of radiologists to discriminate between benign and malignant abnormalities; however this is not supported by all studies. Duijm et al. (2007, 2009) published studies comparing sensitivity and specificity using varying combinations of radiographer and radiologist positive reads, as presented in Table 11. In these studies a
positive read was determined when a reader decided further assessment was needed. It was concluded that recalling all radiographer positive findings would have increased the cancer detection rate whilst maintaining a low recall rate for further assessment, and that for this reason, triple screen reading of a radiologist and two radiographers may replace double radiologist screen reading. The disparate nature of these results is also potentially explained through the individual study rigour variations outlined earlier in the potential reasons for variations in levels of sensitivity.

Table 11- Screen Reading Performance of varying screen reading combinations of Duijm et al. (2007, 2009) studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duijm et al., 2007</td>
<td>77.8%(^a)</td>
<td>98.4%(^a)</td>
</tr>
<tr>
<td></td>
<td>74.2%(^b)</td>
<td>98.9%(^b)</td>
</tr>
<tr>
<td>Duijm et al., 2009</td>
<td>76.9%(^a)</td>
<td>98.5%(^a)</td>
</tr>
<tr>
<td></td>
<td>73.2%(^b)</td>
<td>99.0%(^b)</td>
</tr>
<tr>
<td></td>
<td>75.2%(^c)</td>
<td>98.6%(^c)</td>
</tr>
<tr>
<td></td>
<td>68.6%(^d)</td>
<td>99.1%(^d)</td>
</tr>
<tr>
<td></td>
<td>63.9%(^e)</td>
<td>99.2%(^e)</td>
</tr>
</tbody>
</table>

\(^a\) Recall of all radiographer positive reads
\(^b\) Recall of radiographer positive reads following radiologist reassessment
\(^c\) Recall of all positive reads following single radiologist and double radiographer screen reading
\(^d\) Radiologist double-screen reading
\(^e\) Radiologist single-screen reading

**ROC analysis**

ROC analysis is used in diagnostic accuracy studies to enhance analytical rigour. ROC analysis evaluates accuracy using a combination of sensitivity and specificity at various test positive thresholds. The AUC visually represents accuracy, while numerical AUC values represent accuracy, with 1.0 indicating perfect accuracy, and 0.5 being no better than chance. Professor Thomas Tape of the University of Nebraska (University of Nebraska Medical Center, n.d.) advocates the guide presented in Table 12 for classifying the accuracy of a diagnostic test using ROC (Tape, n.d). ROC analysis is possible with the use of the BI-RADS® classification lexicon as the reporting tool and takes into account each reader’s individual decision threshold (Joy et al., 2005).
Only Pauli et al. (1996b) carried out ROC analysis in the reviewed studies, despite extensive use in radiologist accuracy studies (Goddard, Gilbert, Needham, & Deans, 1998). They reported AUC values of 0.77 (prior to training) to 0.92 (after screen reading 1000 mammograms) (Pauli et al., 1996b). Using Tape’s ROC classification guide (Tape, n.d.), the results of Pauli et al. (1996b) indicate fair accuracy levels improving to excellent levels following screen reading of 1000 mammograms. This indicated a learning effect experienced by the radiographer screen readers. Improving analytical rigour in further studies through ROC analysis and evaluating the effect of learning upon accuracy is essential for increasing understanding of the influences on diagnostic accuracy.

**Lesion accuracy and test positive thresholds**

Lesion accuracy is the analysis of screen reader accuracy when detecting individual lesion types, to enable a comprehensive evaluation of lesion detection difficulties. As described in section 3.2.2, a mammogram is considered as test positive when the resultant screen reading decision is to recall the woman for further assessment. The test positive threshold to recall to assessment may internally vary within the individual screen reader, and it may also be externally varied through a change in the threshold for test positive. Threshold accuracy involves the analysis of accuracy under varying levels of what is considered to be test positive, to explore the resultant effect on sensitivity and specificity. Typically in the case of screen reading, the BI-RADS® classification of ‘3’, as Probably Benign and above, is the threshold for test positive. An example of externally varying the test positive threshold could include increasing the test positive threshold to the BI-RADS® classification of ‘4’, as Probably Malignant and above, and exploring the resultant effect on accuracy. Accuracy is
potentially higher at lower thresholds of test positive. Most studies omitted to analyse both lesion accuracy and threshold accuracy. One exception was a study by Wivell et al. (2003) that reported radiographer ability to detect significantly more calcifications than radiologists. Another study reported accuracy at varying test positive thresholds, potentially enabling ROC analysis, but this was not undertaken (Holt, 2006). Future studies need to include analysis of threshold accuracy and lesion accuracy to maximise the potential to analyse the variables which influence screen reader accuracy.

**Inter-observer accuracy**

Considerable variation in inter-observer accuracy among radiologist screen readers has been identified (Elmore et al., 1994; Gur et al., 2004; Duijm et al., 2009; Pitman et al., 2011) and the level of agreement is typically measured using the Kappa coefficient (Viera & Garrett, 2005). Despite this, the reviewed radiographer accuracy studies did not undertake analysis of inter-observer accuracy or level of agreement. When there is limited agreement due to considerable variation in screen reader accuracy, this limits confidence and applicability of the results.

By reviewing analysis of accuracy evaluation, it can therefore be seen that it is essential to improve study rigour and applicability of results of future accuracy studies by undertaking individual and pooled analysis of sensitivity, specificity, and ROC. Moreover, undertaking analysis of accuracy according to lesion type and varying thresholds, together with inter-observer accuracy analysis, will improve rigour further still.

**3.4. Summary**

Following this review of radiographer screen reading accuracy studies, it is clear there is a need for a large, well-designed Australian study in this area of research. Currently, studies undertaken within the Australian context have been limited to small pilot studies and it is evident that gaps in study methods employed in current international studies, particularly a lack of rigorous accuracy analysis, need improvement. Firstly, it is vital to include a well-
constructed representative image set of adequate cancer prevalence and lesion proportion incorporating varying levels of detection difficulty. Moreover, the representative screen reader sample needs to remain unaware of test set cancer prevalence and read independently and blinded to the opinion of other readers, with the provision of previous comparison images when possible. Secondly, the application of a robust gold standard, consisting of pathology results and the minimum 2-year follow-up of the normal screening interval, is essential in the Australian context. Thirdly, the employment of the BI-RADS® classification lexicon will enable rigorous data analysis. Reporting cancer prevalence, lesion proportion, sample composition, and reader experience is essential. Rigorous data analysis includes individual and pooled analysis, including ROC analysis, lesion accuracy, threshold accuracy and inter-observer accuracy.

Incorporating positive aspects and excluding deficiencies of previous screen reader accuracy studies is essential to maximise rigour in future radiographer screen reader accuracy studies. This provides the rationale for the research design of the study reported in Chapter 4 of this thesis.
Chapter 4: Method
4.1. Aim

The aim of this study was to evaluate the accuracy of radiographers screen reading mammograms within an Australian setting.

4.1.1. Ethics approval

Ethics approval was obtained through Sydney West Area Health Service (SWAHS) Human Research Ethics Committee and the University of Sydney Human Research Ethics Committee. Documentation is included in Appendices B, C, D, E, F and G.

4.2. Samples

4.2.1. Screen reader recruitment

An invitation to participate as a screen reader in this research was sent to all radiographers employed by Westmead Breast Cancer Institute (see Appendix H). Potential participants interested in the project were invited to obtain more information from the project leader who answered any questions and provided each potential participant with a Participant Information and Consent Form (see Appendix I). Potential participants were informed that the image test set consisted of 500 mammograms, but were not informed of cancer prevalence, in order to minimise potential expectation bias. Each of the 10 participants who signed the Participant Information and Consent Form was recruited into the study. They did not receive any formal screen reading training.

4.2.2. Image test set

The image test set was comprised of 500 screening mammograms previously read by two/three radiologists during routine screening. This test set included images demonstrating a combination of normal and abnormal pathology, stratified and systematically selected from screening mammograms of the BreastScreen Sydney West database from the year 2004, to improve the representativeness of the sample across normal and abnormal mammograms (Minichiello, Sullivan, Greenwood, & Axford, 2004; Babbie, 2007). The order was then
randomised to distribute benign and malignant lesions amongst the normal mammograms. This year was selected as the most recent year of interval matching at the time of image test set selection, and ensured known outcomes that could be used as the gold standard. As described in section 3.2.1, interval matching ensures that all breast cancers in the NSW Breast Cancer Registry are recorded in the BreastScreen NSW database.

The method of representative image test set compilation is presented in Figure 6. Firstly, a list in numerical order of all screening mammograms from 2004 in the BreastScreen NSW Sydney West service was generated. A total of 61794 mammograms were identified, comprising 9980 prevalent screens (initial) and 51814 incident (subsequent) screens. Secondly, all 137 interval cancers were removed from this list to ensure that all cancer images had the presence of a visible mammographic abnormality. This minimised the potential bias created when an interval cancer was truly not visible on mammogram at the time of initial screening and had become visible at the subsequent screen during the normal screening interval. Thirdly, the remaining 61657 screening mammograms were stratified into three groups comprising normal mammograms ($N = 59160$); mammograms recalled and assessed as benign ($N = 2134$); and mammograms recalled and confirmed as cancers on the basis of histology ($N = 363$).
To ensure representation of benign and malignant mammograms in the image sample, a systematic selection from these three stratified groups of mammograms formed an image test set with a normal to abnormal ratio of 4:1, and a cancer prevalence of 10% (50 cancers of a total 500 mammograms). It was known that approximately one-third of all normal mammograms had been transferred to an adjoining health service due to boundary realignments; to obtain images across the entire year every 91st normal mammogram of the year 2004 was therefore systematically selected to result in the 400 normal mammograms of the image test set. Every 43rd mammogram of the year 2004 recalled to assessment and found to be benign (and not subsequently identified as an interval cancer) was then systematically selected to result in the 50 benign mammograms of the image test set. Every 7th mammogram of the year 2004 recalled to assessment and confirmed to be a cancer was systematically selected to result in the 50 malignant mammograms of the image test set.

This resulted in the final representative image test set sample consisting of 50 benign lesions, 50 malignant lesions and 400 normal mammograms in a total image test set of 500 images.
There was a representation of all typical lesion types. The distribution of benign and malignant lesion proportions is presented in Table 13.

### Table 13- Distribution of benign and malignant lesion proportions

<table>
<thead>
<tr>
<th>Recalled Lesion</th>
<th>Benign Lesion</th>
<th>% Benign</th>
<th>Malignant Lesion</th>
<th>% Malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcifications</td>
<td>2</td>
<td>4%</td>
<td>10</td>
<td>20%</td>
</tr>
<tr>
<td>Discrete Mass with/calcifications</td>
<td>12</td>
<td>24%</td>
<td>17</td>
<td>34%</td>
</tr>
<tr>
<td>Stellate Lesions</td>
<td>0</td>
<td>0%</td>
<td>17</td>
<td>34%</td>
</tr>
<tr>
<td>Architectural Distortion</td>
<td>7</td>
<td>14%</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>Non-Specific Density</td>
<td>29</td>
<td>58%</td>
<td>4</td>
<td>8%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50</strong></td>
<td><strong>100%</strong></td>
<td><strong>50</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

When a situation arose to prevent the location of the chosen mammogram within the BreastScreen Sydney West facility, the next available mammogram in the same stratified list was selected as a replacement. This occurred in 38 instances and was due to:

- The changing boundary realignments between BreastScreen Sydney West service and adjacent services necessitated re-location of some mammograms to adjoining health services, therefore becoming unavailable for viewing.
- Water damage to some stored mammograms made the images unsuitable for viewing.
- Mammograms in use for routine screening and data cleaning were unavailable for viewing.

Each image demonstrating a malignancy was allocated a confidential mammogram identification number (MID) from 1 to 50; each benign image was allocated a confidential MID number from 51 to 100; and each normal image was allocated a confidential MID number from 101 to 500. These images, numbered from 1 to 500, were then randomly sorted using a random number sorting program located at [http://www.randomizer.org/index.htm](http://www.randomizer.org/index.htm) (Urbaniak, 1997) (see attached disc).

Of the final 500 mammograms in the test set, 3 were unilateral due to women having undergone mastectomy surgery of one of their breasts. Therefore, in total there were 997 (1000 minus 3) individual breast (right and left) images used in this study.
4.3. Procedure

4.3.1. Image presentation

Once these 500 mammograms were randomly sorted, images were placed for viewing (hung) on one of three Mammoviewer 810™ image viewers ("Mammoviewer 810," 1987) (Diversified Diagnostics, Inc.) under identical optimal viewing conditions and in a quiet reporting room as they had been originally viewed. These image viewers were located on Level 4 of Jeffery House, 162 Marsden St., Parramatta. The viewing carousel had an adjustable illumination/dimming dial and normal/fast film travel with hand/foot control. Black film viewer masking was placed in gaps between images to effectively restrict light to exposed areas of film, and side sliding doors were used to minimise glare, as recommended in the BSA NAS standards (BreastScreen Australia, 2008). The readers were provided with a Bausch and Lomb Magnifying Glass and a light-blocking magnifier for viewing the mammograms. The film viewer luminance was set to a minimum of 3,000 candela per square metres (cd/m²), as in normal screen reading practice, and regulated by the minimum BSA NAS (BreastScreen Australia, 2008). The viewing area illuminance was below 50 lux as determined by the BSA NAS (BreastScreen Australia, 2008). These optimal lighting and noise levels, together with optimal desktop, chair and viewer ergonomics, ensured minimal bias potential due to lapse in screen reader concentration. It was noted that just as in consecutive screening, the individual radiographer imaging techniques and client variables contributed to varying image quality within this image test set. Normal variations in breast density were apparent in the image test set.

The total test set of 500 mammograms was hung for viewing in 10 separate batches. The first batch consisted of 30 mammograms to enable readers to become accustomed to viewing and screen reading mammograms and to minimise reader fatigue. Subsequent batch sizes were between 50 and 55 mammograms to continue minimising reader fatigue, and were hung on
one of the three Mammoviewers. As each batch was completely read by all 10 readers, the
next batch was hung, until all 10 batches were completed.

The images were hung for viewing using the same standardised hanging protocol as originally
viewed. That is, the current (2004) round of images was hung directly above the previous
corresponding comparison round, being the round prior to the previous round, when available,
in accordance with normal viewing conditions. In typical screen reading practice, comparison
is made with the screening round prior to the previous round when possible, to ensure any
slow-growing subtle breast changes are noted.

When there was only one previous round available, this was hung below the current round for
comparison. When no previous images were available for comparison, the MLO images were
hung directly above the CC images, previously explained in section 2.2.1., as in normal
viewing conditions. When a previously available image round was currently unavailable, the
entire mammogram was discarded from the test set and replaced by the next mammogram of
that subgroup in the original population, to ensure all images were viewed as per normal
screen reading conditions.

When a client had been recalled or had early review at a previous screening round, these
images were also hung for viewing by the reader, as per the standardised hanging protocol.
The radiographer reporting bag and assessment paperwork (see below in section 4.3.2.
Reporting Documents) of the previous screening round were made available for viewing.
In benign and malignant cases, where the client had been recalled to assessment, only the
screening mammogram images were hung for viewing without the recall images, identically
to when the client originally presented for screening, to prevent any potential bias arising
from the radiographer screen reader knowing that the case had been recalled. This meant that
assessment images were available to the reader for previous rounds but not current rounds, as
in normal screen reading practice.
4.3.2. Reporting documents

An A4 binder folder was placed directly alongside each batch of images for viewing to keep all relevant documents together. The Radiographer Report that assists in informing the screen reader of any reported signs and symptoms (see Appendix J) was placed into this folder alongside 10 standardised Reader Reporting Forms, so there was one for each of the 10 screen readers (see Appendix K). This Reader Reporting Form had been developed specifically for this study and was based on the traditional Radiologist Reporting Form (see Appendix A) used and developed in the Sydney West BreastScreen Service, and similar to forms used in other BreastScreen services. This Radiologist Reporting Form has been in use since 1993, fulfilling the BSA classification requirements of Normal (1) or Suspicious (2) as presented in Table 1, in section 2.2.2. These documents were located in front of the Mammoviewer images (as in normal screen reading practice). Stickers were printed for each client, and placed on each standardised Reader Reporting Form for identification of each mammogram. Information regarding the number of images hung; the frame number of the Mammoviewer; previous images hung for viewing, including previous recall and early review images; and viewer number were recorded on each standardised Reader Reporting Form.

4.3.3. File location communication

To ensure that BreastScreen Sydney West data staff knew the location of all files in use for this research project, a note was left in the space normally taken by each file within the filing compactus. When one of the files being used for the project was needed for normal screening of the BreastScreen Sydney West clients, a note was left to indicate that the file had been taken for screening, and also one attached to the file to ensure it was returned to the research project designated area. These methods allowed for clear communication between data and research staff, and ensured the location of all client files was known.
4.3.4. Participant orientation

Each participant was allocated a confidential radiographer identification number (R30 to R39) to record on each of their completed standardised Reader Reporting Forms. Each participant was individually orientated by the Project Leader and given instructions regarding the requirements for completing the standardised Reader Reporting Form and screen reading of the mammograms within each batch of the image test set. Each participant completed a Confidential Participant Form to ensure participant information, such as age, radiographic experience, and mammographic experience, could be collected and recorded (Appendix L). Instructions were given to each participant regarding security requirements during both normal office hours and after hours to ensure their safety during the flexible hours they read image batches. The participants did not receive any formal screen reading training prior to being recruited into this study, and individually chose the volume and time they read images based upon personal and work commitments. Each participant completed the screen reading of 500 mammograms with no imposed screen reading restraints, when it was personally suitable.

4.3.5. Screen reading process

Screen reading was blind and independent of both the original radiologist reports and those of other screen readers. Previous images were available for comparison. Radiographer screen readers (screen readers) indicated on the standardised Reader Reporting Form whether they considered a lesion to be present or not. The readers indicated on a classification scale, based upon the modified BI-RADS® classification lexicon, whether they considered each breast image (right and left) to be normal or otherwise, as presented in Table 1, in section 2.2.2. The screen readers then drew any potential lesion visualised on a breast diagram in the MLO and CC projections of each breast. The screen readers were told that the circling of a ‘3, 4, or 5’ would be considered as ‘recall to assessment’. This was identical to normal screen reading practice, where the same BSA classification system is used to determine whether the images are normal or suspicious of a malignancy. In routine screen reading the radiologist screen
reader determines whether an abnormality exists which needs further assessment. In this study, the recognition of an abnormality for further assessment was determined. Following screen reading of each batch of images, the participants placed their completed standardised Reader Reporting Forms (and the Confidential Participant Form at the initial sitting) in an envelope, to ensure each screen reader individually and blindly read each mammogram, and they then personally sealed and stapled it to ensure confidentiality. These envelopes were collected by the Project Leader at a later date and data recorded in a Microsoft Excel™ spreadsheet (Microsoft Office, 97-2003). Radiographer screen readers were predominantly self-motivated, though at times needed reminding to make the effort to attend the facility where screen reading took place, as it was not their usual workplace. There were many personal issues which arose for the screen readers that could have prevented the completion of screen reading, i.e. 2 screen readers were caring for children undertaking the Higher School Certificate; 2 screen readers experienced the loss of parents; 2 screen readers experienced pregnancies and subsequent births of daughters; and 1 screen reader was diagnosed with bowel cancer and was undergoing subsequent chemotherapy treatment. Despite these issues, all screen readers completed the screen reading of all 500 mammograms during the period beginning 30th April, 2010 and ending 20th May 2011.

4.4. Data entry

When an entire batch of mammograms had been read by each of the 10 screen readers, data was recorded. The type of data entered for each mammogram is presented in Table 14.
When a file had been borrowed for screening purposes, data was entered when the file was
returned. Additional data was also extracted from client files and the BreastScreen database.
Correlation of lesions to the known assessment outcome was made by viewing the previous
radiologist report, assessment clinic paperwork and the BreastScreen database.
During data entry, if it was noted that a Reader Reporting Form was not completed, the screen
reader was asked to complete it when possible. In the case where a screen reader identified
more than one lesion in a breast, both sets of data were entered. All 10 screen readers were
able to complete all standardised Reader Reporting Forms for each of the 500 mammograms
in the test set.

### 4.5. Comparison to gold standard

Accuracy was determined by comparison of the radiographer read of each breast on the
standardised Reader Reporting Forms with the gold standard of known outcomes, based upon
pathology results, 6-year follow-up and interval matching. A screen reading was considered
positive (and would be considered as recalled to assessment) when a BI-RADS® ‘3’
(probably benign); ‘4’ (probably malignant); or ‘5’ (malignant) classification was circled on

<table>
<thead>
<tr>
<th>BN</th>
<th>Batch Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID</td>
<td>Mammogram Identification Number</td>
</tr>
<tr>
<td>PID</td>
<td>Patient identification Number</td>
</tr>
<tr>
<td>DOS</td>
<td>Date of Service</td>
</tr>
<tr>
<td>DOB</td>
<td>Date of Birth</td>
</tr>
<tr>
<td>RN</td>
<td>Round Number</td>
</tr>
<tr>
<td>RLIN</td>
<td>Radiologist Identification Number</td>
</tr>
<tr>
<td>RLR</td>
<td>Radiologist Read Right Breast</td>
</tr>
<tr>
<td>RLR</td>
<td>Radiologist Read Left Breast</td>
</tr>
<tr>
<td>NMR</td>
<td>Nature of Mammographic Lesion</td>
</tr>
<tr>
<td>RM</td>
<td>Result of Mammography</td>
</tr>
<tr>
<td>ANR</td>
<td>Assessment Needle Result</td>
</tr>
<tr>
<td>AFR</td>
<td>Assessment Final Result</td>
</tr>
<tr>
<td>HML</td>
<td>Histopathology of Malignant Lesion</td>
</tr>
<tr>
<td>GML</td>
<td>Grade of Malignant Lesion</td>
</tr>
<tr>
<td>DCIS</td>
<td>Ductal Carcinoma In Situ (DCIS) size</td>
</tr>
<tr>
<td>ICS</td>
<td>Invasive Carcinoma Size</td>
</tr>
<tr>
<td>RGIN</td>
<td>Radiographer Identification Number</td>
</tr>
<tr>
<td>RGR</td>
<td>Radiographer Read Right Breast</td>
</tr>
<tr>
<td>RGL</td>
<td>Radiographer Read Left Breast</td>
</tr>
</tbody>
</table>
the standardised Reader Reporting Form, as in routine screen reading practice. In the case where there was more than one lesion in a breast, the lesion more likely to be malignant took precedence over the lesion least likely to be malignant for comparison to the gold standard.

4.6. Data analysis

To evaluate the accuracy of the 10 radiographers screen reading 500 mammograms, accuracy for each radiographer was assessed, and then accuracy was pooled across the radiographers. Individual screen reader accuracy levels were evaluated when screen reading two images (CC and MLO) of both the right and the left side of each mammogram (except for the 3 mastectomy clients, where only one breast was imaged) within the test set. Accuracy was determined through comparison of the radiographer screen read to the gold standard of known outcomes as described in section 4.5. A \( p \)-value <0.05 was considered as statistically significant. Because each breast is more similar to the other breast of that same woman than to the other women in the image test set, the observations are not independent, and thus any confidence intervals (CI) that do not account for this correlation will be too small. This image test set had been originally double screen read (triple screen read in the case of discrepancies) by various combinations of 16 radiologists employed as first or second (or third) screen readers by the BreastScreen NSW Sydney West Service, as presented in Table 15.

Table 15- Number of mammograms read by each screen reader

<table>
<thead>
<tr>
<th>Reader ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader ID</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>114</td>
</tr>
<tr>
<td>Reader 1</td>
<td>41</td>
<td>16</td>
<td>11</td>
<td>20</td>
<td>10</td>
<td>114</td>
<td>34</td>
<td>48</td>
<td>137</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>Reader 2</td>
<td>63</td>
<td>6</td>
<td>2</td>
<td>18</td>
<td>6</td>
<td>41</td>
<td>19</td>
<td>194</td>
<td>97</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Reader 3</td>
<td>3</td>
<td>-</td>
<td>8</td>
<td>-</td>
<td>12</td>
<td>-</td>
<td>1</td>
<td>8</td>
<td>14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total Read</td>
<td>107</td>
<td>22</td>
<td>13</td>
<td>46</td>
<td>167</td>
<td>53</td>
<td>243</td>
<td>234</td>
<td>23</td>
<td>39</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reader ID</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader 1</td>
<td>41</td>
<td>16</td>
<td>11</td>
<td>20</td>
<td>10</td>
<td>114</td>
<td>34</td>
<td>48</td>
<td>137</td>
<td>8</td>
<td>13</td>
<td>1</td>
<td>19</td>
<td>21</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Reader 2</td>
<td>63</td>
<td>6</td>
<td>2</td>
<td>18</td>
<td>6</td>
<td>41</td>
<td>19</td>
<td>194</td>
<td>97</td>
<td>15</td>
<td>18</td>
<td>1</td>
<td>9</td>
<td>11</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Reader 3</td>
<td>3</td>
<td>-</td>
<td>8</td>
<td>-</td>
<td>12</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>8</td>
<td>14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Total Read</td>
<td>107</td>
<td>22</td>
<td>13</td>
<td>46</td>
<td>167</td>
<td>53</td>
<td>243</td>
<td>234</td>
<td>23</td>
<td>39</td>
<td>15</td>
<td>1</td>
<td>28</td>
<td>32</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Initial data analysis included the author calculating the total number of true and false positives (TP, FP) and true and false negatives (TN, FN) for each screen reader, using the following test positive threshold. TP was defined as a screen reader indicating BI-RADS® classification of ‘3, 4 or 5’ on the reporting form for an image known to be a cancer according to the gold standard. TN was defined as a screen reader indicating BI-RADS® classification of ‘1 or 2’ on the reporting form for an image known to be normal or benign according to the gold standard. FP was defined as a screen reader indicating BI-RADS® classification of ‘3, 4 or 5’ on the reporting form for a mammogram known to be normal or benign according to the gold standard. FN was defined as a screen reader indicating BI-RADS® classification of ‘1 or 2’ on the reporting form for an image known to be a cancer according to the gold standard. The author used these figures to calculate the sensitivity and specificity of each screen reader, and accuracy according to lesion type. The rate of recall to assessment for each screen reader was determined by dividing each screen reader’s total recalls to assessment by the total of mammograms read, and expressed as a percentage, as represented by the equation:

\[
\text{Recall rate} = \frac{\text{Total recalled to assessment}}{\text{Total mammograms read}} \times 100
\]

Data were also entered into the SAS Version 9.2 (SAS Institute, Inc., Cary, NC, 2008) program for advanced statistical analysis by a statistical collaborator as follows. Individual screen reader accuracy was assessed using sensitivity and specificity for the above-mentioned test positive threshold, and then recalculated at a higher test positive threshold when a screen reader indicated BI-RADS® classification of ‘4 or 5’ on the standardised reporting form (equivalent to the screen reader being more definite of their screen reading decision), to explore the magnitude of the resultant effect upon sensitivity and specificity. Analyses estimating ROC curves were performed.
Pooled accuracy across screen readers was assessed using sensitivity and specificity. A bivariate model was used to calculate the pooled sensitivity and specificity at varying positive thresholds as explained in Table 16, accounting for the within reader correlation.

Table 16- Varying thresholds of positive and negative

<table>
<thead>
<tr>
<th>Varying thresholds of Positive and Negative</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>≥ BI-RADS® ‘1, 2, 3, 4 or 5’</td>
<td>0</td>
</tr>
<tr>
<td>≥ BI-RADS® ‘2, 3, 4 or 5’</td>
<td>BI-RADS® ‘1’</td>
</tr>
<tr>
<td>≥ BI-RADS® ‘3, 4 or 5’</td>
<td>BI-RADS® ‘1 or 2’</td>
</tr>
<tr>
<td>≥ BI-RADS® ‘4 or 5’</td>
<td>BI-RADS® ‘1, 2 or 3’</td>
</tr>
<tr>
<td>≥ BI-RADS® 5</td>
<td>BI-RADS® ‘1, 2, 3 or 4’</td>
</tr>
<tr>
<td>&gt; BI-RADS® 5</td>
<td>BI-RADS® ‘1, 2, 3, 4 or 5’</td>
</tr>
</tbody>
</table>

This model only accounts for within reader correlation and not within woman correlation. To adjust pooled results for within woman correlation, an inflation factor, based on a model of specificity alone that accounts for this correlation, was applied to the standard errors prior to the calculation of CIs. Pooled ROC curves and AUC were then determined to evaluate the overall accuracy across radiographers. Associations between accuracy, and both lesion type and the effect of learning, was assessed in a bivariate model. To analyse if there was a learning effect as the radiographers progressively screen read the images, sensitivity and specificity (including 95% CI) were calculated at two time intervals. A non-parametric test was used to compare differences in the AUC between screen readers (DeLong et al., 1988). This test compared the differences in estimated areas under the curve using a Mann-Whitney U-statistic. A significance test was then conducted to determine whether the differences in curves were non-zero. Results are reported in Chapter 5.
Chapter 5: Results
5.1. Samples

The samples used in this study were comprised of both the screen readers and the test set that they read.

5.1.1. Screen reader sample

The screen reader sample consisted of 10 radiographers aged between 27 and 64 years. They ranged in radiographic experience from 7 to 43 years; mammographic experience from 3 to 28 years; and BreastScreen experience from 3 to 18 years, as presented in Figure 7. Screen reading varied between daytime, evenings and weekends, and participants individually chose to read partial, single, or multiple batches of between 20 and 155 mammograms in each screen reading session, depending upon personal, time and traveling constraints, as presented in Table 17. There were no dropouts of participants over the screen reading period, lasting between April 2010 and May 2011.

![Radiographer screen reader experience](image)
<table>
<thead>
<tr>
<th>Participant</th>
<th>Reader 1</th>
<th>Reader 2</th>
<th>Reader 3</th>
<th>Reader 4</th>
<th>Reader 5</th>
<th>Reader 6</th>
<th>Reader 7</th>
<th>Reader 8</th>
<th>Reader 9</th>
<th>Reader 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch 1 reading date (30 images)</td>
<td>29.5.10</td>
<td>25.5.10</td>
<td>30.4.10</td>
<td>12.6.10</td>
<td>12.6.10</td>
<td>13.5.10</td>
<td>6.5.10</td>
<td>11.5.10</td>
<td>4.6.10</td>
<td>29.5.10</td>
</tr>
<tr>
<td>Batch 2 reading date (50 images)</td>
<td>31.5.10</td>
<td>2.7.10</td>
<td>14.7.10</td>
<td>2.7.10</td>
<td>9.7.10</td>
<td>4.6.10</td>
<td>23.7.10</td>
<td>17.5.10</td>
<td>19.6.10</td>
<td>28.6.10</td>
</tr>
<tr>
<td>Batch 3 reading date (50 images)</td>
<td>10.7.10</td>
<td>2.7.10</td>
<td>14.7.10</td>
<td>9.7.10</td>
<td>16.7.10</td>
<td>28.6.10</td>
<td>5.8.10</td>
<td>7.6.10</td>
<td>16.8.10</td>
<td>9.7.10</td>
</tr>
<tr>
<td>Batch 4 reading date (50 images)</td>
<td>15.10.10</td>
<td>9.7.10</td>
<td>14.7.10</td>
<td>16.7.10</td>
<td>9.7.10</td>
<td>5.8.10</td>
<td>19.7.10</td>
<td>11.10.10</td>
<td>1.10.10</td>
<td>10.7.10</td>
</tr>
<tr>
<td>Batch 5 reading date (55 images)</td>
<td>15.10.10</td>
<td>19.8.10</td>
<td>12.11.10</td>
<td>3.9.10</td>
<td>1.10.10</td>
<td>7.10.10</td>
<td>16.8.10</td>
<td>22.10.10</td>
<td>1.10.10</td>
<td>28.8.10</td>
</tr>
<tr>
<td>Batch 6 reading date (55 images)</td>
<td>11.12.10</td>
<td>28.9.10</td>
<td>12.11.10</td>
<td>3.12.10</td>
<td>1.10.10</td>
<td>7.10.10</td>
<td>20.9.10</td>
<td>20.12.10</td>
<td>12.11.10</td>
<td>27.10.10</td>
</tr>
<tr>
<td>Batch 7 reading date (55 images)</td>
<td>11.12.10</td>
<td>20.12.10</td>
<td>12.11.10</td>
<td>28.1.11</td>
<td>28.1.11</td>
<td>12.11.10</td>
<td>21.2.11</td>
<td>25.10.10</td>
<td>16.2.11</td>
<td>14.1.11</td>
</tr>
<tr>
<td>Batch 8 reading date (55 images)</td>
<td>18.12.10</td>
<td>22.2.11</td>
<td>20.5.11</td>
<td>25.2.11</td>
<td>25.2.11</td>
<td>14.1.11</td>
<td>21.2.11</td>
<td>17.12.10</td>
<td>6.3.11</td>
<td>18.2.11</td>
</tr>
<tr>
<td>Batch 9 reading date (50 images)</td>
<td>25.3.11</td>
<td>15.3.11</td>
<td>20.5.11</td>
<td>1.4.11</td>
<td>11.3.11</td>
<td>18.4.11</td>
<td>24.1.11</td>
<td>6.3.11</td>
<td>11.3.11</td>
<td>25.3.11</td>
</tr>
<tr>
<td>Batch 10 reading date (50 images)</td>
<td>29.3.11</td>
<td>15.3.11</td>
<td>20.5.11</td>
<td>29.4.11</td>
<td>11.3.11</td>
<td>18.4.11</td>
<td>28.2.11</td>
<td>16.5.11</td>
<td>11.3.11</td>
<td>25.3.11</td>
</tr>
</tbody>
</table>
5.1.2. Image test set sample

The representative image test set sample consisted of 500 mammograms comprising 400 normal mammograms; 50 mammograms recalled to assessment and found to be benign; and 50 mammograms recalled to assessment and found to be malignant. A total of 997 individual (right and left) breast images were viewed, due to the inclusion of 3 mastectomy client images within the test set. The 50 benign lesions were comprised of 2 calcifications, 12 discrete masses (with/without calcification), 7 architectural distortions, and 29 non-specific densities. The 50 malignant lesions were comprised of 10 calcifications, 17 discrete masses (with/without calcification), 17 stellate lesions, 2 architectural distortions, and 4 non-specific densities. This distribution of benign and malignant lesions is presented in Table 13, in section 4.2.2. These 500 mammograms were randomly sorted into one image test set, and then separated into 10 batches consisting of between 30 and 55 mammograms for screen reading by the radiographers, in order to minimise potential fatigue levels.

5.2. Individual screen reader accuracy

5.2.1. Sensitivity and specificity

Table 18 presents values for each screen reader of TP, FN, TN, FP, sensitivity, and specificity (including 95% CI). Sensitivity was between 76.0% and 92.0%, and specificity was between 74.8% and 96.2%, for this sample of 10 radiographer screen readers within an Australian setting. These accuracy levels are presented individually in Figures 8 and 9, and because these indicators of accuracy are related, they are also presented as a scatter plot in Figure 10.
Table 18- Individual screen reader accuracy-with BI-RADS® classification of '3, 4 or 5' considered positive

<table>
<thead>
<tr>
<th>Reader</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41</td>
<td>9</td>
<td>887</td>
<td>60</td>
<td>82.0 (71.2, 92.8)</td>
<td>93.7 (92.1, 95.2)</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>12</td>
<td>911</td>
<td>36</td>
<td>76.0 (64.0, 88.0)</td>
<td>96.2 (95.0, 97.4)</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>8</td>
<td>860</td>
<td>87</td>
<td>84.0 (73.7, 94.3)</td>
<td>90.8 (89.0, 92.7)</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>12</td>
<td>881</td>
<td>66</td>
<td>76.0 (64.0, 88.0)</td>
<td>93.0 (91.4, 94.7)</td>
</tr>
<tr>
<td>5</td>
<td>43</td>
<td>7</td>
<td>786</td>
<td>161</td>
<td>86.0 (76.3, 95.7)</td>
<td>83.0 (80.6, 85.4)</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>10</td>
<td>862</td>
<td>85</td>
<td>80.0 (68.8, 91.2)</td>
<td>91.0 (89.2, 92.8)</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>6</td>
<td>815</td>
<td>132</td>
<td>88.0 (78.9, 97.1)</td>
<td>86.1 (83.9, 88.3)</td>
</tr>
<tr>
<td>8</td>
<td>41</td>
<td>9</td>
<td>831</td>
<td>116</td>
<td>82.0 (71.2, 92.8)</td>
<td>87.8 (85.7, 89.8)</td>
</tr>
<tr>
<td>9</td>
<td>38</td>
<td>12</td>
<td>818</td>
<td>129</td>
<td>76.0 (64.0, 88.0)</td>
<td>86.4 (84.2, 88.6)</td>
</tr>
<tr>
<td>10</td>
<td>46</td>
<td>4</td>
<td>708</td>
<td>239</td>
<td>92.0 (84.4, 99.6)</td>
<td>74.8 (72.0, 77.5)</td>
</tr>
</tbody>
</table>

*CI not allowing for correlation at this stage; N = 997 (based on number of breasts screen read)

Figure 8- Screen reader sensitivity
Figure 9- Screen reader specificity

Figure 10- ROC scatter plot of screen readers 1-10
Higher test positive threshold

Further analysis was undertaken at a higher test positive threshold, as described in section 4.6. This involved a screen reader indicating a BI-RADS® classification of ‘4 or 5’ on the standardised reporting form (equivalent to the screen reader being more definite in their decision), and is presented in Table 19. The results in Table 19 confirm the results of Table 18 and indicate that a higher positive threshold is associated with lower sensitivity and higher specificity, as expected. The resultant increase in specificity to between 96.1% and 100.0% was associated with a substantial decrease in sensitivity to between 26.0% and 74.0%. Within a true screen reading situation a BI-RADS® ‘3, 4 or 5’ is considered test positive, and the higher test positive threshold is not used. The screen reader does not determine the nature of a lesion but the presence of an abnormality, so the BI-RADS® classification of ‘3, 4 or 5’ test positive threshold for recall used in this study is appropriate.

The results in Tables 18 and 19 do not allow for within-woman correlation from screen reading two breast images per woman. As described in section 4.6., CIs are conservative because each breast is more similar to the other breast of the same woman, compared to other women in the test set, and are therefore not independent.

Table 19- Individual screen reader accuracy—with BI-RADS® classification of ‘4 or 5’ considered positive

\[ N = 997 \text{ (based on number of breasts screen read)} \]
5.2.2. Recall rates

Recall rates ranged from 14.8% (Reader 2) to 57.0% (Reader 10), and are presented in Figure 11.

![Figure 11- Recall rates](image)

5.2.3. ROC

ROCs for each screen reader are presented in Figure 12 and visually confirm the individual accuracy findings.
5.3. Pooled screen reader accuracy

Pooled screen reader accuracy was determined to evaluate the overall accuracy across the readers, as described in Section 4.6.

5.3.1. Sensitivity and specificity

Pooled sensitivity and specificity is presented in Table 20. As described in Section 4.6., an inflation factor based on a model of specificity alone that accounts for within woman correlation, was applied to the standard errors prior to the calculation of the CIs.

Figure 12- ROC for each screen reader
Table 20- Pooled sensitivity and specificity (including 95% CI) at each threshold

<table>
<thead>
<tr>
<th>Positive Threshold</th>
<th>≥ BI-RADS® '1 2 3 4 or 5'</th>
<th>≥ BI-RADS® '2 3 4 or 5'</th>
<th>≥ BI-RADS® '3 4 or 5'</th>
<th>≥ BI-RADS® '4 or 5'</th>
<th>&gt; BI-RADS® '5'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled Sensitivity</td>
<td>100</td>
<td>85.8 (80.5, 89.8)</td>
<td>82.2 (77.1, 86.3)</td>
<td>53.8 (41.1, 65.9)</td>
<td>8.1 (3.4, 18.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Pooled Specificity</td>
<td>0</td>
<td>82.9 (73.9, 89.2)</td>
<td>89.5 (83.8, 93.3)</td>
<td>99.5 (98.7, 99.8)</td>
<td>100.0 (99.8, 100.0)</td>
</tr>
<tr>
<td>Inflation Factor</td>
<td>1.310</td>
<td>1.187</td>
<td>1.187</td>
<td>1.187</td>
<td>100</td>
</tr>
</tbody>
</table>

5.3.2. ROC

The pooled accuracy results of Table 20 are represented visually in Figure 13 as a pooled ROC curve. This indicates that pooled sensitivity at a positive threshold of BI-RADS® classification of ‘3, 4 or 5’ was 82.2 % (77.1, 86.3) and specificity was 89.5 % (83.8, 93.3).

The data was modeled separately at each threshold to show an implied pooled ROC across each pooled sensitivity and specificity point. The crosses indicate the accuracy of each screen reader, while the red dot is the pooled accuracy with the confidence region shown in red.

Figure 13- Pooled estimates of sensitivity and specificity for threshold of screen reading of BI-RADS® classification of '3, 4 and 5' versus '1 and 2' with estimated ROC
5.3.3. Accuracy according to lesion type

The accuracy of the screen readers in detecting varying lesion types is presented in Table 21. The screen readers’ sensitivity in detecting calcifications and discrete masses is higher than for other lesions. There is no significant difference between detecting discrete masses (sensitivity 88.8%) and calcifications (sensitivity 90.0%) \((p = 0.805)\). The screen readers had lower sensitivity for detecting both architectural distortions and non-specific densities (sensitivity 63.3%), and stellate lesions (sensitivity 77.6%) compared to calcifications as the referent lesion-type \((p = 0.010\) and \(p = 0.064\), respectively).

Table 21- Analysis of accuracy by radiographic appearance of lesion type

<table>
<thead>
<tr>
<th>Radiographic appearance of lesion type</th>
<th>Frequency</th>
<th>Sensitivity (95%CI)</th>
<th>p-value</th>
<th>Specificity (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0=No cancer</td>
<td>947</td>
<td></td>
<td></td>
<td>89.5 (84.9, 92.8)</td>
</tr>
<tr>
<td>1=Calcification</td>
<td>10</td>
<td>90.0 (80.9, 95.0)</td>
<td>reference group</td>
<td></td>
</tr>
<tr>
<td>2=Stellate lesion</td>
<td>17</td>
<td>77.6 (69.6, 84.0)</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>3=Discrete mass with/without calcification</td>
<td>17</td>
<td>88.8 (82.1, 93.2)</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>4=Architectural distortion or non-specific density</td>
<td>6</td>
<td>63.3 (48.5, 76.0)</td>
<td>0.0038</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>997</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lesion type</th>
<th>Frequency</th>
<th>Sensitivity (95%CI)</th>
<th>p-value</th>
<th>Specificity (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0=No cancer</td>
<td>947</td>
<td>90.0 (78.6, 95.7)</td>
<td>reference group</td>
<td></td>
</tr>
<tr>
<td>1=Calcification</td>
<td>10</td>
<td>77.6 (67.9, 85.1)</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td>2=Stellate lesion</td>
<td>17</td>
<td>88.8 (80.5, 93.9)</td>
<td>0.805</td>
<td></td>
</tr>
<tr>
<td>3=Discrete mass with/without calcification</td>
<td>17</td>
<td>63.3 (45.7, 78.0)</td>
<td>0.00996</td>
<td></td>
</tr>
<tr>
<td>4=Architectural distortion or non-specific density</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>997</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.3.4. Learning effect

Table 22 presents accuracy of the first 5 batches \((n = 235\) mammograms), and then the next 5 batches \((n = 265\) mammograms). There was a significant improvement in specificity as participants screen read more images \((p = 0.012)\). There was some evidence that sensitivity improved, nearing statistical significance \((p = 0.056)\).

Table 22- Learning effect- accuracy at two time intervals as more mammograms screen read

<table>
<thead>
<tr>
<th>Learning effect disregarding within woman correlation</th>
<th>Frequency</th>
<th>Sensitivity</th>
<th>(p)-value</th>
<th>Specificity</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0= Batches 1 - 5</td>
<td>469</td>
<td>78.2 (72.1, 83.3)</td>
<td>reference group</td>
<td>88.2 (83.1, 92.0)</td>
<td>reference group</td>
</tr>
<tr>
<td>1 = Batches 6 - 10</td>
<td>528</td>
<td>87.3 (81.3, 91.6)</td>
<td>0.0287</td>
<td>90.5 (86.2, 93.6)</td>
<td>0.0046</td>
</tr>
<tr>
<td>Total</td>
<td>997</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Learning effect adjusting for within woman correlation

<table>
<thead>
<tr>
<th>Learning effect adjusting for within woman correlation</th>
<th>Frequency</th>
<th>Sensitivity</th>
<th>(p)-value</th>
<th>Specificity</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0= Batches 1 - 5</td>
<td>469</td>
<td>78.2 (70.9, 84.1)</td>
<td>reference group</td>
<td>88.2 (82.0, 92.5)</td>
<td>reference group</td>
</tr>
<tr>
<td>1 = Batches 6 - 10</td>
<td>528</td>
<td>87.3 (79.9, 92.2)</td>
<td>0.056</td>
<td>90.5 (85.3, 94.0)</td>
<td>0.012</td>
</tr>
<tr>
<td>Total</td>
<td>997</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.4. Inter-observer accuracy

The AUC measures overall accuracy and takes into account both sensitivity and specificity. The AUC range was 0.842 to 0.923 (and 95% CI) and is reported in Table 23 below. It confirms the visual results of the ROC curve in Figure 12, in section 5.2.3.

Table 23- Comparison of screen reader accuracy using Areas Under the Curve (AUC) (including 95% CI)

<table>
<thead>
<tr>
<th>Reader</th>
<th>AUC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.903 (0.848, 0.958)</td>
</tr>
<tr>
<td>2</td>
<td>0.899 (0.845, 0.953)</td>
</tr>
<tr>
<td>3</td>
<td>0.890 (0.835, 0.945)</td>
</tr>
<tr>
<td>4</td>
<td>0.887 (0.827, 0.947)</td>
</tr>
<tr>
<td>5</td>
<td>0.902 (0.849, 0.956)</td>
</tr>
<tr>
<td>6</td>
<td>0.896 (0.839, 0.954)</td>
</tr>
<tr>
<td>7</td>
<td>0.912 (0.862, 0.961)</td>
</tr>
<tr>
<td>8</td>
<td>0.881 (0.825, 0.936)</td>
</tr>
<tr>
<td>9</td>
<td>0.842 (0.776, 0.909)</td>
</tr>
<tr>
<td>10</td>
<td>0.923 (0.881, 0.966)</td>
</tr>
</tbody>
</table>

There was no evidence that the AUCs were different across the readers. \((\chi^2 = 12.8, df = 9, p\)-value = 0.17). These results are discussed in Chapter 6.
Chapter 6: Discussion
The aim of this thesis was to undertake a large well-designed study to evaluate the accuracy of radiographers screen reading mammograms in an Australian setting.

6.1. Evaluation of sample

Evaluation of this sample of 10 radiographers screen reading this representative test set of 500 images was undertaken by determining individual screen accuracy, pooled screen reader accuracy and inter-observer accuracy. The significance of this evaluation will be discussed.

6.1.1. Screen reader sample

The participant radiographer sample that screen read this image test set was recruited from the Westmead Breast Cancer Institute and is representative of the Sydney West BreastScreen service. An important point to recall is the radiographer screen readers had not received any formal reading training prior to participating in this study. Their wide range of age, mammographic, diagnostic, and BreastScreen experience may be broadly representative of the BSA radiographer population. Individual motivation and experience levels, however, may vary such that this sample of radiographer readers is not necessarily indicative of the entire BSA radiographer population. The radiographers blindly and independently read under optimal viewing conditions to maximise concentration and attention. As per the BSA NAS, lighting conditions were optimised (BreastScreen Australia, 2008). Noise levels were minimised and attention to ergonomics was ensured to maximise accuracy levels. Previous comparison images were made available and no limits of maximum number of mammograms to read were given. This may have contributed to fatigue and loss of concentration when multiple batches of mammograms were read, or when the radiographer screen readers decided to read following a full day’s work as a radiographer (WM Reed et al., 2009). It would be expected that screen readers choosing to view multiple batches may result in potentially lower accuracy levels. As seen in Table 17, screen readers varied in their choice of screen reading mammogram volume. For example, Readers 7, 4 and 3 predominantly chose to read a
maximum of 30, 55 and 165 mammograms each, respectively. Their corresponding AUC values were 0.912, 0.887 and 0.890, respectively (see Table 23). Despite expectations, there was no evidence to suggest individual variations in screen reading volume directly affected accuracy levels in this reported study. This does not consider other variables that may lead to fatigue, so it remains plausible that multiple influences, including experience, fatigue levels and screen reading volume, may simultaneously influence accuracy levels.

6.1.2. Test sets versus consecutive populations

The representative image test set, enriched with lesions, was chosen to represent 2004, the last year that had been interval matched at the time of test set compilation. This was to increase study rigour by ensuring known outcomes as the gold standard. By inflating the image test set with lesions it was possible to calculate accuracy levels without the impractical time-consuming task of screen reading several thousands of consecutive population screening mammograms to produce a sufficiently powered study to calculate meaningful accuracy levels.

Internationally, such as in the UK, where radiographer screen reading is well-established, it is practical and possible for accuracy evaluation studies to take place in an ideal consecutive population setting despite the high volume of normal mammograms that need to be read for a low volume of cancers. Image test sets enriched with lesions have the potential to lead to expectation bias, due to screen readers expecting to find cancers that are less likely in consecutive population screen reading. This may lead to inflated values of false positives when screen readers transfer to consecutive population screen reading, however, no studies were found to support the expected inflation of false positives. As described in section 3.1.1., Pauli et al. and Scott et al. have undertaken studies that report correlation between image test set results and consecutive screen reading practice (Pauli et al., 1996b; Scott et al., 2009). Furthermore, Gur et al. (2008) have reported consecutive screen reading accuracy being better than image test set results. It is therefore reasonable and practical to utilise image sets
enriched with lesions, in an Australian setting, to establish baseline accuracy of radiographers where it is as yet unacceptable for radiographers to screen read in a population screening setting. Radiographer screen readers were not informed of the cancer prevalence or lesion proportion within the image test set in this reported study, in order to minimise expectation bias. It is possible however, that individual readers’ potentially formed their own preconceptions of how many abnormal mammograms they expected to find. In fact, Reader 2 confessed to ‘undercalling’ when screen reading this image test set due to her personal expectation bias, presuming the image test set had a representation of cancers in the same proportion as a true consecutive population, potentially lowering her accuracy levels.

6.2. Individual screen reader accuracy

Individual accuracy was analysed through comprehensive measures that included sensitivity, specificity, recall rates, and AUC from ROC analysis.

6.2.1. Sensitivity and specificity

Sensitivity is a measure of accuracy determined by the true positive rate which is the ability of a reader to accurately identify cases of disease while specificity is a measure of the true negative rate or the ability of a reader to accurately identify cases without disease (North Carolina School of Science and Mathematics Statistics Leadership Institute, 1999, n.d.). Accuracy varied across screen readers. These results, presented in Tables 18 and 19, in section 5.2.1., of sensitivity levels of 76.0% to 92.0%, and specificity levels of 74.8% to 96.2%, still compare favourably with other reported accuracy levels of radiographers and radiologists. Previous studies reported radiographer sensitivity levels of 61.0% to 91.42%, and specificity levels of 45.0% to 99.1%, as presented in section 3.3.1, in Table 10. These same studies reported radiologist sensitivity levels of 63.9% to 100% and specificity levels of 81.0% to 99.2%.

In comparing the range of accuracy values obtained, the accuracy values found in this study demonstrate higher minimum values of both sensitivity and specificity, when compared to
previous studies. This may indicate that a majority of the radiographers in this study had
greater abnormality detection ability in comparison to radiographers in several previous
studies, or that the difference can be explained by sampling variability. The maximum
specificity levels of this reported study were somewhat lower than the maximum specificity
level of one previous study of Duijm et al. (2007). This may indicate that the radiographers in
this reported study had a slightly lower ability to determine the absence of breast disease in
comparison to the radiographers in that study. Another plausible explanation is that readers
may have been operating at a different threshold of test positive, as sensitivity levels in this
reported study were also somewhat higher the findings of Duijm et al. (2007). Variations in
accuracy levels in this reported study, in comparison to previous studies, are also potentially
explained through the findings of Taplin et al. (2002), who report that individual variations in
positioning techniques may contribute to abnormality detection accuracy. Differences in study
size and methods employed may also account for the minor disparities in accuracy levels
between this reported study and previous accuracy studies.

This level of agreement of accuracy levels, from this sample of Australian radiographers to
accuracy levels of previous accuracy studies of radiographers and radiologists, is
encouraging. This is particularly so considering the absence, in this study, of any prior formal
screen reading training that radiographer screen readers from eight previous studies had
undertaken (Haiart & Henderson, 1991; Bassett et al., 1995; Pauli et al., 1996a, 1996b; Wivell
et al., 2003; Duijm et al., 2007; Duijm et al., 2008; Duijm et al., 2009). A further strength of
this reported study is that accuracy was determined by using a robust gold standard, thus
increasing the rigour of the results.

**Formal screen reading training prior to commencing screen reading**

Several studies have recommended the necessity of training radiographers prior to taking on
the role as second screen readers (Bassett et al., 1995; Pauli et al., 1996a, 1996b; Sumkin et
al., 2003; Wivell et al., 2003; Holt, 2006). Conversely, two of those studies (Sumkin et al.,
2003; Holt, 2006) report comparable accuracy to radiologists, without formalised screen
reading training. Bassett et al. (1995) state that obtaining pre-training accuracy levels allows for future training to be effectively tailored to individuals within the group. This reported study analysed accuracy without any formal screen reading training, providing a baseline to measure any potential post-training accuracy differences. This reported study also enables the tailoring of any future training to individual needs, by aiming to improve areas of potential weaknesses and build on known strengths.

**Higher test positive threshold**

In screen reading, both sensitivity and specificity are of major significance to the women being screened, due to the importance of maximising the potential for an accurate breast cancer diagnosis, while minimising anxiety caused by unnecessary recall to assessment. It is therefore of interest to explore the relationship between sensitivity and specificity, when the threshold for test positive is externally varied. In exploring the effect of increasing the test positive threshold in this study, to a BI-RADS® classification of ‘4 or 5’, specificity increased and sensitivity decreased, as expected. The resultant large decrease in sensitivity levels of several screen readers indicates their more frequent selection of a BI-RADS® classification ‘3’ read, in contrast to selecting a more confident BI-RADS® classification of ‘4 or 5’. Such low screen reader confidence would be expected to improve with formal screen reading training and experience. It is possible that when applying a higher test positive threshold, the trade-off between sensitivity decrease and specificity increase would mean some breast cancers would be undiagnosed due to less women being recalled to assessment. In breast screening, therefore, the BI-RADS® ‘3’ classification serves the important purpose of maximising cancer diagnosis without too much more unnecessary recall to assessment. This is because typically, in the case of both screen readers selecting a BI-RADS® classification of ‘3’, the woman is recalled to assessment; if however, only one radiologist screen reader selects a BI-RADS® classification of ‘3’, a third independent radiologist reader makes the final decision. At least two radiologist screen readers therefore need to select a BI-RADS® ‘3’ classification for the woman to be recalled to assessment. It was however, noted
by the researcher when collecting original screen reading data for this study, that in the initial true screen reading when there was a discrepancy of one radiologist screen reader selecting a BI-RADS® classification of ‘3’, while the other radiologist screen reader selected a BI-RADS® classification of ‘1 or 2’, the third deciding radiologist screen reader almost exclusively also selected a BI-RADS® classification of ‘3’, meaning the woman was in fact recalled to assessment. In other words, the third deciding radiologist screen reader preferred to err on the side of caution and recall a woman for further assessment, rather than potentially fail to detect a breast cancer. This is an area that may require future research and formal evaluation.

There is a notable absence of reported accuracy at varying test positive thresholds in previous international studies. Holt (2006) was the only exception, and reported a similar trend of increased specificity and decreased sensitivity with an increased test positive threshold of accuracy. This analysis was possible because of the selection of the BI-RADS® classification lexicon in this study. Other previous studies were unable to undertake this analysis due to their limited screen reading classification, as reported in Table 7, in section 3.2.1. This constitutes one of the strengths in the design of this study.

**Double screen reading**

It is very pertinent to recall at this point, that in typical screen reading practice, mammograms are interpreted by two screen readers. This is undertaken as a result of previous studies reporting an increase in accuracy levels of 10-15% through independent double screen reading (Anderson et al., 1994; Thurfjell et al., 1994; Beam et al., 1996). There has been a suggestion that there may be a psychological order effect for screen readers in knowing they are first or second readers, potentially influencing the effect of accuracy improvement of two screen readers (Williams, Hartswood, & Prescott, 1998). Another study, however, has demonstrated that second screen reading is effective in detecting both the small, more difficult lesions and the larger lesions missed due to reader fatigue and loss of attention (Ciatto et al., 2005). There has been ambiguity surrounding the degree of accuracy correlation between
image test sets and consecutive population screening as outlined in section 3.1.1, however, Gur et al. (2008) have demonstrated that true screen reader performance was significantly better than image test set performance. In potential future true independent double screen reading practice, therefore, the overall accuracy levels can confidently be expected to increase above the levels reported in this study.

**Single and multiple screen reading**

The radiographer screen readers undertook prospective single screen reading of a retrospectively constructed image test set in this reported study. The accepted international practice of independent double screen reading to maximise accuracy (Thurfjell et al., 1994) was confirmed in a previous study where a radiographer/radiologist pair detected more cancers than either working alone (Bassett et al., 1995). Another previous study reported that similar pairing resulted in similar sensitivity to a radiologist pair (Pauli et al., 1996a). Wivell et al. (2003) reported that a radiographer pair detected all abnormalities detected by a radiologist while several studies reported that a radiographer pair improves cancer detection compared to single radiologist screen reading (Hillman et al., 1987; Pauli et al., 1996a; Wivell et al., 2003). A limitation of this reported study is that no further evaluation of radiographer pair or radiographer/radiologist pair accuracy was undertaken. To improve understanding in this area, there is a need for further analysis of radiographer double and multiple screen reading in future studies.

**6.2.2. Recall rates**

There was no evidence that individual recall rates of 14.8% to 57.0% (see Figure 11 in section 5.2.2.) were influenced by levels of radiography, diagnostic mammography, or BreastScreen experience. This may perhaps be explained by much of the radiographer reader experience being focused in the radiographic/mammographic arena (and therefore not in screen reading), and the lack of formal screen reading training. There was a trend in this study indicating higher recall rates were associated with higher sensitivity levels, as would be expected due to the higher number of recalls increasing true positive rates. Screen readers may display varying
levels of caution and select BI-RADS® classification of ‘3, 4, or 5’ more often when unsure, leading to some unnecessary ‘overcall’ of women recalled to assessment. This would potentially lead to increased client anxiety and decreased efficiency. BSA NAS state that recall rates need to be kept below 10% of women screened (BreastScreen Australia, 2008).

Given that this image test set was enriched with lesions, however, it would be expected that there would be a higher recall rate compared to screening populations due to higher cancer prevalence, such that this standard is not applicable to this image test set.

The recall rates of these screen readers may be high for a number of potential reasons, including reader attitudes, inexperience in screen reading, and importantly, the inflated number of cancers in this image test set. The attitudes of the radiographer readers may differ when screen reading an image test set in comparison to consecutive population screen reading. While screen readers are motivated to maximise accuracy and simultaneously minimise unnecessary recall, under test set screen reading conditions there is no true inconvenience to actual clients if overcalled. Thus screen readers may be more willing to overcall in image test sets to maximise true positives, even though they are also increasing false positives. Lack of screen reading experience may also increase false positives. Bennett et al. (2011), however, report that with support and monitoring, the higher recall rates of radiographers screen reading is not likely to have major negative impacts on the UK breast screening program. Further research in the Australian context, following training aimed at minimising recall rates, would be beneficial.

6.2.3. ROC and AUC

Another design strength of this study is that the radiographers’ reading was based on scoring for increasing level of suspicion, allowing the undertaking of ROC analysis to estimate overall accuracy. This analysis combines reader ability to detect disease with ability to recognize normal images, together with reader decision confidence, to give a measure of reader accuracy (Soh et al., 2012). The AUC values ranging between 0.842 to 0.923 (see Figure 12, in section 5.2.3., and Table 23, in section 5.4.) indicate high accuracy levels
(perfect accuracy is equal to 1.0) (Zamora, Muriel, & Abraira, n.d.). It is important to note that these radiographers had not received any formalised screen reading training prior to screen reading, and it can be expected that these accuracy levels would increase with screen reading education and training, as has been reported in previous studies (Bassett et al., 1995; Pauli et al., 1996b). Most previous studies did not report ROC curves, with the exception of Pauli et al. who reported AUC values ranging from 0.77 (prior to training) to 0.92 (after screen reading 1000 mammograms) (Pauli et al., 1996b). This supports the high accuracy levels of this reported study, particularly prior to formal screen reading training. Independent double screen reading can be expected to increase overall accuracy levels by 10-15% (Thurfjell et al., 1994).

6.3. Pooled screen reader accuracy

Pooled statistical analysis was carried out to maximise the ability to infer the study sample results to the population (van Houwelingen et al., 1993; Reitsma et al., 2005). It was comprised of sensitivity, specificity, ROC analysis, and investigation of accuracy by lesion type, and for learning effect.

6.3.1. Pooled sensitivity and specificity

These pooled accuracy results increase the ability to infer the results from the study sample to the population. As stated in section 2.2.2, both sensitivity and specificity are important in screen reading to maximise cancer diagnosis while minimising unnecessary recall to assessment. The aim of breast screening programs is to maximise both sensitivity and specificity simultaneously. Analysis at varying test positive thresholds reflects the varying reader decision thresholds, represented by the ROC curve. Higher test positive thresholds lead to lower sensitivity and higher specificity, just as in the case when individual readers may increase their decision point of what they consider to be test positive (Joy et al., 2005). Pooled sensitivity and specificity levels were reported in four previous studies as presented in Table 9, in section 3.3.1. (Bassett et al., 1995; Holt, 2006; Duijm et al., 2007; Duijm et al., 2009),
however only one study reported pooled accuracy measures at increasing test positive thresholds (Holt, 2006).

6.3.2. Pooled ROC

Pooled ROC curves of the group of screen readers (see Figure 13 in section 5.3.2.) represented further analysis that has not been undertaken in previous radiographer accuracy studies. ROC curves are used extensively in radiologist accuracy studies to analyse performance (Goddard et al., 1998) and this omission is a deficiency of previous radiographer accuracy studies. This work therefore maximises the use of established analytic methods for accuracy through ROC analysis providing a measurement of overall accuracy (combining information for sensitivity and specificity), at varying thresholds. Through the pooling of screen reader accuracy presented in Table 20, in section 5.3.1., and Figure 13, in section 5.3.2., further evidence is provided that even without formal screen reading training, this sample of 10 screen readers in this Australian setting, demonstrated high accuracy levels relative to previous international radiographer accuracy studies.

6.3.3. Accuracy by lesion type

The Duijm et al. (2007, 2008, 2009) studies reported higher proportions of densities and calcifications in their consecutive populations, as reported in Table 5 in section 3.3.1., when compared to this reported constructed image set study enriched with lesions, as reported in Table 13 in section 4.2.2. Duijm et al. (2007, 2008, 2009) also reported lower proportions of asymmetric densities, densities with calcifications and architectural distortions, that is, those abnormalities with increased detection difficulty, in comparison to this reported study. These differences in proportions provide a partial explanation of the comparatively low sensitivity levels and high specificity levels of the radiographers in the Duijm et al. (2007) study, as being the result of individual lesions having varying levels of detection difficulties, explained in section 2.3.2. Unfortunately, Duijm et al. (2007, 2008, 2009) did not undertake analysis of individual lesion detection accuracy, that may have led to further understanding of these accuracy level findings.
Prevalence bias was minimised by ensuring the screen readers in this reported study remained unaware of lesion proportions within the image test set. The analysis in Table 21, in section 5.3.3., indicates that the radiographer screen readers in this reported study are more accurate at detecting calcifications and discrete masses than stellate lesions, architectural distortion and non-specific densities. The difficulties in screen readers detecting stellate lesions, architectural distortion and non-specific densities may be potentially explained by the presence of overlying normal dense breast parenchyma obscuring the abnormalities.

Radiologist studies have reported varying levels of detection difficulty that need to be minimised (Burrell et al., 2001; Tot, 2005; Banik et al., 2012). In addition, Beam at al. (2002) reported that case-related differences account for more disagreement between readers than differences between the readers themselves. This is supported by the findings of this reported study, of varying levels of accuracy detection ability of individual lesions, reported in Table 21 in section 5.3.3., despite minimal inter-observer variability, as reported in Table 23, in section 5.4. Burrell et al. (2001) reported calcifications as representing high levels of false negatives; however Tot et al. (2005) reported some calcifications are detected with high accuracy. The findings of this reported study, with sensitivity levels for detecting calcifications of 90.0%, support the findings of Tot et al. (2005). Tot et al. (2005) also reported that stellate lesions are detected with high accuracy; however the results of this study indicate moderate sensitivity levels of 77.6% for the radiographers detecting stellate lesions. This is potentially due to a proportion of the stellate lesions in mammograms of this image set being of young females with dense breasts. Breast density may also explain the detection difficulty experienced by both radiologists initially screen reading these images, that was noted by the researcher when collecting original screen reading data. This finding is supported by previous studies reporting that dense breasts contribute to lesion detection difficulty (Carney et al., 2003; W Reed et al., 2009). The previous radiologist findings of Burrell et al. (2001) and Banik et al. (2012), reporting architectural distortion detection difficulty, are supported by the sensitivity levels of 63.3% in this present research.
Several previous studies reported that radiographers have skills in detecting breast cancers in mammograms considered normal by radiologists (Haiart & Henderson, 1991; Tonita et al., 1999; Wivell et al., 2003; Duijm et al., 2007; Duijm et al., 2008; Duijm et al., 2009). There is a sparsity of studies in this area of radiographer lesion accuracy, however, the results of this reported study agree with the findings of Wivell et al. (2003), who report that radiographers are able to detect calcifications. This finding can be used in future training and education programs, to increase levels of specificity by focusing on and improving areas of weakness, and increasing sensitivity levels by building upon strengths. Specifically, future education and training needs to focus on improving the recognition of architectural distortion, non-specific densities and stellate lesions. In addition, further improvements can be made to the existing high accuracy levels for detecting calcifications and discrete masses. This could be undertaken through the evaluation of selected teaching images with constructive feedback, followed by further analysis of radiographer breast lesion detection in future studies, to improve understanding in this area.

6.3.4. Learning effect

Weak evidence of an improvement in sensitivity, with a $p$-value of 0.056, and significant specificity improvement, with a $p$-value of 0.012, of the radiographer screen readers became evident as they read more images (see section 5.3.4, Table 22). This implies a learning effect, despite the absence of formal screen reading training. It is possible that as the radiographers viewed more images, they became accustomed to screen reading and were able to establish their own pattern of interpretation, thereby improving their ability to differentiate normal from abnormal screens. Pauli et al. (Pauli et al., 1996b) reported an increase in accuracy levels following a training period, and with further practice these skills were maintained. Radiologist studies report an association between increases in sensitivity with increases in volume of mammograms read (Barlow et al., 2004; Moss, Blanks, & Bennett, 2005). These findings indicate that with further screen reading practice, accuracy levels could be expected to
improve, however, further research evaluating accuracy following further screen reading is needed to support this.

6.4. Inter-observer accuracy

Modest inter-observer variability was found when comparing AUC values between the 10 radiographer screen readers (see section 5.4., Table 23), with a $p$-value of 0.17 indicating that there were no significant differences in accuracy across readers. This is also visually observed in Figure 12 in 5.2.3., which shows that ROC curves for all readers often almost overlapped within the ROC space. In this study, Kappa analysis was not undertaken as it only looks at inter-observer agreement, whereas AUC analysis combines observer agreement, with accuracy compared to the gold standard, in order to determine inter-observer accuracy. Previous studies have reported significant accuracy variability between radiologist screen readers (Elmore et al., 1994; Duijm et al., 2009), however the results of this study do not show significant inter-observer variability with this sample of radiographer screen readers in an Australian setting. This could be due to the lack of formal screen reading training and current employment within the same breast screening facility. This potentially encouraged a levelling of knowledge through occasional clinical conversations. This is unlikely, however, as a result of minimal inter-radiographer clinical contact and previous employment in various radiography clinical situations. It could be argued that the generalisability of these results to the remaining Australian population may be limited by this current employment within the same facility; however, one could also interpret the inter-reader level of agreement as providing confidence in these results. The screen readers varied in their radiographic, mammographic and screening experience (see section 5.1.1., Figure 7); there is no evidence, however to suggest that this variation affected inter-observer variability. The daily work these radiographer screen readers undertake is either predominantly performing screening mammograms by some radiographers or predominantly performing ultrasounds in assessment clinics by other radiographers. All radiographers must undertake some work each year in assessment clinics while employed in BSA (BreastScreen Australia, 2008), and this involves
additional specialised mammographic views and assisting biopsy procedures to help ascertain the nature of abnormalities. These differences in experience would be expected to impact upon accuracy levels; however, the results of this reported study indicate that despite differences in experience, inter-observer variability was minimal.

**Limitations**

The limitations of this reported study related to samples of both image test set and screen readers. The method used to compile the image set meant that it was not possible to calculate meaningful radiologist accuracy. It should be noted that this was not an aim of the study. Other image test set sample limitations included the lack of allowance for within-woman correlation and the use of non-digital images in the test set. There was also limited generalisability of the radiographer screen reader samples to the general radiographer population. These issues are further discussed below.

**Image test sets**

A sample limitation in this study was that it was not possible to evaluate radiologist accuracy because of the minimal volume of between 1 and 243 of the total 500 mammograms read by the 16 radiologists who originally reported these mammograms. This meant it was not possible to analyse meaningful radiologist accuracy and to compare this to radiographer accuracy. Further analysis of varying combinations of radiographer and radiologist screen readers was subsequently not possible. In the future, a study analysing an image test set restricted to a similar image volume (minimum of 500), previously read by a limited number of radiologists would potentially enable meaningful radiologist accuracy analysis.

Another image test set limitation involves the issue of within-woman correlation. The results included in both Tables 18 and 19, in section 5.2.1., do not allow for within-woman correlation. This means that each breast is more similar and correlates more to the breast of that same woman, in comparison to the other women in the mammograms comprising the image test set. The significance of this limitation is that these confidence intervals will be minimally under-estimated due to this correlation between each breast of each woman. A
further limitation involves the image test set being comprised of non-digital mammogram images.

This study was conducted using film-screen images and variations in individual image quality may have contributed to accuracy levels, as outlined in section 2.2.1 (Taplin et al., 2002). Previous studies report overall similar accuracy levels when comparing film-screen images and digital images (Del Turco et al., 2007; Hendrick et al., 2008; Kerlikowske et al., 2011). It would be expected that these results will be transferable to digital images, however further research using digital mammogram images is needed to confirm this. In addition to these image test set sample issues is the limited generalisability of the radiographer screen reader sample.

**Screen readers**

The participant radiographer screen readers recruited in this study were employed in one health service of BSA and may or may not represent the entire BSA radiographer population, potentially limiting generalisability. This group of screen readers was composed of radiographers who volunteered to participate in this study, were highly motivated and had varying levels of confidence and experience. It needs to be considered that radiographers of differing levels of experience and motivation may demonstrate varying accuracy levels, such that it cannot be assumed that all radiographers will possess these encouraging accuracy levels.

**6.5. Clinical implications of results**

The results of this study provide strong evidence of the ability of radiographers to detect abnormalities in screening mammograms in an Australian setting. Overall, the results endorse encouraging radiographer screen reader levels of sensitivity and specificity when compared to the gold standard of known assessment outcomes including pathology, 6-year follow-up and interval matching. The screen readers were more accurate at detecting calcifications and discrete masses, than they were at detecting stellate lesions, architectural distortion and non-
specific densities. They also improved in accuracy as they read more images and displayed minimal inter-observer variation. The screen readers demonstrated high recall to assessment rates, as expected in this image test set given that it was enriched with lesions, when compared to acceptable recall rates of consecutive population screen reading.

Formalised training in screen reading can potentially improve these results even further. This support of ability provides further evidence in addition to previous international accuracy studies, of the ability of radiographers to read screening mammograms.

The practice of radiographer screen reading has been embraced within the UK and the Netherlands to improve the accuracy and efficiency of screening programs. These programs are similar to the BSA program, inferring that this practice can also benefit the Australian accuracy and efficiency issues. The BSA NAS state that “For both medical and legal acceptance of the BreastScreen Australia program, it is necessary that at least one reader be a radiologist”, and “if the need arises…specifically trained non-radiologist readers could be employed” (2008, p. 43). Whilst maintaining this important standard, and to maintain the efficiency and accuracy of the BSA program, a radiographer could take on the role of one of the two screen readers. There are, however, some pertinent issues that remain for discussion before radiographers can be considered as screen readers within the BSA program. These issues include systems of reporting, support to practice, medico-legal implications and the radiographer workforce, discussed below.

**Systems of reporting**

The system of reporting used in screen reading requires the screen reader to indicate the presence or absence of an abnormality in a mammogram image, using a screen-reading reporting form. Many radiographers have experience with the established red-dot system that involves a radiographer indicating the presence of an abnormality by placing a red-dot sticker on a radiograph (Field-Boden, 1997; Hall, Jane, & Egan, 1999). Essentially, radiographer screen reading is an extension of this red-dot system, within the specific area of screening mammograms. The evidence, therefore, of established red dot practice (Smith & Baird, 2007),
together with the positive results of radiographer accuracy studies, including the results of this thesis, support radiographer ability to identify breast cancer abnormalities and take on the role as screen readers.

**Support to practice**

Despite this evidence, however, the view of the Royal Australian and New Zealand College of Radiologists (RANZCR) appears to waiver between reserved support and disapproval of radiographer reading (Sutton & Koenig, 2006). The question remaining, now that evidence of radiographer screen reader ability has been provided, is the appropriateness of radiographers beginning screen reading. It has been suggested that radiographers need to actively participate in change to be in control of their future direction. This change needs to incorporate communication and teamwork between radiographers and radiologists to remain client focused and improve outcomes (Thompson & Pollard, 2007; Australian Institute of Radiography (AIR), 2009). The Inter-Professional Advisory Team Report of the Australian Institute of Radiography (AIR) has recommended Advanced Practice status in Radiography (Freckleton, 2012). Screen reading is a potential role of such an advanced practitioner. Potential advanced roles of radiographers that may lead to improvements to the efficiency and accuracy of the BSA program, inherently carry associated issues.

**Medico-legal implications**

When considering these potential future changes, it will be essential to consider medico-legal implications through the increased responsibilities associated with breast cancer diagnosis, particularly if litigation rates increase. Simultaneously, it is essential to recall that for medical and legal acceptance of the BSA program, one of the two screen readers must be a radiologist reader, and double screen reading must continue to be blind and independent to maximise accuracy of cancer detection.

**Radiographer workforce**

Another important implication to consider is whether there is an adequate workforce of radiographers to take on this role of screen reading. It may, in fact, be possible to minimise
normal staff attrition by advancing the role of radiographers in the area of screen reading and creating a stimulating and more challenging workplace. As a matter of interest, some of the radiographer screen readers in this study commented most favourably upon their enjoyment of the challenge and mental stimulation of screen reading and on the positive effect of their improved critical approach to viewing mammograms, following their participation in this study. They also expressed a desire to continue screen reading in the future and indicated that doing so would contribute to encouraging their continuing employment within BSA, and help alleviate professional boredom.
Chapter 7: Summary and Evaluation of Outcomes
Through undertaking a large, well-designed study, using a representative image test set, a robust gold standard, and extensive analysis, evidence has been provided that even without any formal screen reading training, this sample of radiographers have good accuracy levels when screen reading mammograms. The undertaking of this Australian study employing improved design characteristics provides further evidence to build upon the previous results reported in international radiographer screen reader accuracy studies.

The detailed analysis in this study has additionally provided some essentially new accuracy evaluation knowledge. These radiographers’ demonstrated minimal inter-observer variability and improvement in accuracy as they screen read more images. This was unfortunately associated with high recall rates. Furthermore, this study found that radiographers have higher accuracy levels when detecting calcifications and discrete masses, in comparison to stellate lesions, architectural distortions and non-specific densities. Any future education and training programs need to consider these findings to improve accuracy levels and minimise recall rates.

**Further research**

Further research is recommended in the future to evaluate accuracy under varying conditions. This study was undertaken using an image test set because radiographer screen reading is not yet an accepted practice within the Australian context. Whilst this methodology is robust for this image test set, further rigorous research following the employment of radiographers as screen readers in consecutive Australian populations would be beneficial to confirm these favourable results. Pauli et al. have reported that image test set results are transferable to consecutive screening populations (Pauli et al., 1996b), however this remains to be confirmed within the Australian context. Prior to employment as screen readers, radiographers, who may vary in screening mammogram or ultrasound examination experience, should undertake training aimed at minimising recall rates and increasing detection accuracy. Improving detection levels of stellate lesions, architectural distortion and non-specific densities is essential, while building upon the high accuracy detection levels for calcifications and
discrete masses reported in this study. The BreastScreen Reader Assessment Strategy (BREAST) is a program developed to monitor and assess the performance of BreastScreen image readers in detecting abnormalities. It was designed to target areas of detection accuracy improvement and provides immediate feedback of reader performance. This initiative indicates the need to increase screen reader accuracy, and would be a useful tool to employ for further monitoring and assessment of all screen readers (Brennan, Lee, & Tapia, 2012), whether they be radiologists or radiographers. Moreover, the evaluation of breast density as a potential detractor for lesion detection accuracy would be advantageous. Evaluating radiographer accuracy levels when more screen reading volume has been achieved, and under digital technology conditions, following employment as one of the BSA screen readers, would also be beneficial. Additionally, evaluation of accuracy of radiographer pairs and radiographer/radiologist pairs would be valuable. Further research to investigate the influence of screen reading mammogram volume and fatigue on accuracy levels may also be clinically useful.

**Conclusion**

This Australian study imparts evidence that even prior to any formal reading training radiographers have good accuracy levels when screen reading mammograms. It is expected that with formal screen reading training these accuracy levels will further improve, and that radiographers have the potential to be one of the two screen readers within the BreastScreen Australia program and contribute to the timeliness and accuracy of program outcomes.
References


American College of Radiology. (2003). Illustrated breast imaging reporting and data system (BI-RADS), (Fourth ed.).


86


Forrest, P. (1986). Breast Cancer Screening (pp. 100).


Microsoft Office. (97-2003).


Royal Australian New Zealand College of Radiologists (RANZCR). (2006). New Zealand Radiology Workforce Report, from http://www.google.com.au/#hl=en&rlz=1W1GGIT_en-GB&sclient=psy-ab&q=radiologist+workforce+shortage&oq=radiologist+workforce+shortage&gs_l=serp.1.0.33i21.3588.20479.0.23463.64.31.0.7.7.6.444.6665.0j1j2j3j1.27.0.les%3B..0...1c.1.BjRxfKM41&pbx=1&bav=on.2,or.r_gc.r_pw.r_qf.&fp=a15916df00f85ecd&biw=1440&bih=725


Schmidt, F., Hartwagner, K., Spork, E., & Groel, R. (1998). Medical audit after 26,711 breast imaging studies: improved rate of small breast carcinomas (classified as Tis or T1a,b). *Cancer, 83*(12), 2516-2520.


Appendices
Appendix A - Example of traditional radiologists report

RADIOLOGIST'S REPORT
READER 1

☐ No Previous Films
☐ Old Films Available – BSW ☐ Other ☐
☐ Old Films Not Available
☐ Previous recall
Other: ____________________________

INITIALS: ____________________
FILMS UP BY: ____________ FILMS DOWN BY: ____________
FRAME NUMBER: ____________

RIGHT BREAST

<table>
<thead>
<tr>
<th>Result</th>
<th>(1) N</th>
<th>(2) B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3) PB</td>
<td>Possibly Benign</td>
<td>PB (3)</td>
</tr>
<tr>
<td>(4) PM</td>
<td>Possibly Malignant</td>
<td>PM (4)</td>
</tr>
<tr>
<td>(5) DM</td>
<td>Malignant</td>
<td>DM (5)</td>
</tr>
<tr>
<td>(6) TR</td>
<td>Technical Recall</td>
<td>TR (6)</td>
</tr>
<tr>
<td>(7) MX</td>
<td>Mastectomy</td>
<td>MX (7)</td>
</tr>
</tbody>
</table>

LEFT BREAST

R: ____________________
L: ____________________
Comments: ____________________

RADIOLOGIST (Tick Box)

Cahill ☐ Steinberg ☐ Jones ☐ Lee ☐ Shane Connolly ☐ Hunter ☐ Fitzgerald ☐
Hozan ☐ Hammond ☐ Borecky ☐ William Wong ☐ Scott ☐ Stephens ☐
Wood ☐ Duncombe ☐ Library ☐ ☐ Reason: ____________________
Other ☐ Signature ☐ Film handling problem ☐ ☐ Reason: ____________________

PROPOSED WORK-UP

1 Mag Views
2 Cone Compression
3 ML (90°) Views
4 LM Views
5 Ultrasound
6 Clinical Exam
7 Possible FNB

Data Entered: RD ☐ SM ☐ ES ☐ SK ☐ LP ☐ CW ☐ RR ☐ GM ☐ HH ☐ Other: ____________________

READ10296
Appendix B - SWAHS Human Research Ethics Committee

approval letter

Our Ref:  HREC2009/54.10 (2977)  AU RED09/WMEAD85

6 October 2009

A/Prof John Boyages
NSW Breast Cancer Institute
Westmead Hospital

Dear Professor Boyages

Project title: Evaluation of accuracy of radiographers interpreting screening mammograms

Receipt is acknowledge of Mrs Debono’s email dated 22 September 2009 addressing the matters raised in the HREC’s letter dated 2 June 2009 following single ethical review of the above project at its meeting held on 26 May 2009.

It is also noted that the concerns raised by the SAC have been satisfactorily addressed.

This HREC has been accredited by the NSW Department of Health as a lead HREC to provide the single ethical and scientific review of proposals to conduct research within the NSW public health system. This lead HREC is constituted and operates in accordance with the National Health and Medical Research Council’s National Statement on Ethical Conduct in Human Research and the CPMP/ICH Note for Guidance on Good Clinical Practice.

I am pleased to advise that the HREC has now granted ethical approval of this single site research project to be conducted at:

- NSW Breast Cancer Institute, Westmead Hospital - Chief Investigator A/Professor John Boyages

The following documentation has been reviewed and approved by the HREC:

- Research Project Proposal
- Revised Participant Information and Consent Form, Version 9, dated 30 June 2009
- Confidential Participant Information, Version 4, dated 19 December 2009
- Invitation to Radiographers

Providing health services to the communities of
Auburn • Baulkham Hills • Blacktown • Holroyd • Parramatta • Hawkesbury • Penrith • Blue Mountains • Greater Lithgow
Please note the following conditions of approval:

- The coordinating investigator will immediately report anything which might warrant review of ethical approval of the project in the specified format, including unforeseen events that might affect continued ethical acceptability of the project.
- Proposed amendments to the research protocol or conduct of the research which may affect the ethical acceptability of the project, are provided to the HREC to review in the specific format. A copy of all proposed changes is also provided to the relevant research governance officer.
- The HREC must be notified, giving reasons, if the project is discontinued at a site before the expected date of completion.
- The coordinating investigator must provide an annual report to the HREC and a final report at completion of the study, in the specified format. HREC approval is valid for 12 months from the date of final approval and continuation of the HREC approval beyond the initial 12 month approval period is contingent upon submission of an annual report each year. A copy of the Annual / Final Research Report Form is attached and can be obtained electronically from the Research Office on request.
- It should be noted that compliance with the ethical guidelines is entirely the responsibility of the researcher.

You are reminded that this letter constitutes ethical approval only. You must not commence this research project at a site until separate authorisation from the Chief Executive or delegate of that site has been obtained. A copy of this letter and the approved Participant Information and Consent Forms must be forwarded to the SWAHS Research Governance Officer.

A summary of the HREC Standard Operating Procedures is attached for your reference. Should you have any queries about the HREC’s Terms of Reference, Standard Operating Procedures or membership, please contact the HREC Executive Officer through the Research Office on 9845 8183 or emailing researchoffice@wmi.usyd.edu.au.

In all future correspondence concerning this study, please quote your approval number HREC2009/5/4.10 (2977) AU RED09/WMEAD/85.

The HREC wishes you every success in your research.

Yours sincerely

[Signature]

Mrs Jackie Llewellyn
Acting HREC Executive Officer
SWAHS Human Research Ethics Committee (Westmead Campus)
Appendix C- SWAHS Research Governance Office approval letter

19 October 2009

A/Prof John Boyages
NSW Breast Cancer Institute
Westmead Hospital

Dear A/Prof Boyages

HREC reference number: HREC/09/WMEAD/85
SSA reference number: SSA/08/WMEAD/56
Project title: 'Evaluation of accuracy of Radiographers interpreting screening mammograms'
Protocol number: N/a

Thank you for submitting an application for authorisation of this project. I am pleased to inform you that authorisation has been granted for this study to take place at the following site:

- NSW Breast Cancer Institute, Westmead Hospital

The following conditions apply to this research project. These are additional to those conditions imposed by the Human Research Ethics Committee that granted ethical approval:

1. Proposed amendments to the research protocol or conduct of the research which may affect the ethical acceptability of the project, and which are submitted to the lead HREC for review, are copied to the research governance officer;

2. Proposed amendments to the research protocol or conduct of the research which may affect the ongoing site acceptability of the project, are to be submitted to the research governance officer.

Yours faithfully

Ms Tina Goodenough
SWAHS Acting Research Governance Officer

---

Providing health services to the communities of
Auburn • Baulkham Hills • Blacktown • Holroyd • Parramatta • Hawkesbury • Penrith • Blue Mountains • Greater Lithgow

ABN: 70 667 812 600
Post Office Box 63, Penrith NSW 2751
Telephone: (02) 4734 2120 Facsimile: (02) 4734 3737

LH-001

97
Appendix D- University of Sydney Human Research Ethics
Committee approval

The University of Sydney
Human Research Ethics Committee
Web: http://www.usyd.edu.au/ethics/human

Marietta Coutinho
Deputy Manager
Human Research Ethics Administration

Telephone: +61 2 9357 8176
Facsimile: +61 2 9357 8177
Email: mcoutinho@usyd.edu.au

Mailing Address:
Level 6
Jane Foss Russell Building – G02
The University of Sydney
NSW 2006 AUSTRALIA

Ref: MC/KR
18 November 2009

Associate Professor John Boyages
NSW Breast Cancer Institute
Westmead Hospital – C24
Email: johnb@hci.org.au

Dear Professor Boyages

Title: Evaluation of accuracy of radiographers interpreting screening mammograms (Ref. No. 12389)

Masters student: Mrs Josephine Debono

Your application was reviewed by the Executive Committee of the Human Research Ethics Committee (HREC), and in doing so the Committee has ratified your study to include the Masters student – Mrs Josephine Debono.

The Executive Committee acknowledges your right to proceed under the authority of Sydney West Area Health Service Human Research Ethics Committee (Westmead Campus).

Please note, this ratification has been given only in respect of the ethical content of the study.

Any modifications to the study must be approved by Sydney West Area Health Service Human Research Ethics Committee (Westmead Campus) before forwarding a copy to The University of Sydney Human Research Ethics Committee.

Yours sincerely,

Marietta Coutinho
Deputy Manager
Human Research Ethics Administration

cc Mrs Josephine Debono [Email: rayandi@optusnet.com.au]
Appendix E – SWAHS Human Research Ethics Committee

approval letter

SYDNEY WEST
AREA HEALTH SERVICE

HUMAN RESEARCH ETHICS COMMITTEE (Westmead Campus)
Research Office, Room 2020 Clinical Sciences
Westmead Hospital, Hawkesbury Road, Westmead NSW 2145

Telephone: 02 9646 8183
Facsimile: 02 9646 8322
Email: ResearchOffice@swa.health.nsw.gov.au

JH/kh HREC2010/5/1.10 (2977) AU RED HREC/09/WMEAD/85

19 February 2010

Prof J Boyages
Westmead Breast Cancer Institute
Westmead Hospital

Dear Prof Boyages,

Research Proposal: Evaluation of accuracy of radiographers interpreting screening mammograms

Thank you for your letter dated 8 February 2010 in response to the Committee’s letter dated 3 February 2010 concerning the above study.

As the Committee does not have any further ethical concerns, approval has been given for the following documentation, which was reviewed out of session:

- Participation Information and Consent Form, Version 11, dated 8 February 2010

We thank you for your prompt response to the requested amendment.

Yours sincerely,

Dr Jim Hazel
Secretary
SWAHS Human Research Ethics Committee

[Signature]

ABN: 79 067 512 090
Post Office Box 63, Parramatta NSW 2151
Telephone: (02) 4734 3199 Facsimile: (02) 4734 3737

Providing health services to the communities of:
Auburn • Bankstown • Blacktown • Holroyd • Parramatta • Hawkesbury • Penrith • Blue Mountains • Greater Liverpool

LH-001

99
Appendix F- SWAHS Research Governance approval letter

15 March 2010

A/Prof John Boyages  
NSW Breast Cancer Institute  
Westmead Hospital

Dear A/Prof Boyages

HREC reference number: HREC/09/WMEAD/85  
SSA reference number: SSA/09/WMEAD/96  
Project title: ‘Evaluation of accuracy of Radiographers interpreting screening mammograms’  
Protocol number: N/a

Receipt is acknowledge of Ms. Debono’s letter dated 11 March 2010 submitting the following documents for the above mentioned project which will be held on file:

- SWAHS HREC approval letter dated 19 February 2010  
- SWAHS HREC Request for Amendment/Modification dated 11 January 2010  
- Amendment to Study Version 10 dated 7 December 2009  
- SWAHS HREC letter dated 3 February 2010  
- Local Participant Information Sheet and Consent Form Version 11 dated 8 February 2010

I am pleased to confirm ongoing governance approval for this study.

You are reminded that the conditions outlined in your approval letter dated 19 October 2009 continue apply to this research project.

Yours sincerely

Ms Maggie Piper  
SWAHS Research Governance Officer
Appendix G - Research Project Proposal

NSW BREAST CANCER INSTITUTE
RESEARCH PROJECT PROPOSAL

1. TITLE
Evaluation of accuracy of radiographers interpreting screening mammograms.

2. INVESTIGATORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Title and position</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal investigator*</td>
<td>A/Prof John Boyages</td>
<td>Executive Director BCI</td>
</tr>
<tr>
<td>Principal investigator</td>
<td>Dr. Ann Poulos</td>
<td>Academic Supervisor University of Sydney</td>
</tr>
<tr>
<td>Associate investigator</td>
<td>A/Prof. Nehmat Houssami</td>
<td>Lead, Imaging Research Program, NSW BCI</td>
</tr>
<tr>
<td>Associate investigator</td>
<td>Jo Debono, Radiographer</td>
<td>Project Leader and First Author, NSW BCI</td>
</tr>
</tbody>
</table>

*Must be staff specialist or senior member of staff; cannot be higher-degree candidate, resident or registrar

Support staff on the project team will include a film handler/data entry clerk

3. QUESTIONS OR ISSUES –
How accurate are radiographers in interpreting screening mammograms?

4. JUSTIFICATION –
The current and impending shortage of radiologists in Australia is contributing to delays in providing women with results of their screening mammograms. A significant alleviation to this delay of mammogram reporting could be achieved through the utilisation of the untapped resource of the radiographer. Radiographers have been informally interpreting mammograms since the first mammograms were processed.

In some countries such as the United Kingdom and the United States, there has been a realisation that health care professionals, such as radiographers, have a much greater capacity than their roles have traditionally allowed them to fulfil (Field-Boden, 1997). Radiographers have been shown to have very high accuracy in determining abnormal pathology in previous studies (Pauli et al., 1996; Wivell et al., 2003, Holt, 2006). With further training and education, radiographers have been shown to improve their radiographic interpretation skills (Field-Boden, 1996; Loughran, 1994; Sheft, Jones, Brown, & Ross, 1970).

Of particular significance to the field of mammography is the evidence that after appropriate training and education, radiographers are able to read mammograms at least as well as radiologists and they do not take any longer to do so (Wivell et al., 2003).

The wealth of skills that radiographers hold can be formally analysed in this study by evaluating their accuracy in interpreting screening mammograms. The aim of this research is to evaluate the accuracy of radiographers in detecting abnormalities in screening mammograms, with the intention of alleviating the unsatisfactory lengthy waiting times that are sometimes being associated with reporting of screening.
Of particular significance to the field of mammography is the evidence that after appropriate training and education, radiographers are able to read mammograms at least as well as radiologists and they do not take any longer to do so (Wivell, Denton, Eve, Inglis, & Harvey, 2003).

The wealth of skills that radiographers hold can be formally analysed in this study by determining their accuracy in interpreting screening mammograms. The aim of this research is to determine the accuracy of radiographers in detecting abnormalities in screening mammograms, with the intention of alleviating the unsatisfactory lengthy waiting times that are sometimes being associated with reporting of screening mammograms. This can be achieved through the implementation of radiographers and non-radiologists as one of the two readers who read screening mammograms.

5. METHODS

5.1 This is a baseline prospective applied study of non-radiologist readers interpreting screening mammograms using a retrospectively assembled mammography set with validated outcomes. It is planned that up to 10 radiographers will, independently and blindly view and determine whether an abnormality exists in a test set of screening mammograms. The test set will have been previously reported by two radiologists in routine screening. The gold standard to be used will be a double/triple blind radiological report with assessment clinic outcome (in women recalled for assessment) and subsequent negative mammograms in the case of normal mammograms. The status of cancers included in the test set will be verified on the basis of histology (core needle or excision histology). In the reporting of screening mammograms the reader determines whether an abnormality exists which needs further diagnostic work-up. Therefore the nature of the abnormality is determined at the assessment clinic and not at the initial reading of the screening mammogram. In this study what is being determined is only the recognition of whether an abnormality exists.

5.2 The accuracy of radiographers in determining abnormalities in the test set of screening mammograms will be evaluated using the gold standard defined in 5.1.

5.3 All mammograms to be interpreted will be a combination of normal and abnormal pathology randomly selected from the year 2004 of the BreastScreen Sydney West database. It is anticipated that the random selection of abnormal pathologies will represent varying stages of disease progression and will incorporate differing pathologies routinely detected on imaging such as micro-calcifications, masses, stellate lesions, radial scars, etc. It is also anticipated that varying levels of difficulty of interpretation will be incorporated in the test set through random variation in breast density. The mammograms will all be viewed under exactly the same optimum viewing conditions in one facility with no knowledge of other readers’ opinions. Readers will be given a personal identification reader number to record on each of their reports and will remain de-identified except to the Investigators and the person involved in analysis of the data. The investigators and the person involved in analysis will not be one of the film-readers to ensure independence in evaluation of findings.

5.4 Data will be collected using a standardised reader report data collection form, with each reader personal identification number recorded on it. The standardised reporting form will allow a choice of ‘recall’ or ‘no recall’ to
assessment, and if recall is chosen then the side and site of the abnormality will be indicated using conventional MLO and CC diagrams. An arbiter (one of the Investigators) will view the readers' reports and cross-check with the actual abnormality to ensure correlation (in the case of cancers) before data entry.

5.5 Comparative sensitivity and specificity, and PPV for recall, of all readers will be calculated.

5.6 Sample size will be an initial test set of 500 screening mammograms, with a ratio of abnormal: normal of 1:4. Approximately 50 benign lesions recalled to assessment; and shown to be benign on basis of assessment; (and not identified as interval cancers); 50 confirmed malignant; and 400 normal mammograms.

6. IMPLEMENTATION, RESOURCES AND TIME FRAME

6.1 Scientific approval to be sought following discussion and agreement by all investigators on the exact format of this study. Ethical approval will be sought from the ethics committee of SWAHS and the University of Sydney. This will be sought following approval by the scientific committee. The data to be utilised is already collected in the BSSW database. None of the existing patients will have an alteration to their management as a result of this study. The project will commence immediately following ethics committee approval.

6.2 Access to a PC for purpose of recording database. This database will be secured by password protection. De-identified data will be used in the analysis phase.

6.3 Data to be collected by the above nominated investigators from BSSW database.

6.4 Data analysis to be performed by BCI research unit and Jo Debono with the assistance/verification of a qualified statistician (to be recruited).

6.5 Budget: Funding for basic administration costs (reprints of articles to be used as reference material). Paying participants for reading time. Paying statistician. Paying researcher to collate and correlate test set.

6.6 Time frame – 2.5 years

6.7 Analysis of data – 6 months.

6.8 Initial draft paper to be completed within 4-6 months of completion of data analysis.

7. ASSOCIATED OR FURTHER STUDIES –

The findings from the study will be used to determine whether further evaluation of the accuracy of radiographers participating as one of 2 readers but not the third reader will be undertaken. A further study, depending upon the results from this study, could follow an intervention of further training in image interpretation. The radiographers and non-radiologists could then interpret a subsequent set of screening mammograms. Comparisons could again be made to determine the accuracy of the readers and calculate change, if any, in diagnostic accuracy.
Appendix H- Invitation to participate in study

An invitation from the University of Sydney to radiographers

Are you interested in reading screening mammograms?

Have you worked in mammography screening at a BreastScreen NSW Screening and Assessment Service?

If you answered ‘Yes’ to both questions, then this study is for you!

You are invited to participate in a new Australian study entitled ‘Evaluating the accuracy of radiographers interpreting screening mammograms’.

If this project interests you and you would like some further information, then please contact Jo Debono on 0402 232 511.
Appendix I- Participant Information and Consent Form

Chief Investigators: A/Professor John Boyages and Dr. Ann Poulos.

Associate Investigators: Mrs. Jo Debono and A/Professor Nehmat Houssami.

Breast Cancer Institute and University of Sydney.

Invitation

You are invited to participate in a research study which aims to evaluate the accuracy of radiographers interpreting screening mammograms. The accuracy of interpretation of normal and abnormal pathology in a test set of 500 screening mammograms will be evaluated.

This collaborative study is being conducted by:

Breast Cancer Institute (BCI) and University of Sydney,

and will form the basis for the degree of Master of Applied Science of Mrs. Jo Debono under the supervision of:

Associate Professor John Boyages (Director, BCI),

Dr Ann Poulos (Discipline of Medical Radiation Sciences, University of Sydney), and

Associate Professor Nehmat Houssami (Research Consultant, BCI) and Research Academic (University of Sydney).
Before you decide whether or not you wish to participate in this study, it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully and discuss it with others if you wish.

**What is the purpose of the study?**
The purpose is to investigate and evaluate the accuracy of radiographers interpreting screening mammograms.

**Who will be invited to enter the study?**
You are invited to participate in this study because you are a professional radiographer who has experience in the viewing of screening mammograms.

**Do you have a choice?**
Participation in this study is voluntary. It is completely up to you whether or not you participate. If you decide not to participate, it will not affect your current or future employment. If you wish to withdraw from the study once it has started, you can do so at any time without having to give a reason.

However, if you do decide to withdraw from the study, you will be asked if your de-identified data can still be used in the study. You may decide to withdraw all of your data with no ramifications.

**What will happen in the study?**
If you agree to participate in this study, you will be asked to sign the Participant Consent Form. For this study you will be asked to view a test set of 500 mammograms under optimum viewing conditions and indicate on a standardised reader report data collection form whether an abnormality exists. You will need to attend Parramatta BreastScreen located at
Jeffery House, 158 Marsden St. Parramatta for approximately 2 hours on approximately 10 occasions. This study will be conducted over a period of approximately 6 months.

You will be allocated a personal identification number to record on each of the reader report forms to maintain your anonymity. Data will remain de-identified during all aspects of the study. All data, including results, will be strictly confidential and only the investigators named above, Dr. Greg Heard (Research and Information Manager, BCI) and the person involved in the analysis of the data will have access to information on participants. All data will be secured by password protection. A report of the study will be submitted for publication, but individual participants will not be identifiable in such a report.

**Are there any benefits?**

This study aims to evaluate the diagnostic accuracy of radiographers interpreting screening mammograms; however, it may not directly benefit you.

**Confidentiality / Privacy**

Any information that is collected about you in connection with this study will remain confidential. Only the researchers named above and Dr. Greg Heard, Research and Information Manager, Breast Cancer Institute, will have access to your details and results that will be held securely at the Breast Cancer Institute.

**Will taking part in this study cost me anything, and will I be paid?**

Participation in this study will not cost you anything. You will be reimbursed for your time.
What happens with the results?

If you give us your permission by signing the consent document, we plan to discuss the results with the HREC for monitoring purposes. Results will be published in peer-reviewed journals, a master’s thesis and presented at conferences and other professional forums. In any publication, information will be provided in such a way that you cannot be identified. Results of the study will be provided to you, if you wish.

Complaints

This study has been approved by Sydney West Area Health Service Human Research Ethics Committee. If you have any concerns about the conduct of the study, or your rights as a study participant, you may contact:

The Secretary, SWAHS Human Research Ethics Committee

Telephone No 9845 8183 or email researchoffice@wmi.usyd.edu.au

HREC project number- Code AB/4306/1

Contact details

When you have read this information, the researcher, Mrs. Jo Debono, will discuss it with you and any queries you may have. If you would like to know more at any stage, please do not hesitate to contact her on 0402 232 511. If you have any problems while on the study, please contact

Mrs. Jo Debono

Telephone No – 0402 232 511

Thank you for taking the time to consider this study.

If you wish to take part in it, please sign the attached consent form.

This information sheet is for you to keep.
CONSENT TO PARTICIPATE IN RESEARCH

Name of Researchers: A/Prof. John Boyages;
Dr. Ann Poulos;
Mrs. Jo Debono and
A/Prof. Nehmat Houssami

1. I understand that the researcher will conduct this study in a manner conforming to ethical and scientific principles set out by the National Health and Medical Research Council of Australia and the Good Clinical Research Practice Guidelines of the Therapeutic Goods Administration.

2. I acknowledge that I have read, or have had read to me the Participant Information Sheet relating to this study. I acknowledge that I understand the Participant Information Sheet. I acknowledge that the general purposes, methods, demands and possible risks and inconveniences which may occur to me during the study have been explained to me by ____________________________ (“the researcher”) and I, being over the age of 16 acknowledge that I understand the general purposes, methods, demands and possible risks and inconveniences which may occur during the study.

3. I acknowledge that I have been given time to consider the information and to seek other advice.

4. I acknowledge that refusal to take part in this study will not affect my current or future employment.

5. I acknowledge that I am volunteering to take part in this study and I may withdraw at any time.

6. I acknowledge that this research has been approved by the Sydney West Area Health Service Human Research Ethics Committee.

7. I acknowledge that I have received a copy of this form and the Participant Information Sheet, which I have signed.
Before signing, please read ‘IMPORTANT NOTE’ following.

IMPORTANT NOTE:

This consent should only be signed as follows:

1. Where a participant is over the age of 16 years, then by the participant personally.

Name of participant _________________________________ Date of Birth __________________

Address of participant ________________________________________________________________

Signature of participant ______________________________ Date: ______________________

Signature of researcher ____________________________ Date: ______________________

Signature of witness ______________________________ Date: ______________________
Appendix J- Radiographer Report Form

RADIOGRAPHER'S REPORT

OLD FILMS
☐ No previous Films
☐ Old Films Received
(If first visit)
☐ Old Films Not Received
☐ Old Films Received but Not Relevant - Returned to Client

Radiographer checked -OK
Initials

HISTORY

Excisional Biopsy? (removal of a lump from the breast which proved benign)
No ☐ Yes ☐ If Yes: Total Number of excisional Biopsies: Right ☐ Left ☐

Other Breast Surgery? (eg. reduction, mastectomy)
No ☐ Yes ☐ If Yes: Right ☐ Left ☐ Year of Surgery: Type of Surgery:

Breast Prosthesis?
No ☐ Yes ☐ If Yes: Right ☐ Left ☐ Year of Surgery:

SYMPTOMS

Do You Have A:
Breast Lump? No ☐ Yes ☐ If Yes: Which Side? Right R/L Left
Nipple Discharge? No ☐ Yes ☐ If Yes: Which Side? Right R/L Left
What Colour? Clear ☐ Bloody ☐ Other ☐

Do you have any other symptoms?
No ☐ Yes ☐ If Yes: comment

SIGNs (Indicate all scars, moles and lumps on diagram)
Was a sign present which the client was unaware of?
No ☐ Yes ☐ If Yes: comment

1st Visit Radiographer Initials ☐ Machine No. ☐ Total No. Films Presented ☐

Complexity:
Disabled Client ☐ Frozen Shoulder ☐ Implants ☐ Large Breasts ☐

Repeat At:
Initial Visit ☐ Radiographer Initials ☐ Machine No. ☐ No. Repeat Films ☐
Second Visit ☐ Radiographer Initials ☐ Machine No. ☐ No. Repeat Films ☐

Repeats Due To:
Processor ☐ X-ray Machine ☐ Human Error ☐ Other, specify: ☐

Reason for Repeat:
Positioning ☐ Image Blur ☐ Light Films ☐ Dark Films ☐
Fog ☐ Processing/handling artifacts ☐ Incomplete Client ID ☐
Double Exposure ☐ Other ☐

Film Sorted Administration
Initials

111
Appendix K- Standardised Reader Reporting Form

<table>
<thead>
<tr>
<th>Reader ID</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>No previous films</td>
<td>Old films available - BSW</td>
</tr>
<tr>
<td>Old films not available</td>
<td>Previous recall</td>
</tr>
<tr>
<td>Films up by</td>
<td>Films down by</td>
</tr>
<tr>
<td>Frame number</td>
<td>Viewer number</td>
</tr>
</tbody>
</table>

Circle your interpretation (1, 2, 3, 4 or 5) for each breast. Circling 3, 4 or 5 will be considered as 'recall to assessment'.

Please mark on the diagrams above the area that is abnormal.

Comments

<table>
<thead>
<tr>
<th>DATA ENTERED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initials</td>
</tr>
</tbody>
</table>

BreastScreen
NEW SOUTH WALES
Confidential Participant Information

Name:
Age:

Your Personal Identification Number, to be placed on every data record sheet from this point on is ________.

Radiographic experience in years: (please circle)

<5; 5-10; 10-15; 15-20; 20-25; 25-30; 30-35; 35-40; 40-45; 50+

Attained Certificate of Clinical Proficiency in Mammography: (please circle)

Yes/No

BreastScreen Experience in years: (please circle)

<5; 5-10; 10-15; 15-20; 20-25; 25-30; 30-35; 35-40; 40-45; 50+

Non-BreastScreen Experience as a diagnostic mammographer in years: (please circle)

<5; 5-10; 10-15; 15-20; 20-25; 25-30; 30-35; 35-40; 40-45; 50+

'Evaluation of accuracy of radiographers interpreting screening mammograms' Study.

Page 1 of 1