THE UNIVERSITY OF SYDNEY

Evaluating mammography's screening efficacy using image and reader-based characteristics

Dennis Jay Wong

Bachelor of Medical Science (Hons)

Master of Diagnostic Radiography

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Discipline of Medical Radiation Sciences

Faculty of Medicine and Health

University of Sydney

2023

Candidate's Declaration

I, Dennis Jay Wong, hereby declare that the work contained within this thesis entitled 'Evaluating mammography's screening efficacy using image and reader-based characteristics' is my own and has not been submitted to any other university or institution as a part or a whole requirement for any higher degree.

I, Dennis Jay Wong, hereby declare that I was the principal researcher of all the work included in this thesis, including the work published with multiple authors. I certify that the intellectual content of this thesis is my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

In addition, ethical approval from the University of Sydney Human Ethics Committee was granted for all relevant studies presented in this thesis.

I understand that, if a higher degree is awarded for this thesis, the thesis will be lodged with the Director of University Libraries and made available for immediate use.

30 September 2023

Dennis Jay Wong
Discipline of Medical Imaging Sciences
Faculty of Medicine and Health
The University of Sydney
Date: 30/09/2023

Author attribution statement

Chapter 2 Literature Review 1: Do reader characteristics affect diagnostic efficacy in screening mammography? A systematic review published as *Wong DJ, Gandomkar Z, Lewis S, Reed W, Siviengphanom S, Ekpo E. Do reader characteristics affect diagnostic efficacy in screening mammography? A systematic review. Clinical Breast Cancer. 2023 Jan 26*. I primarily contributed in literature search, methodology development, systemic review, summarizing each review, and discussion and conclusion write up of the manuscript.

Appendix I Artificial Intelligence Literature Review: Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review published as *Wong DJ, Gandomkar Z, Wu WJ, Zhang G, Gao W, He X, Wang Y, Reed W. Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review. Journal of Medical Radiation Sciences. 2020 Jun;67(2):134-42*. I primarily contributed to the literature search, methodology development, summarizing each review, summarizing literature's data and tables and discussion and conclusion write up for the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Dennis Jay Wong, 03/10/2023

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Warren Reed, 03/10/2023

Acknowledgements

I would like to express my sincerest gratitude to several key individuals who have given me the opportunity to pursue a significant passion of mine and have helped me tremendously throughout my research and academic career, enabling the completion of my PhD.

First and foremost, I would like to thank my primary supervisor, Associate Professor Dr. Warren Reed, for advising and guiding me during my candidature from start to finish. Dr. Reed's advice, guidance, and supervision have been stellar, and I am very privileged to have had him as my supervisor throughout my journey.

I would also like to thank my auxiliary supervisor, Dr. Ziba Gandomkar, for her support and guidance.

A special thank you goes to Dr. Phuong Trieu Dung and the BreastScreen Reader Assessment Strategy (BREAST) team for providing the datasets used for the thesis. It was an amazing experience working with the BREAST team during this period.

Table of Contents

Table of Figures

Table of Tables

Publications and Presentations

Peer reviewed journal articles:

Wong DJ, Gandomkar Z, Lewis S, Reed W, Siviengphanom S, Ekpo E. Do reader characteristics affect diagnostic efficacy in screening mammography? A systematic review. Clinical Breast Cancer. 2023 Jan 26.

Wong DJ, Gandomkar Z, Wu WJ, Zhang G, Gao W, He X, Wang Y, Reed W. Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review. Journal of Medical Radiation Sciences. 2020 Jun;67(2):134-42.

Wong DJ, Gandomkar Z, Brennan P, Lewis S, Reed W. Mammographic image characteristics in educational test sets: the effect on expert difficulty ratings and subsequent reader performance. Asia Pacific Journal of Oncology. Under Review.

Wong DJ, Gandomkar Z, Reed W. Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis. Journal of Medical Imaging and Radiation Oncology. Under Review.

Wong DJ, Gandomkar Z, Reed W. Investigating the association of textural radiomic features and cancer characteristics with diagnostic errors in mammography interpretation. European Journal of Radiology. Under Review.

Conference Presentations:

Wong DJ, Gandomkar Z, Reed W. Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis. Royal Australian and New Zealand College of Radiologists (RANZCR) 2022 (Accepted by Peer review process; Oral Presentation; Appendix E).

Wong DJ, Gandomkar Z, Reed W. Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis. Medical Image Perception Society (MIPS) 2022 (Accepted by Peer review process; Oral Presentation; Appendix F).

Wong DJ, Gandomkar Z, Reed W. Assessing the diagnostic difficulty of mammographic cases rated by the expert versus reality. Royal Australian and New Zealand College of Radiologists (RANZCR) 2021, (Accepted by Peer review process; Oral Presentation; Appendix D).

Abstract

Aims

The aim of this thesis was to investigate various characteristics influencing the diagnostic performance of radiologists in mammography interpretation and their implications for improving diagnostic efficacy and early detection of breast cancer. The thesis initially focused on expert determined difficulty and whether this subjective evaluation aligned with an objective measure based on reader performance. After exploring the current practices used to develop test sets, we explored reader-based and image-based characteristics on mammography interpretation, which then further employed quantised textural analysis to assess how image textures impact reader performance.

Methods

The studies primarily used linear mixed modelling to evaluate performance accuracy from a total of 479 Australian radiologists participated in this study on all 120 cancer cases available from BREAST was extracted. The initial study only used 165 readers and one test set, while the other two used all the available six test sets and 479 reader's performance. A multivariate model analysis used in the second study created three models using case-based characteristics, reader-based characteristics, and mixed-characteristics. The final study used radiomics analysis, where the cancer cases were segmented using MatLab with ground truth and annotations provided by BREAST. The pathology was segmented and the mammography's pathology region of interest in the pathology cases.

Results

Expert-predetermined difficulty in breast cancer cases was found to be influenced by breast density classification and whether the cases were normal or cancer-positive. However, there was no significant difference observed between the actual difficulty of cases as determined by reader performance and the expert-predetermined difficulty. The study also indicated that around 10 readers were sufficient to establish a reliable difficulty metric.

In the second study, in normal cases, image-based characteristics played a significant role in determining the expert-determined difficulty, with decreasing performance as the difficulty increased. In the reader-based characteristics model, the interpretation of weekly cases and the

reader's specialization were found to be significant. However, in cancer cases, none of the image-based characteristics were found to be significant. Only the years of experience in the specialty showed significance, with increasing years correlating with improved performance. A mixed model analysis revealed that expert-determined difficulty was significant, with increasing difficulty leading to reduced reader performance. Additionally, the number of cases interpreted per week was also significant in influencing reader performance.

Regarding the detection of different types of cancer using quantised features compared to non-quantised textural features, Haralick quantization method demonstrated the detection of calcifications more easily. However, stellate masses with non-specific densities performed better than spiculated masses alone. Other methods like Compact GLDM and Statistical Feature Matrix supported previous findings, showing better detection rates for spiculated masses compared to stellate masses with non-specific densities. The findings revealed that expert-predetermined difficulty and the number of cases interpreted weekly consistently impacted performance. Higher difficulty levels resulted in reduced performance, whereas an increased number of cases interpreted positively influenced performance.

Conclusion

The thesis demonstrated the significance of expert-predetermined difficulty, cases interpreted weekly, and reader- and image-based characteristics in mammography interpretation. It underscores the importance of these factors in determining performance outcomes and emphasizes the need for continued exploration of texture-based radiomic analysis to unlock further insights into the relationship between cancer characteristics and diagnostic efficacy. These findings contribute to the ongoing efforts to enhance screening and diagnostic practices, ultimately improving patient outcomes in breast cancer detection and management.

Chapter One
Introduction

1.1 Understanding the Research Problem Addressed by this Thesis

1.1.1 Introduction to BreastScreen Australia

Mammography has been shown to be an essential tool in breast cancer detection. Breast cancer is the most frequent cancer in women and the second most common cancer globally(1). Approximately 1.67 million new cases were recorded in 2012, accounting for one-quarter of all cancer diagnoses (1). Furthermore, breast cancer is the most common cancer in women worldwide and ranks fifth in terms of total cancer-related deaths, claiming the lives of 522,000 people. Breast cancer is currently the most common cancer in women in Australia. The number of newly diagnosed cases in the country has more than doubled over the last three decades, rising from 5,310 to 13,567 (2).

Screening mammography initiatives have played a critical role in improving the well-being and healthcare of women globally, especially since the introduction of BreastScreen Australia for Australian women in 1991 (3). This nationwide breast screening project has been extremely beneficial to women because it allows for the early diagnosis of potentially malignant breast anomalies, leading to more successful treatment outcomes. The screening mammography participation rate among eligible Australian women aged 50 to 74 years was 54.8% between 2018 and 2019, maintaining a consistent level with a modest increase from the 53.7% rate reported between 2014 and 2015 (4). The National Accreditation Standards (NAS) of BreastScreen Australia aim to achieve a minimum participation rate of 70% for eligible women in the targeted age range.

Breast cancer has caused fewer deaths among Australian women in recent years. Early identification of breast cancer using mammography has been shown to significantly improve survival chances (5). According to one study, women diagnosed with invasive ductal carcinoma with smaller tumour sizes had a 90% probability of surviving more than five years in 2015, compared to 72% between 1982 and 1987 (6). This increase in survival rates can be attributed to the efficacy of mammography screening and recent advances in treatment techniques. As a result, breast cancer mortality rates have decreased, with the number of deaths per 100,000 women aged 50-69 in the general population falling from 68 in 1991 to 42 in 2012 (7).

Approximately 88% of services in Australia have the capacity to read 2,000 screening cases within the program annually, according to national statistics (8). To ensure high-quality screen reading, all BreastScreen readers are required to read a minimum of 2,000 mammographic screening cases per year within the program. The annual total number of cases read by a reader who screens at BreastScreen Services and/or screening and assessment centres (SCUs) are added to calculate their overall performance.

BreastScreen Australia employs a two-view mammography strategy as part of the screening process, which involves obtaining cranio-caudal and medio-lateral oblique radiographs. Within a nationwide screening program, this screening technique has been demonstrated to be effective in the early diagnosis of breast cancer. Typically, each client has four images taken as part of a routine screening assessment (two cranio-caudal and two medio-lateral oblique). To ensure comprehensive screening, additional projections may be required for women with larger breasts.

1.1.2 Criticisms and Controversies around BreastScreen Program

While mammography can detect early signs of cancer and changes through routine screening, it can also potentially introduce adverse outcomes such as overdiagnosis, false positive results, and false negative findings (9, 10).

1.1.2.1 Ongoing debate: screening mammography and overdiagnosis

Overdiagnosis refers to the detection of cancers that may never progress to become clinically significant. In breast cancer screening, it is estimated that out of every 1,000 women screened over a 10-year period, five healthy women may be over-diagnosed with breast cancer. These women may undergo unnecessary treatments or additional imaging. The rate of overdiagnosis can be as high as 42% in women aged 50-59, raising significant concerns in breast screening programs (9).

Another potential drawback of population-based screening mammography programs is the false positive recall rate. False positive recalls occur when women are called back for further assessment but are ultimately found not to have breast cancer. It has been estimated that the cumulative risk of experiencing a false positive result, over the course of 10 biennial screening tests, ranges from 8% to 21% for women aged 50-69 (11).

These factors, namely overdiagnosis and false positive recall rates, pose challenges and potential harm to the success of screening mammography programs.

1.1.2.2 Criteria in guidelines for certifying the readers

To ensure quality in screening mammography, guidelines often incorporate reader characteristics as a means of certifying individuals for independent reporting of mammograms and maintaining their certification. The primary criterion used for certification is the volume of mammograms interpreted per year. However, the specific minimum volume requirements vary across countries, often driven by the need to deliver adequate services according to the needs of their metropolitan and rural populations. For instance, the United States mandates a minimum of 960 mammogram reads over a span of two years. In the United Kingdom and other European countries, the minimum annual requirement is 5,000 mammogram reads. In Australia, the minimum volume requirement for certification is 2,000 mammogram reads. These variations reflect different national standards and guidelines implemented by each country to ensure the expertise and proficiency of mammography readers (12).

To maintain the effectiveness and standards of the population screening mammography program, BreastScreen Australia has implemented a quality improvement program. This program is designed to comply with the National Accreditation Standards (NAS) and ensure the delivery of high-quality services (8). To assess the program's progress in reducing morbidity and mortality from breast cancer, various performance indicators and measures are utilised. These include sensitivity, specificity, positive predictive value (PPV), and recall rates. By regularly monitoring and evaluating these indicators, BreastScreen Australia can identify areas for improvement and take appropriate measures to enhance breast cancer detection and minimize associated health risks (8).

1.1.3. Solution 1 - Technological advancement in image acquisition

Previous studies has yielded conflicting results regarding the association between case difficulty and breast density. Earlier research focused on film-based mammography and found that as breast density increased, mammogram diagnostic performance decreased (13, 14). This was due to the masking effect generated by radio-opaque structural features potentially overlapping with pathology, resulting in decreased reader sensitivity (15, 16). However, new research suggests that when utilising digital mammography instead of film-based

mammography, the masking effect of thick breasts is greatly reduced (15, 17). Some studies even imply that experienced radiologists who use digital mammography may enhance their performance as breast density increases due potentially to increased vigilance in those areas (18).

1.1.4 Solution 2: Development of educational materials and Self-assessment modules

Radiologists often face cognitive biases and errors that can affect their decision-making process. Meta-cognition, which refers to the awareness and understanding of one's own cognitive processes and underlying reasons for errors, helps radiologists become aware of these biases and consciously counteract them (19).

In the screening process, the radiologist can potentially face bias in their interpretation. These biases include anchoring bias, where they rely too heavily on initial information; confirmation bias, leading to selective interpretation; availability heuristic, based on easily accessible information; satisfaction of search, prematurely ending an evaluation; premature closure, rushing to a conclusion; inattentional blindness, missing unexpected findings; and hindsight bias, viewing outcomes with undue certainty(20). All the aforementioned biases may lead to incorrect diagnoses, resulting in compromised healthcare outcomes such as false negatives, which can result in missing significant pathology, and false positives, which can lead to unnecessary patient recalls or unnecessary treatments.

1.1.4.1 PERFORMS

In the United Kingdom, the Personal Performance in Mammographic Screening (PERFORMS) system serves as a comparable self-assessment platform. It provides radiologists with expert insights, established case pathology, peer opinions, and individualised training and quality assurance (21). The PERFORMS system in the UK collects test set cases from screening centres, along with the original radiologist and pathology reports. A panel of ten expert radiologists scrutinizes key mammographic features in each case, such as breast density and the challenge of determining if the case should be included in PERFORMS (22, 23).

Both the UK PERFORMS, and the Australian BREAST platforms aim to replicate the real clinical setting. However, to enhance the practicality of these platforms, the test sets need to be deliberately oversampled with positive cases.

1.1.4.2 Breast Screen Reader Assessment Strategy (BREAST)

In Australia, a self-assessment tool called the Breast Screen Reader Assessment Strategy (BREAST), created, and developed by medical imaging experts at the University of Sydney and senior radiologists at BreastScreen New South Wales, has been utilised for quality training in BreastScreen services since 2011. The BREAST system operates through an innovative web-based application that provides radiologists with a range of performance scores and immediate, personalised feedback on overlooked cancers and false-positive choices. The BREAST system has gained the participation of 80% of breast screen radiologists throughout Australia and is widely regarded as one of the most effective professional development tools for radiologists (24). Immediate image-based feedback is provided after completing a test set, allowing comparison of the radiologists' selection with the accurate results. There has been reported a significant improvement in lesion sensitivity, ranging from 21% to 31% (24). Furthermore, 83% of radiologists reported enhanced performance in lesion sensitivity after their initial test. Registrars notably improved in the third test compared to the first one, with an average rise in lesion sensitivity of 79% (24).

In Australia, the Breast Screen Reader Assessment Strategy (BREAST) program has been implemented as a training tool within the breast cancer screening program. This program offers multiple test sets to help mammography readers evaluate and enhance their performance in interpreting mammograms.

The test sets provided in the BREAST program typically consist of four digital mammograms, representing both the left and right breasts in the medio-lateral oblique (MLO) and cranial caudal (CC) projections. The mammograms and their details are de-identified to protect patient privacy.

The cases included in the test sets are carefully selected to present a range of difficulties and challenges to the participants. Cancer-positive cases in the test sets are confirmed through biopsy, ensuring their accuracy. Conversely normal cases are determined through the consensus reading of at least two experienced radiologists who review the cases following negative screen reports by two clinical radiologists.

The selection of cases and their level of difficulty in the BREAST program are based on the assessment of an expert radiologist. This ensures that the test sets provide valuable opportunities for self-assessment, education, and training for the participating mammography readers.

To assess and rate the cases, the readers utilize the breast imaging rating system developed by the Royal Australian and New Zealand College of Radiologists (RANZCR). This rating system assigns a grade ranging from 1 to 5 to each case, with each grade representing a different level of suspicion for malignancy. The rating categories are as follows: 1 for "Normal," 2 for "Benign," 3 for "Equivocal," 4 for "Suspicious," and 5 for "Malignant."

The readers assign a score classification between 3 and 5 to indicate increasing confidence in the malignancy of the marked area. If the reader does not identify any abnormalities, they proceed to the next case by clicking "next case" on the platform. In such cases, the platform automatically marks the case as RANZCR score 1, indicating a normal finding.

It is important to note that the readers are not aware of the prevalence or specific types of cancers included in the test sets. This approach helps ensure unbiased evaluation and interpretation of the cases, as the readers' assessments are solely based on the provided images and their clinical expertise.

1.1.5 Solution 3: Using AI Tools

Over the past two decades, significant developments have occurred in computer-aided diagnosis systems and artificial intelligence (AI) tools aimed at assisting radiologists in detecting and diagnosing breast cancer. These tools utilise textural features extracted from mammograms to identify suspicious lesions. In 1998, the FDA granted clearance for computer-aided detection to be used as an additional tool in mammography and it was subsequently approved for reimbursement by the Centres for Medicare & Medicaid Services. However, a notable study nearly a decade later (25) revealed that the older generation of computer-aided detection tools, which relied on traditional machine learning methods, did not improve the accuracy of mammography in routing clinical practice.

In contrast, recent AI tools based on deep learning have emerged, demonstrating significantly higher levels of accuracy. In respective research studies, these advanced AI tools have matched

or even surpassed the average performance of breast radiologists (26). The utilisation of deep learning algorithms has propelled the field forward, enabling more precise and reliable detection of breast cancer.

Additionally, modern research related to artificial intelligence has a primary focus on using parenchymal textural characteristics of the breast for pathology identification in machine-driven studies and computer-aided diagnostic research (27). Although the potential of computer-aided diagnosis and artificial intelligence is encouraging, there is a need for standardization before integrating them into the regular diagnostic process for mammography screening (28).

1.2 Knowledge Gap

Relying solely on technological advancements and the application of AI to improve image accuracy is insufficient in addressing the challenges related to false positive and false negative results (9, 10). Therefore, it is crucial to enhance the performance of readers who interpret screening mammograms. One approach to achieve this is through the establishment of guidelines that define the minimum requirements for becoming screen-readers. Among the factors considered in these guidelines, reading volume, which refers to the number of screening mammograms interpreted by an individual, holds significant importance. However, these guidelines vary across different regions. For example, the minimum annual reading volume is set at 480 cases in the US, 2000 in Australia and 5000 in some European countries. Hence, it is essential to conduct a comprehensive literature review to determine the characteristics that should be included in the guidelines and evaluate the existing evidence supporting these guidelines. The first objective of this thesis – to identify the current literature in evaluating reader performance, aims to address this gap.

Self-assessment modules are widely recognised as a key tool for enhancing the performance of screen-readers. Currently, these modules rely on expert-determined case difficulty to curate the test sets. However, it is uncertain whether expert-determined case difficulty accurately reflects the actual difficulty of the cases. In this thesis, a large-scale study was conducted to investigate whether expert-determined case difficulty is a suitable metric for estimating the actual case difficulty.

To provide an objective assessment of case difficulty, the utilization of quantitative features can be advantageous. While incorrect positioning or exposure is often attributed to missed cancer diagnoses in mammography, previous research suggests that a comprehensive investigation of other features is necessary (25). Among various factors, the impact of breast density on diagnostic performance has been extensively studied. Some studies have indicated that missed cancers on mammograms are more likely to occur in denser breasts (26), while others have suggested that missed cancers tend to have lower density and smaller size compared to easily detectable lesions. Limited data is available regarding the influence of other radiographic appearances of the disease, such as overall lesion morphology and texture (25).

Previous studies have investigated the role of shape and boundary roughness in mass lesions (1, 25, 29, 30). These studies have revealed that mammographic sensitivity may be compromised if spiculation on the boundary and its roughness are not adequately considered. While various quantitative shape-based radiomic features have been included in these studies (31-33), the contribution of textural features to the difficulty of cancerous lesions has not been examined. The final objective of this study, which is to use multi-reader multi-case analysis to validate the impact case textural characteristics has on reader performance, aims to address this gap.

Previous research has explored the use of texture features in mammography and breast MRI to identify the risk of developing breast cancer based on the evaluation of parenchymal textures extracted from these imaging modalities (31-33). Currently, parenchymal textural features of the breast are primarily utilised in machine-driven identification and computer-aided diagnosis studies (27). However, the association between parenchymal textural features and case difficulty has not been explored. Additionally, different test sets may provide the most educational value for different groups of readers based on their characteristics including different levels of breast density combined with different lesions of varying sizes. Therefore, it is necessary to establish the association between readers, case characteristics, and case difficulty.

1.3. Thesis Objectives and Research Goals

The overall aim of this thesis was to explore the current practices in educational materials and quality control test sets for mammography interpretation education in Australia and determine

the reader and image factors that would impact a reader's performance in mammography interpretation and screening. This thesis works to address this aim by establishing four main objectives:

Objective 1- Overview of the literature in evaluating reader performance (Chapter 2)

The first objective was to provide the foundation and context the thesis based on the current literature on screening mammography. Specifically, it examined the current requirements used by accreditation boards for accrediting mammography screening readers, along with other reader characteristics and metrics commonly used in evaluating reader performance.

Objective 2- Relationship between expert-determined case difficulty and actual case difficulty (Chapter 3):

The second objective evaluated the current practices of mammography interpretation training platform by examining how different image characteristics of mammography cases affect perceived difficulty as determined by experts in a test set. Additionally, the study assessed whether the predetermined difficulty assessment aligns with the actual performance of readers. The study analysed various image characteristics such as breast density, lesion size, and lesion type, to determine their impact on the perceived difficulty of the cases. It also compared the experts' difficulty ratings with the actual performance outcomes of the readers to assess the accuracy of the predetermined difficulty assessment. This study also examined the minimum number of readers required to establish a reliable case difficulty measure and conducted simulations to achieve a robust assessment of case difficulty.

Objective 3 - Explore potential reader and case characteristics that impact reader performance (Chapter 4):

The objective of this study was to examine the influence of reader-based and image-based characteristics on reader performance in mammography interpretation. The study employed linear modelling analysis to investigate the relationship between these factors. Reader-based characteristics include factors such as experience, specialization, and training of mammography readers. Image-based characteristics refer to the properties of mammography images themselves, such as breast density, lesion characteristics (e.g., size, type), and image

quality. Linear modelling analysis examined the relationship between reader-based and image-based characteristics and their impact on reader performance. This statistical approach allows for the incorporation of both fixed effects and random effects into the analysis.

Objective 4 - Using multi-reader multi-case analysis to validate the impact case textural characteristics has on reader performance (Chapter 5):

This chapter aimed to explore the relationship between localised extracted textural features in mammography and cancer characteristics and their impact on a radiologist's diagnostic performance. The study involved the analysis of mammography images to extract textural features in specific regions of interest, which may include measures of texture complexity and heterogeneity. The study also accounted for cancer characteristics such as tumour size, type, and region. Ridge regression modelling analysis assessed how specific textural features correspond to different cancer characteristics and their potential interactions in influencing radiologists' diagnostic performance. The study investigated the impact of these features on the accuracy of breast cancer diagnosis.

These objectives collectively contribute to a comprehensive investigation into mammography interpretation and have the potential to inform improvements in reader education and training and quality assurance in mammography screening programs.

1.4. Methodology and Research Approach

1.4.1 Reader Characteristics

As shown in Figure 1, before beginning the test set, participants in the BREAST program were requested to complete a survey about their prior experience and training. This survey collected information on the reader's characteristics, including age, gender, professional role (either a trainee or radiologist), years of specialty experience, their specialization, the number of cases interpreted per week, fellowship training, and whether the reader is employed in a breast screening program. The collected information from these surveys were used to realise the third and fourth objective of this thesis.

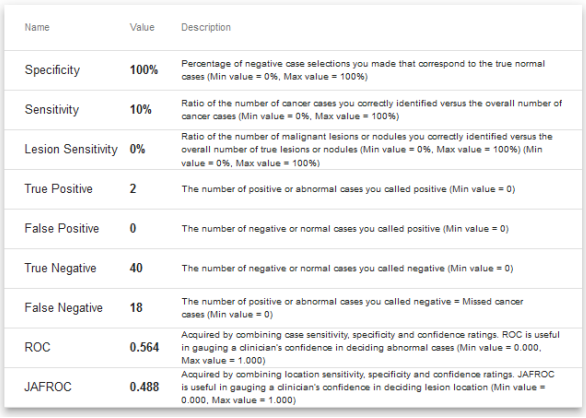
Figure 1. Readers are required to fill in a demographic survey to collect reader characteristics before exploring the test set.

1.4.2 Reader Performance

To assess the cases in the test set, readers were shown the MLO (Mediolateral Oblique) and CC (Cranio-Caudal) views of both left and right breasts per case. The readers had the ability to mark any perceived lesions on the image and utilised the Royal Australian and New Zealand College of Radiologist's (RANZCR) breast imaging rating system to grade the case. The rating system ranges from 1 to 5, where 1 denotes 'Normal', 2 means 'Benign', 3 is 'Equivocal', 4 is 'Suspicious', and 5 signifies 'Malignant'. A score between 3 to 5 was used to demonstrate increasing confidence in the malignancy of the marked area. The interface of the BREAST platform for rating a case is shown in Figure 2.

Figure 2. Test set interface – Readers may select any region of the breast they believe have any malignancy, once selected they are required to provide a RANZCR score, a score above 2 would require the reader to select the lesion type.

If a reader did not find any abnormality, they would proceed to the "Next case", and the system automatically classified such cases as RANZCR score 1, which implies 'Normal'. Only the first attempt by the reader at the test set was utilised. The readers were not informed about the prevalence or types of cancers present in the test sets. Once completed a test set, the reader is given immediate feedback on his/her performance (Figure 3).



Name	Value	Description
Specificity	100%	Percentage of negative case selections you made that correspond to the true normal cases (Min value = 0%, Max value = 100%)
Sensitivity	10%	Ratio of the number of cancer cases you correctly identified versus the overall number of cancer cases (Min value = 0%, Max value = 100%)
Lesion Sensitivity	0%	Ratio of the number of malignant lesions or nodules you correctly identified versus the overall number of true lesions or nodules (Min value = 0%, Max value = 100%) (Min value = 0%, Max value = 100%)
True Positive	2	The number of positive or abnormal cases you called positive (Min value = 0)
False Positive	0	The number of negative or normal cases you called positive (Min value = 0)
True Negative	40	The number of negative or normal cases you called negative (Min value = 0)
False Negative	18	The number of positive or abnormal cases you called negative = Missed cancer cases (Min value = 0)
ROC	0.564	Acquired by combining case sensitivity, specificity and confidence ratings. ROC is useful in gauging a clinician's confidence in deciding abnormal cases (Min value = 0.000, Max value = 1.000)
JAFROC	0.488	Acquired by combining location sensitivity, specificity and confidence ratings. JAFROC is useful in gauging a clinician's confidence in deciding lesion location (Min value = 0.000, Max value = 1.000)

Figure 3. Readers are given immediate feedback on their performance upon completing the test set. If necessary, they may see the answers to compare their selection against the ground truth.

1.4.3 BREAST Test Sets

A single test set from BREAST, comprising 60 mammography cases, was utilised for this study. This included 20 cases confirmed to have cancer and 40 normal cases. The cancer-positive cases exhibited a range of lesion sizes and appearances. Each case contained four mammograms: the cranial-caudal and medial-lateral oblique projections of both left and right breasts.

Additionally, breast density classification data was incorporated, categorised on a scale of 1-4 using the BI-RADS system. The categories were as follows: (1) fatty, (2) fibro-glandular density, (3) heterogeneously dense, and (4) extremely dense.

The test set items were sourced from the Digital Breast Image Library of BreastScreen NSW. Cancer cases were confirmed by surgical pathology, while negative cases were determined by the agreement of at least two senior radiologists after an initial normal screen result by two

readers. Ethics exemption has been granted for the studies in this thesis (Appendix A) and the BAMC and approval for BREAST data's use has been granted (Appendix B and C).

1.4.4 Expert Difficulty

The difficulty level of each case was evaluated by an expert radiologist who previously served as the State Radiologist in BreastScreen New South Wales, Australia. To assign difficulty levels, the expert retrospectively reviewed each case and rated the likelihood of a radiologist potentially overlooking the pathology. The ratings ranged from 1-3, corresponding to 'Easy', 'Intermediate', and 'Difficult' respectively.

1.4.5 Linear and Ridge regression modelling

Linear modelling is a statistical technique for analysing data that accounts for both fixed and random factors in the model. The dependent variable is considered to have a linear relationship with one or more independent variables in linear modelling. Additionally, the presence of random effects allows for the incorporation of individual- or group-level variability.

The fixed effects are the relationships between the independent and dependent variables. These effects, which indicate the average effect of the independent factors on the dependent variable, are evaluated using classic regression methods.

The random effects capture the variation among levels, such as readers with varied features. They account for the variation between individuals within the same group and the variation between different groups is captured by random effects.

Linear modelling can be used to analyse clustered or longitudinal data in which observations are associated in clusters or individuals through time rather than independent. Accounting for random variables, linear models can provide more accurate estimates of fixed effects while also accounting for data uncertainty.

Due to the natural composition of the breast, ridge regression is more effective in handling multicollinearity, which is a situation where independent variables in a regression model are highly correlated i.e., breast density in certain regions of the breast or potential interactions with cancer lesions. High multicollinearity can make it challenging to identify the individual

effect of each predictor variable. Ridge regression can help in stabilize coefficient estimates and making them less sensitive to multicollinearity.

Given the nature of the BREAST platform, which includes test sets and cases with varying image properties and reader performance with varying characteristics, linear modelling is an ideal statistical approach. The code for the linear modelling can be referred to in Appendix G and Ridge Regression's in Appendix H.

1.5. Significance and Contribution of the Thesis

This thesis holds significant importance and makes substantial contributions to the field of mammography education. These include:

1) Large scale study: This research stands out as one of the largest studies conducted in evaluating readers and test set images in mammography. With a dataset spanning several years and involving a substantial number of radiologists and case readings, the study ensures robustness and generalizability of the findings. The dataset used includes the reader performance between 2015-2021. In the dataset, 479 radiologists have completed the selected test set. All images and radiologist's data provided were anonymous and de-identified.
In terms of mammography test sets, A total of six test sets with a total of 60 mammography cases each from the BREAST platform were used. Each test set consisted of 20 cancer-positive cases and 40 normal cases. The cancer-positive cases had a variety of lesion sizes and appearances confirmed by biopsy. Normal cases were confirmed as normal based on a two-year follow-up. Each case consisted of four mammograms: the left and right craniocaudal projections and the mediolateral oblique projections.

2) Rigorous Methodology: The research employed advanced statistical techniques, particularly linear modelling, and ridge regression, to analyse the data. This approach accounts for the variability introduced by multiple readers and cases, leading to more accurate and reliable results with higher statistical power.

3) Application of radiomic features, describing the mammographic texture: This thesis introduces the application of radiomic features in predicting case difficulty. While these features have been used in computer-aided detection and survival analysis, their potential in assessing case difficulty has been largely underexplored. This research sheds light on the

factors influencing the complexity of cancerous lesions, contributing to a deeper understanding of their characteristics and implications for diagnostic accuracy.

4) Enhancing radiologists' meta cognition: The findings of this research have implications for improving radiologists' performance and meta-cognitive abilities. By informing radiologists about the specific textural features analysed in this study, they can develop an awareness of their cognitive processes and recognize potential errors. This heightened meta-cognition can empower radiologists to counteract biases and improve their diagnostic accuracy, ultimately leading to better patient outcomes.

5) AI integration and human observer performance: This study explores the complex interactions between textural features and radiologists' accuracy, which has broader implications for the integration of AI in radiology. Understanding how these features influence human observer performance can guide the development of AI systems that complement and assist radiologists, enhancing overall diagnostic accuracy and patient care.

6) Applicability beyond mammography: The methodologies employed in this study for test set curation, assessment, quality assurance, and continuing education have broader applicability beyond mammography. They can serve as a foundation for research and advancements in other medical disciplines, such as pathology and other diagnostic fields.

In conclusion, this thesis makes a significant contribution to the field of mammography education by conducting a large-scale study, utilizing advanced statistical techniques, proposing new applications for radiomic features, promoting meta-cognition among radiologists, highlighting the importance of AI integration, and considering the applicability of the study beyond mammography. The insights gained from this research have the potential to drive further advancements in the field of medical imaging and analysis.

1.6 Thesis Structure and Organization

The thesis structure and organization are presented visually in Figure 4, and a brief description of each chapter is provided below.

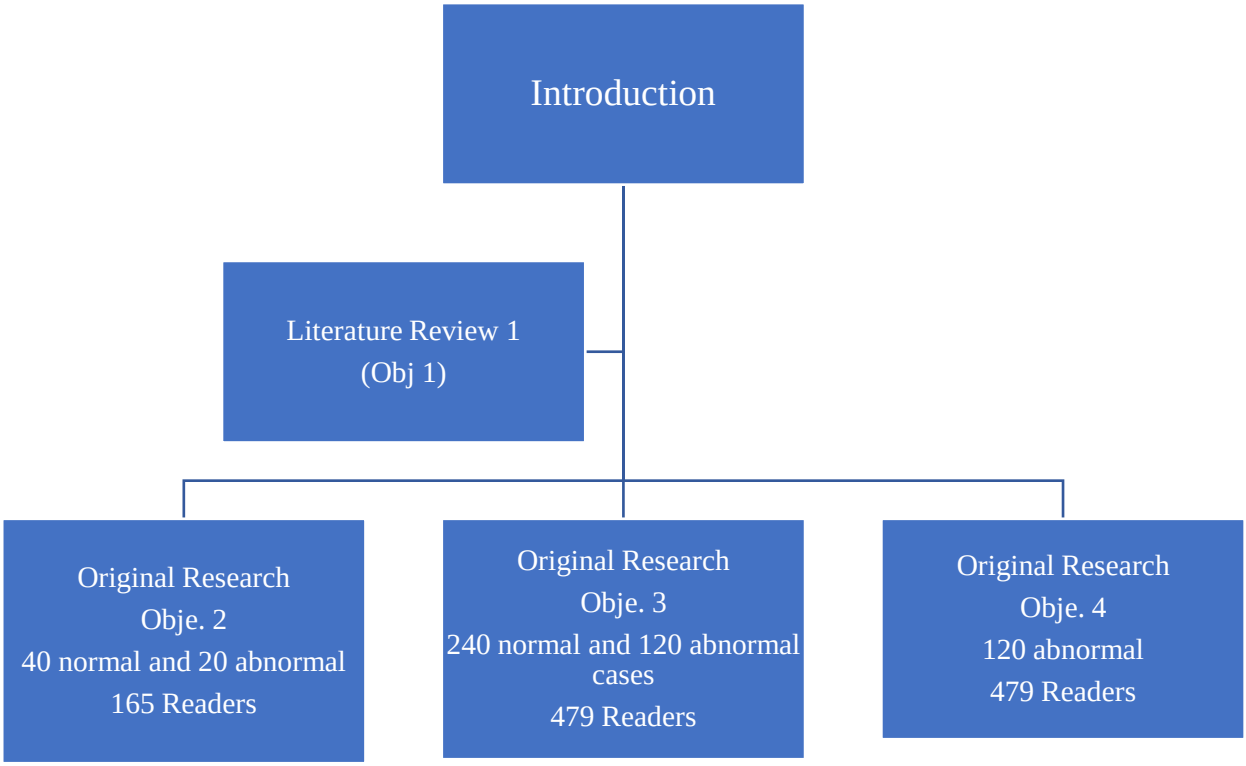


Figure 4. Thesis Structure and Organization

1.6.1 Introduction chapter

The introduction chapter sets the context for the thesis by providing background information on key aspects of the topic. It explores the rationale and need for the study, discussing mammography, breast screening in Australia, common screening challenges such as over-diagnosis, accreditation standards for participating in dappbreast screening, and the training platforms designed to improve breast interpretation. The chapter highlights current practices in training services and accreditation standards, serving as a foundation for investigating potential alternatives to address previous research findings.

1.6.2 Literature review

This systematic review investigated how reader characteristics influence the diagnostic efficacy in interpreting screening mammograms. It begins with a comprehensive literature search conducted using databases such as Cochrane, Scopus, Medline, Embase, Web of Science, and PubMed.

1.6.3 The First original Research Paper (Objective 2)

This chapter investigates the existing approaches employed in the BREAST Training and Learning platform for developing training test sets. The focus is on understanding the criteria and characteristics that experts utilise to determine the difficulty of mammography cases for the average mammography reader. By examining these current practices, the chapter aims to gain insights into how the difficulty level of cases is assessed and how it may impact the education and performance of mammography readers using the B REAST platform.

1.6.4 The Second original Research Paper (Objective 3)

This chapter aims to identify the reader characteristics and mammogram image characteristics that influence a radiologists' diagnostic performance. A comprehensive analysis was conducted on a larger set of case images and reader performances, disregarding their specific test set groupings. The results of this analysis revealed differing findings in comparison to the previous chapter. By examining the influence of various factors on diagnostic performance, this chapter provides valuable insights into the complex relationship between reader characteristics, mammogram image characteristics and diagnostic accuracy.

1.6.5 The Third original Research Paper (Objective 4)

This chapter builds on the findings from the previous two chapter by emphasising the evaluation of the effects of cancer characteristics has on reader performance. In addition to utilising metadata characteristics of the mammograms, textural analysis of each mammographic image was conducted using radiomics This analysis also aimed to explore the effects of ant pathology on reader performance. By integrating cancer characteristics and radiomic analysis, this chapter provides further insights into the relationship between pathology features, image textures and their impact on reader performance.

1.6.6 Discussion and conclusion

This chapter provides a comprehensive synthesis of the thesis's major findings, underscores their clinical implications and importance, and highlights avenues for future research and development in the field of mammography.

1.6.7 Bridging sections

Additionally, to enhance the flow and coherence of the thesis, introduction sections have been presented at the start of each chapter. These introductions serve as bridging sections that provide a brief overview of the upcoming chapter, establish a connection with the preceding chapter and set the context for the subsequent discussion. Their aim is to maintain a logical structure that facilitates the readers' understanding of the research progression and the key themes addressed in each chapter.

1.6.8 References

The Vancouver referencing technique, has been used consistently throughout the thesis, ensuring a uniform approach.

References

1. Pohlman S, Powell KA, Obuchowski NA, Chilcote WA, Grundfest-Broniatowski S. Quantitative classification of breast tumors in digitized mammograms. Med Phys. 1996;23(8):1337-45.

2. Health AIo, Welfare. BreastScreen Australia monitoring report 2021. Canberra: AIHW; 2021.

3. Health AIo, Welfare. Cancer screening programs: quarterly data. Canberra: AIHW; 2023.

4. Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. Cochrane Database of Systematic Reviews. 2013(6).

5. van Schoor G, Moss SM, Otten JD, Donders R, Paap E, den Heeten GJ, et al. Increasingly strong reduction in breast cancer mortality due to screening. Br J Cancer. 2011;104(6):910-4.

6. AIHW, AACR. Cancer in Australia: an overview 2012. Canberra: AIHW; 2012.

7. Health AIo, Welfare. BreastScreen Australia monitoring report 2012–2013. Canberra: AIHW; 2015.

8. Health AIo, Welfare. BreastScreen Australia data dictionary: version 1.2. Canberra: AIHW; 2019.

9. Morrell S, Barratt A, Irwig L, Howard K, Biesheuvel C, Armstrong B. Estimates of overdiagnosis of invasive breast cancer associated with screening mammography. Cancer Causes Control. 2010;21(2):275-82.

10. Brodersen J, Jørgensen KJ, Gøtzsche PC. The benefits and harms of screening for cancer with a focus on breast screening. Pol Arch Med Wewn. 2010;120(3):89-94.

11. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. J Med Screen. 2012;19 Suppl 1:57-66.

12. Wong DJ, Gandomkar Z, Lewis S, Reed W, Suleiman Ma, Siviengphanom S, et al. Do Reader Characteristics Affect Diagnostic Efficacy in Screening Mammography? A Systematic Review. Clinical Breast Cancer. 2023;23(3):e56-e67.

13. Holland K, van Gils CH, Mann RM, Karssemeijer N. Quantification of masking risk in screening mammography with volumetric breast density maps. Breast Cancer Res Treat. 2017;162(3):541-8.

14. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000;92(13):1081-7.

15. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. Radiographics. 2015;35(2):302-15.

16. DS ALM, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? Clin Radiol. 2014;69(4):333-41.

17. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology. 2008;246(2):376-83.

18. Al Mousa DS, Mello-Thoms C, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Mammographic density and cancer detection: does digital imaging challenge our current understanding? Acad Radiol. 2014;21(11):1377-85.

19. Itri JN, Patel SH. Heuristics and Cognitive Error in Medical Imaging. American Journal of Roentgenology. 2018;210(5):1097-105.

20. Busby LP, Courtier JL, Glastonbury CM. Bias in Radiology: The How and Why of Misses and Misinterpretations. Radiographics. 2018;38(1):236-47.

21. Gale A, Chen Y. A review of the PERFORMS scheme in breast screening. Br J Radiol. 2020;93(1112):20190908.

22. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA. Radiology resident mammography training: interpretation difficulty and error-making patterns. Acad Radiol. 2014;21(7):888-92.

23. Grimm LJ, Zhang J, Lo JY, Johnson KS, Ghate SV, Walsh R, et al. Radiology Trainee Performance in Digital Breast Tomosynthesis: Relationship Between Difficulty and Error-Making Patterns. J Am Coll Radiol. 2016;13(2):198-202.

24. Trieu PD, Tapia K, Frazer H, Lee W, Brennan P. Improvement of Cancer Detection on Mammograms via BREAST Test Sets. Academic Radiology. 2019;26(12):e341-e7.

25. Rawashdeh MA, Bourne RM, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses. Acad Radiol. 2013;20(5):576-80.

26. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology. 1992;184(3):613-7.

27. Anandarajah A, Chen Y, Colditz GA, Hardi A, Stoll C, Jiang S. Studies of parenchymal texture added to mammographic breast density and risk of breast cancer: a systematic review of the methods used in the literature. Breast Cancer Research. 2022;24(1):101.

28. Wong DJ, Gandomkar Z, Wu WJ, Zhang G, Gao W, He X, et al. Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review. J Med Radiat Sci. 2020;67(2):134-42.

29. Rangayyan RM, Mudigonda NR, Desautels JEL. Boundary modelling and shape analysis methods for classification of mammographic masses. Medical and Biological Engineering and Computing. 2000;38(5):487-96.

30. Shi J, Sahiner B, Chan HP, Ge J, Hadjiiski L, Helvie MA, et al. Characterization of mammographic masses based on level set segmentation with new image features and patient information. Med Phys. 2008;35(1):280-90.

31. Wang J, Kato F, Oyama-Manabe N, Li R, Cui Y, Tha KK, et al. Identifying Triple-Negative Breast Cancer Using Background Parenchymal Enhancement Heterogeneity on Dynamic Contrast-Enhanced MRI: A Pilot Radiomics Study. PLoS One. 2015;10(11):e0143308.

32. Gastounioti A, Conant EF, Kontos D. Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. Breast Cancer Research. 2016;18(1):91.

33. Daye D, Keller B, Conant EF, Chen J, Schnall MD, Maidment AD, et al. Mammographic parenchymal patterns as an imaging marker of endogenous hormonal exposure: a preliminary study in a high-risk population. Acad Radiol. 2013;20(5):635-46.

Chapter Two
Literature Review

2.1 Chapter 2 Introduction

Chapter 2 serves to address the first aim of the thesis: which is to provide an overview of the current literature on mammography accreditation and its determinants.

The literature review presents an overview of the literature on screening mammography accreditation and its requirements. It explores the various reader characteristics that have been considered in relation to diagnostic efficacy and mammography interpretation. The review reveals that no individual characteristic has been consistently associated with diagnostic efficacy. However, it highlights that annual reading volume and specialization in breast imaging have shown associations with increased sensitivity and cancer detection rates, while not significantly affecting other performance metrics.

It is important to note that the literature review presented in this chapter have been previously published as:

Wong DJ, Gandomkar Z, Lewis S, Reed W, Suleiman M, Siviengphanom S, Ekpo E. Do Reader Characteristics Affect Diagnostic Efficacy in Screening Mammography? A Systematic Review. Clin Breast Cancer. 2023 Apr;23(3):e56-e67. doi: 10.1016/j.clbc.2023.01.009. Epub 2023 Jan 26. PMID: 36792458.

The review serves as valuable resource to establish a foundation of knowledge on mammography accreditation, and reader characteristics. They provide a comprehensive understanding of the current state of research in these areas and lay the groundwork for further exploration and investigation in the subsequent chapters of this thesis.

Review Article

Check for updates

Do Reader Characteristics Affect Diagnostic Efficacy in Screening Mammography? A Systematic Review

Dennis Jay Wong, Ziba Gandomkar, Sarah Lewis, Warren Reed, Mo'ayyad Suleiman, Somphone Siviengphanom, Ernest Ekpo

Abstract

To examine reader characteristics associated with diagnostic efficacy in the interpretation of screening mammograms. A systematic search of the literature was conducted using databases such as Cochrane, Scopus, Medline, Embase, Web of Science, and PubMed. Search terms were combined with "AND" or "OR" and included: "Radiologist's characteristics AND performance"; "radiologist experience AND screening mammography"; "annual volume read AND diagnostic efficacy"; "screening mammography performance OR diagnostic efficacy." Studies were included if they assessed reader performance in screening mammography interpretation, breast readers, used a reference standard to assess the performance, and were published in the English language. Twenty-eight studies were reviewed. Increasing reader's age was associated with lower false positive rates. No association was found between gender and performance. Half of the studies showed no association between years of reading mammograms and performance. Most studies showed that high reading volume was more likely to be associated with increased sensitivity, cancer detection rates (CDR), lower recall rate, and lower false positive rates. Inconsistent associations were found between fellowship training in breast imaging and reader performance. Specialization in breast imaging was associated with better CDR, sensitivity, and specificity. Limited studies were available to establish the association between performance and factors such as time spent in breast imaging (n = 2), screening focus (n = 1), formal rotation in mammography (n = 1), owner of practice (n = 1), and practice type (n = 1). No individual characteristics is associated with versatility in diagnostic efficacy, albeit reading volume and specialization in breast imaging appear to be associated with with increased sensitivity and CDR without significantly affecting other performance metrics.

Clinical Breast Cancer, Vol. 23, No. 2, e56–e67
Keywords: Breast cancer, Observer performance, Quality assurance, Diagnostic efficacy, Radiologists' characteristics

Introduction

Breast cancer is the most common cancer in women worldwide.[1,2] Early detection of the disease through screening mammography has been credited with a reduction in mortality from the disease.[1] However, mammography has some limitations that can cause errors, including the failure to detect cancer when present and the incorrect diagnosis of cancer when no cancer is in fact present. Therefore, for screening mammography to be effective, the images produced must be interpreted accurately. Radiological image interpretation involves search, perception, and decision-making processes.[3,4] Errors in each of these processes may lead to a diagnostic error and consequently negatively impact upon the goal of early detection and treatment of breast cancer.[4,5] Readers demonstrate differences in search, perceptual and decision-making abilities. This can lead to a wide variability in classification and diagnostic tasks in mammographic image interpretation. Inter-reader variability in mammography interpretation suggests that human errors remain a major limitation to the early detection of breast cancer.[3]

To account for reader limitations and differences in human perceptual and decision-making abilities, technological interventions such as computer-assisted diagnostic systems have been integrated into the image interpretation framework to improve these processes with mixed success.[6,7] In some countries, single reading with computer-aided algorithms has also been explored to improve diagnostic efficacy, but this strategy has been met with high false-positive results in mammography screening.[7,8] To reduce false-positive rates, some of the areas identified by computer-assisted devices as having a cancer are dismissed by human interpreters, even when a cancer is present.[9] Therefore, despite the incorporation of

Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Sciences, Faculty of Medicine and Health, University of Sydney, Sydney, NSW 2006, Australia

Submitted: Apr 5, 2022; Revised: Jan 10, 2023; Accepted: Jan 21, 2023; Epub: 26 January 2023

Address for correspondence: Ernest Ekpo, Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Sciences, Faculty of Medicine and Health, University of Sydney, Western Ave, Camperdown, Sydney, NSW 2006, Australia
E-mail contact: ernest.ekpo@sydney.edu.au

1526-8209/$ - see front matter
https://doi.org/10.1016/j.clbc.2023.01.009

technology into screening mammography interpretation, diagnostic errors, and variability still persist because the final decision regarding the diagnosis of cancer is made by humans.[7] Therefore, technology is not the complete solution to mitigating diagnostic errors. Instead, we must explore strategies to reduce the intrinsic limitations of the image interpreter. In countries such as Australia and the Netherlands, independent double reading strategy, where 2 readers independently interpret a mammogram to increase the chance that if a cancer is missed by 1 reader, it can be detected by another reader has been explored.[10,11] Although, double reading has reduced recall rates and increased sensitivity compared to single reading,[10,11] an inappropriate pairing of readers may reduce its diagnostic efficacy.[10] Identifying reader characteristics that improve diagnostic efficacy is critically important to optimize the paring of readers.

For quality assurance purposes, many screening guidelines use reader characteristics to certify readers for independent reporting of screening mammograms and maintenance of certification.[12] The volume of mammograms interpreted per year is the criterion mostly used for certification,[12] albeit, the minimum volume read requirements does vary across countries. Whilst the USA requires a minimum of 960 mammogram reads over 2 years, it is 5000 per year in the UK[13] and other European countries,[14] and 2000 in Australia.[15] The differences in reading volume requirements across the world suggest that the impact of reading volume on diagnostic efficacy is poorly understood. In addition, the literature shows wide variability in the association between diagnostic efficacy and reader characteristics such as experience,[16,17] training,[17–21] and practice type.[16–23] The reasons for these differences are unclear and the best markers of diagnostic performance requires further investigation. If reader-related factors that impact diagnostic efficacy are identified, informed strategies to mitigate diagnostic discrepancies and optimize the performance of breast cancer screening programs can be designed. Therefore, this review aims to identify the reader characteristics associated with the diagnostic efficacy of screening mammography so that these can be used to inform interventions to improve the performance and quality assurance guidelines of screening programs.

Patients and Methods

Search Strategy

The preferred reporting items for a systematic review and meta-analysis of diagnostic testing studies (PRISMA-DTA) statement was used to guide the literature search for relevant articles.[24] The search was conducted using the following databases: Cochrane Central Register of Controlled Trials, Scopus, Medline, Embase, ScienceDirect, Web of Science, and PubMed. A Google crossline search was also conducted, and reference lists of published studies were manually searched to capture articles not identified on database search. Search terms included: "Breast cancer," "screening mammography," "Radiologists characteristics," "Diagnostic performance," "radiologist experience," "screening mammography," "annual volume read," "diagnostic efficacy," "screening mammography performance," "diagnostic efficacy," "Radiologist training," "screening mammography accuracy," "feedback," "training," "breast cancer detection," "practice characteristics," "screening performance," "Breast readers," "Breast screen readers," "Breast physician," "Breast imaging reporter," "radiographers," "psychometric factors," and "observer performance in mammography." A detailed search strategy is presented in Supplemental Table 1. Two researchers were tasked with the database search, filtering, and collecting relevant articles according to the keywords. The articles were then screened to meet the eligibility criteria and were responsible for extracting the relevant data and conducting the quality assessment and analysis. Any potential disagreements between articles between the 2 researchers were mediated and discussed with the initial 2 researchers until the disagreements were settled.

Eligibility Criteria

Articles identified were assessed for eligibility using the population, intervention, comparator, and outcome (PICO) system (Table 1). Articles were included if they assessed reader performance in screening mammography interpretation, included trained readers (radiologists, breast physicians, or trained radiographers), used a reference standard to assess the performance of the readers, performed statistical analysis including at least one of the following outcomes: sensitivity, specificity, receiver operating characteristics curve (ROC) analysis, and location ROC, free-response operating characteristic (FROC) analyses such as Jackknife alternative free-response receiver operating characteristic (JAFROC), alternative free response receiver operating characteristic (AFROC), and were published in English. Both clinical and laboratory-based studies were considered for analysis, and there was no restriction on the date of publication. Only studies that nested their analysis on the radiologists or facility-level analysis of radiologist characteristics were considered. Studies that did not meet the criteria above including those that assessed performance in the diagnostic setting and those that did not examine radiologists' characteristics associated with performance were excluded, as were conference proceedings, posters, letters to the editor, and literature reviews. For studies that assessed both screening and diagnostic mammograms, we included only the results of the screening mammography analysis.

Quality Assessment

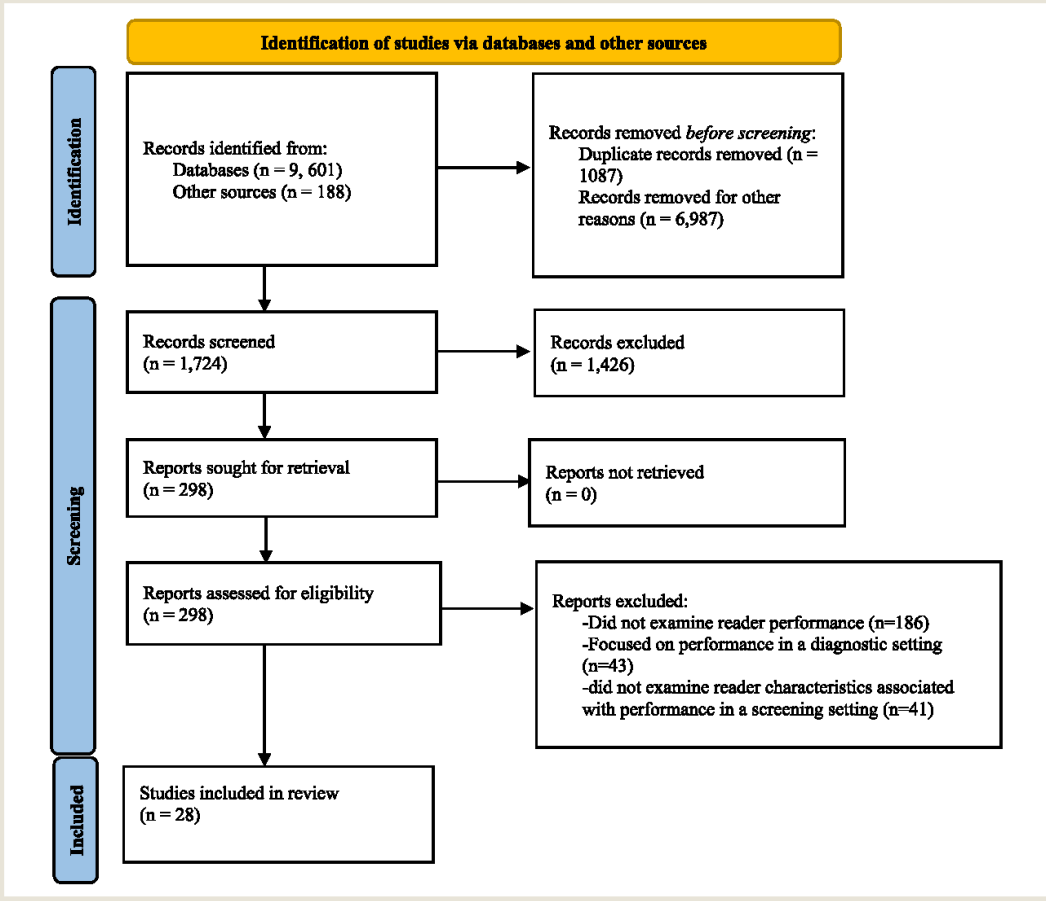
The assessment of study quality assessment was performed independently by 2 authors (E.E. and D.W.) based on the quality assessment of diagnostic accuracy studies (QUADAS-2) tool.[25] Each assessor independently assessed the risk of bias in Response each article based on 4 domains: (1) "could the patient selection have introduced bias?" (Patient selection); (2) "could the conduct or interpretation of the index test have introduced bias?" (index test); (3) "could the reference standard, its conduct, or its interpretation have introduced bias?" (reference standard); and (4) "could the patient flow have introduced bias?" (flow and timing). Each assessor independently used these 4 domains to assess study quality. If there was no concern regarding a domain, a low risk of bias was assigned. If the assessor had concerns regarding a domain, a high risk of bias was assigned. For studies that did not provide sufficient information to establish bias, the risk of bias was rated as unclear. A weighted kappa analysis was conducted to assess interassessor agreement, and discrepancies between assessors were resolved via discussion.

Table 1 The PICO Method Detailing Eligibility Criteria

Element	Characteristics
Population	Screening readers (Radiologists, breast physicians, breast reporting radiographers)
Intervention	Screening mammography
Comparators	Reader characteristics (age, gender, years reading mammograms, reading volume, specialisation in breast imaging, academic medical center affiliation, time spent in breast imaging, screening focus, formal rotation in mammography, owner of practice, feedback, lifetime mammograms read, practice type, certification in breast radiology, and year of certification in radiology
Outcome	Cancer detection rate, recall rate, sensitivity, specificity, PPV, False positive rates, location sensitivity, JAFROC, ROC AUCs, abnormal interpretation rates,

Abbreviations: AUC = area under the receiver operating characteristic curve; JAFROC = jacknife free-response receiver operating characteristic curve; PPV = positive predictive value; ROC = receiver operating characteristics curve.

Figure 1 Flow diagram showing the search results.



Data Extraction

Using the PICO system, data pertaining to the characteristics of the studies (eg, study design, adjustment to protocols, and reader recruitment strategies), population characteristics (eg, sample size, disease prevalence, lesion type and sizes, and breast composition information), reader characteristics (qualification, clinical experience, number of mammograms read per week, workload, practice-related information), and interpretation protocol were extracted. Data related to the reference standard used and reader diagnostic performance metrics such as area under receiver

operating characteristic curve, FROC analysis, location sensitivity, specificity, accuracy, and odds ratios were collected.

Data Analysis

Outcomes of the studies were summarized using descriptive statistics. Due to the different metrics used to quantify reader performance and the statistical tests to assess the association between reader characteristics and performance (relative risk, odds ratio, Pearson's correlation, Spearman's correlation), weighted averages of performance metrics and meta-analysis could not be performed.

Results

A total of 9789 articles were identified on search. Of these 1724 articles with titles related to the topic were considered for the initial screening. Of these articles, 1087 were duplicates and 298 were considered eligible for full-text review based on the eligibility criteria detailed above. After the full-text review, only 28 of the 298 studies considered eligible filled inclusion criteria. The remaining 270 articles were excluded because they did not examine reader performance (n = 186), were focused on performance in a diagnostic setting (n = 43) and did not examine reader characteristics associated with performance in a screening setting (n = 41) as shown in Figure 1.

The characteristics of these studies are summarized in Table 2 . Ten of these were data linkage studies, 10 were observer performance studies, and 6 were retrospective observational studies. There was one each of cross-sectional and clinical assessment studies. Most of these studies were based on US radiologists or physicians (n = 16) followed by Australian and New Zealand radiologists (n = 4) and Canadian radiologists (n = 3). The image sample size used in these studies ranged from 50 to 2,373,433 (mean = 301,314) and the sample of readers varied from 4 to 1067 (mean = 142). For data linkage studies the sample size ranged from 8734 to 2,234,947 (mean = 768,829.6) and the sample of readers ranging from 24 to 231 (mean = 129), but the number of cancers in the samples were sometimes not reported. For observer performance studies, the sample of images varied across studies from 50 to 1000 (mean = 350) with the number of cancer cases ranging from 12 to 36 (mean = 19) and reader sample ranging from 20 to 194 (mean = 128). The sample of images in the retrospective observational studies varied from 27,394 to 2,373,433 (mean = 696,279) with the number of cancers often not reported, and reader sample ranging from 9 to 1067 (mean = 304). Only 6 of the 28 studies adjusted for some of the patient-level factors that may confound the association between reader characteristics and performance (Table 2).

The reader characteristics varied across studies and included age, gender, years reading mammograms, number of images read per year, specialization in breast imaging, academic medical center affiliation, percentage time spent in breast imaging, hours spent in breast imaging, screening focus, formal rotation in mammography, owner of a private mammography practice (self-reported), frequency of immediate workup, feedback, lifetime mammograms read, number of CME credit, practice type, certification in breast radiology, and year of certification in radiology.

There were differences in the metrics used to assess the association between reader characteristics and performance. These metrics included sensitivity, specificity, recall rate, cancer detection rate (CDR), ROC area under receiver operating characteristic curve, JAFROC, location sensitivity, recall rates, abnormal interpretation rates (the percentage of examinations with an abnormal final interpretation), and positive and negative predictive values of recall. Furthermore, different statistical tests such as relative risk, odds ratio, Pearson's correlation, and Spearman's correlation were used to compare reader groupings.

Quality Assessment

Most the 28 studies included in the review demonstrated low risk of bias in terms of patient selection (n = 23) and index and comparator test interpretations (n = 27). In the flow and timing domains, the risk of bias was low in 16 studies, high in 4 studies, and unclear in 8 studies. In the reference standard domain, most of the studies (n = 18) demonstrated low bias due to the interpretation of the images without prior knowledge of the patient's diagnosis and the risk of bias was high in 5 studies (Appendix 1). The interassessor agreement showed a weighted kappa of k = 0.80, 95% CI, 0.68-0.92.

Age and Gender

Only 5 studies examined the association between reader's age and performance. Four of these studies showed that older readers have lower false-positive rates[26,27] or better specificity.[28,29] Two studies reported a decrease in sensitivity with ageing,[28,30] but no association was observed between age and either sensitivity or specificity in another study.[31] Seven studies assessed the association between gender and performance. Five of these studies reported no association between gender and either sensitivity, specificity, recall rate, and CDR[26,28,31–33]; however, 2 studies reported that women show high false-positive rates,[32] with another study showing higher uncertainty scores for female readers.[34]

Number of Years Reading Mammograms

Twelve studies assessed the association between number of years reading mammograms and performance. There were wide differences in the categorization of readers into different levels of experience.[28,31,34, 35, 18] Some studies used mean years of reading mammograms without specifying the cut-off for inexperienced and experienced readers,[35–37] others studies considered readers with $\leq$9 or $\leq$10 years as low experience readers.[28,32,34,38] Six of these studies reported no association between number of years and either sensitivity, specificity, CDR, accuracy, false-positive rate, recall rate, or positive predictive value.[31–33,35,18,39] Other studies reported that number of years reading mammograms increased specificity,[28,36] as well as sensitivity and location sensitivity.[38,40] Years of reading mammograms was also found to be associated with increase in JAFROC[36] and ROC[30] values.

Reading Volume and Performance

Nineteen studies investigated the association between reading volume and reader performance. Different thresholds were used to classify readers into between 2 and 7 levels of reading volumes. Low

Table 2 Summary of the Characteristics of the Studies Included in the Review

Author (y)	Study Design	Study Population	Type of Mammograms	Reader Characteristics/Intervention Assessment	Outcomes Assessed	Major Significant Results	Adjustment for Confounding Factors
Barlow et al.[28]	Data linkage & survey (BCSC registries)	US Radiologists n = 124	Screening (n = 469,512) Cancers (n): 2402	Age; gender; Reading volume; Years of experience reading mammograms; practice type; affiliation with academic medical center	Sensitivity Specificity ROC AUC Recall rate	Age: decreased sensitivity (OR, 0.52; 0.37-0.74); increased specificity (OR, 1.48; 1.16-1.88); increased recall rates (younger: 12.1 vs. older: 8.5) **Full-time work:** increased sensitivity(0.60; 0.44-0.82) but similar specificity with part-time work; **Years interpreting mammograms:** decreased sensitivity (OR, 0.50; 0.37-0.74), reduced recall rates (11.8 vs. 8.6) & increased specificity (1.51; [1.17-1.95]); no association with AUC **Reading volume**: increased sensitivity (OR, 1.89, 95% CI, 1.36-2.63) & decreased specificity (OR, 0.76, 95% CI, 0.60-0.96); increased recall rates (low 7.6 vs. high 10.4); no association with accuracy **Affiliation, gender, and percentage screening mammogram interpretation:** No association with sensitivity and specificity.	Patient's age; breast density; previous mammographic history; radiologists' years of experience
Beam et al.[35]	Observer performance study	US Radiologists n =110	Screening (n = 148)	Reading volume; years of reading mammograms; number of years since residency; formal rotation in mammography; owner of practice;	Sensitivity Specificity AUC	**Reading volume**: 5% mean increase in sensitivity; no association with ROC AUC (+0.01 (−0.07 to 0.05) **Years of reading mammograms:** No association with AUC (+0.19 [−0.05 to 0.44]) **Years since residency: Decrease in AUC (**−0.30 [−0.60 to 0.09]) **Formal rotation in mammography:** Decrease in AUC (−0.55 [−2.75 to 0.00]) **Owner of practice:** increase in AUC (+0.59 [0.02-2.46]). change in specificity not reported.	Radiologist-level factors; facility-level factors
Buist et al.[42]	Data linkage Data linkage & survey (BCSC registries)	US radiologists (n = 96)	Screening (n = 651,671)	Total annual volume read; Volume of diagnostic work-up (feedback)	Sensitivity FPR CDR	**Work-up:** Increasing radiologist workups of own recalled cases (from 0-25 mammograms to >50) increases: sensitivity (from 81.1 [77.4, 84.3] to 87.0 [84.9, 88.8]); FPR (from 6.7 [5.6, 8.0] to 10.3 [8.7, 12.1]) CDR (from 3.1 [2.7, 3.5] to 4.5 [4.1, 5.0])	Age at screening; first degree family history; time since past mammogram
Buist et al.[48]	Data linkage Data linkage & survey (BCSC registries)	US radiologists (n = 120)	Screening (n = 783,965)	Screening focus; Annual volume of screening mammograms; Annual volume of diagnostic mammograms;	Sensitivity FPR CDR	**Screening focus (<75 vs. 90+):** associated with lower sensitivity (86.3 [83.5, 88.7] vs. 81.8 [78.1, 85.0]), FPR (9.9 [7.4, 13.2] vs. 5.6 [4.4, 7.0]); CDR (4.5 [3.8, 5.4] vs. 3.4 [2.7, 4.2]). **Reading volume <480 vs. 5000+**: Decrease in sensitivity (83.9 [72.5, 91.1] vs. 84.0 [81.7, 86.1]); no association with FPR (7.7 [4.8, 12.1] vs. 9.5 [7.0, 12.7] and CDR 3.4 [2.4, 4.7] vs. 3.6 [3.1, 4.2]) **Diagnostic focus**: Low diagnostic volume (<100) associated with lower FPR (6.7 [5.4, 8.4] vs. 9.8 [7.4, 12.8]) and CDR (3.3 [2.6, 4.1] vs. 4.1 [3.6, 4.7]) but no association with sensitivity 82.5 (75.0, 88.1) vs. 85.7 (83.0, 88.1)	Age at screening; interpretive volume; time since past mammogram;

(*continued on next page*)

Table 2 (*continued*)

Author (y)	Study Design	Study Population	Type of Mammograms	Reader Characteristics/Intervention Assessment	Outcomes Assessed	Major Significant Results	Adjustment for Confounding Factors
Burnside et al.[45]	Clinical assessment/observer performance	US physicians (n = 4)	Screening (n = 30,363)	Annual reading volume	CDR Recall rates	**Reading volume:** Volume of 2770 readings is adequate to achieve CDR of 4.4/1000; volume lower than 120 is needed to achieve benchmark recall rate of 9.7%	None
Carney et al.[34]	Data Linkage	US radiologists (n = 124)	Screening (sample not reported)	Gender; reading volume; years of experience	Uncertainty score for RR and specificity	**Gender**: Females have lower uncertainty score **Interpretive volume:** associated with lower uncertainty scores **Years interpreting mammograms**: associated with lower scores Uncertainty score not associated with sensitivity, specificity or RR (R^2 values ranged from 0.02 to 0.001).	None
Ciatto et al.[43]	Observer performance	Italian radiologists (n = 117)	Screening (n = 150 including 17 cancer cases	Mammographic practice; reading volume per year; Total volume read per year	Sensitivity Recall rate	**Mammographic practice:** associated with better sensitivity and reduced RR **Reading volume:** associated with sensitivity ($X^2 = 7.215$; $P = 0.027$) and reduced RR.	None
Cornford et al.[39]	Data linkage	UK Radiologists, Breast Physician, and Radiographers (n = 37)	Actual sample not reported. Only range of cases reported by each reader was provided 1474-12,877	Reading volume; Years of reading mammograms (experience)	CDR, PPV, RR	**Reading experience:** no association with CDR, PPV, and RR High vs. low experience (CDR: 7.4/1000 vs. 7.0/1000); RR (6.2 vs. 3.6); PPV (15.0 vs. 20.9) **Reading volume**: High vs. low volume (CDR: 6.9/1000 vs. 7.5/1000); RR (3.1 vs. 6.1); PPV (14.2 vs. 19.1)	None
Elmore et al.[32]	Data linkage	US radiologists (n = 205)	Screening (n = 1,036,155 cancers: 4961	Gender; Breast fellowship training; years reading mammograms, academic medical center affiliation, percent time spent in breast imaging, hours spent in breast imaging	RR, FPR, sensitivity, PPV of recall (PPV1)	**Male vs Female:** RR (8.4; 5.6-12.5 vs. 11.4; 8.5-13.5), FPR (8.1 (5.3-12.1 vs. 11.2 (8.2-13.1), sensitivity(82.4 (70.7-91.3) vs. 87.0 (77.4-94.6), and PPV1(4.1; 2.6-6.1 vs. 3.9 (2.7-5.4) **Fellowship training in breast imaging:** increased sensitivity 83.3 (72.7-91.3) vs. 87.7 (80.9-100.0); RR 9.1 (6.0-13.1) vs. 11.6 (9.3-14.7), FPR 8.6 (5.7-12.8) vs. 11.0 (8.8-14.3), and PPV1 (3.9 [2.6-5.8]) vs. 5.2 (3.1-7.4); **Years interpreting mammograms:** decreased RR 13.1 (7.8-16.6) vs. 7.4 (4.9-10.7) and FPR (12.8 [7.4-16.3] vs. 7.1 [4.8-10.2]) **Academic medical center affiliation:** Lower RR (9.5 [6.3-13.5] vs. 7.6 [6.3-13.0]) and FPR 9.1 (6.0-13.2) vs. 7.2 (6.3-12.6)	Radiologists random effect. No adjustment for patient-related factors
Elmore et al.[26]	Data linkage	US radiologists (n = 24)	Screening (n = 8734)	Age, gender, year of graduation, years reading mammograms, facility type	FPR	Younger and recent graduates show high FPRs (11.7 [8.8-16.1]) compared to older radiologists (6.7 [5.2-9.1])	patient and testing characteristics, but the exact characteristics not reported

(*continued on next page*)

Table 2 (*continued*)

Author (y)	Study Design	Study Population	Type of Mammograms	Reader Characteristics/Intervention Assessment	Outcomes Assessed	Major Significant Results	Adjustment for Confounding Factors
Elmore et al.[53]	Observer performance	US radiologists (n = 10)	Screening (n = 150; cancer: 27)	Experience; frequency of immediate workup; feedback; lifetime mammograms read; number of CME credit; practice type	sensitivity	**Experience:** reader experience quantified by feedback, lifetime mammograms read, number of CME credit, practice type is associated with better sensitivity ($P < 0.05$ for all).	None
Esserman et al.[44]	Observer performance	UK radiologists (n = 194) US radiologists (n = 60)	Screening test set (n = 60)	Reading volume	Sensitivity and specificity	**Reading volume:** associated with increased sensitivity (low volume: 0.648; high volume: 0.756: and specificity (low volume: 83.6%; high volume: 88.0%)	None
Geller et al.[51]	Observer performance (self-reported interpretive ability).	US radiologists (n = 119)	Screening test set (n = 109); Cancer: n = 15-36	Reading volume read per week; specialisation; Fellowship in breast imaging	PPV and NPV of recall;	**Reading volume:** Increase in NPV for high volume readers PPV: nonexperts: 55.9 (45.8-68.1); experts: 58.8 (48.7-71.1) NPV: nonexperts: 83.5 (77.0-90.6); experts: 83.0 (76.4-90.1).	None
Hoff et al.[47]	Retrospective observational study	Norway Radiologists (n = 121)	Screening (n = 2,373,433)	Annual and cumulative reading volume	Sensitivity; screen-detected cancer; FPR	**Reading volume** 100 vs. 4000: sensitivity (87% vs. 90%), FPR (5.3% vs. 4.0%); CDR (4.9/1000 vs 4.7/1000)	None
Kan et al.[46]	Data linkage?	Canada radiologists (n = 35)	Not reported	reading volume	SAIR; CDR	**Reading volume:** A minimum of 2500 volume read per year is associated with lower abnormal interpretation rates and average or improved cancer detection rates. Mean standardised AIR for <2000 (1.15 ± 0.31) vs. 4000 + (1.09 ± 0.38 Mean standardised CDR for <2000 (0.89 ± 0.47) vs. 4000 + (1.07)	None
Kim et al.[18]	Observer performance	Korean radiologists (n = 12)	Screening test set (1000 cases; 12 cancers)	Reading volume; years of experience; fellowship	CDR, PPV, sensitivity, specificity, FPR, RR	RR CD. PPV. Sen. SPC. FPR. **Experience** <10 y 7.2. 10.4. 15.1. 86.9. 93.3. 6.3 ≥10 y 7.7. 10.8. 16.9. 90.0. 93.7. 6.7 **Fellowship** Yes 8.0. 10.0. 12.9. 83.3. 92.9. 7.1 No. 7.0. 11.0. 18.1. 91.7. 94.0. 6.0 **Reading volume** <3000. 6.9 10.0 17.3. 83.3. 94.0. 6.0 ≥3000. 8.2. 11.4. 14.0. 95.0. 92.8. 7.2	None
Leung et al.[63]	Retrospective observational study	US radiologists (n = 9; 3 breast radiologists and 6 general radiologists)	Screening n = 106,405	Specialisation in breast imaging	CDR; Recall rates	No significant difference between specialist and general radiologists in CDR and RR CDR: Specialist (115) vs. generalist (116)	None

(*continued on next page*)

Table 2 (*continued*)

Author (y)	Study Design	Study Population	Type of Mammograms	Reader Characteristics/Intervention Assessment	Outcomes Assessed	Major Significant Results	Adjustment for Confounding Factors
Miglioretti et al.[21]	Data linkage	US radiologists (n = 231)	Screening (n = 1,599,610)	Fellowship training in breast imaging; years of experience as mammography interpreter; first year of practice	Sensitivity; RR; FPR	**Nonfellowship trained radiologists:** higher recall but lower FPR; No association between experience is and sensitivity. **For fellowship trained:** sensitivity increased with experience; no difference in FPR and RR at first year of practice but FPR declines over time. No learning curve for fellowship trained readers 25 y (5-year change): FPR (1.23 [1.12, 1.35]; sensitivity: 1.16 [0.51, 2.64]; PPV 0.72 [0.48, 1.07])	None
Molins et al.[29]	Observer performance	Spanish radiologists (n = 21)	Screening test set (n = 200 with 30 cancers)	Focus on breast radiology; feedback; time spent in breast radiology per day; reading volume; age; feedback	Sensitivity and specificity	**Age:** ageing (>45 years) is associated with lower sensitivity (OR, 0.56, 0.31-1.01) and better specificity (OR, 1.33, 1.12-1.59) **Time in breast radiology:** not associated with sensitivity (OR, 0.70; 0.36-1.37) but improves specificity (OR, 1.49; 1.18-1.89). **Feedback & consulting with colleagues**: not associated with sensitivity (OR, 0.89; 0.43-2.27) but increases specificity (OR, 1.37; 1.03-1.85). **Reading volume:** no association with sensitivity (OR, 0.93, 0.56-1.54) or specificity. (OR, 1.15; 0.98-1.35) **Years reading mammograms**. No association with either sensitivity (OR, 1.40; 0.84-2.34) or specificity (OR, 1.04; 0.89-1.21)	None
Rawashdeh et al.[36]	Observer performance	Australian and New Zealand radiologists (n = 116)	Screening test-set (n = 60 including 20 cancers)	Years since qualification as radiologist; years reading mammograms; reading volume; hours per week reading mammograms	JAFROC, Location sensitivity, specificity	**Years since qualification:** associated with higher JAFROC (r = 0.22) and specificity (r = 0.18) **Years reading mammograms**: associated with higher JAFROC (r = 0.19) and specificity (r = 0.19) **Reading volume (<1000 vs 5000+):** associated with JAFROC (0.67[0.62-0.73] vs. 0.86[0.83-0.88]), location sensitivity 0.53(0.45-0.60) vs. 0.59(0.54-0.64) and specificity 0.60 (0.51-0.70) vs. 0.83 (0.80-0.87)	None
Reed et al.[37]	Observer performance	Australian and New Zealand radiologists (n = 69)	Screening test-set (n = 50 including 15 cancers)	years of certified, years of experience, reading volume, hours reading mammograms, previous experiences, breast fellowship, and satisfaction levels	ROC, sensitivity, specificity	**Number of years certified**: associated with improved AUC (r = 0.32) **Years of experience:** associated with improved AUC (r = 0.43) **Reading volume (>2000 per year):** associated with improved AUC (r = 0.34) **Hours reading mammograms per week:** associated with improved AUC: 0.75 (<1000) vs. 0.85 (>5000)	None
Rickard et al.[49]	Data linkage	Australian radiologists (n = 134)	Screening (n = 903,702 including 3819 cancers)	Reading volume (<500 cases per year)	CDR	Reading volume: CDR increases with reading volume but plateaus after 1000 reads per year. No difference in CDR above 875 reads per year (RR = 0.79, 95% CI, 0.63-0.99).	None

(*continued on next page*)

Table 2 (*continued*)

Author (y)	Study Design	Study Population	Type of Mammograms	Reader Characteristics/Intervention Assessment	Outcomes Assessed	Major Significant Results	Adjustment for Confounding Factors
Sickles et al.[40]	Retrospective observational study	US radiologist (n = 10 including 3 specialist and 7 general)	Screening (n = 47,798)	Specialisation in breast radiology	Abnormal interpretation rates; CDR	Abnormal interpretation rates lower for specialist (4.9%) compared to general radiologists (7.1%). CDR for specialist (6.0/1000) was double that of general radiologists (3.4/1000)	None
Suleiman et al.[27]	Observer performance	Australian and US radiologists (n = 41; 21 Australian and 20 US	Screening test-set (n = 60 including 20 cancers)	Country of practice; Reading volume; years of experience.	JAFROC), inferred ROC, sensitivity, specificity, and location sensitivity	**Country of practice:** No difference in JARFOC, inferred ROC, sensitivity, and specificity, but higher location sensitivity for Australian radiologists **Years of experience (more vs less experienced):** associated with higher sensitivity (0.95 vs. 0.89) and location sensitivity (0.86 vs. 0.79). **Reading volume (more vs less experienced:** associated with higher sensitivity (0.94 vs. 0.89) and location sensitivity (0.86 vs. 0.79).	None
Tan et al.[50]	Retrospective observational study	US radiologists (n = 1,067)	Screening (n = 27,394)	Gender; period since graduation; age, type of practice; volume read per year; certification in breast radiology	FPR	**Gender:** Female radiologists higher FPR (female: 7.9 vs. male: 5.94) **Age:** younger radiologists have higher FPR: <40 years (7.55) vs. 60 + (5.27); **Period since graduation:** associated with lower FPR: <10 years (7.92) vs. 30+ (5.27) **Reading volume:** No association with FPR (low: 6.34 vs. high: 6.13) **Certification in breast radiology**: No association with FPR: Yes (6.23) vs. No (6.62)	age, race, family history of breast cancer
Taplin et al.[41]	Cross-sectional study	US radiologists (n = 128)	Screening (n = 360,149 and 1972 cancers)	Specialisation in breast radiology	sensitivity, specificity, and PPV	Specialisation in breast radiology: associated with better AUC (0.932 vs. 0.905), but no association with specificity (OR = 1.26, 95% CI, 0.97-1.63) or PPV1 (OR = 1.23, 95% CI, 1.01-1.50)	None
Theberge et al.[33]	Retrospective observational study	Canadian radiologists (n = 340)	Screening (n = 1,315,327	Reading volume	Sensitivity, FPR, and accuracy	**Reading volume:** associated with 32% better accuracy <1000 vs. 4000 (false-positive rate, 1.32; 95% CI, 1.13-1.54) and 24% reduced FPR (false-positive rate ratio = 0.76; 95% CI, 0.65-0.89	None
Theberge et al.[31]	Retrospective observational study	Canadian radiologists (n = 275)	Screening (n = 307,314)	Reading volume	CDR and FPR	**Reading volume:** No association with CDR but inverse association with FPR 4000 vs. 2000: CDR = 1.28 (95% CI, 1.07-1.52); False-positive rate ratio: 0.53 (95% CI, 0.35-0.79)	gender, age, BMI, breast density and previous breast aspiration, biopsy or implant

Abbreviations: AIR = abnormal interpretation rate; AUC = area under the receiver operating characteristic curve; BCSC = breast cancer surveillance consortium; CD = cancer detected; CDR = cancer detection rate; Exp = experience; Fellow = fellowship in breast imaging; FPR = false positive rate; JAFROC = jacknife free-response receiver operating characteristic curve; NPV = negative predictive value; PPV = positive predictive value; ROC = receiver operating characteristic curve; RR = recall rate; SAIR = standardized abnormal interpretation rate.

volume reading thresholds varied considerably and included ≤500 cases,[34,41] ≤1000 cases,[36,37] ≤3000 cases,[18] or <5000 cases.[29] Eight out of 13 studies reported that high reading volume was associated with better sensitivity[28–35, 18, 36–44] and 4 studies showed no association between reading volume and sensitivity. Three studies reported that higher reading volume increased CDR[37,45,46]; 2 studies reported that high reading volume decreased CDR,[39,47] and 1 study reported no association.[48] High volume readers also demonstrated better location sensitivity[36,40] and diagnostic accuracy,[30,36] lower recall rates,[39,45] and lower false positive rates.[33,47]

Fellowship, Specialization in Breast Radiology, and Academic Medical Center Affiliation

Seven studies examined the association between fellowship in breast imaging and different performance metrics. Breast fellowship, which requires physicians to undertake 3 to 12 months training in breast image interpretation depending on a country's requirement, was associated with lower abnormal interpretation rates[49] and lower or no effect on recall rate (RR).[21,18] Completion of a breast fellowship did not show any association with false positive rates,[18,50] but increased the positive predictive value of recall[51] and accuracy.[41] Only 4 studies examined the association between specialization in breast imaging and performance. Three of these studies reported that specialist readers had better CDR, sensitivity, and specificity, as well as higher positive predictive value of recall and lower abnormal interpretation rates.[49–51] Affiliation with an academic medical center showed no association with either sensitivity or specificity[28,31,32] as well as RR, FPR, and PPV of recall.[32]

Other Reader Characteristics

Limited studies were available to establish the association between reader performance and factors such as time spent in breast imaging (n = 2), screening focus (n = 1), and formal rotation in mammography (n = 1), being owner of a private mammography practice (n = 1), frequency of immediate workup or feedback (n = 2), lifetime mammograms read (n = 1), and practice type (n = 1). Rotation in mammography showed a decrease in accuracy, sensitivity, and specificity.[35] Two studies showed no association between time spent in mammography and reader performance in terms of sensitivity, FPR, and RR.[31,32] Increasing radiologist workup or feedback increased sensitivity, specificity, CDR, and FPR,[30,42] and 1 study showed that being an owner of a practice improved accuracy.[35]

Discussion

The evidence presented above shows mixed findings for the association between reader characteristics and screening performance. The number of years reading mammograms did not show an overall association with diagnostic efficacy, but reading volume appeared to improve performance metrics such as CDR, sensitivity, location sensitivity in a larger number of studies. Data also showed that older readers make fewer false-positive errors, but performance is not affected by reader gender. Whilst affiliation with an academic medical center is not associated with reader performance, specialization in breast imaging appears to be associated with better reader performance.

Among the workload-related characteristics examined, years of reading mammograms, and volume of mammograms read have been the most studied factors to-date. The literature generally shows that radiologists with many years of reading mammograms demonstrate lower false-positive rates and lower sensitivity.[27,28,32,38] These findings have been attributed to older readers having a higher threshold for classifying mammograms as abnormal, due to their ability to identify and dismiss lesion types where experience has shown that these lesions are more likely to be benign.[52] It has also been suggested that the old breast imaging training curriculum, which balances sensitivity and specificity, progressive skill loss over time, and the less detailed interaction with mammograms, may be the reasons why readers with many years of practice demonstrate lower sensitivity and lower false positive rates.[35] Years of reading mammograms is commonly used to quantify experience; however, our findings show that years of reading mammograms does not necessarily reflect reader's ability to identify and characterize cancer lesions. A combination of reader characteristics such feedback, lifetime mammograms read, number of CME credit, practice type has been shown to provide a better measure of experience, and reader performance.[53] Therefore, it is important that no single reader characteristics be used to quantify experience in mammography interpretation.

Data presented in Table 2 show that high reading volume is significantly associated with sensitivity without affecting specificity. It should be noted that many of the published studies used self-reported reading volumes,[19,20,36,52,54,55] which is subject to recall bias. It has also been suggested that lifetime reading volume may better capture the association between reading volume and performances since it combines both experience and reading volume.[56,57] Improved performance with higher reading volume may be related to increased exposure of high volume readers to normal image features as well as features of cancer. It has also been suggested that the better diagnostic efficacy of high volume readers may be due to their ability to discriminate normal mammograms.[36] Therefore, it is understandable why many guidelines require readers to read a certain number of cases per year to maintain the certification for independent screening mammography interpretation; however, it is unclear why there are differences in minimum numbers needed to achieve optimum performance across the literature and in the reading volume requirement across countries. These differences suggest that other reader-related factors may confound the relationship between reading volume and performance. Therefore, quality assurance guidelines for improving the performance of screening programs should consider other factors that may confound the relationship between reading volume and performance when setting minimum reading volume thresholds.

Breast fellowship and specialization in breast imaging is designed to develop expertise in breast image interpretation. The literature on the relationship between beast fellowship and performance is limited and inconsistent. It is important to note that training curriculum, duration and quality training, and mentorship vary across countries. These differences are expected to cause variability in the association between breast fellowship and performance between countries. However, the published studies on breast fellowship examined different performance metrics: abnormal interpretation rates,[49] recall rate,[21,18] false-positive rates,[27,18] positive predictive value of recall,[51] and accuracy.[50] The limited number of studies examining each of these performance metrics makes it difficult to establish the relationship between breast fellowship and reader performance. Interestingly, 3 quarters of the few published studies on specialization show that breast specialists had better CDR, sensitivity, and specificity, as well as higher positive predictive value of recall and lower abnormal interpretation rates.[49–51] These findings are reasonable since specialization leads to task repetition and refinement and suggest that screening mammography readers require a lot of practice and dedication to achieve high performance regardless of their training and affiliation.

Older adults consistently showed lower false positive rates. Psychological studies show that older people are more risk averse in decision-making, particularly when involved in activities requiring learning and memory and are more positive and capable of controlling their response to emotional and pictorial stimuli.[58] In addition, people's motivational priorities change with aging through greater knowledge acquisition and emotion regulation.[59] Thus, it is possible that the lower FPRs observed for older readers may be related to factors such as older adults' better understanding of features that elicit false positive errors, ability to regulate their perception of malignant features, and motivation to reduce false positive errors. It is also possible that older readers have more years of experience reading mammograms, which has been shown to be associated with higher threshold for classifying mammograms as abnormal.[27,28,32,38] However, further studies are needed to understand whether cognitive factors associated with ageing impact upon FPRs in screening mammography interpretation.

There is limited data to establish the relationship between reader factors such as time spent in breast imaging, screening focus, and formal rotation in mammography, owner of a practice, frequency of immediate workup or feedback, lifetime mammograms read, and practice type. However, feedback on performance characteristics such as sensitivity, specificity, and false positives or negatives, and abnormal findings appeared to improve CDR but showed an inconsistent relationship with false-positive rates. The mechanisms for providing feedback varied across studies[54,56,60] and maybe the underlying reason for the observed variations in FPRs. Although feedback on performance characteristics provides an indication of a reader's cancer reading outcomes, it is unlikely to improve their perceptual and discriminative skills since it does not tell the reader the features that were missed or misinterpreted. Educational programs such as Personal Performance in Mammographic Screening (PERFORMS)[58] and Breast Reader Assessment Strategy (BREAST), which identify the types of errors made by readers and provide visual feedback, have been shown to be useful for tailoring reader training[58] and improved reader performance by 24% to 34%.[61] In addition, these improvements due to visual feedback were not affected by reader characteristics. Therefore, visual feedback on the types and imaging features of lesions missed and normal image appearances the illicit false positive errors is needed to improve discriminative skills. Without such perceptual and discriminative feedback, readers may not change their perceptual and decision-making techniques and will continue to make similar diagnostic errors.

It is important to acknowledge methodological differences among the studies reviewed and their potential impact on the findings discussed. Whilst differences in study methodologies are not uncommon, the literature shows that certain patient characteristics such as age at screening, breast density, implants, family history of breast cancer, and breast size, which affect diagnostic accuracy of mammography were not considered in 79% of the studies reviewed. Unlike observer performance studies that tested multiple readers per case, data linkage, and retrospective studies had a maximum of 2 readers per case depending on the reading strategy used (single or double reading of mammograms). Therefore, such studies, even with large sample of readers and images, may not completely account for case-based variation in performance between readers and the impact that patient and lesion characteristics may have on the association between reader characteristics and diagnostic efficacy. It is often argued that experimental observer performance studies have limited external validity due to the laboratory effect[62]; however, we did not find any consistent pattern in the association between reader characteristics and performance according to study setting (data linkage, retrospective analysis, clinical studies, or observer performance studies), suggesting that large scale reader performance studies that account for patient-related characteristics may be used to identify and model reader characteristics that improve the performance of mammography screening programs.

A limitation of the review includes that, only studies published in English language were reviewed. In addition, differences in performance metrics and statistical tests used for comparing reader groupings limited our ability to compute weighted averages of performance or compare results across studies.

Conclusion

No specific reader characteristics is associated with versatility in the interpretation of screening mammograms, albeit reading volume, and specialization in breast imaging improved sensitivity and CDRs without significantly affecting other performance metrics. Therefore, multivariate models, which consider the interaction between reader characteristics and the confounding effects of patient and facility-level characteristics as well as tailored feedback interventions such as follow-up of recalled cases may help to inform strategies to improve the performance and quality assurance guidelines of screening programs.

Disclosure

The authors have stated that they have no conflicts of interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dbc.2023.01.009.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136:E359–E386.
2. Fitzmaurice C, Dicker D, Pain A, et al. The global burden of cancer 2013. *JAMA Oncol*. 2015;1:505–527.
3. Ekpo EU, Alakhras M, Brennan P. Errors in mammography cannot be solved through technology alone. *Asian Pac J Cancer Prev*. 2018;19:291–301.
4. Waite S, Grigorian A, Alexander RG, et al. Analysis of perceptual expertise in radiology - current knowledge and a new perspective. *Front Hum Neurosci*.. 2019;13:213.
5. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*. 2015;35:1668–1676.
6. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
7. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175:1828–1837.
8. Bargalló X, Santamaría G, del Amo M, et al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. *Eur J Radiol*. 2014;83:2019–2023.
9. Cole EB, Zhang Z, Marques HS, et al. Impact of computer-aided detection systems on radiologist accuracy with digital mammography *Am J Roentgenol*. 2014;203:909–916.
10. Brennan PC, Ganesan A, Eckstein MP, et al. Benefits of independent double reading in digital mammography: a theoretical evaluation of all possible pairing methodologies. *Acad Radiol*. 2018;26:717–723.

11. Duijm LE, Groenewoud JH, Fracheboud J, et al. Introduction of additional double reading of mammograms by radiographers: effects on a biennial screening programme outcome. *Eur J Cancer*. 2008;44:1223–1228.
12. Hofvind S, Bennett R, Brisson J, et al. Audit feedback on reading performance of screening mammograms: an international comparison. *J med screening*. 2016;23:150–159.
13. Committee NBSPNRQAC-o. *Quality assurance guidelines for radiologists*. NHS Breast Screening Programme; 1997.
14. Perry N, Broeders M, Schouten J, et al. *European guidelines for quality assurance in mammography screening*. Office for official publications of the European Communities; 2001.
15. Committee NA. *National Programme for the Early Detection of Breast Cancer. National Accreditation Requirements: March, 1994*. Canberra: Australian Commonwealth Department of Human Services and Health; 1994.
16. Miglioretti DL, Smith-Bindman R, Abraham L, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. *J Natl Cancer Inst*. 2007;99:1854–1863.
17. Beam CA, Conant EF, Sickles EA. Association of volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst*. 2003;95:282–290.
18. Kim SH, Lee EH, Jun JK, et al. Interpretive performance and inter-observer agreement on digital mammography test sets. *Korean J Radiol*. 2019;20:218–224.
19. Reed WM, Lee WB, Cawson JN, et al. Malignancy detection in digital mammograms. Important reader characteristics and required case numbers. *Acad Radiol*. 2010;17:1409–1413.
20. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*. 2009;253:641–651.
21. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology*. 2009;253:632–640.
22. Woodard D, Gelfand A, Barlow W, et al. Performance assessment for radiologists interpreting screening mammography. *Stat Med*. 2007;26:1532–1551.
23. Jackson SL, Abraham L, Miglioretti DL, et al. Patient and radiologist characteristics associated with accuracy of two types of diagnostic mammograms. *AJR Am J Roentgenol*. 2015;205:456–463.
24. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA*. 2018;319:388–396.
25. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529–536.
26. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *JNCI*. 2002;94:1373–1380.
27. Tan A, Freeman Jr DH, Goodwin JS, et al. Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. *Breast Cancer Res Treat*. 2006;100:309–318.
28. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst*. 2004;96:1840–1850.
29. Molins E, Macià F, Ferrer F, et al. Association between radiologists' experience and accuracy in interpreting screening mammograms. *BMC Health Serv Res*. 2008;8:91.
30. Reed WM, Lee WB, Cawson JN, et al. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. *Acad Radiol*. 2010;17:1409–1413.
31. Elmore JG, Cook AJ, Bogart A, et al. Radiologists' interpretive skills in screening vs. diagnostic mammography: are they related? *Clin imaging*. 2016;40:1096–1103.
32. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*. 2009;253:641–651.
33. Théberge I, Hébert-Croteau N, Langlois A, et al. Volume of screening mammography and performance in the Quebec population-based breast cancer screening program. *CMAJ*. 2005;172:195–199.
34. Carney PA, Elmore JG, Abraham LA, et al. Radiologist uncertainty and the interpretation of screening. *Med Decis Making*. 2004;24:255–264.
35. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst*. 2003;95:282–290.
36. Rawashdeh MA, Lee WB, Bourne RM, et al. Markers of good performance in mammography depend on number of annual readings. *Radiology*. 2013;269:61–67.
37. Rickard M, Taylor R, Page A, et al. Cancer detection and mammogram volume of radiologists in a population-based screening programme. *Breast*. 2006;15:39–43.
38. Miglioretti DL, Smith-Bindman R, Abraham L, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. *J Natl Cancer Inst*. 2007;99:1854–1863.
39. Cornford E, Reed J, Murphy A, et al. Optimal screening mammography reading volumes; evidence from real life in the East Midlands region of the NHS breast screening programme. *Clin Radiol*. 2011;66:103–107.
40. Suleiman WI, Lewis SJ, Georgian-Smith D, et al. Number of mammography cases read per year is a strong predictor of sensitivity. *J Med Imaging*. 2014;1.
41. Théberge I, Chang SL, Vandal N, et al. Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. *J Natl Cancer Inst*. 2014;106:djt461.
42. Buist DS, Anderson ML, Smith RA, et al. Effect of radiologists' diagnostic work-up volume on interpretive performance. *Radiology*. 2014;273:351–364.
43. Ciatto S, Ambrogetti D, Catarzi S, et al. Proficiency test for screening mammography: results for 117 volunteer Italian radiologists. *J Med Screen*. 1999;6:149–151.
44. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. *JNCI*. 2002;94:369–375.
45. Burnside ES, Lin Y, Munoz del Rio A, et al. Addressing the challenge of assessing physician-level screening performance: mammography as an example. *PLoS One*. 2014;9:e89418.
46. Kan L, Olivotto IA, Warren Burhenne LJ, et al. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. *Radiology*. 2000;215:563–567.
47. Hoff SR, Myklebust T, Lee CI, et al. Influence of mammography volume on radiologists' performance: results from breast screen Norway. *Radiology*. 2019;292:289–296.
48. Buist DSM, Anderson ML, Haneuse SJPA, et al. Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology*. 2011;259:72–84.
49. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*. 2002;224:861–869.
50. Taplin S, Abraham L, Barlow WE, et al. Mammography facility characteristics associated with interpretive accuracy of screening mammography. *J Natl Cancer Inst*. 2008;100:876–887.
51. Geller BM, Bogart A, Carney PA, et al. Is confidence of mammographic assessment a good predictor of accuracy? *AJR Am J Roentgenol*. 2012;199:W134–W141.
52. Woodard DB, Gelfand AE, Barlow WE, et al. Performance assessment for radiologists interpreting screening mammography. *Stat Med*. 2007;26:1532–1551.
53. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? *J Womens Health*. 1998;7:443–449.
54. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst*. 2004;96:1840–1850.
55. Suleiman WI, Lewis SJ, Georgian-Smith D, et al. Number of mammography cases read per year is a strong predictor of sensitivity. J Medic Imaging. 2014;1: 015501-015503-6.
56. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? *J Womens Health*. 1998;7:443–449.
57. Ciatto S, Ambrogetti D, Catarzi S, et al. Proficiency test for screening mammography: results for 117 volunteer Italian radiologists. *J Med Screen*. 1999;6:149–151.
58. Scott HJ, Gale AG. Breast screening: PERFORMS identifies key mammographic training needs. *Br J Radiol*. 2006;79(Spec No 2):S127–S133.
59. Mather M. *When I'm 64. National Research Council (US) Committee on Aging Frontiers in Social Psychology, Personality, and Adult Developmental Psychology*. A review of decision-making processes: weighing the risks and benefits of aging. Washington, DC: National Academies Press; 2006 eds..
60. Molins E, Macià F, Ferrer F, et al. Association between radiologists' experience and accuracy in interpreting screening mammograms. *BMC Health Serv Res*. 2008;8.
61. Suleiman WI, Rawashdeh MA, Lewis SJ, et al. Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance. *J Med Imaging Radiat Oncol*. 2016;60:352–358.
62. Gur D, Bandos AI, Fuhrman CR, et al. The prevalence effect in a laboratory environment: changing the confidence ratings. *Acad Radiol*. 2007;14:49–53.
63. Leung JW, Margolin FR, Dee KE, et al. Performance parameters for screening and diagnostic mammography in a community practice: are there differences between specialists and general radiologists? *AJR Am J Roentgenol*. 2007;188:236–241.

Chapter Three

Mammographic image characteristics in educational test sets: the effect on expert difficulty ratings and subsequent reader performance

3.1 Chapter Introduction

Chapter 3 manuscript is currently under review at the *Asia Pacific Journal of Oncology*.

The previous chapter's aim was to provide the context and background of the current literature of the standards and accreditation requirements for screening mammography that are currently implemented in practice. The evidence presented above shows mixed findings for the association between reader characteristics and screening performance. The number of years reading mammograms did not show an overall association with diagnostic efficacy, but reading volume appeared to improve performance metrics such as cancer detection rates, sensitivity, location sensitivity in a larger number of studies.

Educational programs such as Personal Performance in Mammographic Screening (PERFORMS) and Breast Reader Assessment Strategy (BREAST), which identify the types of errors made by readers and provide visual feedback, have been shown to be useful for tailoring reader training and improved reader performance.

Chapter 2 ultimately found the standards for mammography screen to be inconsistent across different boards and countries and no specific reader characteristics were associated with versatility in mammography interpretation. Therefore, it was suggested that multivariate models accounting for interactions between reader and image characteristics may help to develop strategies to improve the performance and quality assurance in screening programs.

Chapter 3 is the starting point to the three-staged studies that first evaluated the current development of test sets in training programs (Chapter 3), the development of a multivariate model to determine the interactions between reader and case image characteristics in mammography interpretation (Chapter 4) and the specific textural features that may most influence reader's performance in mammography reading (Chapter 5). Chapter 3 now scrutinizes the current practices of developing mammography tests sets and aimed to find the requirements of developing a mammography training test set.

Chapter 3 focuses on the second objective of the thesis: Evaluating the current practice in evaluating mammography cases and its difficulty and presents the introduction, methodology, results, discussion and conclusion for the current practices of using experts to determine a mammography case's difficulty, what characteristics were used for consideration to its

difficulty and whether the difficulty determine is accurately reflected from the reader's actual performance from the mammography cases. An Ethical approval exemption was obtained for this retrospective study and informed consent was collected from the radiologists. All images and radiologist's data were anonymous and de-identified.

In the following study in Chapter 3, both the Kruskal-Wallis H test and the Chi-squared test was used to determine the significance between objective case difficulty and various independent variables. By using the capabilities of both tests, the study aimed to unveil the relationship between objective case difficulty and a variety of independent variables, offering a multifaceted view of the factors influencing diagnostic accuracy.

The Kruskal-Wallish H test is a non-parametric statistical test used to compare more than two groups to ascertain if there is a statistically significant difference in their medians. Differing from traditional ANOVA, it does not assume a normal distribution of the data, providing a more flexible approach to understanding ordinal or continuous data that is not normally distributed. The study first compares objective case difficulty with various independent variables, such as expert difficulty, breast density, and case type, Secondly, it explores how different characteristics of the case images, such as breast density and case type, affect the reader's performance in terms of objective case difficulty.

The Chi-squared test is used for categorical data to assess the likelihood that an observed distribution is due to chance. The association between expert-determined difficulty and breast density for normal and abnormal cases was analysed to determine if the distribution of difficulty levels across different breast densities is random or if there is a significant pattern.

In summary, the study's methodical use of the Kruskal-Wallis H test and the Chi-squared test provides a rigorous statistical framework to dissect the layers of complexity within medical image interpretation. The calculated interplay between objective case difficulty and variables like expert-determined difficulty, breast density, and case type offers invaluable insights into the factors shaping diagnostic accuracy. These insights have the potential to inform better practices, enhance training, and ultimately lead to improved patient outcomes.

Mammographic image characteristics in educational test sets: the effect on expert difficulty ratings and subsequent reader performance

Abstract

Rationale and Objectives: To investigate the impact that case image characteristics have on expert-determined difficulty in a mammography test set and whether the predetermined difficulty assessment reflects reader performance.

Materials and Methods: Retrospective data was collected from one hundred and sixty-five radiologists who participated in the BreastScreen Reader Assessment Strategy (BREAST) platform between 2015 and 2021. Data collected was from one single test set containing the same 60 digital mammograms. A total of 20 cases were cancer-positive and 40 cases were normal. Case image characteristics included breast density, expert-determined difficulty, objective case difficulty (percentage of incorrect reports per case) and lesion characteristics. Statistical significance was tested using a Kruskal Wallis H test and a Chi square test and a simulation model was used to determine the number of readers needed for a reliable difficulty metric.

Results: Expert-predetermined difficulty was significantly associated with breast density classification (Kruskal-Wallis H = 7.675 p=0.022) and the case's abnormality (Kruskal-Wallis H = 13.279 p=0.001). However, there was no significant difference found between actual case difficulty as determined by reader performance and expert-predetermined difficulty (p=0.750). The simulation predicated around 10 readers were sufficient for a reliable difficulty metric.

Conclusion: Overall, our results showed that experts consider cases as more difficult that have a greater breast density and when they are abnormal. Our data did not show any significant association between expert-determined difficulty and actual case difficulty experienced by the reader's as determined by their performance in this test set.

Keywords: Mammography, quality assurance, training, expert, difficulty

Introduction

Test sets are widely used for quality assurance and educational purposes to review and report reader performance in breast screening [1-3]. In Australia, the Breast Screen Reader Assessment Strategy (BREAST) program has been implemented as a self-assessment training tool throughout the breast cancer screening program and provides several test sets for readers to evaluate and improve their performance in mammography interpretation [4]. A similar self-assessment platform in the United Kingdom, the Personal Performance in Mammographic

Screening (PERFORMS)'s , provides readers with expert opinions, known case pathology, peer opinions as well as personal training and quality assurance [5].These platforms attempt to imitate the real clinical setting, although the test sets are oversampled with positive cases for practical purposes.

In BREAST, similar to real-world screening examinations, test set cases typically consist of four mammograms: images of the left and right breasts in the medio-lateral oblique (MLO) and cranial caudal (CC) projections and all images are digital and de-identified. Cancer-positive cases are biopsy proven, while normal cases are evaluated and determined by the consensus reading of at least two senior expert radiologists following negative screen reports based on at least a two-year follow-up. The cases in BREAST are carefully selected to be challenging to participants to provide value in self-assessment, education, and training. The range of case difficulty and inclusion is determined by an expert radiologist [6]. For PERFORMS, cases are collected from the UK's screening centres with the original radiologist and pathology report. Similar to BREAST, based on mammographic features of the cases, expert radiologists determine whether the case is suitable to be included in PERFORMS [7]. Therefore, in both BREAST and PERFORMS, the current practice of test set development largely depends on expert-determined case difficulty and quantifying of case difficulty is largely subject to the expert's opinion [8, 9]. However, one could argue that an expert's understanding of difficulty is not necessarily correlated with the actual case difficulty as usually ground truths (i.e., cancer location and pathology report) are provided to experts while making decision about the case difficulty and this could cause bias. Moreover, expert's perception of difficulty could differ from the actual appearance of difficulty for less experienced colleagues. Determining if the expert-determined case difficulty is strongly correlated with the actual case difficulty is essential for developing training and quality assurance programs such as BREAST and PERFORMS as well as for developing the examinations.

Grimm et al. [8] and Mazurowski's [10] studies explored the relevance of expert-determined case difficulty with actual case difficulty for radiology trainees. Both studies found that an increase in reader's self-assessed and expert-assessed rated difficulty was associated with a decrease in resident performance in mammography by evaluating the relationship between the error rate of radiology trainees (an objective measure of difficulty) and the assessment of case difficulty by expert radiologists (subjective measures of difficulty) [8][10]. It was shown that an expert's assessment of difficulty was associated with the trainee's increase in false negative errors but not false positive errors. However, their studies focused on radiology trainees, and it

is unclear if the findings can be extended to radiologists, in general. Additionally, Tapia's et al. [11] found that readers tend to perform better in cases with lower breast density in comparison to cases with higher breast density. It is unknown if the expert-determined case difficulty is correlated with breast density and whether the expert-determined case difficulty could add any complementary information to the breast density for predicting case difficulty.

In all previous known studies using mammography test sets as a method of evaluating and training radiologist's performance the cases and their difficulty are pre-curated and determined by experts. However, no known studies have evaluated how prospective grading of case difficulty by experts is decided and how this correlates with actual measures of retrospective case difficulty recorded by the participants. The percentage of incorrect reports can be used as an actual or objective measure of difficulty.

Ideally, if unlimited resources were available to create a test set, one would recruit a cohort of radiologists to assess a bank of cases. If the cohort is large enough, based on their performance, an objective and reliable case difficulty could have been calculated. However, as in practice, recruiting a large cohort of readers is not possible, determining the minimum number of required radiologists to produce a reliable objective case difficulty is highly desirable. Currently this minimum number is unknown. As this study included a very large number of radiologists (i.e., 165 participants), we also conducted a set of simulations to identify this minimum number.

Therefore, the purpose of this study was (1) to measure the image characteristics used to determine case difficulty by experts, (2) whether this prospective expert determined difficulty is an accurate predictor of the actual case difficulty using an objective measure of difficulty produced by reader performance in the datasets, and (3) to determine by simulation the minimum number of radiologists required to provide a reliable objective case difficulty.

Methodology

Reading conditions

The BREAST test set was read in reporting rooms in the clinical setting or in conference venues in rooms designed to emulate environments typically used to interpret radiology images. In our data set, out of 165 readers, 111 readers were clinic-based, while 54 readers were from conferences. The readers at conferences viewed the images using the Sectra Breast Imaging PACS (Sectra, Linköping, Sweden; Hologic, Bedford, Mass), which provided post-processing tools for the reader such as zooming, windowing, panning. These rooms used ambient lighting below 10 lux [12] and in terms of hardware, workstations were either linked to two MFGD

5621 monitors (Barco, Kortrijk, Belgium), or two RadiForce G51 monitors (Eizo, Ishikawa, Japan) and were calibrated to the Digital Imaging and Communication in Medicine grayscale standard display function, had a contrast ratio of 365:1, and displayed a maximum luminance within 5% of 475 cd. m^{-2}, minimum luminance of 1.3 cd.m^{-2}.

Data collection

Ethical approval was obtained for this retrospective study and informed consent was collected from the radiologists. Between 2015-2021, 165 radiologists completed this selected test set. All images and radiologist's data were anonymous and de-identified.

Test set

A single test set with a total of 60 mammography cases from BREAST was used. It consisted of 20 cancer-positive cases and 40 normal cases. The cancer-positive cases had a variety of lesion sizes and appearances. Each case consisted of four mammograms: the left and right cranial caudal projections and the medial lateral oblique projections. Breast density classification data was included and was categorised from 1-4 using BI-RADS and consisted of the following categories: (1) fatty, (2) fibro glandular density, (3) heterogeneously dense, and (4) extremely dense. For the purposes of this study, breast density was classified as either "Dense" breasts consisting of category 3 and 4 cases, and "non-dense" breasts consisting of category 1 and 2 cases.

The test set items were selected from BreastScreen NSW's Digital Breast Image Library. Cancer cases are all been obtained from surgical pathology, while negative cases have been determined by the consensus of at least two senior radiologists, following a normal screen result by two readers (4).

Expert-determined difficulty

The case difficulty was determined by an expert radiologist who had held the position of State Radiologist in BreastScreen New South Wales, Australia. To determine the difficulty of each case, the expert examined each case retrospectively and rated the difficulty based on the probability of a radiologist potentially missing the pathology. The range was defined from 1-3 ('Easy', 'Intermediate', and 'Difficult', respectively). In this test set, there were 12 'Easy' cases (9 non-dense cases and 3 dense cases), 38 'Intermediate' cases (17 non-dense cases and 21 dense cases), and 10 'Difficult' cases (1 non-dense case and 9 dense cases). It is important to note that the 165 readers who participated were unaware of the difficulty of the case items as previously determined by the expert radiologist during the reading session.

Reader performance data

To evaluate the cases in the test set, readers were provided the MLO and CC views of the left and right breast per case. The reader could mark lesions anywhere on the image and used the Royal Australian and New Zealand College of Radiologist's (RANZCR) breast imaging rating system to grade the case. The rating system ranges from 1- 5 (1- 'Normal'; 2-'Benign'; 3-'Equivocal'; 4-'Suspicious'; 5-'Malignant'). A score classification between 3-5 was used to determine increasing confidence in malignancy of the marked area. Should the reader not find any abnormality, they clicked "Next case" and the platform automatically marked the cases as RANZCR score 1 (normal). Only the reader's first attempt at the test set was used. Readers were not aware of the prevalence and types of cancers included in the test sets.

Objective case difficulty

The objective difficulty for each case was calculated at the case level by determining the percentage of incorrect interpretations per case. This calculation involved dividing the total number of incorrect readings by a reader per case (including false positives and false negatives) by the total number of interpretations for that case (n=165). A lower percentage indicates an easier case, and vice versa. For example, case number 1 was classified by experts as a normal dense case with an intermediate difficulty. Given that the total number of incorrect readings (false positives) was n=115, the objective measure of difficulty is [100 – (115/165) * 100], equating in a 30% difficulty level, classified as easy.

Statistical analysis

For the statistical analysis, the Kruskal-Wallis H test was used to determine the significance between objective case difficulty and the groups of independent variables. For the first analysis, the independent variables included expert difficulty (Easy, Intermediate, and Difficult), breast density (dense and non-dense) and case type (Normal and Cancer-positive). To measure which case image characteristics impact reader performance, the Kruskal Wallis H test was performed to determine the significance of breast density and case type for objective (or actual) case difficulty.

To determine factors impacting the difficulty determined by experts, the case image characteristics were correlated with the expert-determined difficulty. For normal cases, the association between expert-determined difficulty with the breast density was measured using the chi-square test. Whereas for abnormal (cancer-positive) cases, in addition to the breast density, the relationship between the case difficulty and lesion characteristics, including lesion type, side of lesion, site of lesion and size of lesion, against expert determined case difficulty was investigated using the chi-square test.

Simulation

Finally, to determine the minimum number of required radiologists to produce a reliable case difficulty, for each case and each value for *k* ranging from 1 to 165, we conducted a simulation, where *k* readers were randomly selected, and the case difficulty is calculated based on the assessment of these selected readers. To create a distribution, we repeated the random sampling process 1000 each time.

Results

Test set characteristics

The details of the test set in terms of expert-determined difficulty and breast density are shown in Table 1. Out of the 20 cancer cases, there was 7 dense cases and 13 non-dense cases. In terms of expert-determined difficulty, no dense cancer cases were classified as easy, 3 dense cancer cases were classified as intermediate and 4 were classified as difficult. For non-dense cancer cases, there was only 1 case classified as easy, 8 classified as intermediate and 4 were classified as difficult. The number of cancer cases categorised on their characteristics ('lesion type', 'side of cancer' and 'site of cancer') is shown in Table 1, lesion size ranges from 3mm to 25mm. Both cancer-positive dense and non-dense cases had 4 difficult cases each. Regarding the normal cases, most of the dense cases were classified as intermediate cases (n=16), while non-dense normal cases have around the same easy and intermediate cases (10,11, respectively), and non-dense cases had no difficult cases.

Factors determining difficulty.

The Kruskal Wallis H test did not show a significant difference between the expert-determined difficulty (Easy, Intermediate, and Difficult) and objective case difficulty (p=0.750). However, Chi-Squared test results showed the expert radiologist tended to assign higher difficulty to abnormal (cancer-positive) cases, compared to the normal cases (P-0.001). Furthermore, the expert radiologist also assigned higher difficulty ratings to dense breasts compared to non-dense breasts (p= 0.022).

For the chi-square test conducted for cancer cases, the case image characteristics (breast density, lesion size, lesion type, side of cancer, and site of cancer) against expert-determined difficulty, none of the variables showed significance (p= 0.445, 0.076, 0.425, 0.580, 0.154 respectively). Conversely, using the Chi-square test showed significance for expert-determined difficulty against breast density for normal cases (p=0.006).

Table 1. Number of cases per characteristic according to expert determined difficulty.

Characteristic	Easy cases	Intermediate cases	Difficult cases
Normal			
Dense	1	16	2
Non-Dense	10	11	0
Cancer			
Dense	0	3	4
Non-Dense	1	8	4
Lesion Type			
Discrete Mass	0	3	1
Non-specific density	1	2	2
Stellate/Spiculated Mass	0	4	5
Calcification	0	2	0
Side of Cancer			
Left	1	5	4
Right	0	6	4
Site of Cancer			
Central	1	3	0
Lower	0	3	2
Upper	0	5	6

Reader performance

Notably, case type (defined as either a normal or abnormal case), breast density and expert-determined difficulty did not have any significance in affecting actual case difficulty (p= 0.701, 0.186, 0.750, respectively).

Figure 1 illustrates the percentage of correct reports (calculated by 100%-objective difficulty) between normal and cancer-positive groups measured against their case's accuracy. Notably, cases 1 and 4 were outliers in normal cases with an average accuracy of 35.15% and 42.42%, respectively.

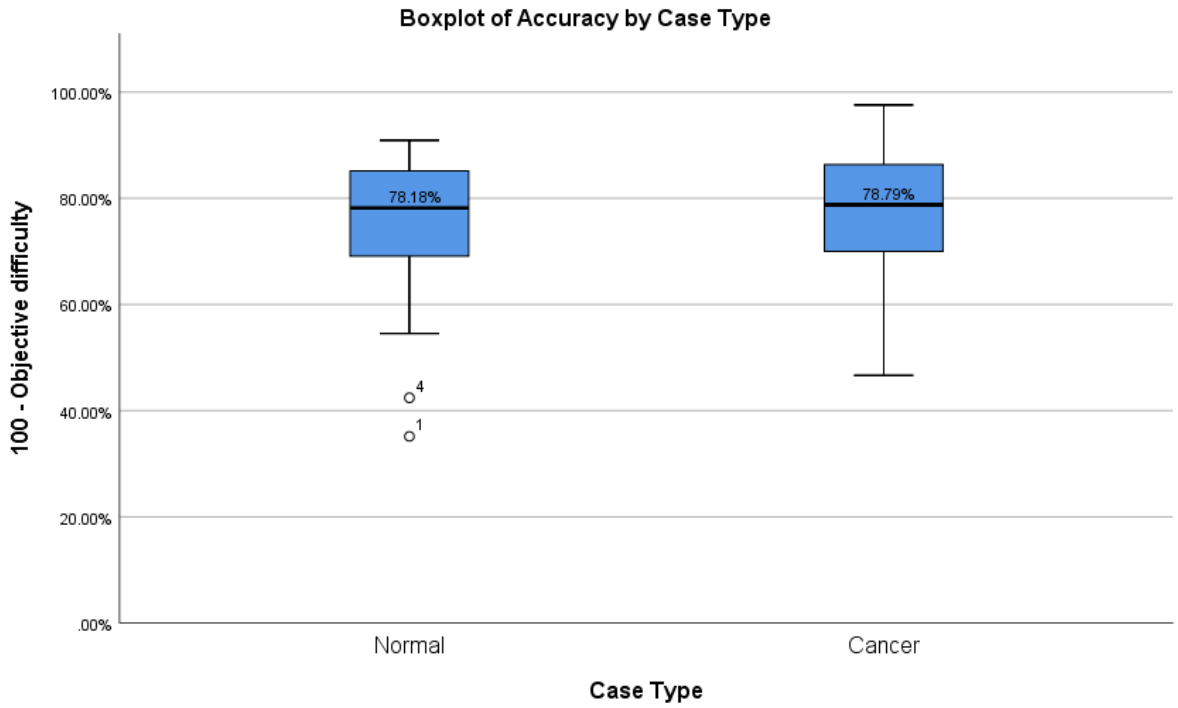


Figure 1. Boxplot between normal and cancer cases against 100% – objective difficulty, i.e., percentage of correct reports per case.

Figure 2 shows the percentage of correct reports between the three groups of difficulty determined by experts against the reader's case accuracy.

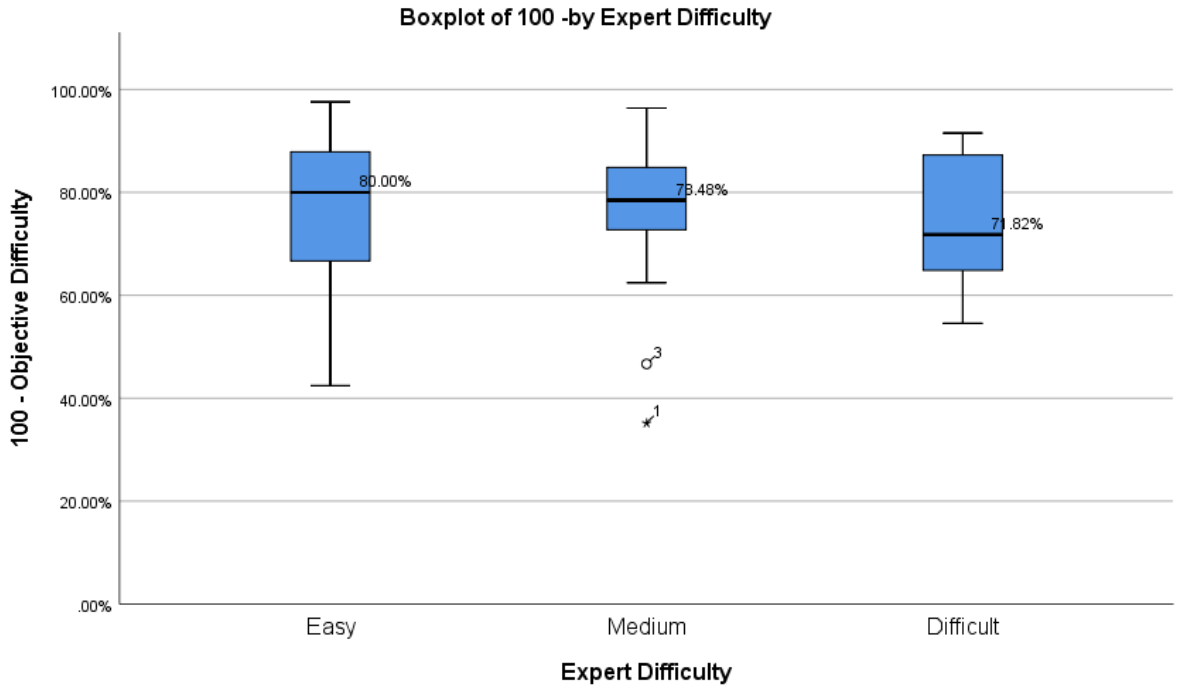


Figure 2. Boxplot between Overall Expert-determined difficulty against 100% – objective difficulty.

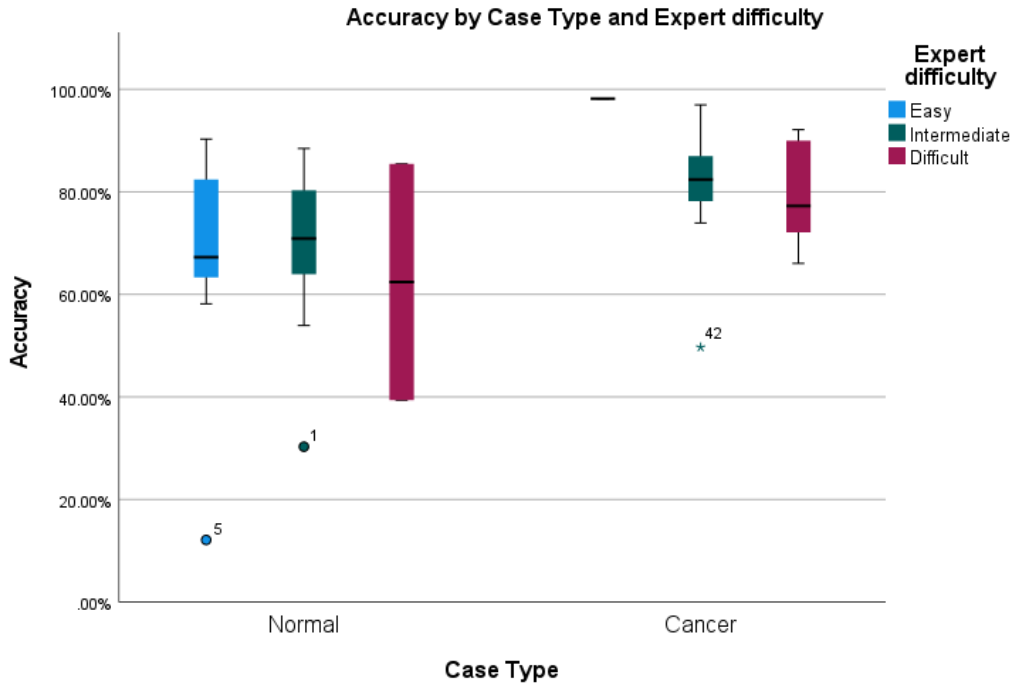


Figure 3. Boxplot between Expert-determined difficulty relative to the case's abnormality type against 100% – objective difficulty.

Figure 3 illustrates that accuracy varies for "Normal" case assessments, with a broad range in the easy cases, indicating divided reader performance. In contrast, "Cancer" cases are more consistently assessed, especially for easy and intermediate difficulties. However, accuracy dips for difficult cancer cases, as shown by a lower median and a narrow IQR, despite high consistency. Outliers in both normal and cancer categories highlight exceptionally divergent accuracy rates. Overall, while experts show consistent assessments in cancer cases across difficulties, difficult cases generally result in lower accuracy, and the variation in normal cases suggests differing levels of challenge for the experts.

Simulation

To simulate case difficulty, a boxplot similar to the one shown in Figure 3 can be created for each case. The x-axis represents the different values for k, which is the number of readers whose assessments were used to calculate the case difficulty. Each boxplot shows the distribution, based on 1000 repetitions, of the case difficulty.

As expected, the interquartile range (IQR) is wider when k is smaller, but with larger values of k (i.e., when the difficulty score is calculated based on the opinions of more readers), the IQR becomes smaller. For each case and for each possible value of k, we calculated the IQR and

identified the k value where the IQR is less than 0.2. This value represents approximately ±0.1 (±10%) from the median value (the red line in the middle of the boxplots in Figure 4), which is stable across all possible values of k and is equal to the difficulty score from all 165 readers.

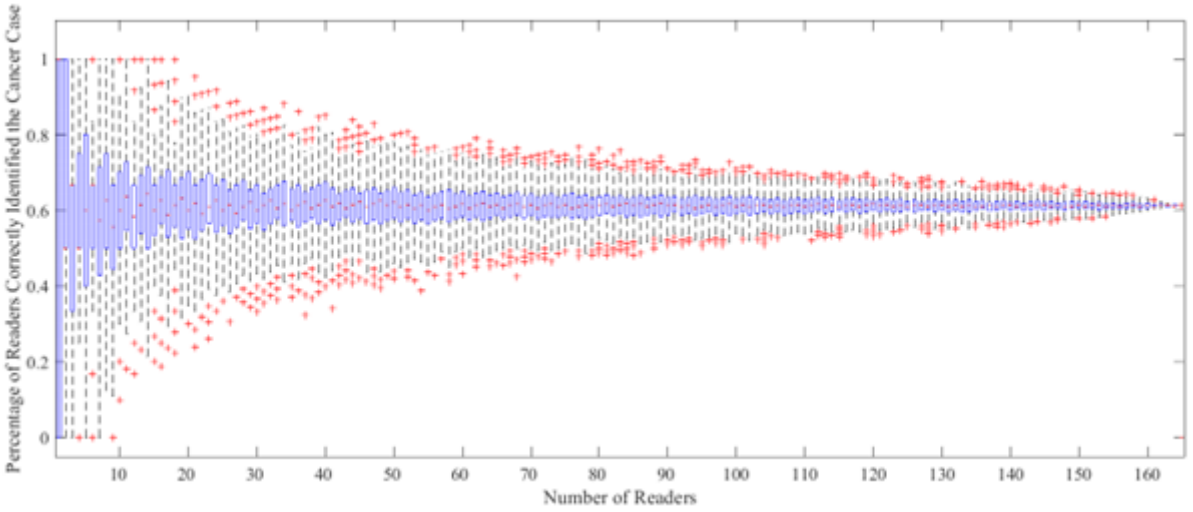


Figure 4. Boxplot illustrating the simulation used to calculate case difficulty.

Minimum number of readers for having a reliable case difficulty metric

We calculated the minimum number of required radiologists to achieve a reliable objective difficulty measure for each case. For all cases, relying on the assessments of 13 or fewer readers would result in a reliable difficulty metric. For 75% of the cancer cases (15 out of 20), assessments from less than 10 readers were sufficient to produce a reliable difficulty metric. For four of the normal cases, assessments from 12 readers were required, while for the remaining normal cases, relying on 11 or fewer readers was sufficient to reach a satisfactory level of reliability (Figure 5). Overall, for 78% of normal cases (31 out of 40), access to less than 10 readers was adequate. The average number of required readers was 8.0±3.9 for cancer cases and 8.1±2.8 for normal cases (p=0.36)

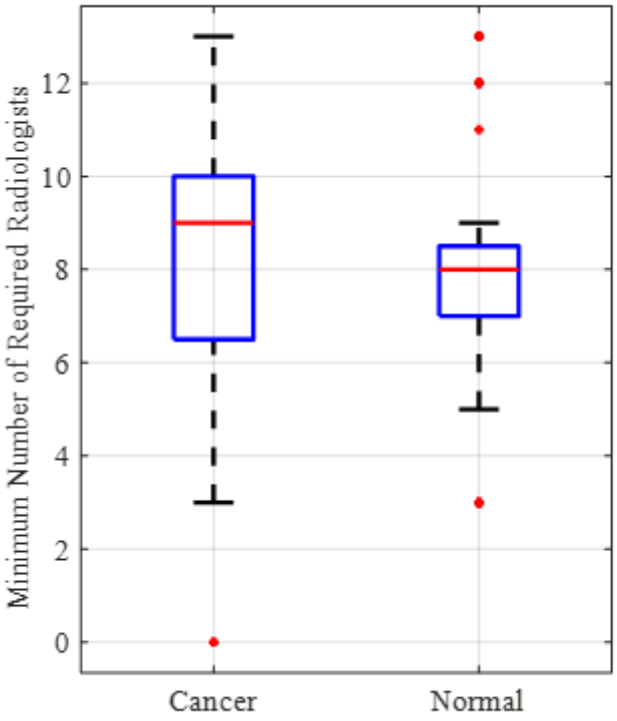


Figure 5 - The distribution of the minimum number of required radiologists for producing a reliable metric for the case difficulty. The median values are 9 and 8 radiologists for cancer and normal cases, respectively.

Discussion

The aim of this study was to determine which case image characteristics were used by experts to determine case difficulty in mammography test sets and whether this rating accurately reflected reader performance. Also, to determine by simulation the minimum number of radiologists required to determine an objective assessment of difficulty. It was hypothesised that an expert radiologist's difficulty rating would be reflected in the performance of the readers in our study, but our data did not support this hypothesis. From our study's findings, retrospective grading of mammography case difficulty by an expert was shown to increase primarily determined by increased breast density and if the case was abnormal (cancer positive. With regards to its reliability as an accurate predicator to objective reader performance, our study showed no significant association between expert (subjective) determined case difficulty and actual (objective) case difficulty, defined as percentage of incorrect report for a case. With regards to the simulation, to our knowledge, this is one of the first papers which used simulations to predict the number of readers needed to predict difficulty as a metric.

The mammogram's case image type (normal or abnormal cancer-positive cases) and breast

density (dense and non-dense) were the main factors determining case difficulty rating by these experts and neither of these lesion characteristics were shown to be a predictor of reader performance difficulty. The reader's objective performance in this test set was not associated with the expert-determined difficulty and contrary to expert-opinion, readers performed marginally better in cancer-positive cases in comparison to normal cases. Figure 2 illustrates that expert-determined difficulty did not accurately reflect on readers performance as they showed similar performance levels regardless of difficulty level.

Previous studies have presented varied conclusions regarding the association of case difficulty and breast density. Some studies, primarily focused on film-based mammography, have indicated that diagnostic efficacy in mammograms decreased when breast density increased (13, 14). Specifically, breast density reduced reader sensitivity due to the masking effect as radio-opaque anatomic structures were potentially superimposing pathology. Similar findings were observed in studies related to digital mammography (15-18). However, more recent findings have suggested that the masking effect of dense breasts was greatly reduced in the use of digital mammography compared to film-based mammography (15, 19, 20). Al Mousa et al. (21) even suggested that experienced radiologists using digital mammography may improve in performance with increasing mammography density. Our results suggest that the experts who determined the case difficulty in this test set tend to follow previous findings, in that, increasing breast density increases the difficulty of the case. This corresponds to Ang et al.'s findings where they found that mammographic breast density has an impact on determining the case difficulty in normal cases(22). However, in actual reader performance breast density did not appear to have an impact ($p=0.186$). It is likely that as digital-based mammography offers post-processing tools for readers to manipulate the mammograms (such as changing contrast, windowing, and colour inversion) these tools may have electronically mitigated some of the issues radiologists used to face when interpreting dense breasts using previous film-based mammography.

In terms of clinical implications, our findings show that, where possible, the development of test sets as educational or quality control materials should ideally use objective reader performance where this is available instead of expert opinion as a measure of difficulty in mammographic cases to provide a comprehensive test set to improve and monitor reader performance. Notably, our simulation suggests that using around 10 readers was sufficient to determine the difficulty of a test set image, which may be implemented as a guideline for test set development. Furthermore, our results emphasise the need to enable readers to recognise

normal cases better and more confidently, as our reader performance results suggested that they perform better in identifying pathology in mammograms in comparison to identifying normal cases.

With regards to the test set curation, it is important to note that the curation process may have impacted the case difficulty. The expert who selected the mammography cases to build the test set considered the educational value each case could offer to the readers. As a result, the test set may inherently be more 'difficult' compared to what readers typically encounter in a screening or diagnostic environment.

The prevalence effect could limit the generalisability of our findings to real-world practice and other settings. In comparison to the clinical setting and in line with Gur et al.'s study, test sets usually exhibit a higher number of positive cases. This may contribute to the laboratory effect, as our readers tended to perform better in cancer-positive cases compared to normal ones (23). However, a previous study has demonstrated a moderate level of agreement between clinical performance and test set interpretation (24). Another potential limitation of our data is the comparability of our user data, as most of our readers used the BREAST platform in the clinical setting, while around a third performed their BREAST interpretations either at conferences or workshops. However, in the conferences and workshops, BREAST has accounted for the reading conditions and replicated the condition to be as close to the clinical setting as possible, so the impact the reading environment has on the reader should be minimised.

Another limitation of this study involves a discrepancy between the number of readers in the BREAST platform (n=165) and the number of experts who determined the test set difficulty (n=1). Currently the BREAST platform uses only used one expert to grade difficulty of each test set and it would be useful to investigate the effect of an increased number of experts in future studies. Additionally, from our simulation, it may be beneficial to find readers to curate to the test set to determine the difficulty of each image set in the development of future test sets.

Conclusion

In summary our findings show that our expert primarily used breast density and case type (normal or abnormal cancer-positive) to determine case difficulty in our case test set. However, this expert-determined difficulty does not appear to mirror the case's objective difficulty as demonstrated by reader performance. The results emphasise the importance of examining reader performance and using it as an indicator in the development of test sets where available.

As such, it would be useful to further explore and identify what case characteristics the readers found most difficult to add further to the construction of test sets to develop their educational value in the future.

References

1. Chen Y, James JJ, Cornford EJ, Jenkins J. The Relationship between Mammography Readers' Real-Life Performance and Performance in a Test Set-based Assessment Scheme in a National Breast Screening Program. Radiol Imaging Cancer. 2020;2(5):e200016-e.
2. Trieu PD, Tapia K, Frazer H, Lee W, Brennan P. Improvement of Cancer Detection on Mammograms via BREAST Test Sets. Academic Radiology. 2019;26(12):e341-e7.
3. Suleiman WI, Rawashdeh MA, Lewis SJ, McEntee MF, Lee W, Tapia K, et al. Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance. (1754-9485 (Electronic)).
4. Brennan PC, Trieu PD, Tapia K, Ryan J, Mello-Thoms C, Lee W, editors. BREAST: A Novel Strategy to Improve the Detection of Breast Cancer2014; Cham: Springer International Publishing.
5. Gale A. Maintaining quality in the UK breast screening program: SPIE; 2010.
6. Brennan P, Tapia K, Ryan J, Lee W. BREAST: a novel method to improve the diagnostic efficacy of mammography: SPIE; 2013.
7. Gale A, Chen Y. A review of the PERFORMS scheme in breast screening. The British journal of radiology. 2020;93(1112):20190908.
8. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA. Radiology resident mammography training: interpretation difficulty and error-making patterns. Acad Radiol. 2014;21(7):888-92.
9. Grimm LJ, Zhang J, Lo JY, Johnson KS, Ghate SV, Walsh R, et al. Radiology Trainee Performance in Digital Breast Tomosynthesis: Relationship Between Difficulty and Error-Making Patterns. J Am Coll Radiol. 2016;13(2):198-202.
10. Mazurowski M. Difficulty of mammographic cases in the context of resident training: preliminary experimental data: SPIE; 2013.
11. A Tapia K, Rickard MT, McEntee MF, Garvey G, Lydiard L, C Brennan P. Impact of breast density on cancer detection: observations from digital mammography test sets. International Journal of Radiology & Radiation Therapy. 2020;7(2):36-41.
12. Brennan PC, McEntee M Fau - Evanoff M, Evanoff M Fau - Phillips P, Phillips P Fau - O'Connor WT, O'Connor Wt Fau - Manning DJ, Manning DJ. Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist. American Journal of Roentgenology. 2007;188(2):177-80.
13. Holland K, van Gils CH, Mann RM, Karssemeijer N. Quantification of masking risk in screening mammography with volumetric breast density maps. Breast Cancer Res Treat. 2017;162(3):541-8.
14. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000;92(13):1081-7.
15. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. RadioGraphics. 2015;35(2):302-15.
16. DS ALM, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? Clin Radiol. 2014;69(4):333-41.
17. Devolli-Disha E, Manxhuka-Kërliu S, Ymeri H, Kutllovci A. Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density. Bosn J Basic Med Sci. 2009;9(2):131-6.
18. Vachon CM, van Gils CH, Sellers TA, Ghosh K, Pruthi S, Brandt KR, et al. Mammographic density, breast cancer risk and risk prediction. Breast Cancer Res. 2007;9(6):217.

19. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology. 2008;246(2):376-83.
20. Olinder J, Johnson K, Åkesson A, Förnvik D, Zackrisson S. Impact of breast density on diagnostic accuracy in digital breast tomosynthesis versus digital mammography: results from a European screening trial. Breast Cancer Res. 2023;25(1):116.
21. Al Mousa DS, Mello-Thoms C, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Mammographic density and cancer detection: does digital imaging challenge our current understanding? Acad Radiol. 2014;21(11):1377-85.
22. Ang ZZ, Rawashdeh MA, Heard R, Brennan PC, Lee W, Lewis SJ. Classification of normal screening mammograms is strongly influenced by perceived mammographic breast density. Journal of Medical Imaging and Radiation Oncology. 2017

61(4):461-9.
23. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. Radiology. 2008;249(1):47-53.
24. BaoLin PS, Warwick ML, Peter LK, Warren MR, Mark FM, Patrick CB, editors. Test set readings predict clinical performance to a limited extent: preliminary findings. ProcSPIE; 2013.

Chapter Four

Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis

4.1 Chapter Introduction

Chapter 4 manuscript is currently under review at the *Journal of Medical Imaging and Radiation Oncology*. Additionally, a shorter version of the paper was accepted and presented in RANZCR 2021 (Appendix D).

Chapter 3 evaluated the existing practice of developing mammography test sets for the training platform BREAST by collecting data from a single test set consisting of 60 digital mammograms, containing 20 cancer-positive cases and 40 normal cases. The findings revealed that expert-determined difficulty was primarily influenced by breast density classification for both normal and cancer-positive cases. However, there was no significant difference observed between the expert's determination of difficulty and the actual difficulty experienced by the readers based on their performance in the test set.

Furthermore, the simulation suggested that around 10 readers would be sufficient to establish a reliable difficulty metric. The study indicated that experts tend to consider cases with higher breast density and abnormality as more difficult. However, there was no significant correlation between expert-determined difficulty and the actual difficulty experienced by the readers in this particular test set. This approach may have potentially introduced a limitation in Chapter 3's results as the protocol followed BREAST's training platform's practices of individual test sets. As such, only the image characteristics presented by the cases in the test set was considered and the only reader performances that were extracted were related to this test set, which may potentially skew the results. Therefore, in the methodology in this Chapter would effectively 'merge' all of the test sets available on BREAST and use all of the reader's performance from there.

Continuing on this project, Chapter 4's aim was to explore potential reader and case characteristics impacting reader performance. As Chapter 2's literature review proposed that the use of multivariate models that consider the interactions between reader and image characteristics could be beneficial in devising strategies to enhance performance and quality assurance in screening programs.

Chapter 4 focused on the third objective of the thesis and presents the introduction, methodology, results, discussion, and conclusion of the multivariate model for determining reader and imaging characteristics that was most influential to reader performance.

Linear Mixed Models (LMM) are a class of statistical models that are particularly useful for analysing data with complex structures, such as those arising from hierarchical or longitudinal designs. In the context of Chapter 4, LMM is used due to its ability to handle both fixed and random effects within the same model. Random effects include reader characteristics (like years of experience, specialization, etc.) and case characteristics (such as breast density and difficulty rating).

LMMs distinguish themselves by their adeptness in handling multilevel data structures. These models are designed to address the multifaceted nature of datasets that are not adequately represented by flat, two-dimensional table structures. Hierarchical data, for instance, can arise in medical studies where patients are nested within different treatment groups or in educational research where students are grouped within classes. Longitudinal designs, where the same subjects are measured multiple times over a period, also benefit from the layered approach of LMMs.

The analytical power of LMMs in this study is primarily attributed to their ability to simultaneously accommodate both fixed and random effects. Fixed effects are consistent and predictable influences across all individuals or units in the study—factors that are of primary interest to researchers. In contrast, random effects represent the unexplained variability within these units. They could stem from inherent differences among subjects or experimental units that cannot be directly measured or controlled.

The strength of LMMs lies in their capacity to incorporate random effects, which account for variations that are not explained solely by the fixed effects. By including these as random effects, LMMs allow for individual differences and do not force the assumption that all readers or cases respond in a homogenous manner. Another significant aspect of LMMs is their ability to manage the potential correlation within grouped data. For instance, multiple interpretations made by the same reader are likely to be more similar to each other than to interpretations made by other readers. This intra-group correlation can lead to biased estimates and incorrect conclusions if not properly accounted for.

The versatility of LMMs is further showcased in their treatment of random effects. Unlike fixed-effects models, LMMs do not operate under the restrictive assumption that all subjects or cases are drawn from a homogenous population. This is particularly relevant when considering reader characteristics in medical studies. Two radiologists with the same number

of years in practice may differ significantly in their diagnostic approaches due to factors like sub-specialization or the specific nature of their clinical experiences. LMMs account for such individual differences, thereby providing a more realistic and individualized analysis.

Another salient feature of LMMs highlighted in the chapter is their adept handling of correlated group data. This is crucial when the same reader may interpret multiple cases, introducing the possibility that their interpretations could be more similar to each other than to those made by different readers. Such intra-group correlations, if unaccounted for, can skew results and lead to false inferences. By incorporating random effects that capture these correlations, LMMs ensure that estimates remain unbiased, and conclusions drawn from the analysis stand on firm ground.

Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis

Abstract

Rationale and Objectives: To investigate the impact of reader-based and image-based characteristics on reader performance on mammography interpretation using linear mixed modelling analysis.

Materials and Methods: A total of 479 Australian radiologists participated in this study, where six test sets were used. Readers completed a survey collecting reader-based characteristics. These test sets contained 60 cases (20 cancer cases, 40 normal cases) and reader and image-based characteristics were collected. Case difficulty was determined prospectively by an expert radiologist classifying cases as easy, intermediate, and difficult. For statistical analysis, linear mixed modelling analysis was used. Three models were created: case-based characteristics; reader-based characteristics, and a combination of both. Normal cases and cancer cases were evaluated separately.

Results: For the normal case model, image-based characteristics showed expert-determined difficulty was significant ($P<0.05$) exhibiting decreasing performance as difficulty increased. The reader-based characteristics model found weekly cases interpreted and reader's specialization significant, P-values were $=<0.01$ in normal cases. The cancer case's model found none of the image-based characteristics significant. For reader-based characteristics only years in specialty was significant with increasing years improving performance. The mixed model found expert-determined difficulty (P values<0.05) significant with increasing difficulty reducing reader performance and number of cases interpreted per week (P values <0.05).

Conclusion: Expert-predetermined difficulty and cases interpreted weekly were consistently significant factors impacting performance, with increasing difficulty reducing performance and increasing cases interpreted improving performance. However, the cancer cases' characteristics were inconclusive and require future texture-based radiomic analysis for further investigation.

Keywords: Mammography, quality assurance, training, expert, characteristics

Introduction

Screening mammography remains the gold standard for the early detection of breast cancer (1). However, sensitivity of mammography-based screening for cancer detection remains varied across different countries around the world, ranging from 50% to 99% (2-4). To ensure an acceptable sensitivity level, many guidelines in different countries require their mammography

readers to interpret a specific number of annual cases to maintain their screening mammography certification (5-7). However, the minimum required interpretative volume varies among countries and jurisdictions and the existing literature does not provide adequate evidence to necessarily support different minimum requirements. These discrepancies in the requirements and this variable reader performance makes it difficult to standardize quality assurance processes and determine the suitable educational materials needed to improve screening standards.

Annual reading volumes are the most common reader-based characteristic used to estimate the reader performance in mammography interpretation (3, 8-13). However, some studies suggest that reading volume alone is insufficient(14) and other metrics such as lifetime reading volume may be more suitable in evaluating reader performance as it combines both experience and reading volume (15-17). Previous studies have reported mixed findings regarding the relationship between annual reading volume and reader performance, in Australia the annual reading volume is 2000. Most studies have found a significant positive relationship between annual screening volume and improved reader performance (18-23) , however two studies did not find any association between these two factors (24, 25). Also accounting and adjusting for image-based features was not commonly performed in previous studies, as their methodologies mostly focused on evaluating the reader's predictive performance such as cancer detection rate, sensitivity, lesion sensitivity, ROC and JAFROC.

Using a multi-case multi-reader analysis framework can further increase the power of these studies. In such analysis rather than considering the pooled reader performance (8), the information provided by each data point i.e., each reader's assessment for each case is considered. This multivariate analysis using several reader factors including image-based characteristics is useful to further interrogate the underlying primary determinants of reader performance as an alternative to purely evaluating reader characteristics individually. Therefore, the purpose of this study was to conduct a multivariate investigation of the impact of reader-based and image-based characteristics in reader performance in mammography interpretation to investigate which characteristics are associated with the diagnostic efficacy of screening mammography.

Methodology

Reading conditions

The BREAST test sets were interpreted by reader mainly conference venues, where the rooms were designed to emulative the clinical environment typically used to interpret radiology

images, or in actual reporting rooms in the clinical setting. In various conferences, the readers viewed images using Sectra Breast Imaging PACS (Sectra, Linköping, Sweden; Hologic, Bedford, Mass), which provided post-processing tools for the reader such as zooming, windowing, panning. The rooms used ambient lighting below 10 lux and for the workstation's hardware, they were either linked to two MFGD 5621 monitors (Barco, Kortrijk, Belgium), or two RadiForce G51 monitors (Eizo, Ishikawa, Japan) and were calibrated to the Digital Imaging and Communication in Medicine grayscale standard display function, had a contrast ratio of 365:1, and displayed a maximum luminance within 5% of 475 cd. m−2, minimum luminance of 1.3 cd.m^{-2}.

Data collection

Ethical approval was given for this retrospective study and informed consent was collected from the radiologists. Between 2015-2021, 479 radiologists completed the selected test set. All images and radiologist's data were anonymous and de-identified.

Test sets

A total of six test sets with a total of 60 mammography cases each from BREAST were used. Each test sets consisted of 20 cancer-positive cases and 40 normal cases. The cancer-positive cases had a variety of lesion sizes and appearances. Normal cases remained normal based on a two-year follow-up. Each case consisted of four mammograms: the left and right cranial caudal projections and the medial lateral oblique projections.

Case characteristics

Mammography cases were categorised as either normal or abnormal cases. Normal case characteristics consisted of only breast density. Breast density categorisation initially used the BIRADS breast density categorisation as follows: A – Fatty (0-25% glandular), B – Fibroglandular (25-50% glandular), C – Heterogenous (51-75% glandular), D – Extremely dense (76-100% glandular). For the statistical purposes of this study, breast density was then grouped together classified as either "Dense" breasts consisting of category 3 and 4 cases, and "non-dense" breasts consisting of category 1 and 2 cases.

For abnormal cases, in addition to breast density, the pathology characteristics were included and consisted of lesion side (left or right breast), lesion type (calcification, discrete mass, stellate mass, non-specific density, architectural distortion, and spiculated mass), lesion length measured in millimetres via biopsy.

Both normal and abnormal cases had a difficulty rating determined by an expert who had held a very senior position in BreastScreen Australia. This expert rated the difficulty based on the

probability of a radiologist potentially missing the pathology. The range was defined from 1-3 (‘Easy’, ‘Intermediate’, and ‘Difficult’, respectively).

Reader data

To evaluate the cases in the test set, readers were provided the MLO and CC views of the left and right breast per case. The reader could mark lesions anywhere on the image and used the Royal Australian and New Zealand College of Radiologist’s (RANZCR) breast imaging rating system to grade the case. The rating system ranges from 1- 5 (1- ‘Normal’; 2- ‘Benign’; 3- ‘Equivocal’; 4- ‘Suspicious’; 5- ‘Malignant’, respectively). A score classification between 3-5 was used to determine increasing confidence in malignancy of the marked area. Should the reader not find any abnormality, they clicked “next case” and the platform automatically marked the cases as RANZCR score 1 (normal). Only the reader’s first attempt at the test set was used. Readers were not aware of the prevalence and types of cancers included in the test sets.

Reader Characteristic

Readers who participated in the BREAST program were asked to complete a survey regarding their experience and training prior to starting the test set. The survey collected the reader’s characteristics including age, gender, role (either a trainee or radiologist), years in speciality, the reader’s specialisation, number of cases interpreted per week, fellowship training, and whether the reader works in a breast screening program.

Statistical analysis

For the statistical analysis, the first part of this study was to evaluate how the reader/case characteristics affected the reader’s accuracy via linear mixed modelling. For reader accuracy, the dependent variable used in the linear mixed model is an accuracy score calculated as a proportion of correct interpretations. Linear mixed modelling was selected to evaluate how the fixed effects such as a reader characteristic (i.e., number of cases interpreted per week) or case characteristic (i.e., breast density) affects the reader’s accuracy, measured by either correctly marking the lesions or by their confidence score. To account for multiple readers and cases, these two factors were classified as random effects in the linear mixed model. The modelling was done in the R statistical programming environment (version 4.1.2) and used the lme4 and lmerTest packages.

The linear mixed models were divided from into three categories (1. case-based characteristics, 2. reader-based characteristics, and 3. mixed-characteristics) and explores normal and cancer cases separately.

Results

In this study, we selected readers who completed one specific test set. Between 2015-2021, there has been 479 Australian radiologists who had participated in the six test sets used for this study, totalling to 38160 individual case readings. To ensure minimal bias in test set interpretation, only the reader's first attempt at the test set is used in this study. The demographics of readers who completed the survey and participated in interpreting the BREAST test sets are listed in Table 1.

The details of the six test sets used are shown in Table 2 and 3 for normal and cancer cases, respectively. Each test set contained 20 cancer cases and 40 normal cases. For cancer cases' classification, Table 3 shows additional cancer case details categorised based on the lesion type, side of cancer, site of cancer, and lesion size measure in millimetres.

Table 1. Radiologists' characteristics and experience during test set completion

Reader characteristic	**n=479**
Mean age (years old)	46.22
Gender	
Female	191
Male	239
Non-specified	49
Mean number of years in specialisation	4.0489
Reader Specialisation	
Breast	301
Breast | Lung | Other	2
Breast | Other	14
Lung	1
Other	149
Not available	12
Number of mammograms interpreted weekly	
0-20	301
21-60	56
61-100	46
101-150	29
151-200	21
over 200	26
Fellowship trained in breast	94
Working in a breast screen program	120
Readers using digital images for interpretation	418

Table 2. Normal case characteristics

Normal Case characteristic	n=240
Breast density	
Non-dense	124
Dense	116
Case difficulty	
Easy	117
Intermediate	109
Difficult	14

Table 3. Cancer case characteristics

Cancer Case characteristic	n=120
Breast density	
Non-dense	57
Dense	63
Case difficulty	
Easy	19
Intermediate	38
Difficult	70
Cancer Type	
Architectural Distortion, Stellate	62
Calcification	21
Discrete Mass	15
Nonspecific density	22
Lesion Size	
<=15mm	101
>15mm	24
N/A	2

Table 4. Linear mixed model results separated by reader and case characteristics and normal and cancer case type

		Normal			Cancers		
	Characteristic	Number of data points	Model 1 & 2	Combined Model (3)	Number of data points	Model 1 & 2	Combined Model (3)
Case	Difficulty						
	Difficult (Intercept)	14	0.634*	1.5797*	67	0.692*	-1.0047*
	Easy	117	0.174*	0.17512	18	0.169*	0.522*
	Intermediate	109	0.119**	0.26644**	35	0.0298	0.0404*
	Breast density						
	Dense	116	-0.016	-	57	0.00393	-
	Non-dense	124	-0.049	-	63	0.0401	-
	Lesion Characteristics						
	Calcification	-	-	-	21	-0.0213	0.15345
	Discrete Mass	-	-	-	15	0.0553	0.109
	Nonspecific density	-	-	-	22	-0.0501	-0.0946
	Architectural distortion, spiculated mass and stellate	-	-	-	62	0.0242	0.183
	Lesion size	-	-	-	120	-0.00169	-
Reader	Reader Specialty						
	Breast (Intercept)	317	1.6968*	0.54829*	317	0.7454*	3.0509*
	Non-Breast	162	-0.05675	-0.062**	162	-0.02502	-0.176*
	Breast Screen Reader						
	No	359	-0.04837	-	359	0.038867	-
	Yes	120	-0.06038	-	120	-0.03585	-
	Breast fellowship						
	No	384	0.27387	-	384	0.041077	-
	Yes	94	0.017557	-	94	-	-

					0.0527**	
Number of mammograms interpreted per week						
0-20	301	-0.09835*	-0.08832*	301	0.0229	0.15205*
21-60	56	-0.19212*	-0.20419*	56	0.0678*	0.21397*
61-100	46	-0.24791*	-0.26502*	46	0.0482	0.26168*
101-150	29	-0.32152*	-0.34073*	29	0.0674	0.2916*
151-200	21	-0.0848	-0.09766	21	0.111*	0.39452*
>200	26	-0.46454*	-0.45984*	26	0.0480	0.23064**

"*" for <0.01 and ** for P-values between 0.01-0.05.

Models 1 and 2 represent case-only and reader-only models, respectively while the combined model included both case and reader-based features.

Linear Mixed Model – Case-based characteristics

For our first linear mixed model, we studied the impact case-based characteristics had on the reader's accuracy and the results are shown in Table 4. For both normal and cancer cases, the model shows that the reader accuracy decreases as expert-determined difficulty increases (P <0.05).

For breast densities, readers showed better accuracy in non-dense cases, compared to the dense cases, however, breast density was not a significant characteristic in impacting reader accuracy (P >0.05). Notably, although lesion-based characteristics were not significant, lesions classified as stellate and discrete mass demonstrated that readers were more likely to mark these cases correctly in comparison to calcifications. Furthermore, all lesion types with non-specific densities (NSD) tend to perform worse compared to the other lesion types.

Linear Mixed Model – Reader-based characteristics

For reader-based characteristics, the results for normal and cancer case's model differed greatly. For the reader characteristics relative to reader accuracy in normal cases, the results in Table 4 shows that the most significant reader characteristic is the number cases interpreted per week, excluding 151-200 (p=0.236), the others (0-20, 21-60, 61-100. 101-150, and >200) were all significant (P= <0.01). Another reader characteristic of note is the reader's specialization, as the results for readers who have a specialization relevant to breast were significant, the results for readers without a specialization in breast were not significant and in comparison, to breast specialised readers, non-breast specialists had a lower intercept, which suggests lower reader accuracy.

In terms of mammograms interpreted per week, the mean estimate decreases with increasing number of mammograms interpreted per week. This suggests that readers who interpreted more mammograms per week were more accurate in their confidence rating.

Conversely, for the cancer case's linear mixed model, the number of mammograms interpreted per week showed a similar trend in as the normal cases' trend, where the estimate steadily increased until 151-200 cases interpreted per week. However, the significance was not consistent, only 21-60 and 151-200 were significant (P= <0.01). Notably, years in specialization was significant with increasing years slightly improving reader accuracy.

Combined Model

When we combined significant reader and image characteristics into our mixed model, our results show that number of mammograms interpreted per week and case difficulty both retain a similar trend as they did from the previous two linear mixed models.

In terms of the mixed model where we combined case characteristics and reader characteristics for the function for reader's confidence, the notable significant results found are the case difficulty and number of mammograms interpreted per week and years, respectively.

Discussion

This study explored the impact of reader-based and image-based characteristics on reader accuracy in screening mammography. After the readers undertook a test set, we compiled their accuracy and survey details and processed their data in a series of linear mixed models. Our results show that expert-predetermined difficulty and the number of cases interpreted weekly were the consistently significant factors for image-based and reader-based characteristics, respectively. The results we obtained from this study were largely expected both anecdotally and from previous studies. However, we have provided this additional evidence to reinforce these findings. Additionally, in comparison to previous studies, we have established a trend between reader and image characteristics and their relationship with reader accuracy using a large sample size for both readers and cases.

The purpose of using a linear mixed model was to find the relationship between multiple variables with coefficients that could potentially vary with respect to one or more grouping variables. Linear mixed modelling is composed of two parts: fixed effects and random effects. The fixed effects are in essence, conventional linear regression analysis to predict the value of a variable based on the value of another. The random effect factors are values that can be considered a random sample from a larger population and are used to justify excess variability per variable i.e., readers and test set items with varying characteristics between each record.

By using this approach, we have provided stronger statistical power and establish a better relationship between predictors and outcome variables. The estimate co-efficient represents the strength of association between the variables relative to the determined intercept in the model. By evaluating the difference in co-efficient values from the intercept and between other variables, we could establish the characteristics that have the greatest impact on affecting reader accuracy.

Previous studies have found that readers with an annual mammogram interpretation volume of over 1000 perform better than those reading less than 1000 per year, with varying levels of accuracy improvement among the previous studies(6, 11, 23, 26-29). Notably, our findings on reading volume correlate closely with Kan's study, where the cancer detection rate improves proportionally with the number of mammograms interpreted, until a certain upper threshold is reached where the reader's cancer detection rate improvement rate is negligible. For Kan et al.'s study, improvement in reader accuracy were not statistically significant beyond 2,500 annual screening mammograms interpreted (6). This suggests that while the reader's accuracy can be improved by increasing the interpretation volume, there is an upper limit to which readers do not significantly improve in accuracy relative to the increased number of mammograms read.

Regarding expert-determined difficulty, our findings show that with increasing expert-predetermined difficulty, the reader's accuracy is reduced. The current development for test sets for education and training largely relies on using expert opinion for the curation test sets and quantifying each case's difficulty(30, 31). Notably, two studies found that reader accuracy in mammography interpretation and expert-determined difficulty were associated. Specifically, as expert-predetermined difficulty increases, the reader's false negative errors increase as well, but false positive errors were not associated (30, 31). In contrast in our finding's expert-determined difficulty impacted in the reader's false positive errors in the normal images and mixed model, but not exclusively in the cancer model for false negative errors.

Although from our results breast density was not a significant factor in impacting reader accuracy, interestingly the readers in this study performed better in dense breasts in comparison non-dense breasts for normal cases. Conversely, for cancer cases, readers performed better in non-dense cases compared to dense breast cases. This finding may suggest that breast density and the presence of malignancy together have an impact on the readers 'abilities to screen mammography. This finding is mixed compared to previous studies, while older studies related to film-based mammography suggested that mammography interpretation in dense breasts

encounter difficulty due to the masking effect super imposing potential malignancies (32, 33). Modern digital-based mammography has partially resolved this issue as digital-based mammography offers post-processing tools to manipulate the mammogram, reducing the issue encountered the issue in film-based mammography(34). Therefore, this finding may suggest that although digital mammography may have significantly reduced the negative impacts of the masking effect in normal cases, the masking effect remains prevalent in malignant cases in mammography. However, as the result is not significant further investigation is required.

For lesion characteristics, although the results were not significant; the cases categorised as primarily stellate and discrete mass demonstrated that readers were more likely to detect this pathology correctly in comparison to calcifications and non-specific densities. Incidentally, all the cases that were classified primarily as non-specific densities (NSD) tended to perform inferiorly when compared to the other lesion types. This correlates with Norsuddin et al.'s findings which reported cancers with stellate masses were more easily detected while cancers with non-specific density were more likely to be missed (35).

In terms of clinical implications, our findings reinforce that using annual reading volumes required to maintain certification in mammography interpretation is a valuable method to evaluate a reader's ability. However, one problem with the current implementation of the mammography's guidelines is the variable requirement across different countries. Currently, the reading volume required to maintain certification greatly varies between different regions, In North America, the mean annual mammographic interpretive volume was 1,777, with most readers interpreting less than 1,000 per year(36). In the UK and several other European countries, the recommended annual minimum reading volume is 5,000 mammograms per reader(37).

Another clinical implication to consider is screening for non-specific densities and architectural distortions. This type of breast cancer is the third most common appearance, however, it is often a subtle finding in mammography(38). Compared to calcifications and discrete masses, non-specific densities are often more difficult to diagnose because of their variability in appearance (39). Recent studies show that compared to mammography, digital breast tomosynthesis is better in detecting non-specific densities and architectural distortion in breasts with readers often having lower interobserver variability and higher specificity in the detection non-specific densities(40, 41). However, digital breast tomosynthesis is not widely implemented in screening programs, therefore, we need to implement alternative methods to improve the diagnosis of non-specific densities until the modality is introduced more widely in

breast screening. One novel method of improving non-specific density diagnosis was suggested by Pereira Junior et al, who used a combination of computer-aided detection and radiomics to detect different differences in pixel values and their surrounding mean pixel value (42).

A potential limitation of our study may be the prevalence effect, where our results are effectively created in a lab setting, and therefore our findings may not entirely reflect on real practice and other settings. In comparison to what the reader commonly sees in the clinical setting, the BREAST test sets have a higher number of positive cases and this may have resulted in a laboratory effect as our readers tend to perform better in cancer-positive cases compared to normal (43). However, a previous study has demonstrated a moderate level of agreement between clinical accuracy and test set interpretation (44). Another potential limitation is the potential comparability of the data from BREAST, as readers may have completed the test sets in either a clinical setting or in a conference or workshop. However, in conferences and workshops, BREAST has accounted for the reading conditions and simulated the conditions as close as possible to actual clinical settings, so the impact of the conference reading environment has been minimised.

Conclusion

In summary, our findings show that expert-determined difficulty and cases interpreted weekly were consistently significant factors impacting reader accuracy using advanced analytical methods. For case characteristics, increasing the case difficulty determined by experts reduces the reader's accuracy which has implications for experts curating future test sets for educational and audit purposes. While for reader characteristics, increasing the number of cases interpreted weekly improves the reader accuracy up to a certain threshold. Our results show that linear mixed modelling provides extra insights into the impact of reader- and image-based factors on reader accuracy in mammography interpretation. Future texture-based radiomic analysis would be useful to further investigate, the cancer image cases' characteristics.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359-86.
2. Demchig D, Mello-Thoms C, Lee WB, Khurelsukh K, Ramish A, Brennan PC. Mammographic detection of breast cancer in a non-screening country. Br J Radiol. 2018;91(1091):20180071.
3. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009;253(3):641-51.
4. Zhu C, Wang L, Du LB, Li J, Zhang J, Dai M, et al. [The accuracy of mammography screening for breast cancer: a Meta-analysis]. Zhonghua Liu Xing Bing Xue Za Zhi. 2016;37(9):1296-305.
5. Quality assurance guidelines for radiologists. NHS Breast Screening Programme. 2011.
6. Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology. 2000;215(2):563-7.
7. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. Ann Oncol. 2008;19(4):614-22.
8. Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings. Radiology. 2013;269(1):61-7.
9. Reed WM, Lee WB, Cawson JN, Brennan PC. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. Acad Radiol. 2010;17(11):1409-13.
10. Suleiman WI, Lewis SJ, Georgian-Smith D, Evanoff MG, McEntee MF. Number of mammography cases read per year is a strong predictor of sensitivity. J Med Imaging (Bellingham). 2014;1(1):015503.
11. Hoff SR, Myklebust T, Lee CI, Hofvind S. Influence of Mammography Volume on Radiologists' Performance: Results from BreastScreen Norway. Radiology. 2019;292(2):289-96.
12. Morrone D, Giordano L, Artuso F, Bernardi D, Fedato C, Frigerio A, et al. Factors associated with breast screening radiologists' annual mammogram reading volume in Italy. Radiol Med. 2016;121(7):557-63.
13. Walker MJ, Hartman K, Majpruz V, Leung YW, Fienberg S, Rabeneck L, et al. The Impact of Radiologist Screening Mammogram Reading Volume on Performance in the Ontario Breast Screening Program. Can Assoc Radiol J. 2021:8465371211031186.

14. Gandomkar Z, Lewis SJ, Li T, Ekpo EU, Brennan PC. A machine learning model based on readers' characteristics to predict their performances in reading screening mammograms. Breast Cancer. 2022.
15. Burnside ES, Lin Y, Munoz del Rio A, Pickhardt PJ, Wu Y, Strigel RM, et al. Addressing the Challenge of Assessing Physician-Level Screening Performance: Mammography as an Example. PLOS ONE. 2014;9(2):e89418.
16. Ciatto S, Ambrogetti D, Catarzi S, Morrone D, Rosselli Del Turco M. Proficiency test for screening mammography: results for 117 volunteer Italian radiologists. J Med Screen. 1999;6(3):149-51.
17. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Womens Health. 1998;7(4):443-9.
18. Hoff SR, Myklebust TÅ, Lee CI, Hofvind S. Influence of mammography volume on radiologists' performance: Results from breastscreen Norway. Radiology. 2019;292(2):289-96.
19. Théberge I, Hébert-Croteau N, Langlois A, Major D, Brisson J. Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. CMAJ. 2005;172(2):195-9.
20. Theberge I, Chang SL, Vandal N, Daigle JM, Guertin MH, Pelletier E, et al. Radiologist interpretive volume and breast cancer screening accuracy in a canadian organized screening program. Journal of the National Cancer Institute. 2014;106(3).
21. Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings. Radiology. 2013;269(1):61-7.
22. Suleiman WI, Lewis SJ, Georgian-Smith D, Evanoff MG, McEntee MF. Number of mammography cases read per year is a strong predictor of sensitivity. Journal of Medical Imaging. 2014;1(1).
23. Kim SH, Lee EH, Jun JK, Kim YM, Chang YW, Lee JH, et al. Interpretive Performance and Inter-Observer Agreement on Digital Mammography Test Sets. Korean J Radiol. 2019;20(2):218-24.
24. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009;253(3):641-51.
25. Molins E, Macià F, Ferrer F, Maristany MT, Castells X. Association between Radiologists' Experience and Accuracy in Interpreting Screening Mammograms. BMC Health Services Research. 2008;8.
26. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology. 2002;224(3):861-9.
27. Buist DS, Anderson ML, Smith RA, Carney PA, Miglioretti DL, Monsees BS, et al. Effect of radiologists' diagnostic work-up volume on interpretive performance. Radiology. 2014;273(2):351-64.

28. Buist DSM, Anderson ML, Haneuse SJPA, Sickles EA, Smith RA, Carney PA, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology. 2011;259(1):72-84.
29. Haneuse S, Buist DSM, Miglioretti DL, Anderson ML, Carney PA, Onega T, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. Radiology. 2012;262(1):69-79.
30. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA. Radiology resident mammography training: interpretation difficulty and error-making patterns. Acad Radiol. 2014;21(7):888-92.
31. Grimm LJ, Zhang J, Lo JY, Johnson KS, Ghate SV, Walsh R, et al. Radiology Trainee Performance in Digital Breast Tomosynthesis: Relationship Between Difficulty and Error-Making Patterns. Journal of the American College of Radiology : JACR. 2016;13(2):198-202.
32. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. RadioGraphics. 2015;35(2):302-15.
33. DS ALM, Ryan EA, Mello-Thoms C, Brennan PC. What effect does mammographic breast density have on lesion detection in digital mammography? Clin Radiol. 2014;69(4):333-41.
34. Al Mousa DS, Mello-Thoms C, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Mammographic density and cancer detection: does digital imaging challenge our current understanding? Acad Radiol. 2014;21(11):1377-85.
35. Mohd Norsuddin N, Mello-Thoms C, Reed W, Rickard M, Lewis S. An investigation into the mammographic appearances of missed breast cancers when recall rates are reduced. The British journal of radiology. 2017;90(1076):20170048-.
36. Smith-Bindman R, Miglioretti DL, Rosenberg R, Reid RJ, Taplin SH, Geller BM, et al. Physician workload in mammography. AJR American journal of roentgenology. 2008;190(2):526-32.
37. Moss SM, Blanks RG, Bennett RL. Is radiologists' volume of mammography reading related to accuracy? A critical review of the literature. Clin Radiol. 2005;60(6):623-6.
38. Shaheen R, Schimmelpenninck CA, Stoddart L, Raymond H, Slanetz PJ. Spectrum of diseases presenting as architectural distortion on mammography: multimodality radiologic imaging with pathologic correlation. Semin Ultrasound CT MR. 2011;32(4):351-62.
39. Gaur S, Dialani V, Slanetz PJ, Eisenberg RL. Architectural Distortion of the Breast. American Journal of Roentgenology. 2013;201(5):W662-W70.
40. Dibble EH, Lourenco AP, Baird GL, Ward RC, Maynard AS, Mainiero MB. Comparison of digital mammography and digital breast tomosynthesis in the detection of architectural distortion. Eur Radiol. 2018;28(1):3-10.
41. Bahl M, Lamb LR, Lehman CD. Pathologic Outcomes of Architectural Distortion on Digital 2D Versus Tomosynthesis Mammography. AJR Am J Roentgenol. 2017;209(5):1162-7.

42. Junior OP, Oliveira HCR, Ferraz CT, Saito JH, Vieira M, Gonzaga A. A Novel Fusion-Based Texture Descriptor to Improve the Detection of Architectural Distortion in Digital Mammography. J Digit Imaging. 2021;34(1):36-52.
43. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. Radiology. 2008;249(1):47-53.
44. BaoLin PS, Warwick ML, Peter LK, Warren MR, Mark FM, Patrick CB, editors. Test set readings predict clinical performance to a limited extent: preliminary findings. ProcSPIE; 2013.

Chapter Five

Investigating the association of textural radiomic features and cancer characteristics with diagnostic errors in mammography interpretation.

5.1 Chapter Introduction

Chapter 5 manuscript is currently under review at the *European Journal of Radiology*. Additionally, this study was accepted and presented at both RANZCR 2022 and MIPS2022 (Appendix E and F).

Chapter 5 finalises the project and builds on the findings from Chapter 4, The study found that both expert-predetermined difficulty and the number of cases interpreted weekly had a consistent impact on the reader's performance in mammography interpretation. Increasing difficulty levels were associated with reduced performance, while an increase in the number of cases interpreted led to improved performance. However, the impact of cancer cases' characteristics on performance was inconclusive.

While the multivariate model has established the primarily reader characteristics that affect reader performance was the reading volume, image characteristics remains uncertain. As previously mentioned, the findings of Chapter 3 and 4 may become contradictory due to the inconsistent number of cases and readers between the two studies when it came to expert-determined difficulty. Regardless, if we consider the findings of Chapter 3; specifically, the characteristics that experts used to determine case difficulty, these findings may contrast the findings from Chapter 4.

In Chapter 3, while experts primarily use breast density and presence of cancer as the indicator of case difficulty, Chapter 4's found none of the image characteristics to be significant. However, there was a trend that readers consistently performed better in dense breasts compared to non-dense breasts and specific malignancies such as stellate discrete masses were more likely detected. As such, Chapter 5's aim was to further investigate the influence cancer case characteristics has on reader performance using texture-based radiomic analysis.

Chapter 5 focuses on the fourth objective of the thesis: Validating the impact case characteristics has on reader performance and presents the introduction, methodology, results, discussion, and conclusion for the impact textural radiomics in cancer cases has in reader performance in mammography interpretation. After evaluating the performance metrics and reader characteristics, this chapter would utilize textural quantization to quantify mammography textures to evaluate the impact breast textures has on reader performance, as

there may be interactions between anatomical textures and pathological textures influencing reader performance.

Ridge Regression is a linear regression model that accounts for multicollinearity among predictor variables. In multi-reader multi-case dataset, Ridge Regression is advantageous as it permits the inclusion of all data points from multiple readers without the need of averaging their performance. This approach enhances the statistical power by leveraging the full dataset. The model treats cancer characteristics as random effects to account for variability across multiple readers and cases. Fixed effects like breast density, cancer type, and radiomic features are included to predict the reader's performance in lesion detection.

The model's approach to cancer characteristics is particularly noteworthy. Treating these characteristics as random effects, Ridge Regression adeptly accounts for the inherent variability across different readers and cases. This variability is a cornerstone of real-world medical scenarios, where factors such as lesion size, shape, and location can influence a reader's ability to detect them accurately. By recognizing and adjusting for these random effects, Ridge Regression provides a more accurate reflection of the diagnostic process.

Furthermore, fixed effects like breast density, cancer type, and radiomic features are integral to the model. These variables, known for their predictive value in lesion detection, are incorporated to enhance the model's ability to forecast reader performance accurately. The inclusion of these fixed effects allows for a nuanced exploration of how different characteristics influence diagnostic accuracy, offering valuable insights into areas that might benefit from focused training or technological enhancement.

Ridge Regression applies a penalty to the regression coefficients variables which shrinks them towards zero, but not precisely zero. This penalty is the L2 norm of the coefficients, and it helps in reducing model complexity. However, the limitations of Ridge Regression include its inability to reduce the coefficients to absolute zero, which means it does not perform feature selection and can result in a model that includes all predictors, thereby possibly retaining irrelevant features. Additionally, the interpretation of the coefficients can become more complex due to the shrinkage, especially when trying to understand the impact of interactions in models that combine different types of characteristics, such as expert-determined and radiomic features.

However, it's important to acknowledge the limitations inherent to Ridge Regression. One significant constraint is its inability to reduce coefficients to absolute zero. Unlike some other regression techniques, such as Lasso Regression, Ridge Regression does not perform feature selection. Consequently, the model may include all predictors, potentially retaining irrelevant or less impactful features. This characteristic can lead to a more complex model that, while robust, may include variables that do not significantly contribute to the predictive power of the model.

Additionally, the interpretation of the coefficients in Ridge Regression can become challenging due to the shrinkage effect. The reduced coefficients can be difficult to interpret, especially in models that combine various types of characteristics, such as expert-determined and radiomic features. Understanding the impact of interactions between these diverse variables becomes more complex under the influence of coefficient shrinkage.

Investigating the association of textural radiomic features and cancer characteristics with diagnostic errors in mammography interpretation.

Abstract

Rationale and Objectives: This study investigated the association between localised extracted textural features and cancer characteristics with the difficulty of a screening mammogram.

Materials and Methods: Using radiological and pathological reports, two expert radiologists determined malignancy characteristics using features such as cancer type (e.g., architectural distortions, calcification, etc.), cancer site, and cancer size (<15mm or > 15mm). The lesions were also segmented and a comprehensive set of radiomic features, including first, second, and higher order statistical features, features from Gabor multi-scale decomposition, features from MR8 Filter Bank, and fractal dimension were extracted. Data from a total of 479 Australian radiologists interpreting 120 cancer cases were collected. The analysis was conducted using ridge regression model with least absolute shrinkage and selection operator. Three models were created to establish the characteristics describing the appearance of difficult-to-interpret breast cancer: one based on expert-determined cancer characteristics, another on extracted radiomics characteristics, and a combination of both to explore potential interactions.

Results: The model based on expert-determined cancer characteristics suggests that the cancer type, site, and size have the most significant impact on reader accuracy. The radiomic model indicates that specific histogram and Haralick textural features affect diagnostic accuracy and the atypical cancerous appearance (e.g., lesions with a higher information measure of correlation1 (IMC1), indicating a more structured texture, resembling the normal breast parenchyma) hinders lesion detection.

Conclusion: Using quantised textural features, the results of this study have identified specific lesion characteristics associated with the texture of malignancy that may contribute to reduced cancer detection.

Keywords mammography, radiomics, textural features, pathology

Introduction

Mammography is a crucial diagnostic tool for the early detection of breast cancer. However, the interpretation of mammograms frequently presents challenges, owing to the intricate and heterogeneous nature of breast tissue, consequently giving rise to both inter-observer and intra-observer variability (1). To ensure optimal performance in breast cancer detection and to mitigate these variabilities, it is imperative to devise effective strategies. One such strategy could involve the identification of specific characteristics within lesions or images that render cancer detection more challenging (2). During the development of educational materials, particular attention should be paid to including cases featuring these distinct characteristics. Moreover, equipping radiologists with knowledge about these features can contribute to enhancing their diagnostic accuracy.

Radiologists often grapple with cognitive biases and errors that can substantially influence their decision-making process. Meta-cognition, which denotes awareness and understanding of one's own cognitive processes and underlying causes of errors, empowers radiologists to recognise these biases and consciously counteract them (3). Over the past two decades, computer-aided diagnosis systems and artificial intelligence (AI) tools have been developed to assist radiologists in the detection and diagnosis of breast cancer. These tools utilise textural features extracted from mammograms to identify suspicious lesions. Further understanding the intricate interactions between these textural features and radiologists' diagnostic accuracy can help delineate cases on which AI might offer substantial benefits, given the heightened potential for human observer errors.

While incorrect positioning or exposure is often cited as a cause of missed cancers in mammography, previous research suggests that a comprehensive examination of other features is necessary (2). Among various factors, the impact of breast density on diagnostic performance has garnered significant attention. While some studies indicate that missed cancers on mammograms are more likely to occur in denser breasts (2), others posit that missed cancers tend to exhibit lower density and smaller size compared to easily detectable lesions. Notably, limited data exists regarding the impact of other radiographic appearances of disease, such as overall lesion morphology and texture (2). Some previous studies have examined shape and boundary roughness for mass lesions and found that mammographic diagnostic sensitivity may be adversely affected without appropriate attention to spiculation on the boundary and its roughness (4-6). Despite the inclusion of a wide range of quantitative shape-based radiomic features, textural features contributing to the difficulty of diagnosing cancerous breast lesions

have not been investigated (4-6). According to the current literature previous studies have explored the use of texture features in mammography and breast MRI but to assess the risks of developing breast cancer (4-6) and parenchymal textural features of the breast are primarily used in machine-driven identification and computer-aided diagnosis studies (7). Finally, none of the earlier studies (4-6) included a very large cohort of radiologists to establish the case difficulty to make the diagnosis.

Therefore, the purpose of this study was to explore the interactions between localised textural radiomic features and cancer characteristics in mammography and how these factors influence a radiologist's final diagnosis. By doing so, this research aimed to identify specific areas for intervention to enhance interpretive accuracy, develop quality assurance measures and create training materials designed to improve reader performance, ultimately advancing the early detection of breast cancer. The identification of meaningful associations between texture features and diagnostic errors can also offer valuable insights into the effectiveness of future AI tools, enabling the determination of cases in which radiologists may require more assistance than others.

Methodology

Reading conditions

The BREAST test sets were read either in conference venues in rooms designed to simulate environments typically used for interpreting radiology images, or in actual reporting rooms within the clinical setting of the participants. In conference venues, the readers viewed images using the Sectra Breast Imaging PACS (Sectra, Linköping, Sweden; Hologic, Bedford, Mass), which provided post-processing tools for the readers such as zooming, windowing, and panning. These rooms were illuminated with ambient lighting levels below 10 lux. In terms of hardware, workstations were either connected to two MFGD 5621 monitors (Barco, Kortrijk, Belgium), or two RadiForce G51 monitors (Eizo, Ishikawa, Japan). These monitors were calibrated to conform to the Digital Imaging and Communication in Medicine grayscale standard display function, with a contrast ratio of 365:1, and displaying a maximum luminance within 5% of 475 cd. m−2, minimum luminance of 1.3 cd.m^{-2}.

Test sets

Six test sets, each consisting of 60 mammography cases from BREAST were used. Each test sets comprised of 20 cancer-positive cases and 40 cancer-negative cases. Each case consisted of four mammograms (left and right cranial caudal (CC) projections and the medial lateral

oblique (MLO) projections). The cancer-positive cases were biopsy proven, with a variety of lesion sizes and appearances.

Reader Performance

Ethical approval for this retrospective study was granted by the University of Sydney and informed consent was obtained from the radiologists. To evaluate the cases in the test set, readers were provided the MLO and CC views of the left and right breast for each case. The readers were able to mark lesions anywhere on the image and used the RANZCR breast imaging rating system to assess and grade each case. A RANZCR classification between 3-5 was used to indicate increasing confidence in the malignancy of the marked area. If the reader did not identify any abnormalities, they clicked "next case" and the platform automatically marked the case as RANZCR score of 1 (normal).

Case characteristics

Breast density was initially classified using the BIRADS breast density categorisation (A – Fatty, B – Fibroglandular, C – Heterogenous, D – Extremely dense). However, for statistical purposes, it was grouped into either "dense" (category 3 and 4 cases), and "non-dense" (category 1 and 2 cases).

Variables describing cancer type (architectural distortions, calcification, discrete mass, non-specific density, spiculated mass, stellate mass and cases with a site that had both stellate masses and non-specific density), site of cancer (upper inner, upper outer, upper only, retro areolar, outer half, lower outer, lower only, and lower inner), Royal Australian and New Zealand College of Radiologist's (RANZCR) rating, size (<15mm or > 15mm) measured via biopsy, expert-determined case difficulty (easy, intermediate, and difficult) and availability priors were all also included.

The RANZCR rating system ranges from 1- 5 (1- 'Normal'; 2- 'Benign'; 3- 'Equivocal'; 4- 'Suspicious'; 5- 'Malignant', respectively) and the rating was provided by two highly experienced radiologists.

Cases were also rated by an expert who held a very senior position in BreastScreen Australia. This expert rated the difficulty based on the probability of a radiologist potentially missing the pathology. The range was defined from 1-3 ('Easy', 'Intermediate', and 'Difficult', respectively).

The characteristics of the cases used are listed in Table 1, please note that some cases have multiple cancer lesions which resulted in additional lesion size measurements and case difficulties.

Table 1. Cancer case characteristics

Cancer Case characteristic	n=120
Breast density	
Non-dense	57
Dense	63
Case difficulty	
Easy	19
Intermediate	38
Difficult	70
Cancer Type	
Architectural Distortion, Stellate	62
Calcification	21
Discrete Mass	15
Nonspecific density	22
Lesion Size	
<=15mm	101
>15mm	24
N/A	2

Radiomic features

For the radiomics analysis, the cancer areas were segmented using Matlab, with ground truth and annotations provided by two highly experienced radiologists. A comprehensive set of radiomic features were extracted from each segmented area, categorised into four main groups:

1. First-order statistics based on histograms, which describe the distribution of grey level intensities.
2. Second-order texture features or Haralick texture features, quantifying the spatial relationships between pixels and overall texture.
3. Higher-order statistics features and features based on image transforms, obtained from decomposed images using a set of filters, providing additional information about the interrelationship among pixels.
4. Fractal features.

Specifically, the features include:

- 28 histogram-based statistics,
- 88 features are derived from the grey level co-occurrence matrix (GLCM) (4),
- seven from the grey level run length matrix (GLRLM) (5),
- six from the grey level sharpness measure (GLSM) (6),
- 15 from grey level difference statistics (GLDS) (7),

- 15 from the neighbourhood grey tone difference matrix (NGTDM) (8)
- eight from the statistical feature matrix (SFM) (9),
- 18 laws texture energy measures (10), and
- two from fractional dimension texture analysis (11).

Additionally, there are 16 transform-based features, including:

- six from Gabor-filtered images
- eight from RFS-filtered images (12, 13)
- two features from the Fourier-transformed images (9).

For a complete list of the radiomic features used in this study, please refer to the Supplementary Materials.

Statistical analysis

The study includes readings from multiple readers on various cases. When dealing with multi-reader multi-case datasets, Ridge Regression enables us to leverage all readers' data without averaging their performance thereby enhancing the statistical power of the study (14). In this context, to account for multiple readers and cases, cancer characteristics were classified as random effects and account for collinearity, Ridge Regression was used to assess how fixed effects such as breast density, cancer type, and radiomic features affected the reader's performance, measured by their ability to correctly identify the lesions. Ridge Regression was performed using the least absolute shrinkage and selection operator (15) within the R statistical programming environment (version 4.1.2). The model output could range from 0 to 1, where 0 represented the most difficult lesion, missed by all readers and 1 represented a lesion detected by all readers. Prior to feeding to the model, all radiomic features were normalised.

Given the inclusion of a large set of radiomic features, Ridge Regression was chosen as it can effectively handle cases when the study sample size is very small or when the number of parameters exceeds the number of samples.

Three models were created:

1. The first model incorporated the expert-determined cancer characteristics.
2. The second model featured extracted radiomics characteristics.
3. The third model combined expert-determined cancer characteristics and radiomic features to investigate potential interactions.
4.

Results

Model 1: Cancer Characteristics

A total of 479 Australian radiologists participated in this study. The initial model employed all cancer characteristics to assess their influence on reader accuracy. The variables and their significance, as determined by the Ridge Regression Model are presented in Table 2. The results indicate that the size of the cancer has the most significant impact on reader accuracy.

Table 2. Case characteristics and their significance and impact on reader performance

Characteristic	Estimate	P-Value
Significant Features		
Cancer type		
Stellate and Non-specific Density	-0.018	0.01577
Calcification	-0.010	<0.0001
Non-specific Density	-0.007	<0.0001
Spiculated Mass + Non-specific density	-0.002	<0.0001
Site		
Lower	0.005	<0.0001
Difficulty (relative to intermediate)		
Easy	0.030	<0.0001
Difficult	-0.015	<0.0001
RANZCR		
RANZCR 4	-0.003	<0.0001
RANZCR 5	0.003	<0.0001
Other		
Size	0.733	<0.0001
Non-Significant Features		
Non-specific density and Stellate	0.001	0.07
Discrete Mass	0.020	0.08
Non-specific density and Architectural Distortion	-0.019	0.10
Spiculated Mass	-0.001	0.75
Stellate Mass	0.011	0.20
Architectural Distortion	-0.002	0.22
Breast Density		
Non-Dense	0.005	0.66
Dense	-0.005	0.25
Site		

Central	-0.012	0.30
Upper	0.004	0.43
Side		
Right	0.002	0.79
Left	-0.002	0.13
Prior		
Yes	-0.006	0.49
No	0.006	0.97

As shown in Table 2, specific cancer types demonstrated, substantial negative impacts on reader accuracy. For example, stellate masses with non-specific density (-0.018, p <0.0001) and non-specific density (-0.007, p <0.0001), were identified as factors associated with increased diagnostic difficulty. Conversely, larger pathological masses (0.733, p <0.0001) exhibited a significant positive impact on accuracy, suggesting that their size facilitates more precise diagnosis. Calcification (-0.010, p <0.0001) was found to significantly hinder accuracy, underscoring its role as a challenging diagnostic factor. Spiculated masses with non-specific density (-0.002, p <0.0001) also posed diagnostic challenges.

Regarding the cancer's relative site, cancers located in the lower region (0.005, p <0.0001) had a significant positive impact on reader accuracy. However, none of the other mammogram regions yielded significant results.

In terms of the mammogram's case difficulty level, as determined by experts, it had a substantial influence, with easy cases (0.030, p <0.0001) enhancing accuracy and difficult cases (-0.015, p <0.0001) impeding it. Lastly, RANZCR scores played a significant role, with RANZCR 4 (-0.003, p <0.0001) negatively affecting accuracy and RANZCR 5 (0.003, p <0.0001) having a positive effect.

Model 2: Extracted radiomic features.

The second model assessed how the reader's performance was influenced by radiomic features. The significant variables are presented in Table 3.

Table 3. Significant textural quantization

Coefficient	Estimate	P-Value
Histogram-based features		
Mean	0.0427	0.03
25th percentile	0.0427	0.04

Second order textural Features (Haralick)		
Sum average	-0.195	<0.0001
Information measure of correlation 1	-0.756	<0.0001
Difference variance	0.414	<0.0001
Contrast	0.731	<0.0001
Maximum probability	0.366	<0.0001
Sum of squares	0.122	<0.0001
Difference entropy	0.117	<0.0001
Sum variance	0.449	0.03

Histogram-Based Features

Among the histogram-based features, the "Mean" intensity value led to a positive coefficient of 0.0427 (P-value: 0.03), signifying a positive association between higher mean values of the histogram and improved diagnostic accuracy. Similarly, the "25^{th} percentile" feature also showed a positive coefficient of 0.0427 (P-value: 0.04), implying that higher values of the 25th percentile of the histogram correlated with enhanced diagnostic accuracy.

Haralick Features

In Haralick features, the "Sum Average" feature displayed a negative coefficient of -0.195 (P-value: <0.0001), indicating an association between higher values of this feature and decreased diagnostic accuracy. Similarly, the "Information Measure of Correlation 1 (IMC1)" exhibited a negative coefficient of -0.756 (P-value: <0.0001), suggesting that higher values were associated with reduced accuracy. Conversely, "Difference Variance" showed a positive coefficient of 0.414 (P-value: <0.0001), implying that higher values were correlated with improved diagnostic accuracy.

Additionally, Haralick's "Contrast" feature demonstrated a negative coefficient of -0.731 (P-value: <0.0001), suggesting that higher values were negatively associated with diagnostic accuracy. "Maximum Probability" exhibited a positive coefficient of 0.366 (P-value: <0.0001), indicating that higher maximum probability values were linked to enhanced accuracy. "Sum of Squares: Variance" displayed a positive coefficient of 0.122 (P-value: <0.0001), suggesting that higher values of this feature were positively associated with better diagnostic accuracy. "Difference Entropy" had a positive coefficient of 0.117 (P-value: <0.0001), implying that higher values were linked to improved accuracy. Similarly, "Sum Variance" exhibited a positive coefficient of 0.449 (P-value: 0.03), indicating that higher values were associated with better diagnostic accuracy.

Model 3 Interactions with cancer type.

The third model was used to examine the interactions between significant cancer characteristics and significant textural quantised features. However, the model's findings revealed no significant interactions between cancer characteristics and radiomic features.

Discussion

This study investigated the association between reader performance in cancer detection and cancer characteristics, as well as local textural features in mammography interpretation. The findings indicate that a mass, which is characterised as stellate and stellate with non-specific densities, had the greatest impact on reader performance in mammography interpretation. However, breast density, lesion size and the relative location of the pathology on the breast did not significantly affect reader performance. Previous studies have supported these findings, indicating that stellate and non-specific densities often have a subtle appearance that can mask signs of malignancy, leading readers to mis-interpret these cases positives (16-18).

Based on a previous study (19), when non-specific densities are present, readers may identify irregularities, but they often interpret them as benign, leading to fewer recalls. Although previous research has indicated that readers are better at recognizing certain cancer types, such as discrete masses and calcifications, compared to stellate masses and NSD, these differences were not significant in our study (2, 20).

Our study's findings highlighted the importance of lesion size and location in detecting cancer as that lesion size and site were found to be significant, which aligns with previous studies as it was expected that larger lesions are typically easier to detect (13.14). Based on our study, considering the coefficients of the first model, relying on lesion size to estimate the case difficulty for the educational test set curation appears to be a reasonable approach and this parameter is more important than any other factor for estimating the case difficulty. Additionally, Trieu et al. (2021) found that cancers located in the lower inner, upper outer, or mixed regions of the breast were more easily detected than in other areas (14). However, there are currently no known studies examining the relationship between false-positive rates in lesions locations and whether they contribute to increased detection rates or thresholds, given that current incorrect points and features marking are not saved in DICOM images. Further investigation is needed in this area.

The positive association between mean grey level intensity and the case difficulty suggests that lesions in mammographic images have higher brightness or intensity, making them less subtle

and more prominent. Consequently, radiologists may find it easier to identify these brighter lesions, resulting in a positive association between mean intensity and average accuracy. Similarly, the positive association with the 25th percentile of intensity suggests that the lower end of intensity values in mammograms is higher, indicating a greater prevalence of brighter areas in the image. This increased brightness can make lesions more conspicuous.

Among the normalised Haralick feature, the coefficient for the IMC1 exhibited the highest absolute value. A higher IMC1 value is typically associated with more structured and uniform textures in an image. Cancerous lesions often disrupt the ordered breast parenchymal pattern, resulting in a reduced IMC1 value. Our data suggests that when a lesion exhibits some level of structured pattern (which resembles non-cancerous breast parenchyma rather than a typical cancerous areas), correctly categorizing it as a cancerous lesion becomes more challenging. Therefore, in our study, a higher IMC1 hinders differentiation and raise difficulty. The normalised Haralick's contrast also shows strong association. As anticipated, a lower contrast value makes subtle texture variations harder to discern, increasing the perceived case difficulty.

Two other important texture features are Difference variance and Maximum Probability. Difference variance measures the local variations in pixel intensity differences within the lesion area. Higher Difference variance values indicate greater irregularities and fluctuations in pixel intensity differences. Based on our dataset, when lesions are more difficult to detect, there tends to be less variation and irregularity in pixel intensity differences, as measured by the Difference variance. It is reasonable to assume lesions that are challenging to identify may have characteristics that blend more seamlessly with the surrounding breast tissue. This results in less pronounced variations in pixel intensity differences, which can make these lesions less conspicuous and more difficult to distinguish from normal tissue.

Maximum Probability is a texture feature that represents the probability of the most frequently occurring gray-level pair in the co-occurrence matrix. It tends to be higher when a specific gray-level pair is dominant in the texture. Our model implies that in more difficult-to-interpret lesions, the dominance of a specific gray-level pair is less common in the texture. Hence, in challenging cases where lesions are less distinct, there may be a more diverse range of gray-level pairs in the texture, and a single pair does not dominate. This diversity can contribute to the difficulty of lesion detection, as the texture lacks a clear, dominant pattern that stands out.

The results of this study do not provide evidence that breast density has impacted reader performance in mammography interpretation. Previous studies on the impact of breast density on reader performance have produced mixed results. Older studies using film-based mammography have reported that denser breasts have a negative effect on reader performance (21, 22). However, more recent studies using more modern digital mammography suggest that non-dense breasts may have a greater negative impact on reader performance compared to dense breasts due to the use of post processing tools for image manipulation (23-25). Nevertheless, in our current study, breast density was found to have no significant effect on reader performance.

Previous studies on textural quantization have reported mixed findings regarding the association of breast cancer risk factors with breast density. Some studies have reported it dependent on breast density (26, 27) while others independent (28) of breast density. Due to the heterogenous nature of breast tissue, it has been suggested that quantised textural features of mammography should be evaluated as an independent risk factor for breast cancer (29). Our results demonstrated that only certain types of radiomic features had a significant impact on reader performance.

Our study has clinical implications suggesting that radiologists can consider incorporating these texture features into their reporting. Highlighting the presence of dominant texture patterns (higher maximum probability) or complex textures (higher difference variance or sum of squares variance) in their reports can provide additional context for more targeted follow-up actions. While there is controversy surrounding the implementation of machine learning as a predictor of breast cancer in the clinical setting, it could be used as a preliminary categorization to integrate these texture features to assist radiologists in their work. AI can flag images with textural features for closer review, potentially reducing the risk of missed diagnoses and suggesting cases of higher risks to the reader.

The limitations of our study concern the generalizability of our findings to real-world practice highlighted by the prevalence effect. This occurs when a higher proportion of positive cases are present in a test data set compared to what would be expected in the clinical setting, which can lead to the ‘laboratory effect’s (30). However, a previous study has shown a moderate agreement between clinical and test set reader performance (31). Another potential limitation is the comparability of our user data. While most readers used the BREAST platform in the clinical setting, around one-third of interpretations were performed at either conferences or workshops. Nonetheless, reading conditions at these events were carefully controlled to be as

close to the clinical setting as possible, minimizing the impact of the reading environment on the reader.

Regarding the curation of the test sets in BREAST, it is important to acknowledge that the selection method could have influenced the complexity of the cases. The expert responsible for selecting the mammography cases for the test set took into consideration the potential learning and educational benefits for the learners and readers on the BREAST platform. Consequently, the test set might inherently present a higher level of 'difficulty' than what readers might encounter in a typical screening or diagnostic setting.

In summary, our study utilised a very large dataset to investigate the impact of textural quantization on reader performance in mammography interpretation. The results suggest that predicting reader performance based on textural quantification of cancer characteristics is a feasible method for produce educational materials, quality assurance and training. Also, textural quantification could be valuable in assisting readers in determining the likelihood of cancer presence in a mammogram, using machine learning used as a preliminary categorisation method to suggest cases of higher risk. Future research should focus on implementing machine learning algorithms that use textural quantification as a predictive tool to assist radiologists in their interpretation of mammograms. Additionally, exploring the integration of these findings into radiology practice and education can further enhance the earlier detection of breast cancer for our patients.

References

1. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol. 2000;174(6):1769-77.

2. Rawashdeh MA, Bourne RM, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses. Acad Radiol. 2013;20(5):576-80.

3. Itri JN, Patel SH. Heuristics and Cognitive Error in Medical Imaging. American Journal of Roentgenology. 2018;210(5):1097-105.

4. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics. 1973(6):610-21.

5. Galloway MM. Texture analysis using gray level run lengths. Computer graphics and image processing. 1975;4(2):172-9.

6. Yao Y, Abidi B, Doggaz N, Abidi M, editors. Evaluation of sharpness measures and search algorithms for the auto-focusing of high-magnification images. Visual Information Processing XV; 2006: SPIE.

7. Weszka JS, Dyer CR, Rosenfeld A. A comparative study of texture measures for terrain classification. IEEE transactions on Systems, Man, and Cybernetics. 1976(4):269-85.

8. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Transactions on systems, man, and Cybernetics. 1989;19(5):1264-74.

9. Wu C-M, Chen Y-C. Statistical feature matrix for texture analysis. CVGIP: Graphical Models and Image Processing. 1992;54(5):407-19.

10. Laws KI, editor Rapid texture identification. Image processing for missile guidance; 1980: Spie.

11. Wu C-M, Chen Y-C, Hsieh K-S. Texture features for classification of ultrasonic liver images. IEEE Transactions on medical imaging. 1992;11(2):141-52.

12. Fogel I, Sagi D. Gabor filters as texture discriminator. Biological cybernetics. 1989;61(2):103-13.

13. Litimco C, Villanueva M, Yecla N, Soriano M, Naval P, editors. Coral Identification Information System. 2013 IEEE International Underwater Technology Symposium (UT); 2013: IEEE.

14. Liu W, Pantoja-Galicia N, Zhang B, Kotz RM, Pennello G, Zhang H, et al. Generalized linear mixed models for multi-reader multi-case studies of diagnostic tests. Stat Methods Med Res. 2017;26(3):1373-88.

15. Jinming Z, Jennifer MC, Shona CF, Marc GW, Xihong L, Murray AM, et al. Application of linear mixed-effects model with LASSO to identify metal components associated with cardiac autonomic responses among welders: a repeated measures study. Occupational and Environmental Medicine. 2017;74(11):810.

16. Duncan KA, Needham G, Gilbert FJ, Deans HE. Incident round cancers: what lessons can we learn? Clin Radiol. 1998;53(1):29-32.
17. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. Radiographics. 2003;23(4):881-95.
18. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Brenner RJ, Sickles EA. Analysis of 172 subtle findings on prior normal mammograms in women with breast cancer detected at follow-up screening. Radiology. 2003;226(2):494-503.
19. Cherel P, Becette V, Hagay C. Stellate images: anatomic and radiologic correlations. Eur J Radiol. 2005;54(1):37-54.
20. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology. 1992;184(3):613-7.
21. Holland K, van Gils CH, Mann RM, Karssemeijer N. Quantification of masking risk in screening mammography with volumetric breast density maps. Breast Cancer Res Treat. 2017;162(3):541-8.
22. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000;92(13):1081-7.
23. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. RadioGraphics. 2015;35(2):302-15.
24. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology. 2008;246(2):376-83.
25. Al Mousa DS, Mello-Thoms C, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Mammographic density and cancer detection: does digital imaging challenge our current understanding? Acad Radiol. 2014;21(11):1377-85.
26. Huo Z, Giger ML, Olopade OI, Wolverton DE, Weber BL, Metz CE, et al. Computerized analysis of digitized mammograms of BRCA1 and BRCA2 gene mutation carriers. Radiology. 2002;225(2):519-26.
27. Li H, Giger ML, Olopade OI, Margolis A, Lan L, Chinander MR. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. Acad Radiol. 2005;12(7):863-73.
28. Nielsen M, Karemore G, Loog M, Raundahl J, Karssemeijer N, Otten JD, et al. A novel and automatic mammographic texture resemblance marker is an independent risk factor for breast cancer. Cancer Epidemiol. 2011;35(4):381-7.
29. Zheng Y, Keller BM, Ray S, Wang Y, Conant EF, Gee JC, et al. Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Med Phys. 2015;42(7):4149-60.

30. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The “Laboratory” Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. Radiology. 2008;249(1):47-53.
31. BaoLin PS, Warwick ML, Peter LK, Warren MR, Mark FM, Patrick CB, editors. Test set readings predict clinical performance to a limited extent: preliminary findings. ProcSPIE; 2013.

Chapter Six

Discussion and Conclusion

6.1 Chapter Introduction

The primary goal of this thesis was to investigate the existing approaches in educational materials and quality control test sets for mammography interpretation education in Australia. Its aim was to identify the factors related to readers and images that influence their performance in mammography interpretation and screening. To accomplish these objectives, the thesis set out to achieve the following:

1. An overview of the current literature in mammography accreditation and its determinants

2. Evaluating the current practice in evaluating mammography cases and its difficulty

3. Exploring potential reader and case characteristics impacting reader performance

4. Validating the impact case characteristics has on reader performance

This chapter serves as the discussion section of this thesis, where the key ideas and findings explored throughout this thesis are presented. The chapter is divided into four main sections each addressing specific aspects of the study:

Section 1 provides a comprehensive summary of the major findings of the thesis focusing on the objectives outlined in the previous chapters. This section specifically examines the results obtained from the original studies conducted in Chapter 4, Chapter 5, and Chapter 6. specifically in the three original studies from chapter 4,5,6.

Section 2 delves into the clinical implications and impact of the findings on the development of education in mammography. It explores how he results of the thesis can be applied in practical settings to enhance diagnostic efficacy and improve patient outcomes.

Section 3 critically evaluates the limitations of the work conducted in the thesis. It acknowledges any potential constraints that may have influenced the results and interpretations.

Section 4 concludes the discussion by exploring the future directions and implications of the findings. It identifies potential areas for further research and development, considering the gaps

and opportunities for improvement uncovered by this thesis. This section provides insights into how the findings can be expanded upon and applied in future investigations.

6.2 Section 1: Summary and Major findings of the thesis

6.2.1 Mammography case difficulty determined by experts

The initial study aimed to investigate and scrutinize the current practices of using experts to determine case difficulty in mammography. The study examined whether the expert's difficulty rating matches the actual performance of readers for each case, as well as the image characteristics used by the experts used to determine case difficulty. Additionally, the study aimed to determine the minimum number of radiologists needed to obtain an objective assessment of case difficulty through simulation, the simulation was created by randomly selecting a reader and the case difficulty is calculated based on the assessment of these selected readers. This process was repeated a thousand times to estimate the minimum number of radiologists needed. However, the data did not support the hypothesis that expert radiologists' difficulty ratings would correspond to reader performance.

The findings of the study indicated that experts mainly based their retrospective grading of mammography case difficulty on increased breast density and the presence of abnormalities. The original linear mixed modelling used to identify difficult determining characteristics can be found in Chapter 4. However, these characteristics did not serve as predictors of reader performance difficulty. Furthermore, the findings also revealed no significant association between expert-determined case difficulty and objective case difficulty, which was measured as the percentage of incorrect reports for a case. In contrast to expert opinions, readers performed slightly better in cancer-positive cases compared to normal cases.

Additionally, our results demonstrated that expert-determined difficulty did not accurately correspond to reader performance, as the performance levels of readers remained similar regardless of the difficulty level.

In summary, the initial study's findings revealed that expert radiologists' assessments of mammography case difficulty based on breast density and abnormality did not align with reader performance. The simulation-based approach provided insights into the number of readers needed to determine case difficulty objectively. The results emphasise the importance of

examining reader performance and using it as an indicator in the development of test sets where available.

A key discovery in the initial study was that the current practice of using experts to retrospectively determine a case's difficulty may not necessarily be reflective of the case's actual difficulty level as indicated from our results. The simulation provided can estimate the number of readers needed to evaluate case difficulty. Based on these findings, the next step of the study was to further explore and identify the characteristics that would have the most impact on performance to enhance the construction of test sets and develop their educational value.

6.2.2 Characteristics influencing diagnostic accuracy in mammography

The next stage of the thesis aimed to examine the influence of reader-based and image-based characteristics on reader performance in screening mammography. In contract to the initial study, the results of this stage indicated that expert-determined difficulty and the number of cases interpreted weekly were consistently significant factors for image-based and reader-based characteristics, respectively. These findings were largely expected based on anecdotal evidence and previous studies, where mammogram interpretation volume of over 1000 perform better than those reading less than 1000 per year, with varying levels of performance improvement among the previous studies(1-7). However, this study provided additional evidence to reinforce these conclusions. Furthermore, by utilizing a large sample size for both readers and cases, the researchers established a trend between reader and image characteristics and their impact on reader performance.

While breast density was not found to be a significant factor impacting reader performance, interesting observations were made regarding the readers' performance in dense and non-dense breasts for different types of cases. For normal cases, readers performed better with dense breasts compared to non-dense breasts. Conversely, for cancer cases, readers performed better in non-dense breasts compared to dense breasts. This suggests that breast density, combined with the presence of abnormalities, may have an influence on readers' abilities to interpret mammography images.

Regarding lesion characteristics, although the results were not significant, there were observed trends. Cases primarily classified as stellate and discrete masses were more likely to be correctly detected by readers compared to cases with calcifications and non-specific densities.

Interestingly, cases classified primarily as non-specific densities tended to result in inferior reader performance compared to other lesion types.

In summary, while breast density did not significantly affect reader performance, observations emerged regarding the interaction between breast density and malignancy, suggesting that the masking effect may still play a role in mammography interpretation, particularly for malignant cases. Although not found to be statistically significant, trends indicated that certain lesion characteristics influenced reader performance, with stellate and discrete masses being more easily detected compared to calcifications and non-specific densities.

With the primary reader characteristic – number of cases interpreted weekly established as the characteristic that influences the reader performance the most, further insights are required to investigate the case characteristics that influence reader performance. Therefore, the next step was to investigate how cases would affect reader performance at a parenchymal level via radiomics.

6.2.3 Radiomics and image texture's influence on reader performance

Among the various image characteristics analysed, stellate masses and stellate masses with non-specific densities had the most significant impact on reader performance. However, breast density, lesion size, and relative location of the pathology did not significantly affect reader performance. A comparison between quantised features with cancer type and non-quantised textural features revealed slight differences. While it was expected that Haralick's quantization method would be better at detecting calcifications than stellate masses, the unexpected finding was that our readers performed better at detecting stellate masses with non-specific densities than detecting spiculated masses alone. Other methods, such as Compact GLDM and Statistical Feature Matrix, were more consistent with previous findings, where reader demonstrated better detection rates for spiculated masses compared to stellate masses with non-specific densities (8).

The presence of non-specific densities can lead readers to interpret irregularities as benign, resulting in fewer recalls. In contrast, discrete masses and calcifications did not show significant differences, compared to stellate masses and non-specific densities in this study. Regarding breast density, the results did not provide evidence that breast density the results did

not provide evidence of its significant effect on reader performance. Similarly, lesion size and location were not found to be significant factors, contrary to previous research (9, 10).

Regarding textural quantization, the heterogenous nature of breast tissue suggests that quantised textural features should be investigated as an independent risk factor for breast cancer. The results of this study indicated that only Haralick's quantization methods had a significant impact on reader performance, as it is commonly adopted for characterizing tissue properties to detect heterogeneity (11).

Overall, this study provided insights into the relationship between reader performance, cancer characteristics, and local textural features in mammography interpretation. The results emphasise the impact of stellate and non-specific densities on reader performance and highlight the need for further investigation into the role of quantization methods and textural analysis in breast cancer risk assessment.

6.3 Comparison with Previous Research

In Chapter 3, our first original study aimed to evaluate the current practice in evaluating mammography cases and its difficulty. Our study has established that expert primarily used breast density as an indicator to case difficulty, specifically increasing difficulty with increasing breast density. Previous research has inconsistent results regarding the relationship between case difficulty and breast density. Earlier studies focusing on film-based mammography indicated that as breast density increased, the diagnostic effectiveness of mammograms decreased (12, 13). Our findings indicate that the experts associate increasing breast density with increased case difficulty. This observation is consistent with the findings of Ang et al., who also noted the impact of mammographic breast density on determining case difficulty in normal cases (14). However, when considering actual reader performance, breast density did not seem to have a significant influence. As discussed in Chapter 4's introduction, this discrepancy may be attributed following the limitation of following BREAST's standard of using one test set's findings, as the study followed the standard practice of analysing the results of one test set.

For the findings in the second of the original research, Chapter 4 aimed to explore the potential reader and case characteristics impacting reader performance. The findings of the study were similar to previous studies, as they have consistently demonstrated that readers who interpret

more than 1000 mammograms annually tend to outperform those who interpret fewer than 1000 per year. The extent of performance improvement has varied across different studies(1-7). Our findings regarding reading volume closely align with Kan et al.'s study, they found that the improvement in reader performance was proportional to the number of mammograms interpreted, up to a certain threshold. However, beyond the threshold of 2,500 annual screening mammograms interpreted, the improvement in reader performance became negligible and statistically insignificant (2). This suggests that increasing the interpretation volume can enhance reader performance, but there is a limit beyond which additional mammograms read do not significantly contribute to improved performance.

In terms of expert determined difficulty, two studies found that reader performance in mammography interpretation and expert-determined difficulty were associated. Specifically, as expert-predetermined difficulty increases, the reader's false negative errors increase as well, but false positive errors were not associated (15, 16). In contrast in our finding's expert-determined difficulty impacted in the reader's false positive errors in the normal images and mixed model, but not exclusively in the cancer model for false negative errors.

For the final goal of this thesis, the impact case characteristics has on reader performance was evaluated, compared to previous studies have suggested that readers demonstrate better recognition of certain cancer types, such as discrete masses and calcifications, compared to stellate masses and NSD(9, 10). However, in the current study, these differences in recognition were not statistically significant. Contrary to previous studies emphasizing the significance of lesion size and location in cancer detection, our study did not find evidence supporting the importance of these factors. Specifically, our findings did not indicate a significant relationship between lesion size and cancer detection, which contrasts with the expectation that larger lesions are generally easier to detect based on previous research (9, 10).

Previous studies examining textural quantization has inconsistent findings regarding the association between breast cancer risk factors and breast density. Some studies have reported risk factors associated to breast density (17, 18), while others have suggested independent risk factors from breast density (19). Given the heterogeneous nature of breast tissue, it has been proposed that quantised textural features in mammography should be assessed as independent risk factors for breast cancer (20) Our study demonstrated that Haralick's quantization methods had a primary impact on reader performance. This finding suggests that while quantization

methods can be employed by machine learning algorithms to evaluate breast cancer risk factors, further investigation is needed to understand the correlation between textural analysis and human readers.

A consistent finding across the project was that breast density's impact had no significant impact on reader's performance in mammography interpretation regardless of dense or non-dense breast. Previous research also yielded mixed findings, as older studies found that using film-based mammography have reported that denser breasts have a negative effect on reader performance due to the masking effect (12, 13). However, more recent studies using more modern digital mammography suggest that non-dense breasts may have a greater negative impact on reader performance compared to dense breasts due to the use of post processing tools for image manipulation (21, 22).

6.4 Clinical Implications of the findings of this thesis

The findings of this thesis have several clinical implications. The development of test sets for educational and quality control purposes should incorporate objective reader performance rather than relying solely on expert opinion to determine case difficulty. Ideally, around 10 readers can provide a comprehensive assessment of case difficulty, serving as a guideline for future test set development.

Furthermore, there is a need to enhance readers' ability to confidently recognize normal cases in mammograms. The first study revealed that readers performed better in identifying abnormalities compared to identifying normal cases. Improving the recognition of normal cases can contribute to more accurate interpretations, reduce false positives, and unnecessary recalls.

The annual reading volume needed to maintain certification in mammography interpretation are valuable in evaluating a reader's ability. However, there is a variability in reading volume requirements across different countries. Standardizing the reading volume requirements would provide a more consistent approach to assessing reader competence (23).

Screening for non-specific densities and architectural distortions, which are subtle findings in mammography, should be emphasised. Digital breast tomosynthesis has shown promise in detecting these abnormalities with lower interobserver variability and higher specificity compared to mammography. Until it is widely implemented, alternative methods such as

computer-aided detection and radiomics can be explored to improve the diagnosis of non-specific densities.

Regarding textural quantization, it may not currently reliably predict a reader's performance based on cancer characteristics in mammograms. However, it can be used as a supplementary tool to assist readers in assessing the likelihood of cancer presence in a case. Machine learning approaches, while controversial in their implementation for breast cancer prediction, can serve as preliminary categorization method to suggest higher-risk cases to the reader. However, it is important to remember that technological improvement is not a substitute for improving training and education as these improvements do not help in resolving errors associated with false positive and false negative diagnostics (9, 10).

The clinical implications of this study highlight the importance of incorporating objective reader performance in test set development, improving recognition of normal cases, standardizing reading volume requirements, screening for non-specific densities and architectural distortions, and utilizing textural quantization and machine learning as supplementary tools in mammography interpretation. These findings can contribute to enhancing reader performance and improving the accuracy of breast cancer detection.

6.5 Limitations of the thesis

6.5.1 Prevalence effect

The primary limitation of this thesis is that it primarily uses retrospective data from an education and training platform, which introduces a potential prevalence effect as a significant factor in the reader’s performance. The prevalence effect arises due to the higher proportion of positive cases in test sets compared to what is typically encountered in the clinical setting. In the case of the BREAST test set used in this study, it consisted of a total of 60 cases per test set, with 40 normal cases and 20 cancer cases. This may have influenced our results and created a laboratory effect, as readers tended to perform better in cancer-positive cases compared to normal cases. However, previous research has shown moderate agreement between clinical performance and test set interpretation, providing some reassurance (24).

6.5.2 Single expert used to determine case difficulty

Furthermore, the first study's reliance on a single expert to determine the test set difficulty represents another potential limitation. However, it is important to note that this was not a flaw in the study's design but rather the current practice used in the BREAST platform. Our findings highlight the need for future studies to explore the effects of involving multiple expert readers in grading the difficulty of test sets. Additionally, our simulation results indicate that involving readers in curating the test sets to determine the difficulty of each image set could be valuable in developing future test sets.

In summary, potential limitations include the prevalence effect, comparability of user data from different settings, the reliance on a single expert for determining test set difficulty, and the discrepancy between the number of readers and experts involved.

6.6 Recommendations for future work

The findings and investigations presented in this thesis have primarily been based on BREAST's efforts to develop a training platform aimed at improving the diagnostic efficacy of BreastScreen's readers in screening mammography. The thesis has investigated several key areas:

1. Examining the characteristics experts use to determine the retrospective case difficulty in the mammography cases in the BREAST platform and assessing whether the platform's reader's performance accurately reflects this difficulty level.
2. Evaluating which reader and case characteristics would most significantly impact the reader's ability to interpret mammography accurately.
3. Investigating the impact did various quantitative textures of cancer lesions on reader performance in identifying lesions in mammograms.

6.6.1 Potential directions for future research

These areas of study have provided valuable insights, and for future research in this field, further exploration and expansion is recommended upon these topics to continue enhancing the field on mammography interpretation. findings from the results of the thesis have highlighted several potential directions that future research derived from this thesis to address the

limitations mentioned in the previous section and further advance the field. The potential directions future research include:

1. **Test set development**: Future studies should explore alternative methods for determining the difficulty of test sets, such as involving multiple expert readers to define a case's difficulty and incorporating previous reader performance data to update the ratings of existing test sets more accurately. Additionally, the findings in Chapter 5 have identified the potential textures and patterns that readers tend to perform better against. These findings can be incorporated to the development of more comprehensive and representative test sets for education, training, and quality assurance purposes.

2. **Investigating the association between expert-determined difficulty and reader accuracy:** While this thesis has established an association between expert-determined difficulty and reader accuracy, the underlying reason remains to be confirmed. Experts primarily used breast density as a determining factor for the case's difficulty, but our findings suggest that although the reader's performance in each case reflects the expert-determined difficulty, breast density was not an associated factor. Further investigation into the relationship between the case difficulty and reader performance is needed to understand the underlying reasons. Discovering this association can lead to the curation of training sets of varying difficulty for improved training and quality assurance for mammography readers.

3. **Normal case recognition**: Further research should focus on enhancing readers' ability to confidently recognize normal cases in mammography. This can contribute to reducing false positives and unnecessary recalls, improving the accuracy of mammography interpretations.

4. **Standardization of Reading Volume Requirements;** Efforts should be made to attempt the standardization of the reading volume requirements for maintaining certification in mammography interpretation across different countries where circumstances allow for the maintenance of services. This approach can help to ensure a more consistent approach to evaluate reader competence and maintain a high level of quality in mammography screening.

5. **Refinement of Textural Quantization and Machine Learning:** Continued exploration and refinement of textural quantization methods and machine learning approaches are critical. This includes evaluating their effectiveness in predicting reader performance based on cancer characteristics and utilizing them as supplementary tools in mammography

interpretation. Ethical considerations surrounding the implications of machine learning for breast cancer diagnosis and prediction should also be carefully addressed.

6. **Application of textural quantization in digital breast screening:** Currently, BreastScreen Australia uses full-field digital mammography as the gold standard for the screening program for the early detection of breast cancer. The application of radiomics and other textural quantization methods can be employed to this imaging modality to investigate how readers perform in this modality and whether specific parenchyma would impact reader performance in this modality.

7. **Texture Features and Histopathological Correlation**: The final study in this thesis primarily focused on evaluating the local effects of lesion textures on reader performance. However, this can be expanded to explore how normal anatomical structures of the breast potentially impact reader performance. For example, the thesis found that readers better identified lesions that are in the lower regions of the breast. Further investigation into how the textural characteristics of normal breast tissue would affect reader performance and whether the normal anatomical structure interacts with the cancer lesion's structure would be valuable.

6.6 Conclusion

This project has undertaken a comprehensive review of the current literature on screening mammography requirements and the development of artificial intelligence in mammography interpretation. Throughout this research, it has identified existing issues in mammography interpretation standards and requirements, as well as crucial knowledge gaps in the development of mammography test sets. Through these investigations, valuable insights have been gained into the factors influencing expert-determined difficulty and reader performance in mammography.

One of the key recommendations arising from this study is the need where available to incorporate of objective reader performance metrics in the development of test sets designed for educational and quality control purposes. By considering reader performance as a fundamental factor in test set design, these test sets can be better tailored to assess and monitor reader performance more effectively.

Furthermore, future research in this field should place a strong emphasis on refining textural quantization methods and exploring the vast potential of machine learning approaches. These

technological advancements hold the promise of significantly improving the accuracy of mammography interpretation and enhance the overall performance of readers. Embracing these innovations is essential to meeting the ever-increasing demands for precision in breast cancer detection.

The findings of this thesis have not only shed light on the current state of mammography interpretation but have also paved the way for promising future directions in this vital area of healthcare. These directions encompass the development of more sophisticated test sets and the ongoing refinement of textural quantization, radiomics and machine leaning techniques. By addressing these critical areas of study, we can assist radiologists to further enhance their performance and ultimately advance the outcomes for our patients participating in mammography screening programs. This thesis serves as a steppingstone recommending continued exploration and innovation in this field to help bring about transformative improvements in breast cancer detection and patient care.

Reference

1. Hoff SR, Myklebust T, Lee CI, Hofvind S. Influence of Mammography Volume on Radiologists' Performance: Results from BreastScreen Norway. Radiology. 2019;292(2):289-96.

2. Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology. 2000;215(2):563-7.

3. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology. 2002;224(3):861-9.

4. Buist DS, Anderson ML, Smith RA, Carney PA, Miglioretti DL, Monsees BS, et al. Effect of radiologists' diagnostic work-up volume on interpretive performance. Radiology. 2014;273(2):351-64.

5. Kim SH, Lee EH, Jun JK, Kim YM, Chang YW, Lee JH, et al. Interpretive Performance and Inter-Observer Agreement on Digital Mammography Test Sets. Korean J Radiol. 2019;20(2):218-24.

6. Buist DSM, Anderson ML, Haneuse SJPA, Sickles EA, Smith RA, Carney PA, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology. 2011;259(1):72-84.

7. Haneuse S, Buist DSM, Miglioretti DL, Anderson ML, Carney PA, Onega T, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. Radiology. 2012;262(1):69-79.

8. Mohd Norsuddin N, Mello-Thoms C, Reed W, Rickard M, Lewis S. An investigation into the mammographic appearances of missed breast cancers when recall rates are reduced. The British journal of radiology. 2017;90(1076):20170048-.

9. Rawashdeh MA, Bourne RM, Ryan EA, Lee WB, Pietrzyk MW, Reed WM, et al. Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses. Acad Radiol. 2013;20(5):576-80.

10. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology. 1992;184(3):613-7.

11. Kang H, Kim EE, Shokouhi S, Tokita K, Shin HW. Texture Analysis of F-18 Fluciclovine PET/CT to Predict Biochemically Recurrent Prostate Cancer: Initial Results. Tomography. 2020;6(3):301-7.

12. Holland K, van Gils CH, Mann RM, Karssemeijer N. Quantification of masking risk in

screening mammography with volumetric breast density maps. Breast cancer research and treatment. 2017;162(3):541-8.

13. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. Journal of the National Cancer Institute. 2000;92(13):1081-7.

14. Ang ZZ, Rawashdeh MA, Heard R, Brennan PC, Lee W, Lewis SJ. Classification of normal screening mammograms is strongly influenced by perceived mammographic breast density. Journal of Medical Imaging and Radiation Oncology. 2017
61(4):461-9.

15. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA. Radiology resident mammography training: interpretation difficulty and error-making patterns. Acad Radiol. 2014;21(7):888-92.

16. Grimm LJ, Zhang J, Lo JY, Johnson KS, Ghate SV, Walsh R, et al. Radiology Trainee Performance in Digital Breast Tomosynthesis: Relationship Between Difficulty and Error-Making Patterns. Journal of the American College of Radiology : JACR. 2016;13(2):198-202.

17. Huo Z, Giger ML, Olopade OI, Wolverton DE, Weber BL, Metz CE, et al. Computerized analysis of digitized mammograms of BRCA1 and BRCA2 gene mutation carriers. Radiology. 2002;225(2):519-26.

18. Li H, Giger ML, Olopade OI, Margolis A, Lan L, Chinander MR. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. Acad Radiol. 2005;12(7):863-73.

19. Nielsen M, Karemore G, Loog M, Raundahl J, Karssemeijer N, Otten JD, et al. A novel and automatic mammographic texture resemblance marker is an independent risk factor for breast cancer. Cancer Epidemiol. 2011;35(4):381-7.

20. Zheng Y, Keller BM, Ray S, Wang Y, Conant EF, Gee JC, et al. Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Med Phys. 2015;42(7):4149-60.

21. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. RadioGraphics. 2015;35(2):302-15.

22. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology. 2008;246(2):376-83.

23. Wong DJ, Gandomkar Z, Lewis S, Reed W, Suleiman Ma, Siviengphanom S, et al. Do

Reader Characteristics Affect Diagnostic Efficacy in Screening Mammography? A Systematic Review. Clinical Breast Cancer. 2023;23(3):e56-e67.
24. BaoLin PS, Warwick ML, Peter LK, Warren MR, Mark FM, Patrick CB, editors. Test set readings predict clinical performance to a limited extent: preliminary findings. ProcSPIE; 2013.

Appendix

Appendix A. Ethics Exemption Notice

THE UNIVERSITY OF SYDNEY

Research Integrity & Ethics Administration
HUMAN RESEARCH ETHICS COMMITTEE

Friday, March 19, 2021

Dr Warren Reed
Sydney School of Health Sciences Discipline of Medical Imaging Science
warren.reed@sydney.edu.au

Dear Dr Reed,

RE: The impact of breast density in test set difficulty for online mammographic self-assessment tools

Many thanks for submitting the above titled ethics application to The University of Sydney HREC for consideration.

The NHMRC National Statement on Ethical Conduct in Human Research (2007) outlines the circumstances in which research can be exempted from review:

5.1.22 Institutions may choose to exempt from ethical review research that:

(a) is negligible risk research (as defined in paragraph 2.1.7); and

(b) involves the use of existing collections of data or records that contain only non-identifiable data about human beings.

5.1.23 Institutions must recognise that in deciding to exempt research from ethical review, they are determining that the research meets the requirements of this National Statement and is ethically acceptable.

Based on email communication with the ethics office dated 09/03/2021, it has been determined that the research proposal submitted 25/02/2021 is exempt from ethical review.

Should any future work not comply with any of the above criteria, the project must be submitted for ethical review prior to commencing any research activity. Please note that retrospective ethical approval of research cannot be granted by the HREC.

Kind Regards,

EMILY OURO
Ethics Officer | Research Integrity and Ethics Administration
Level 3, F23 Administration Building
The University of Sydney | NSW | 2008
T +61 2 9036 9161 | **E** human.ethics@sydney.edu.au | **W** http://sydney.edu.au/ethics

Appendix B. BAMC application for BREAST Data

APPLICATION FOR ACCESS TO RESOURCE
Submit completed forms to breastaustralia@sydney.edu.au

BREAST

1. PROJECT TITLE					
Identifying case characteristics impacting expert opinion and reader performance					
2. INVESTIGATORS' DETAILS					
Chief Investigator's Name (full name and title)	Warren Reed				
Institution	University of Sydney				
Applicant's Name (full name and title)	Dennis Jay Wong				
Institution	University of Sydney				
Postal Address	Susan Wakil Health Building, The University of Sydney				
City/Suburb	Camperdown	State	NSW	Postcode	2000
Phone		Mobile	0481799708		
Email	Dwon6926@uni.sydney.edu.au				

3. PROJECT INFORMATION

3.1 Summary of project in lay language (maximum 200 words)

To goal of this project is to identify the features on a mammogram experts use to determine case difficulty and compare it to reader performance.

3.2 Please provide a brief research/project plan (*maximum length 3 pages*)

3.2.1 Aim(s)

To

1. To identify anatomical features/landmarks or regions of interest experts use to determine difficulty of a mammography case.
2. To identify regions readers marked on false positive cases.
3. Comparing the results from (1) and (2) to identify potential agreements and disparities between experts and readers.
4. Based on the previous aims, identify features on a mammogram that determines case difficulty

3.2.2 Project methodology

Reader's heatmap/hotspot

First, to identify the which regions readers marked in both true positive and false positive images, the co-ordinates the readers marked will be compiled together to generate a heatmap to identify the common/unique regions readers mark on the test set item. This can be used to identify which areas were commonly marked as false positives in negative images to identify which anatomical region(s) of the mammogram was difficulty for the readers AND/OR identify regions the lesions are in that readers frequently marked correctly as making cases easy.

Expert opinion

Using the meta data such as breast density and lesion characteristics such as size of lesion, site of lesion, and lesion type; to perform a statistical analysis to identify which of these aspects' experts factor in to determine case difficulty.

To identify the disparity between experts and readers, the reader heatmap will compared against the expert factors to see if they agree or disagree in what define difficulty in each case item.

3.2.3 Number of readers from whom data are requested *and/or* number and type of cases requested *and/or* equipment for which access is requested.
All cases from all Australian test sets
Readers who work in Australia
Readers defined as radiologists and trainees/registrars

3.3 Planned start and end date of the project	Start date:07/6/2021		End date:	31/8/2021

3.4 Please list all other researchers collaborating on this project, including their name, title, appointment and institution	***Name (full name and title)***	***Department***	***Institution***
	1. Dennis Jay Wong	*Medical radiation Sciences*	*University of Sydney*
	2. Ziba Gandomkar	*Ageing Work and Health Unit*	*University of Sydney*
	3.		
	4.		
	5.		

4. ETHICS

4.1 Do you have Human Research Ethics Committee (HREC) approval to conduct this research project?	Yes ☐ No ☐ Pending ☐ Not applicable ☐ **If No**, we advise you to seek ethical approval. Access to any resource **will not** be facilitated until proof of ethical approval is submitted to us. If **Not applicable**, please specify why.

5. RESOURCES REQUIRED

5.1 What BREAST resources do you require for your project? [Specific details should be discussed with BREAST project staff at the EOI stage]	**Please tick all that apply:** **A. READER DATA** ➢ **NUMBER OF PARTICIPANTS:**

	➢ **READER TYPE** • *General radiologist* ☐ • *BreastScreen radiologist* ☐ • *Registrar / Fellow* ☐ • *Other reader* ☐ *Please specify:* ➢ **QUESTIONNAIRE DATA** ☐ ➢ **SCORES FROM TEST SET** ☐ ➢ **RAW DATA FROM TEST SET** ☐ ➢ **ADDITIONAL INFORMATION FROM PARTICIPANTS** ☐ * *[E.g. If you need to contact participants to complete an additional survey or for clarification. If Yes to 'additional information' required, please provide details at 6.2.] **B. IMAGES** ➢ **IMAGES IN EXISTING TEST SET(S)** ☐ • *Name(s) of test set(s):* *Canberra, Darwin, Gold Coast, Leura, Melbourne, Northern Territories, Queenstown, Sydney, Western Australia* • *File type (DICOM, jpeg):* *Both DICOM and JPEG* ➢ **IMAGES NOT IN ANY EXISTING TEST SET** ☐ • *Number of images:* • *Case details:* • *File type (DICOM, jpeg):* • *Specific criteria:* **C. SOFTWARE AND ONLINE TEST SETS** ☐ • *Name(s) of test set(s):* **D. MAMMOGRPAHY PACS WORKSTATIONS** ☐ • *Number of workstations:* • *Dates and times required:*
5.2 Any additional questionnaires or surveys intended for distribution to participants by BREAST MUST be supplied for review by the Management team	None.

6. PROJECT OUTPUTS
Please list all the proposed project outputs or research outputs. Journal publication, conference proceeding.

7. CERTIFICATION BY APPLICANT (STUDENT / RESEARCHER)
In signing this page, you certify that, - **All details given in this application are correct, and** - **You agree to carry out the project according to the terms and conditions as described by the BREAST Access Policy**

Appendix C. BREAST Data Application Approval

From: Phuong Trieu
To: Dennis Jay Wong
Cc: BreastScreen Reader Assessment Strategy
Subject: RE: BAMC application of BREAST data
Date: Wednesday, 16 June 2021 9:37:01 AM
Attachments: Tes set details_Dennis.xlsx

Dear Dennis,

We are delighted to let you know that your application for extra data has been approved. Please find in the attachment for the download links of Dicom and Jpeg images (the links will be expired on 30-Jun-21). The folder IDs are also included in the file for your convenience to find out the name of folders containing the images that you look for.

Good lucks to your study.

Kind regards,
Yun

From: Phuong Trieu
Sent: Tuesday, June 8, 2021 9:08 AM
To: Dennis Jay Wong <dwon6926@uni.sydney.edu.au>
Cc: BreastScreen Reader Assessment Strategy <breastaustralia@sydney.edu.au>
Subject: RE: BAMC application of BREAST data

Hello Dennis,

Thank you. We will process your application and let you know soon.

Kind regards,
Yun

From: Dennis Jay Wong <dwon6926@uni.sydney.edu.au>
Sent: Tuesday, June 8, 2021 11:43 AM
To: Phuong Trieu <phuong.trieu@sydney.edu.au>
Cc: BreastScreen Reader Assessment Strategy <breastaustralia@sydney.edu.au>
Subject: RE: BAMC application of BREAST data

Hello Yun,

Yes, I would like to use the DICOM/JPEG from the test sets of the first application e.g. Western Australia, Leura etc.

Kind regards,

Dennis

From: Phuong Trieu <phuong.trieu@sydney.edu.au>
Sent: Tuesday, 8 June 2021 9:45 AM

Appendix D. RANZCR 2021 Presentation

Sections

- Topic
- Purpose
- Methodology
- Results
- Conclusion

- Include buzzwords – implication of AI

insert logo of organisation here

The Royal Australian and New Zealand College of Radiologists

Topic

- Assessing the diagnostic difficulty of mammographic cases rated by the expert versus reality

The Royal Australian and New Zealand College of Radiologists

Purpose

To investigate the relationship between the subjective measure of case difficulty based on expert's prospective ratings and the objective difficulty based on retrospective actual reader performance

The Royal Australian and New Zealand College of Radiologists

Methodology

- A test set from BreastScreen Reader Assessment Strategy (BREAST) was used
- The case difficulty was determined prospectively by two expert radiologists, the range is defined from 1-3 (easy, intermediate, and difficult, respectively)
- The reader performance of 184 radiologists was used to measure each case's objective difficulty compared to the subjective difficulty determined by two expert radiologists
- For the statistical analysis, the Kruskal -Wallis H test was used to determine the significance between the groups of independent variables
- Furthermore, linear modelling with interaction has been applied to relate the objective case difficulty with density, subjective case difficulty and case type

The Royal Australian and New Zealand College of Radiologists

Results

- For the Kruskal Wallis test, the significance of the reader's accuracy level between the difficulty variables (easy, intermediate, and difficult) and the case's breast density showed a lack of significance
- Comparing the reader performance between normal and abnormal cases, suggested that reader's overall accuracy level for abnormal cases was higher than normal cases, indicating that readers have better specificity than sensitivity
- Linear modelling suggests that the only significant p -value interactive factor related to objective case difficulty was when readers comparing normal to abnormal cases
- The results indicate that when experts assigned difficulty labels to a case, there was a drop of 32.85% in performance only seen in the normal group. (p-value – 0.047).

The Royal Australian and New Zealand College of Radiologists

Conclusion

- The subjective expert ratings of case difficulty was not predictive of actual reader performance which highlights the importance of collecting reader performance data as an objective measure of actual case difficulty for educational and research purposes.

The Royal Australian and New Zealand College of Radiologists

Appendix E. RANZCR 2022 Presentation

Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis

Dennis Wong, Grad Cert (CE), MDR, BMSc
Ziba Gandomkar, PhD, MSc, BSc
Warren Reed PhD, Grad Cert (T&L), BSc
Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney

Introduction

- Annual reading volumes are the most common reader-based characteristic (1-3)
- Reading volume alone is insufficient as a sole measure to evaluate reader performance (4)
- Previous studies have reported mixed findings regarding the relationship between annual reading volume and reader performance (5-6)

Purpose of this study

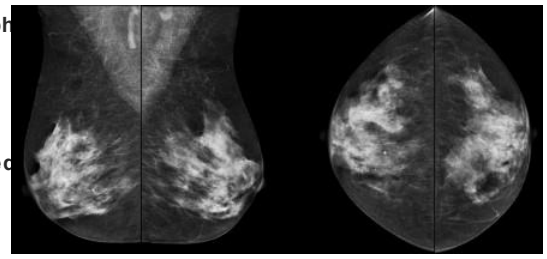
- Using a multi-case multi-reader analysis framework can further increase the power of these studies.
- This multivariate analysis used several reader factors including image-based characteristics to further interrogate the underlying primary determinants of reader performance
- The aim of this study was to conduct a multivariate investigation of reader-based and image-based characteristics in reader performance in mammography interpretation.

Methodology – Test sets

A total of six test sets with a total of 60 mammograph cases each from BREAST
Images were capture using GE machines and were from NSW BreastScreen
Each test sets consisted of 20 cancer-positive cases and 40 normal cases.
Breast density was then grouped together classified either "Dense" or "non-dense" breasts.
For abnormal cases, the pathology characteristics were included and consisted of

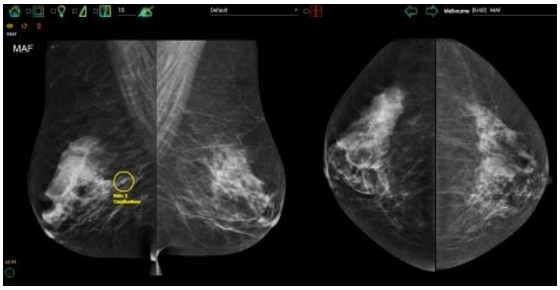
- lesion side
- lesion type
- lesion length

Both normal and abnormal cases had a difficulty rating determined by an expert



Methodology – Reader data

- The reader could mark lesions anywhere on the image
- A score classification between 35 was used to determine increasing confidence
- Only the reader's first attempt at the test set was used.
- Readers were not aware of the prevalence and types of cancers included in the test sets.



Methodology– Statistical analysis

For the statistical analysis, Linear mixed modelling was selected to evaluate how the fixed effects such as a reader characteristic affects the reader's performance

To account for multiple readers and cases, these two factors were classified as random effects in the linear mixed model.

The modelling was done in the R statistical programming environment (version 4.1.2) and used the lme4 and lmerTest packages.

The linear mixed models were divided from into three categories (1. case-based characteristics,2. reader-based characteristics,and 3. mixed-characteristics)and explores normal and cancer cases separately.

Results– LMM Case-based characteristics

Significant findings

- For both normal and cancer cases, reader performance decreases as expert-determined difficulty increases (P <0.05).

Non-significant findings

- Readers showed better performance in non-dense cases, compared to the dense cases
- Lesions classified as stellate and discrete mass demonstrated that readers were more likely to mark these cases correctly.

Results– LMM Reader-based characteristics

- **The most significant reader characteristic is the number cases interpreted per week, excluding 15-200 per week.**
- **Mean estimate decreases with increasing number of mammograms interpreted per week.**
- **Conversely, for the cancer case's linear mixed model, the number of mammograms interpreted per week showed a similar trend**
- **However, the significance was not consistent.**

Results – LMM Combined model

- The third model combined the characteristics found in the previous models
- For reader-based characteristics, mammograms interpreted per week remains significant
- For case-based characteristics, case difficulty determined by experts remains significant
- These findings were similar in both normal and cancer cases

Discussion

Our results show that expert predetermined difficulty and the number of cases interpreted weekly were the consistently significant factors.

The purpose of using a linear mixed model was to find the relationship between multiple variables with coefficients.

This novel methods provides more power and the ability to observer smaller effects

Previous studies have found that readers with an annual mammogram interpretation volume of over 1000 perform better than those reading less than 1000 per year [6,7]

Notably, our findings on reading volume correlate closely with Kan's study (9):
- Cancer detection rate improves proportionally with the number of mammograms interpreted,
- Up to a certain upper threshold

This suggests an upper limit to which readers do not significantly improve in performance

Discussion– Clinical Implications

Our findings reinforce that using annual reading volumes required to maintain certification in mammography interpretation is a valuable method to evaluate a reader's ability.
However, one problem with the current implementation of the mammography's guidelines (10,11)
Another clinical implication to consider is screening for non-specific densities and architectural distortions (12).
Compared to calcifications and discrete masses, non-specific densities are often more difficult to diagnose because of their variability in appearance (13).

Conclusion

Our findings show that expert-determined difficulty and cases interpreted weekly were consistently significant factors impacting reader performance
For case characteristics, increasing the case difficulty determined by experts reduces the reader's performance

For reader characteristics, increasing the number of cases interpreted weekly improves the reader performance up to a certain threshold.
Our results show that linear mixed modelling provides extra insights into the impact of reader- and image-based factors on reader performance in mammography interpretation.

References

1.Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings. Radiology. 2013;269(1):61-7.

2.Reed WM, Lee WB, Cawson JN, Brennan PC. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. Acad Radiol. 2010;17(11):1409-13.

3.Suleiman WI, Lewis SJ, Georgian-Smith D, Evanoff MG, McEntee MF. Number of mammography cases read per year is a strong predictor of sensitivity. J Med Imaging (Bellingham). 2014;1(1):015503.

4.Gandomkar Z, Lewis SJ, Li T, Ekpo EU, Brennan PC. A machine-learning model based on readers' characteristics to predict their performance in reading screening mammograms. Breast Cancer. 2022.

5.Hoff SR, Myklebust TÅ, Lee CI, Hofvind S. Influence of mammography volume on radiologists' performance: Results from breast screen Norway. Radiology. 2019;292(2):28996.

6.Théberge I, Hébert-Croteau N, Langlois A, Major D, Brisson J. Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. CMAJ. 2005;172(2):1959.

7.Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology. 2002;224(3):861-9.

8.Buist DS, Anderson ML, Smith RA, Carney PA, Miglioretti DL, Monsees BS, et al. Effect of radiologists' diagnostic work-up volume on interpretive performance. Radiology. 2014;273(2):351-64.

9.Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology. 2000;215(2):563-7.

10.Smith-Bindman R, Miglioretti DL, Rosenberg R, Reid RJ, Taplin SH, Geller BM, et al. Physician workload in mammography. AJR American journal of roentgenology. 2008;190(2):526-32.

11.Moss SM, Blanks RG, Bennett RL. Is radiologists' volume of mammography reading related to accuracy? A critical review of the literature. Clin Radiol. 2005;60(6):623-6.

12.Shaheen R, Schimmelpennink CA, Stoddart L, Raymond H, Slanetz PJ. Spectrum of diseases presenting as architectural distortion on mammography: multimodality radiologic imaging with pathologic correlation. Semin Ultrasound CT MR. 2011;32(4):351-62.

13.Gaur S, Dialani V, Slanetz PJ, Eisenberg RL. Architectural Distortion of the Breast. American Journal of Roentgenology. 2013;201(5):W662-W70.

Appendix – Reader characteristics

Reader characteristic	n=479
Mean age (years old)	46.22
Gender	
Female	191
Male	239
Non-specified	49
Mean number of years in specialisation	4.0489
Reader Specialisation	
Breast	301
Breast | Lung | Other	2
Breast | Other	14
Lung	1
Other	149
Not available	12
Number of mammograms interpreted weekly	
0-20	301
21-60	56
61-100	46
101-150	29
151-200	21
over 200	26
Fellowship trained in breast	94
Working in a breast screen program	120
Readers using digital images for interpretation	418

Appendix – Case Characteristics

Normal Case characteristic	n=240
Breast density	
Non-dense	124
Dense	116
Case difficulty	
Easy	117
Intermediate	109
Difficult	14

Cancer Case characteristic	n=120
Breast density	
Non-dense	57
Dense	63
Case difficulty	
Easy	19
Intermediate	38
Difficult	70
Cancer Type	
Architectural Distortion, Stellate	62
Calcification	21
Discrete Mass	15
Nonspecific density	22
Lesion Size	
<=15mm	101
>15mm	24
N/A	2

Appendix– Linear Mixed Modeling results

		Normal			Cancers		
	Characteristic	Number of data points	Model 1 & 2	Combined Model (3)	Number of data points	Model 1 & 2	Combined Model (3)
Case	Difficulty						
	Difficult (Intercept)	14	0.634*	1.5797*	67	0.692*	-1.0047*
	Easy	117	0.174*	0.17512	18	0.169*	0.522*
	Intermediate	109	0.119**	0.26644**	35	0.0298	0.0404*
	Breast density						
	Dense	116	-0.016	-	57	0.00393	-
	Non-dense	124	-0.049	-	63	0.0401	-
	Lesion Characteristics						
	Calcification	-	-	-	21	-0.0213	0.15345
	Discrete Mass	-	-	-	15	0.0553	0.109
	Nonspecific density	-	-	-	22	-0.0501	-0.0946
	Architectural distortion, spiculated mass and stellate	-	-	-	62	0.0242	0.183
	Lesion size	-	-	-	120	-0.00169	-
Reader	Reader Specialty						
	Breast (Intercept)	317	1.6968*	0.54829*	317	0.7454*	3.0509*
	Non-Breast	162	-0.05675	-0.062**	162	-0.02502	-0.176*
	Breast Screen Reader						
	No	359	-0.04837	-	359	0.038867	-
	Yes	120	-0.06038	-	120	-0.03585	-
	Breast fellowship						
	No	384	0.27387	-	384	0.041077	-
	Yes	94	0.017557	-	94	-0.0527**	-
	Number of mammograms interpreted per week						
	0-20	301	-0.09835*	-0.08832*	301	0.0229	0.15205*
	21-60	56	-0.19212*	-0.20419*	56	0.0678*	0.21397*
	61-100	46	-0.24791*	-0.26502*	46	0.0482	0.26168*
	101-150	29	-0.32152*	-0.34073*	29	0.0674	0.2916*
	151-200	21	-0.0848	-0.09766	21	0.111*	0.39452*
	>200	26	-0.46454*	-0.45984*	26	0.0480	0.23064**

Thank you very much for your attention!

Email address: dwon6926@uni.sydney.edu.au

Appendix F. MIPS 2022 Presentation

Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using multivariate mixed analysis

Dennis Wong, Grad Cert (CE), MDR,BMSc
Ziba Gandomkar, PhD, MSc, BSc
Warren Reed PhD, Grad Cert (T&L), BSc

Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney

THE UNIVERSITY OF SYDNEY

Introduction

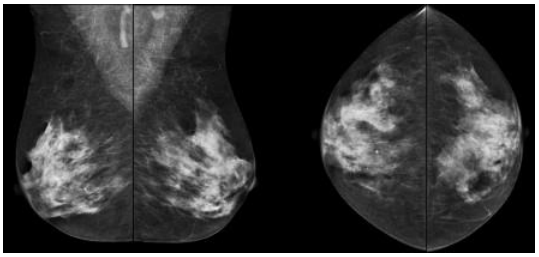
- Annual reading volumes are the most common reader-based characteristic (1-3)
- Reading volume alone is insufficient as a sole measure to evaluate reader performance (4)
- Previous studies have reported mixed findings regarding the relationship between annual reading volume and reader performance(5 -6)

Purpose of this study

- Using a multi-case multi-reader analysis framework can further increase the power of these studies.
- This multivariate analysis used several reader factors including image-based characteristics to further interrogate the underlying primary determinants of reader performance
- The aim of this study was to conduct a multivariate investigation of reader-based and image-based characteristics in reader performance in mammography interpretation.

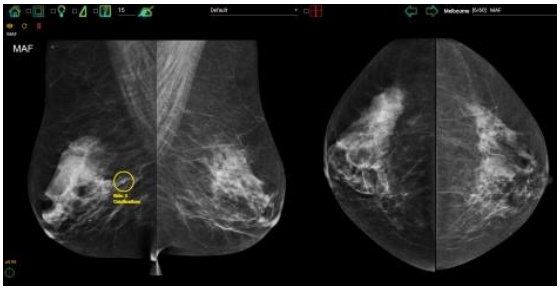
Methodology – Test sets

- A total of six test sets with a total of 60 mammography cases each from BREAST
- Images were capture using GE machines and were from NSW BreastScreen
- Each test sets consisted of 20 cancer-positive cases and 40 normal cases.
- Breast density was then grouped together classified as either "Dense" or "non-dense" breasts.
- For abnormal cases, the pathology characteristics were included and consisted of:
  - lesion side
  - lesion type
  - lesion length
- Both normal and abnormal cases had a difficulty rating determined by an expert



Methodology – Reader data

- The reader could mark lesions anywhere on the image
- A score classification between3-5 was used to determine increasing confidence
- Only the reader's first attempt at the test set was used.
- Readers were not aware of the prevalence and types of cancers included in the test sets.



Methodology – Statistical analysis

- For the statistical analysis, Linear mixed modelling was selectedto evaluate how the fixed effectssuch as a reader characteristicaffectsthe reader's performance.
- To account for multiple readers and cases, these two factors were classified as random effects in the linear mixed model.
- The modelling was done in the R statistical programming environment (version 4.1.2) and used the lme4 and lmerTest packages.
- The linear mixed models were divided from into three categories (1. case-based characteristics, 2. reader-based characteristics, and 3. mixed-characteristics) and explores normal and cancer cases separately.

Results – LMM Case-based characteristics

Significant findings

- For both normal and cancer cases, reader performance decreases as expert-determined difficulty increases ($P < 0.05$).

Non-significant findings

- Readers showed better performance in non-dense cases, compared to the dense cases
- Lesions classified as stellate and discrete mass demonstrated that readers were more likely to mark these cases correctly.

Results – LMM Reader-based characteristics

- The most significant reader characteristic is the number cases interpreted per week, excluding 151-200 per week.
- Mean estimate decreases with increasing number of mammograms interpreted per week.
- Conversely, for the cancer case's linear mixed model, the number of mammograms interpreted per week showed a similar trend
- However, the significance was not consistent.

Results – LMM Combined model

- The third model combined the characteristics found in the previous models
- For reader-based characteristics, mammograms interpreted per week remains significant
- For case-based characteristics, case difficulty determined by experts remains significant
- These findings were similar in both normal and cancer cases

Discussion

- Our results show that expert-predetermined difficulty and the number of cases interpreted weekly were the consistently significant factors
- The purpose of using a linear mixed model was to find the relationship between multiple variables with coefficients.
- This novel methods provides more power and the ability to observer smaller effects
- Previous studies have found that readers with an annual mammogram interpretation volume of over 1000 perform better than those reading less than 1000 per year (7 -8)
- Notably, our findings on reading volume correlate closely with Kan's study (9):
  - Cancer detection rate improves proportionally with the number of mammograms interpreted,
  - Up to a certain upper threshold
- This suggests an upper limit to which readers do not significantly improve in performance

Discussion – Clinical Implications

- Our findings reinforce that using annual reading volumes required to maintain certification in mammography interpretation is a valuable method to evaluate a reader's ability.
- However, one problem with the current implementation of the mammography's guidelines (10,11)
- Another clinical implication to consider is screening for non-specific densities and architectural distortions (12).
- Compared to calcifications and discrete masses, non-specific densities are often more difficult to diagnose because of their variability in appearance (13).

Conclusion

- Our findings show that expert-determined difficulty and cases interpreted weekly were consistently significant factors impacting reader performance
- For case characteristics, increasing the case difficulty determined by experts reduces the reader's performance
- For reader characteristics, increasing the number of cases interpreted weekly improves the reader performance up to a certain threshold
- Our results show that linear mixed modelling provides extra insights into the impact of reader- and image-based factors on reader performance in mammography interpretation.

References

- 1.Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings Radiology 2013;269(1):61-7.
- 2.Reed WM, Lee WB, Cawson JN, Brennan PC. Malignancy detection in digital mammograms : important reader characteristics and required case numbers. Acad Radiol 2010;17(11):1409-13.
- 3.Suleiman WI, Lewis SJ, Georgian-Smith D, Evanoff MG, McEntee MF. Number of mammography cases read per year is a strong predictor of sensitivity J Med Imaging (Bellingham) 2014;1(1):015503
- 4.Gandomkar Z, Lewis SJ, Li T, Ekpo EU, Brennan PC. A machine learning model based on readers' characteristics to predict their performances in reading screening mammograms . Breast Cancer. 2022
- 5.Hoff SR, Myklebust TÅ, Lee CI, Hofvind S. Influence of mammography volume on radiologists' performance: Results from breast screen Norway. Radiology 2019;292(2):289-96.
- 6.Théberge I, Hébert-Croteau N, Langlois A, Major D, Brisson J. Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. CMAJ. 2005;172(2):195-9.
- 7.Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists Radiology 2002;224(3):861-9.
- 8.Buist DS, Anderson ML, Smith RA, Carney PA, Miglioretti DL, Monsees BS, et al. Effect of radiologists' diagnostic work-up volume on interpretive performance. Radiology 2014;273(2):351-64.
- 9.Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program Radiology 2000;215(2):563-7.
- 10.Smith-Bindman R, Miglioretti DL, Rosenberg R, Reid RJ, Taplin SH, Geller BM, et al. Physician workload in mammography. AJR American journal of roentgenology 2008;190(2):526-32.
- 11.Moss SM, Blanks RG, Bennett RL. Is radiologists' volume of mammography reading related to accuracy? A critical review of the literature Clin Radiol 2005;60(6):623-6.
- 12.Shaheen R, Schimmelpennink CA, Stoddart L, Raymond H, Slanetz PJ. Spectrum of diseases presenting as architectural distortion on mammography: multimodality radiologic imaging with pathologic correlation Semin Ultrasound CT MR. 2011;32(4):351-62.
- 13.Gaur S, Dialani V, Slanetz PJ, Eisenberg RL. Architectural Distortion of the Breast. American Journal of Roentgenology 2013;201(5):W662-W70.

Appendix– Reader characteristics

Reader characteristic	n=479
Mean age (years old)	46.22
Gender	
Female	191
Male	239
Non-specified	49
Mean number of years in specialisation	4.0489
Reader Specialisation	
Breast	301
Breast | Lung | Other	2
Breast | Other	14
Lung	1
Other	149
Not available	12
Number of mammograms interpreted weekly	
0-20	301
21-60	56
61-100	46
101-150	29
151-200	21
over 200	26
Fellowship trained in breast	94
Working in a breast screen program	120
Readers using digital images for interpretation	418

Appendix – Case Characteristics

Normal Case characteristic	n=240
Breast density	
Non-dense	124
Dense	116
Case difficulty	
Easy	117
Intermediate	109
Difficult	14

Cancer Case characteristic	n=120
Breast density	
Non-dense	57
Dense	63
Case difficulty	
Easy	19
Intermediate	38
Difficult	70
Cancer Type	
Architectural Distortion, Stellate	62
Calcification	21
Discrete Mass	15
Nonspecific density	22
Lesion Size	
<=15mm	101
>15mm	24
N/A	2

Appendix – Linear Mixed Modeling results

		Normal			Cancers		
	Characteristic	Number of data points	Model 1 & 2	Combined Model (3)	Number of data points	Model 1 & 2	Combined Model (3)
Case	Difficulty						
	Difficult (Intercept)	14	0.634*	1.5797*	67	0.692*	-1.0047*
	Easy	117	0.174*	0.17512	18	0.169*	0.522*
	Intermediate	109	0.119**	0.26644**	35	0.0298	0.0404*
	Breast density						
	Dense	116	-0.016	-	57	0.00393	-
	Non-dense	124	-0.049	-	63	0.0401	-
	Lesion Characteristics						
	Calcification	-	-	-	21	-0.0213	0.15345
	Discrete Mass	-	-	-	15	0.0553	0.109
	Nonspecific density	-	-	-	22	-0.0501	-0.0946
	Architectural distortion, spiculated mass and stellate	-	-	-	62	0.0242	0.183
	Lesion size	-	-	-	120	-0.00169	-
Reader	Reader Specialty						
	Breast (Intercept)	317	1.6968*	0.54829*	317	0.7454*	3.0509*
	Non-Breast	162	-0.05675	-0.062**	162	-0.02502	-0.176*
	Breast Screen Reader						
	No	359	-0.04837	-	359	0.038867	-
	Yes	120	-0.06038	-	120	-0.03585	-
	Breast fellowship						
	No	384	0.27387	-	384	0.041077	-
	Yes	94	0.017557	-	94	-0.0527**	-
	Number of mammograms interpreted per week						
	0-20	301	-0.09835*	-0.08832*	301	0.0229	0.15205*
	21-60	56	-0.19212*	-0.20419*	56	0.0678*	0.21397*
	61-100	46	-0.24791*	-0.26502*	46	0.0482	0.26168*
	101-150	29	-0.32152*	-0.34073*	29	0.0674	0.2916*
	151-200	21	-0.0848	-0.09766	21	0.111*	0.39452*
	>200	26	-0.46454*	-0.45984*	26	0.0480	0.23064**


Thank you very much
for your attention!
Email address: dwon6926@uni.sydney.edu.au

Appendix G. Chapter 3's Linear Mixed Model Code (RStudio)

```
library(tidyverse)
library(dplyr)
library(inspectdf)
library(moments)
casenormal <- read_csv('P2casedatanormal.csv')
readernormal <- read_csv('P2readerdatanormal.csv')
casecancer <- read_csv('P2casedatacancer.csv')
readercancer <- read_csv('P2readerdatacancer.csv')

#convert to correct data type - data is mainly categorical
casenormal$cancer <-as.factor(casenormal$cancer)
casenormal$density <-as.factor(casenormal$density)
casenormal$side <-as.factor(casenormal$side)
casenormal$site <-as.factor(casenormal$site)
casenormal$size <-as.factor(casenormal$size)
casenormal$prior <-as.factor(casenormal$prior)
casenormal$difficulty <-as.factor(casenormal$difficulty)
casenormal$ranzcr <-as.factor(casenormal$ranzcr)

readernormal$age <-as.factor(readernormal$age)
readernormal$gender <-as.factor(readernormal$gender)
readernormal$specialyr <-as.factor(readernormal$specialisationyr)
readernormal$weeklymammo <-as.factor(readernormal$weeklymammo)
readernormal$fellowship <-as.factor(readernormal$fellowship)
readernormal$bscreen <-as.factor(readernormal$bscreen)
readernormal$digitalmammo <-as.factor(readernormal$digitalmammo)
readernormal$specialtype <-as.factor(readernormal$specialtype)

casecancer$cancer <-as.factor(casecancer$cancer)
casecancer$density <-as.factor(casecancer$density)
casecancer$side <-as.factor(casecancer$side)
casecancer$site <-as.factor(casecancer$site)
casecancer$size <-as.factor(casecancer$size)
casecancer$prior <-as.factor(casecancer$prior)
casecancer$difficulty <-as.factor(casecancer$difficulty)
casecancer$ranzcr <-as.factor(casecancer$ranzcr)

readercancer$age <-as.factor(readercancer$age)
readercancer$gender <-as.factor(readercancer$gender)
readercancer$specialyr <-as.factor(readercancer$specialisationyr)
readercancer$weeklymammo <-as.factor(readercancer$weeklymammo)
readercancer$fellowship <-as.factor(readercancer$fellowship)
readercancer$bscreen <-as.factor(readercancer$bscreen)
readercancer$digitalmammo <-as.factor(readercancer$digitalmammo)
readercancer$specialtype <-as.factor(readercancer$specialtype)


CaseNormalModel <- lm(accuracy ~ density+ difficulty , data = casenormal)

summary(CaseNormalModel)

CaseCancerModel <- lm(accuracy ~ cancer+ density + side + site+ size + prior + difficulty
+ ranzcr, data = casecancer)

summary(CaseCancerModel)

ReaderNormalModel <- lm(accuracy ~ age + gender +specialtype+ specialyr + weeklymammo +
fellowship + bscreen + digitalmammo, data = readernormal)

summary(ReaderNormalModel)
#removing insignificant variables
ReaderNormalModelv2 <- lm(accuracy ~ specialtype + weeklymammo, data = readernormal)

summary(ReaderNormalModelv2)
```

```
ReaderCancerModel <- lm(accuracy ~ age + gender +specialtype+ specialyr + weeklymammo +
fellowship + bscreen + digitalmammo, data = readercancer)

summary(ReaderCancerModel)
#removing insignificant variables
ReaderCancerModelv2 <- lm(accuracy ~ specialtype + weeklymammo, data = readercancer)

summary(ReaderCancerModelv2)

mixednormal <- read_csv('mixednormal.csv')
mixednormal$weeklymammo <-as.factor(mixednormal$weeklymammo)
mixednormal$difficulty <-as.factor(mixednormal$difficulty)
mixednormal$specialtype <-as.factor(mixednormal$specialtype)

MixedModelNormal <- lm(accuracy ~ difficulty *specialtype + difficulty *weeklymammo  ,
data = readernormal)

summary(MixedModelNormal)

mixedcancer <- read_csv('mixednormal.csv')
mixedcancer$weeklymammo <-as.factor(mixedcancer$weeklymammo)
mixedcancer$difficulty <-as.factor(mixedcancer$difficulty)
mixedcancer$specialtype <-as.factor(mixedcancer$specialtype)

MixedModelCancer <- lm(accuracy ~ difficulty *specialtype + difficulty *weeklymammo  ,
data = readercancer)

summary(MixedModelCancer)
```

Appendix H. Chapter 4's Ridge Regression Model Code (RStudio)

```
# Load required libraries
library(glmnet)
library(caret)
# Load data from the "P3datav7.csv" file (make sure it's in your working directory)
data <- read.csv("P3datav7.csv")

# Assuming your CSV file contains a column 'y' for the response variable
y <- data$accuracy

# Extract predictor variables from the CSV file (adjust column names as needed)
# Include both numeric and categorical predictor columns
X_numeric <- data[, c("size")]
X_categorical <- data[, c("cancer", "density","side", "site","prior", "Difficulty",
"ranzcr")]

# Create a formula for the dummyVars function
formula <- as.formula("~ .")

# Create dummy variables for categorical predictors
X_categorical_dummy <- predict(dummyVars(formula, data = X_categorical), newdata =
X_categorical)


# Combine numeric and dummy variables
X <- cbind(X_numeric, X_categorical_dummy)

# Normalize the predictor variables
X_normalized <- scale(X)

# Fit LASSO regression model
lasso_model <- cv.glmnet(X_normalized, y, alpha = 1)

# Fit Ridge regression model
ridge_model <- cv.glmnet(X_normalized, y, alpha = 0)

# Find optimal lambda for LASSO and Ridge models
optimal_lambda_lasso <- lasso_model$lambda.min
optimal_lambda_ridge <- ridge_model$lambda.min

# Get coefficient paths for LASSO and Ridge models
lasso_coefficients <- coef(lasso_model, s = optimal_lambda_lasso)
ridge_coefficients <- coef(ridge_model, s = optimal_lambda_ridge)

# Extract significant variables and their p-values
lasso_significant_vars <- lasso_coefficients[-1, "s1"] != 0
ridge_significant_vars <- ridge_coefficients[-1, "s1"] != 0

# Print significant variables and their coefficients along with p-values
cat("Ridge Model:\n")
cat("Significant Variables:", paste(names(ridge_significant_vars)[ridge_significant_vars],
collapse = ", "), "\n")
cat("Coefficients:", ridge_coefficients[ridge_significant_vars, "s1"], "\n")
cat("P-values:", ridge_model$glmnet.fit$beta[, optimal_lambda_ridge]
[ridge_significant_vars], "\n")

# Fit an OLS regression model
ols_model2 <- lm(y ~ X_normalized)

# Obtain p-values for the coefficients
p_values <- summary(ols_model2)$coefficients[, "Pr(>|t|)"]

# Print the p-values
cat("P-values:", p_values, "\n")
```

Appendix I. Artificial Intelligence Literature Review

Journal of Medical Radiation Sciences

Open Access

REVIEW ARTICLE

Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review

Dennis Jay Wong, MDR, BMSc (Hons), Ziba Gandomkar, PhD, MSc, BSc, Wan-Jing Wu, MDR, Guijing Zhang, MDR, Wushuang Gao, MDR, Xiaoying He, MDR, Yunuo Wang, MDR, & Warren Reed, PhD, Grad Cert (T&L), Bsc (Hons)

Discipline of Medical Imaging Sciences, The University of Sydney, Lidcombe, New South Wales, Australia

Keywords
Artificial intelligence, breast cancer, breast density, convolutional neural network, mammography

Correspondence
Dennis Jay Wong, M504, Block M, 75 East Street, Lidcombe, 2141, New South Wales, Australia. Tel: +61 02 9351 9586; E-mail: dwon6926@uni.sydney.edu.au

Received: 3 October 2019; Revised: 18 January 2020; Accepted: 11 February 2020

J Med Radiat Sci **67** (2020) 134–142

doi: 10.1002/jmrs.385

Abstract

Studies have shown that the use of artificial intelligence can reduce errors in medical image assessment. The diagnosis of breast cancer is an essential task; however, diagnosis can include 'detection' and 'interpretation' errors. Studies to reduce these errors have shown the feasibility of using convolution neural networks (CNNs). This narrative review presents recent studies in diagnosing mammographic malignancy investigating the accuracy and reliability of these CNNs. Databases including ScienceDirect, PubMed, MEDLINE, British Medical Journal and Medscape were searched using the terms 'convolutional neural network or artificial intelligence', 'breast neoplasms [MeSH] or breast cancer or breast carcinoma' and 'mammography [MeSH Terms]'. Articles collected were screened under the inclusion and exclusion criteria, accounting for the publication date and exclusive use of mammography images, and included only literature in English. After extracting data, results were compared and discussed. This review included 33 studies and identified four recurring categories of studies: the differentiation of benign and malignant masses, the localisation of masses, cancer-containing and cancer-free breast tissue differentiation and breast classification based on breast density. CNN's application in detecting malignancy in mammography appears promising but requires further standardised investigations before potentially becoming an integral part of the diagnostic routine in mammography.

Introduction

Screening mammography is the recommended tool for the early detection of breast cancer in women experiencing no symptoms and has been shown to decrease breast cancer mortality rate by 40–63%.[1] However, current diagnostic frameworks for assessing mammograms such as the breast imaging reporting and data system (BI-RADS), developed by the American College of Radiology (ACR) used to report findings into a number of well-defined categories, can be limited by 'detection errors' (pathology missed) and 'interpretation errors' (pathology misinterpreted).[2] One possible solution to reduce the errors is utilising artificial intelligence (AI) tools, which automatically find abnormalities (detection) and categorise them either as normal or as abnormal (interpretation).

In general, AI can be defined as a sophisticated computer program able to carry out tasks usually requiring human intellect in areas such as visual perception and decision-making. Over the past few years, deep learning has dramatically transformed the AI industry. Deep learning is a subset of machine learning methods based on artificial neural networks. A convolutional neural network (CNN) is a class of deep neural networks, commonly applied to analyse image data. Similar to conventional machine learning methods, CNN allows the machine to perform specific tasks by relying on the inference of decision boundaries rather than explicit instructions by a user.

Furthermore, CNNs learn relevant features from the data, with limited input from a human expert, and make predictions via heuristics when new data are entered.[3] CNNs can be used to assist with the detection of breast cancer on mammographic images with a high degree of accuracy as measured by sensitivity and specificity.[4] It can also reduce the time required to assess a large volume of mammograms.[5] These potential benefits have emphasised the importance of exploring CNN's ability to accurately diagnose abnormalities on mammograms promptly.[4,5] CNN is faster with similar accuracy when compared to radiologists without other limiting factors such as a lack of qualified personnel.[6,7]

Considering recent advances in AI tools in mammography, it is essential for the imaging and the wider health community to better understand current trends and developments. However, there is a lack of comparisons in the literature between similar studies that this literature review aims to address. The purpose of this study therefore was to compare the recent developments of CNN in the diagnosis of mammographic images to enable medical imaging professionals to gain greater insight on how this technology can assist our patients in future.

Methodology

The focus of this review was to identify the current developments and uses of CNN in mammography diagnostics. Due to the review's narrative nature, results were described by description and exploration with qualitative elements rather than systematically.

Using the keywords listed in Table 1, ScienceDirect, PubMed, MEDLINE, British Medical Journal and Medscape databases were searched for original English research studies published between 2011 and 2019. Studies which used imaging modalities other than mammography were excluded. A total of 33 relevant peer-reviewed articles were found.[8 40] To compare the results among studies, relevant features were extracted: (1) author and publication year; (2) study aim; (3) utilised database; (4) number of images used; and (5) performance measures including sensitivity, specificity and area under the curve (AUC).

Results

A total of 33 out of 65 articles were selected under the inclusion criteria; 32 studies were rejected due to lack of direct relevance. In the accepted literature, there were various recurring themes, so articles were further categorised into four different salient categories:

Category 1. Differentiation between benign and malignant tissues;
Category 2. Localisation of mass in breast tissue;
Category 3. 'Cancer-containing' and 'cancer-free' breast tissue differentiation;
Category 4. Breast classification based on breast density.

Table 1. List of keywords used to search for relevant topics related to the review.

Topic	Keywords
Artificial Intelligence	Artificial Intelligence; Deep learning; Artificial Neural Networks; Convolutional Neural Network
Anatomy; Pathology of Interest	Breast; Breast Cancer; Breast Lesion; Breast Carcinoma; Breast Neoplasm
Imaging Modality	Mammogram; Mammography
Others	Image Reading; Diagnostics; BI-RADS; Breast Screening

A total of 12 studies focused exclusively on Category 1, primarily testing CNN's ability to distinguish between benign and malignant masses. Category 2 had seven studies that focused on mass localisation. Category 3 had seven studies which focused on screening images as 'cancer-free' and 'cancer-containing'; Category 4 had seven studies that used CNN to categorise mammograms by breast density.

Twenty-two of the 33 articles used the ACR's BI-RADS system to report mammograms assigning breast imaging studies to one of six assessment categories (0 – incomplete study, 1 – negative, 2 – benign, 3 – probably benign, 4 – suspicious for malignancy, 5 – highly suggestive of malignancy and 6 – known biopsy-proven malignancy).[41]

In all 33 articles, four primary image databases were used:

1. Digital Database for Screening Mammography (DDSM) (Digitised Images): This database from the University of South Florida contains 2500 studies, with each study containing two images of each breast and patient information such as age, when the image biopsy proven, and ACR breast density classification. All cases are female.[42]
2. MIAS Mini Mammographic Database (Digital Images): The mini-MIAS provided by the Mammographic Image Analysis Society contains 161 pairs of films and has commonly encountered pathologies and normal cases.[43]
3. INbreast database (Digital images): This database has 410 images from 115 cases: 90 cases have images of both breasts and 25 cases from mastectomy patients. All patients are females. Different pathologies include masses, calcification, asymmetries and distortions.[44]
4. Breast Cancer Digital Repository (BCDR) (Digital and Digitised images): BCDR is a public repository

composed of breast cancer patient's case studies hosted by the IMED Project in March 2009.[8] Data classification uses BI-RADS and is subdivided into two repositories: the Film Mammography Repository and Full Field Digital Mammography Repository. The BCDR has been used to train CAD schemes and is still in development.[45] Digitised images contain 1010 patients (998 female and 12 male) ranging from ages 20 to 90, 1125 studies and a total of 3703 images. The digital database is in development with 724 patients (723 female and 1 male) ranging from ages 27 to 92, 1042 studies and a total of 3612 images.[46]

Discussion

Category 1: Differentiation between a benign and malignant mass

The differentiation between benign and malignant masses in CNN breast screening was the most prominent category of study. Akselrod-Ballin's CNN accurately diagnosed mammograms classified as BI-RADS 1; however, it found difficulty diagnosing higher graded images.[19] Importantly, they noted that CNN systems have issues differentiating benign and malignant masses in BI-RADS 3 images.[19] When BI-RADS 3 images were excluded, their CNN system improved detection AUC from 0.60 to 0.72. Aside from this study, other studies did not mention BI-RADS 3 exclusion.

Furthermore, Carneiro suggested that their CNN's inability to diagnose BI-RADS images >1 was due to training images being annotated for either microcalcifications or masses, but not for both in a single image, potentially affecting heuristics.[15] Additionally, automated segmentation made by CNN generally resulted in lower accuracy, since detection of false-positive masses impacts the system's performance in mass classification.[15]

Category 2: Localisation of masses in breast tissue

Although many studies focused on benign and malignant differentiation, tumour localisation was another common topic. Five studies showed detecting a mass in dense breast regions and pectoral muscles to be difficult due to high intensities.[21 24,36]

Al-Masni et al.[47] employed a You Only Look Once (YOLO) system that uses a singular network for diagnostics by dividing the input image into multiple subregions, compared to traditional neural networks requiring multiple networks for each extracted region. As a result, their system operated quicker compared to traditional neural networks and successfully identified masses in pectoral muscles and dense tissues; they suggested that its ability to localise masses in radiopaque areas was due to augmented training by digitally manipulating images to produce more training images.[47]

Hwang et al.[21] experimented with CNN's ability to self-learn via a weakly supervised CNN system, meaning that images used for training were minimally annotated with a weak description. Compared to other studies, the results were poor for AUC, sensitivity, specificity and accuracy with no results exceeding 0.77.[21] This finding suggests that current CNN's need highly annotated images for training for accurate results.

Overall, the results from the studies in Category 2 appear promising in highly supervised training methods, but weak in lowly supervised systems.

Category 3: 'Cancer-containing' and 'cancer-free' breast tissue differentiation

In 2018, a study using a CNN system was designed to differentiate between normal and abnormal images via biomarkers such as breast density, presence of mass and microcalcification.[28] Their results indicated that abnormal images were more accurately identified when microcalcifications and benign cases were excluded. The authors suggested that due to the training data having an uneven number of cases of normal and cancer cases, this caused overfitting and database dependency. This creates difficulty distinguishing between benign and malignant cases due to overreliance on heuristics and an inability to learn and adapt from new cases.[28]

Al-Antari modified Al-Masni et al.'s YOLO system to differentiate between cancer-containing and cancer-free images, their study demonstrated an accuracy of 95.64% and noted that a robust CNN system heavily relies on the deep learning model in segmenting specific region of masses to reduce potential false positives from surrounding tissues.[30]

Category 4: Breast classification based on breast density

Breast tissue density may increase the likelihood of developing breast cancer, as increased breast density and glandular tissue may make it difficult to detect early signs of breast cancer such as microcalcifications.[47] The importance of classifying breast density for breast cancer risk must be stressed as the relative risk of developing breast cancer in women with heterogeneously dense breasts is 1.2 times more compared to the average women, and for women with extremely dense breasts, the relative risk is 2.1 times higher. This is likely due to the masking effect, excess glandular tissue and breast density as an independent risk factor.[48]

Between 2018 and 2019, there were studies that tested CNN's capabilities in categorising mammograms by breast density.[37,38,40,49] Ha's preliminary system demonstrated an accuracy of 72% in predicting 'high'-risk and 'low'-risk mammograms by generating heat maps on the mammograms where 'red' areas had overlapping mammographic features suggesting high density, high and low risk defined as dense and non-dense, respectively.[37] They noted that while the red areas on the heatmaps suggest areas of high density, the pixel map generated does not necessarily indicate specific areas where breast cancer may develop.[37] Notably, Mohammed's study demonstrated a high accuracy on differentiating between BI-RADS category 'scattered density' and 'heterogeneously dense' images, however, noted that their study was a retrospective study and the images used had little variation in mammographic vendors and imaging protocol, suggesting that their system may not guarantee the same results from different manufacturers and imaging standards.[38]

Fonseca et al conducted a preliminary study to diagnose breast density using the ACR's breast composition categories (a – entirely fatty; b – scattered areas of fibro-glandular density; c – heterogeneously dense tissue, potentially obscuring small masses; and d – breasts extremely dense, lowering mammography sensitivity)[41] and achieved a mean accuracy of 0.7305, similar to a radiologist when categorising breast density.[35] The study's weakness was that the images used were sourced locally and not from a mainstream database. Additionally, a study found that inter-reader agreement regarding breast density classification was only at 49% and agreement was usually on either fatty breasts or extremely dense breasts.[50] Therefore, there is a great potential in improving breast density classification in assessing breast cancer risk using CNN's capabilities.

Diniz in 2018[36] classified breast tissue by density and regions with or without a mass; by combining them, they localised and detected masses in dense (Cooper's ligaments, mammillary glands and ducts) and non-dense tissue (adipose tissue). Sensitivity was slightly higher in non-dense tissue (0.9156 vs. 0.9036), yet dense tissue had significantly higher specificity (0.9635 vs. 0.9073), meaning that their CNN was better at recognising negative non-dense images but identified abnormalities more accurately in dense tissues.[36]

Sensitivity and specificity, total accuracy and AUC of CNNs

Basic statistical analyses are employed for simple comparisons between a study's AUC, sensitivity, specificity and total accuracy gathered from Table 2. The overall mean of CNN's AUC, specificity, sensitivity and total accuracy is listed in Table 3. It is important to note that data sets used in the radiologist studies are not standardised and images are from patients with vastly different clinical history.

For sensitivity, minimum and maximum value disparity is greater for CNN, but the lowest and highest sensitivity range is more significant. The lowest sensitivity and specificity are from a study by Hwang and Kim; their results are lowest across all parameters, reinforcing that a highly supervised learning requirement is required. They argued that the training database (DDSM) contained low-quality images with many artefacts. However, this issue did not occur for other studies.

Of note, some studies omitted sensitivity and specificity results, so performance comparisons between the CNN and the radiologist are rudimentary. Furthermore, every study used a different number of images to train and test their systems, making exact comparisons difficult.

From these four parameters, the current development of CNN appears promising in mammographic diagnostics but requires highly supervised learning (images that are highly annotated with regions of interests to learn and be segmented), sufficient images for training and adjustment of learning algorithms to ensure overfitting and database dependency do not occur to achieve good diagnostic results.

In 2018, a retrospective study compared the performance of a CNN system to 101 radiologists and found that the CNN system slightly outperformed the radiologists with an AUC of 0.840 and 0.814, respectively.[51] However, due to the study's retrospective nature, results may suffer from the 'laboratory effect' and may not be applicable in a clinical setting.[51]

Limitations of this review and current literature

Although some studies have used the same databases for training, many have used a different combination of databases and various numbers of images for training. Hence, valuable comparisons between studies are difficult to make as each database uses mammograms from different hospitals, image parameters and manufacturers, resulting in a lack of reference standards between data sets. Since the study outcomes depend significantly on the database used for training, the variability between these databases can affect each AI's performance algorithm. This variability is called 'inter-database variability', causing optimisation of algorithm and standardised comparison between studies extremely challenging.[52]

Additionally, there are several 'cases' per database and studies may not use the same cases. For example, Rouhi

Table 2. 33 studies listed by category, author, year, database used, number of images, AUC, specificity, sensitivity and accuracy.[8–40]

Author	Year	Database	#Cases (Images)	AUC	Specificity	Sensitivity	Accuracy
Category 1 studies							
Ramos-Pollán et al.[8]	2011	BCDR	286	0.996	0.77	0.95	0.97
Rouhi et al.[9]	2014	DDSM	170 (170)	0.951	0.96	0.97	0.96
Arevalo et al.[10]	2015	BCDR	344 (736)	0.86			
Arevalo et al.[11]	2016	BCDR	344 (736)	0.7			
Dhungel et al.[12]	2016	INbreast	115 (410)	0.87			0.91
Huynh et al.[13]	2016	Custom	( )	0.81			
Jiao et al.[14]	2016	DDSM	( )				0.97
Carneiroet al.[15]	2017	INbreast	115 (410)	0.72	0.66	0.69	
Carneiro et al.[15]	2017	INbreast	115 (410)	0.87	0.92	0.69	
Carneiro et al.[15]	2017	DDSM	172 (680)	0.91	0.97	0.94	
Kooi et al.[27]	2017	Custom	956 (1804)	0.8			
Teare et al.[17]	2017	DDSM	0	0.92	0.91	0.8	
Chougrad et al.[18]	2018	DDSM	1329 (5316)	0.98			0.97
Chougrad et al.[18]	2018	INbreast	50 (200)	0.97			0.96
Chougrad et al.[18]	2018	BCDR	300 (600)	0.96			0.97
Chougrad et al.[18]	2018	MD***	1529 (6116)	0.99			0.99
Chougrad et al.[18]	2018	MIAS	113 (113)	0.99			0.98
Akselrod-Ballin et al.*[19]	2016	DDSM, INbreast	850 (850)				0.78
Akselrod-Ballin et al.**[19]	2016	DDSM	850 (850)				0.77
	2016	INbreast (without BI RADS 3)	( )				
Akselrod-Ballin et al.*[19]	2016	DDSM, INbreast	~850	0.6			
Akselrod-Ballin et al.**[19]	2016	DDSM, INbreast (without BI-RADS 3)	~850	0.72			
Ribli et al.[20]	2018	INbreast	115 (115)	0.85	0.9		
Category 2 studies							
Hwang et al.[21]	2016	DDSM and MIAS	10,363 (322)	0.68	0.76	0.58	0.7
Hwang et al.[21]	2016	DDSM and MIAS	10,363 (322)	0.54	0.66	0.44	0.66
Al-masni et al.[47]	2017	DDSM	~600				0.9633
Sampaio[23]	2011	DDSM	566 (566)	0.87	0.86	0.8	0.85
Sun et al.[24]	2016	Custom	~1874	0.8818	0.72	0.81	0.8243
Shen et al.[25]	2019	DDSM	2223	0.922		0.9643	
Savelli et al.[26]	2019	INbreast	115 (410)			0.763	
Kooi et al.[16]	2017	Custom	44,090	0.941			
Category 3 studies							
Kim et al.[28]	2018	Custom	29,107	0.91	0.89	0.76	
Jadoon et al.[29]	2017	DDSM and MIAS	( )	0.85	0.82	0.88	0.82
Al-antari et al.[30]	2018	INbreast	115 (410)	0.9478	0.9241	0.9714	0.9564
Kaur et al.[31]	2019	MIAS	20		0.88	0.9	0.9
Anitha et al.[32]	2017	DDSM	300			0.925	
Anitha et al.[32]	2017	MIAS	170			0.935	
Wichakam et al.[33]	2016	INbreast	216				0.9727
Suzuki et al.[34]	2016	DDSM	198			0.899	
Category 4 studies							
Fonseca et al.[53]	2015	Custom	~1157				0.73
Bandeira Diniz et al.[36]	2018	DDSM	Non-dense; ~ (1004)		0.91	0.92	0.91
Bandeira Diniz et al.[36]	2018	DDSM	Dense; (1482)		0.96	0.9	0.95
Ha et al.[37]	2019	Custom	1474				0.72
Mohamed et al.[49]	2018	Custom	15,415	0.95			
Ionescu et al.[39]	2019	Custom	67,520				
Mohamed et al.[38]	2018	Custom	22,000	0.926			
Trivizakis et al.[40]	2019	DDSM	2500 (10,239)	0.548			
Trivizakis et al.[40]	2019	MIAS	161 (322)	0.798			

*Exp1.
**Exp2.
***MD = DDSM+ INbreast + BCDR.

Table 3. Comparison between mean, median, minimum and maximum values of AUC, specificity, sensitivity and total accuracy from the 33 CNN studies[8–40] and four radiologists studies.[54–57]

	Mean		Median		Minimum		Maximum	
	CNN	Radiologist	CNN	Radiologist	CNN	Radiologist	CNN	Radiologist
AUC	0.851		0.876		0.540		0.996	
Specificity	0.851	0.905	0.890	0.808	0.660	0.889	0.970	0.988
Sensitivity	0.833	0.795	0.899	0.916	0.440	0.600	0.971	0.899
Total accuracy	0.883		0.930		0.660		0.990	

et al.'s[9] model aimed to differentiate malignant or benign lesions, while Diniz et al.'s[36] design aimed to localise mass and categorise images based on breast tissue density. Although both used the same database, the cases used were different, meaning that direct comparisons are difficult. Furthermore, current data sets are archaic and lack realism; many images still use digitised rather than digital images; the data sets lack clinical history such as risk factors, ethnicity and age make thorough analysis difficult. Also, in real-world screening approximately 0.5–0.8% cases contain cancer, meaning that data sets used had a high number of malignancies.

Likewise, the number of images used to pre-train and test the CNN varies between studies and this potentially affects its calibration. A CNN extensively trained with small training schemes may lead to overfitting and database dependency which causes the CNN to be over-reliant on heuristics taught in training, resulting in weak future predictions, analysis and adaptation to new images.[3] To improve comparisons between CNN and radiologists, their performance must be examined on the same data set and investigations to identify overlapping errors between AI and humans are recommended.

To solve these issues, mainstream databases should be selected and standardised with the same cases and images used so meaningful statistical analysis can be performed between studies such as ROC curve analysis and homogeneity.

Conclusion

CNN diagnosis in mammography has been demonstrated as an expanding field in artificial intelligence diagnostics. The most significant limitation of the current studies is the lack of standardisation between studies. Despite this limitation, the mean accuracy across the studies is promising, showing the potential reliability of CNN in mammography diagnostics.

Current innovation in CNN diagnostics for mammography includes the differentiation of benign and malignant masses, the localisation of masses, cancer-containing and cancer-free breast tissues differentiation and breast classification based on density. Literature indicates that the differentiation of benign and malignant masses has the most developments, and the consensus suggests that CNN is comparable in performance to a radiologist. However, there are still some significant weaknesses within the 33 studies due to the inter-database variability, inadequate original data reports and other aspects of AI training such as unsupervised learning and training schemes. Therefore, as promising as the introduction of a CNN application in detecting malignancy in breasts seems, it does not currently appear ready for the real clinical environment and requires further standardised investigations before it can be a part of the diagnostic routine for mammography imaging. With the current developments, instead of solely investigating CNN performance versus radiologists, the real opportunity is to examine CNN's potential to support radiologist's performance both for training purposes and as a computer aid to help them to further improve medical image diagnosis.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001; **91**(9): 1724–31.
2. Boyer B, Canale S, Arfi-Rouche J, Monzani Q, Khaled W, Balleyguier C. Variability and errors when applying the BIRADS mammography classification. *Eur J Radiol* 2013; **82**(3): 388–97.
3. Lee J-G, Jun S, Cho Y-W, et al. Deep learning in medical imaging: general overview. *Korean J Radiol* 2017; **18**(4): 570.
4. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; **2**(4): 230–43.
5. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. *Eur J Intern Med* 2018; **48**: e13–4.

6. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018; **15**(11): e1002686.
7. Hanhirova J, Kämäräinen T, Seppälä S, Siekkinen M, Hirvisalo V, Ylä-Jääski A. Latency and throughput characterization of convolutional neural networks for mobile computer vision. In: Proceedings of the 9th ACM Multimedia Systems Conference on - MMSys '18 [Internet]. ACM Press, New York, 2018 [cited 2019 May 5]; 204–15. Available from: http://dl.acm.org/citation.cfm?doid=3204949.3204975.
8. Ramos-Pollán R, Guevara-López MA, Suárez-Ortega C, et al. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *J Med Syst* 2012; **36**(4): 2259–69.
9. Rouhi R, Jafari M, Kasaei S, Keshavarzian P. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Syst Appl* 2015; **42**(3): 990–1002.
10. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Convolutional neural networks for mammography mass lesion classification. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. 2015.
11. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed* 2016; **127**: 248–57.
12. Dhungel N, Carneiro G, Bradley AP.The automated learning of deep features for breast mass classification from mammograms [Internet], 2019 [cited 2019 May 5]. Available from: https://cs.adelaide.edu.au/~neeraj/mass_classification_miccai_16.pdf.
13. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging* 2016; **3**(3): 34501.
14. Jiao Z, Gao X, Wang Y, Li J. A deep feature based framework for breast masses classification. *Neurocomputing* 2016; **197**: 221–31.
15. Carneiro G, Nascimento J, Bradley AP. Automated analysis of unregistered multi-view mammograms with deep learning [Internet]. 2017 [cited 2019 May 5]. Available from: https://cs.adelaide.edu.au/~carneiro/publications/tmi_multimodal.pdf.
16. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal* 2017; **35**: 303–12.
17. Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. *J Digit Imaging* 2017; **30**(4): 499–505.
18. Chougrad H, Zouaki H, Alheyane O. Deep convolutional neural networks for breast cancer screening. *Comput Methods Programs Biomed* 2018; **157**: 19–30.
19. Akselrod-Ballin A, Karlinsky L, Alpert S, Hasoul S, Ben-Ari R, Barkan E.A region based convolutional network for tumor detection and classification in breast mammography, 2016 [cited 2019 May 5]; 197–205. Available from: http://link.springer.com/10.1007/978-3-319-46976-8_21.
20. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* 2018; **8**(1): 4165.
21. Hwang S, Kim H-E.Self-transfer learning for fully weakly supervised object localization, 2016 Feb 4 [cited 2019 May 5]. Available from: http://arxiv.org/abs/1602.01625.
22. Al-masni MA, Al-antari MA, Park JM, et al. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. IEEE; 2017 [cited 2019 May 5]; 1230–3. Available from: https://ieeexplore.ieee.org/document/8037053/.
23. Borges Sampaio W, Moraes Diniz E, Corrêa Silva A, Cardoso de Paiva A, Gattass M. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Comput Biol Med* 2011; **41**(8): 653–64.
24. Sun W, Tseng T-L, Zhang J, Qian W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput Med Imaging Graph* 2017; **57**: 4–9.
25. Shen R, Yan K, Tian K, Jiang C, Zhou K. Breast mass detection from the digitized X-ray mammograms based on the combination of deep active learning and self-paced learning. *Futur Gener Comput Syst* 2019; **101**: 668–79.
26. Savelli B, Bria A, Molinara M, Marrocco C, Tortorella F. A multi-context CNN ensemble for small lesion detection. *Artif Intell Med* 2019; **103**: 101749.
27. Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Med Phys* 2017; **44**(3): 1017–27.
28. Kim E-K, Kim H-E, Han K, et al. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci Rep* 2018; **8**(1): 2762.
29. Jadoon MM, Zhang Q, Haq IU, Butt S, Jadoon A. Three-class mammogram classification based on descriptive CNN features. *Biomed Res Int* 2017; **2017**: 1–11.
30. Al-antari MA, Al-masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection,

segmentation, and classification. *Int J Med Inform* 2018; **117**: 44–54.
31. Kaur P, Singh G, Kaur P. Erratum: Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. *Informat Med Unlocked* 2019; **16**: 100151.
32. Anitha J, Dinesh Peter J, Immanuel Alex Pandian S. A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms. *Comput Methods Programs Biomed* 2017; **138**: 93–104.
33. Wichakam I, Vateekul P. Combining deep convolutional networks and SVMs for mass detection on digital mammograms. In 2016 8th Int Conf Knowl Smart Technol KST 2016, 2016; 239–44.
34. Suzuki S, Zhang X, Homma N, et al. Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. In 2016 55th Annu Conf Soc Instrum Control Eng Japan, SICE 2016. 2016;1382–6.
35. Fonseca P, Mendoza J, Wainer J, Ferrer J, Pinto J, Guerrero J, Castaneda B. Automatic breast density classification using a convolutional neural network architecture search procedure. In: *Proc. SPIE 9414, Medical Imaging 2015: Computer-Aided Diagnosis*, 941428. 2015 https://doi.org/10.1117/12.2081576
36. Bandeira Diniz JO, Bandeira Diniz PH, Azevedo Valente TL, Corrêa Silva A, de Paiva AC, Gattass M. Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. *Comput Methods Programs Biomed* 2018; **156**: 191–207.
37. Ha R, Chang P, Karcich J, et al. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Acad Radiol* 2018; **26**(4): 544–9.
38. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys* 2018; **45**(1): 314–21.
39. Ionescu GV, Fergie M, Berks M, et al. Prediction of reader estimates of mammographic density using convolutional neural networks. *J Med Imaging* 2019; **6**(3): 1.
40. Trivizakis E, Ioannidis GS, Melissianos VD, et al. A novel deep learning architecture outperforming 'off-the-shelf' transfer learning and feature-based methods in the automated assessment of mammographic breast density. *Oncol Rep* 2019; **42**(5): 2009–15.
41. Sickles E, D'Orsi C, Bassett L. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, 5th edn. Reston, VA: American College of Radiology, 2013.
42. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. In: Yaffe MJ (ed). Digital Mammography 2000: Proceedings of the Fifth International Workshop on Digital Mammography. Medical Physics, Madison, 2000; 431-434.
43. Suckling J, Parker J, Dance D, et al.Mammographic Image Analysis Society (MIAS) database v1.21. 2015 Aug 28 [cited 2019 May 5]. Available from: https://www.repository.cam.ac.uk/handle/1810/250394.
44. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol* 2012; **19**(2): 236–48.
45. Suarez Ortega C, Es CS, Franco Valiente JM, et al.Improving the breast cancer diagnosis using digital repositories [Internet]. 2019. [cited 2019 May 5]. Available from: http://iwbbio.ugr.es/papers/iwbbio_094.pdf.
46. BOUME. Breast Cancer digital mammography [Internet]. [cited 2019 Jul 3]. Available from: https://bcdr.eu/information/about.
47. Al-masni MA, Al-antari MA, Park JM, et al. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. IEEE; 2017 [cited 2019 May 5]; 1230–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29060098.
48. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. *Radiographics* 2015; **35**(2): 302–15.
49. Mohamed AA, Luo Y, Peng H, Jankowitz RC, Wu S. Understanding clinical mammographic breast density assessment: a deep learning perspective. *J Digit Imaging* 2018; **31**(4): 387–92.
50. Nicholson BT, LoRusso AP, Smolkin M, Bovbjerg VE, Petroni GR, Harvey JA. Accuracy of assigned BI-RADS breast density category definitions. *Acad Radiol* 2006; **13**(9): 1143–9.
51. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst.* 2019; **111**(9): 916–22.
52. Wirth M, Lyon J, Fraschini M, Nikitenko D. The effect of mammogram databases on algorithm performance. In: Proceedings 17th IEEE Symposium on Computer-Based Medical Systems [Internet]. IEEE Comput. Soc [cited 2019 May 5]; 15–20. Available from: http://ieeexplore.ieee.org/document/1311684/.
53. Fonseca P, Mendoza J, Wainer J, et al.Automatic breast density classification using a convolutional neural network architecture search procedure [Internet], 2019 [cited 2019 May 5]. Available from: http://metis.pucp.edu.pe/~pfonseca/cnn_breast_density_pfonseca.pdf
54. Kavanagh AM, Giles GG, Mitchell H, Cawson JN. The sensitivity, specificity, and positive predictive value of screening mammography and symptomatic status. *J Med Screen* 2000; **7**(2): 105–10.

55. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015; **175**(11): 1828.
56. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology* 2002; **225**(1): 165–75.
57. Kerlikowske K, Hubbard RA, Miglioretti DL, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States. *Ann Intern Med* 2011; **155**(8): 493.