RESEARCH ARTICLE

Editorial Process: Submission:04/02/2018 Acceptance:02/05/2019

Breast Cancer Diagnostic Efficacy in a Developing South-East Asian Country

Rhianna L Jackson[1], Callan R Double[1], Hayden J Munro[1], Jessica Lynch[1], Kriscia A Tapia[2], Phuong Dung Trieu[2,3], Maram Alakhras[2], Aarthi Ganesan[2], Thuan Doan Do[4], Baolin Pauline Soh[5], Patrick C Brennan[2], Louise Puslednik[1,2]*

Abstract

Background: Breast cancer, is increasing in prevalence amongst South East (SE) Asian women, highlighting the need for high quality, early diagnoses. This study investigated radiologists' detection efficacy in a developing (DC) and developed (DDC) SE Asian country, as compared to Australian radiologists. **Methods:** Using a test-set of 60 mammographic cases, 20 containing cancer, JAFROC figures of merit (FOM) and ROC area under the curves (AUC) were calculated as well as location sensitivity, sensitivity and specificity. The test set was examined by 35, 15, and 53 radiologists from DC, a DDC and Australia, respectively. **Results:** DC radiologists, compared to both groups of counterparts, demonstrated significantly lower JAFROC FOM, ROC AUC and specificity scores. DC radiologists had a significantly lower location sensitivity than Australian radiologists. DC radiologists also demonstrated significantly lower values for age, hours of reading per week, and years of mammography experience when compared with other radiologists. **Conclusion:** Significant differences in breast cancer detection parameters can be attributed to the experience of DC radiologists. The development of inexpensive, innovative, interactive training programs are discussed. This non-uniform level of breast cancer detection between countries must be addressed to achieve the World Health Organisation goal of health equity.

Keywords: Detection- experience- geographical location- mammography- radiologists

Asian Pac J Cancer Prev, **20 (3)**, 727-731

Introduction

Breast cancer is one of the most common cancers and the leading cause of death for women world-wide (WHO, 2017). In 2012, the disease affected 1.7 million women and caused over 522,000 deaths (WHO, 2017). Historically, it has been reported that the incidence rates of breast cancer in South East (SE) Asia are lower than that of western countries (Bray et al., 2004). However, recent studies suggest an increasing prevalence of this disease in SE Asia (WHO, 2017), with breast cancer becoming the second most common malignancy among women, making up to 22.4% of all cancers (Ferlay et al., 2015). For example, in 2012, approximately 11,060 cases of female breast cancer were diagnosed in one DC (Vuong, 2010), indicating a 30% increase in the number of cases compared with 10 years ago (Ferlay et al., 2015). Furthermore these cases are generally more aggressive than Australian cases and occur at a younger age with 89% of cancers detected having local recurrence and distant metastases and 64.7% of the cases occurring in women below the age of 50 years (Nguyen, 2009; Vuong et al., 2010; Trieu et al., 2015). Given that mortality and morbidity is highly reliant on early detection of small lesions, investigations to ensure optimum diagnostic efficacy are required.

Cancer detection depends on an individual reader's interpretation, with perceptual errors accounting for 60% of all diagnostic errors in radiology. The minimisation of perceptual errors is usually dependent on the radiologist's characteristics, such as specialisation and experience (Rawashdeh et al., 2013). It is also well-established that the effect of incorrect diagnosis of normal images (false positives) can result in significant emotional trauma, even when normality is subsequently shown (Rawashdeh et al., 2013). The use of test-set mammograms designed to test diagnostic efficacy has previously been instrumental in providing a deeper knowledge of the factors that influence the performance of radiologists world-wide (Rawashdeh et al., 2013; Mello-Thoms et al., 2014; Suleiman et al., 2014; Soh et al., 2016). However, despite the increase in breast cancer prevalence and the aggressive nature of these cancers in younger women (Trieu et al., 2015),

*[1]St Matthews Catholic School, Mudgee, New South Wales, [2]Faculty of Health Sciences, The University of Sydney, Australia, [3]Department of Medical Imaging, Ho Chi Minh City University of Medicine and Pharmacy, [4]Department of Diagnostic Imaging, Vietnam National Cancer Hospital, Vietnam, [5]Singapore Eye Research Institute, Singapore. *For Correspondence: l.puslednik@bth.catholic.edu.au*

application of such test-set programs have not occurred to date in SE Asia.

The aim of this study is therefore to develop an understanding of the efficacy of DC radiologists involved in breast cancer diagnosis by using a test set methodology. Performance of DC radiologists will be compared with that of Australian readers as well as those reporting in a DDC.

Materials and Methods

A test set containing 60 mammographic examinations, 40 of which were normal and 20 demonstrating cancers was used in this study. Cancer cases were verified using biopsy and normal cases were identified after a two year follow up. Each examination consisted of a two-view mammogram (cranio-caudal (CC) and medio-lateral oblique projections (MLO)) of both breasts. All Breastscreen Reader Assessment Strategy (BREAST) test images were acquired using digital mammography and were de-identified of all health record data. Cases with visible post-biopsy markers or surgical scars were excluded from the study.

The same test set was examined by 35, 15 and 53 radiologists from a DC, a DDC and Australia respectively. Each radiologist, without being informed of the prevalence of disease within the case set, read all images in his/her own native country.

Reading conditions were standardised through the following measures: reporting rooms had ambient lighting of no greater than 20 lux; mammograms were displayed using two high fidelity workstations, each driving two 5MP reporting monitors calibrated to the Greyscale Standard Display Function (DICOM GSDF). Online software, developed by University of Sydney, was used to present the test set images at full native resolution. Demographic information was collected from each participant at the start of each reading session via a questionnaire, which included information of reader's age, years of experience reading mammograms, number of mammograms read per week and number of hours of mammographic reading per week. This information did not include any participant identifiers.

Radiologists were asked to localise all detected lesions and give each marked location a score of 1-5 indicating their level of confidence that a lesion was present: a rating of 1 indicated complete confidence that the case was normal and 5 complete confidence that a cancer was present. Post-processing tools and unlimited time were provided to each radiologist. All performance data was anonymised and no link between the performance data and individual readers was made.

Institutional and ethics approvals were granted for the study and informed consent was obtained from each reader. The need for obtaining informed consent from patients whose mammograms used was waived by the New South Wales Cancer Institute.

Data gained from each individual radiologist was used to calculate jackknife free response operating characteristic figure-of-merit (JAFROC FOM), receiver operating characteristic curve area under the curve (ROC AUC), sensitivity, location sensitivity and specificity.

The ROC AUC is based upon whether a mammogram does or does not have cancer. ROC curves plot false positives against true positives, thus the area under the curve represents the overall accuracy of a test with a perfect test of 1.0 indicating high sensitivity and specificity. A ROC AUC score of 0.5 represents zero discrimination (Lalkhen and McCluskey, 2008). However, ROC does not take into account lesion location and therefore JAFROC is used as a measure of radiologists performance in detecting lesion location (Chakraborty and Yoon, 2009). These performance values as well as demographic data were then compared across radiologists' groups using the non-parametric two-tailed Kruskall Wallis test followed by Dunn's Multiple Comparisons test to compare pairs of data (e.g. DC vs Australian data). GraphPad© PRISM software was used for all statistical comparisons and a P-value of <0.05 was considered to be significant.

Results

Results are shown in Tables 1 and 2, whilst the significant findings are summarised below.

Performance Metrics

As shown in Table 1, the JAFROC analysis demonstrated significant lower scores for DC radiologists compared with their counterparts in both the DDC and Australia (P< 0.0001). The DC radiologists' ROC scores were significantly lower than the Australian (P= 0.0003) and DDC (P= 0.01) radiologists.

Whilst no significant differences were seen for the sensitivity values, the location sensitivity analysis yielded several statistically significant results: Australia demonstrated higher scores than both the DC (P< 0.0001) and DDC (P= 0.01) SE Asian country scores. Radiologists' scores in the DDC was higher than that of their DC

Table 1. Comparison of the Performance Metrics between Across the Three Groups of Radiologists (i.e. DC SE Asian, DDC SE Asian and Australian)

Metric	Radiologists		
	DC SE Asian	DDC SE Asian	Australian
JAFROC	0.47*,**	0.74 *	0.80 **
	(0.40 - 0.54)	(0.61 - 0.82)	(0.74 - 0.85)
ROC	0.81 *,**	0.87 *	0.87 **
	(0.74 - 0.84)	(0.81 - 0.92)	(0.80 - 0.92)
Sensitivity	0.8	0.8	0.85
	(0.65 - 0.95)	(0.70 - 0.90)	(0.78 - 0.95)
Location Sensitivity	0.48 *,**	0.67 *,***	0.81 **,***
	(0.38 - 0.57)	(0.48 - 0.76)	(0.67 - 0.88)
Specificity	0.70 *	0.85 *	0.8
	(0.5 - 0.88)	(0.83 - 0.93)	(0.73 - 0.88)

Median values are shown as well as the interquartile range displayed in brackets; *, Scores of radiologists from the DC SE Asian are significantly different (P<0.05) from radiologists from the DDC SE Asian country; **, Scores of radiologists from the DC SE Asian are significantly different (P<0.05) from radiologists from Australia; *** , Scores of radiologists from the DDC SE Asian are significantly different (P<0.05) from radiologists from Australia.

Table 2. Demographic Data for Each of theThree Groups of Radiologists

Radiologists	Age	Experience in reading mammograms		
		No. of years	Hours/ week	Cases per week
DC SE Asian				
Minimum	24	0	1	1
Median	31*,**	2*,**	1*,**	1**
Maximum	53	16	4	4
IQR	27-34	1-3	1-2	1-2
Mean	32.63	2.49	1.4	1.51
DDC SE Asian				
Minimum	34	2	1	1
Median	47*	15*	2*	2
Maximum	65	20	6	3
IQR	39-57	5-15	2-5	2
Mean	47.67	11.47	3.07	2
Australia				
Minimum	35	0	1	1
Median	51**	8**	2**	3**
Maximum	76	30	6	4
IQR	43-59	3-18.5	1-3	2-3
Mean	51.96	10.45	2.4	2.68

IQR, Interquartile Range; *, radiologists from the DC SE Asian are significantly different (P<0.05) from the DDC SE Asian country; **, radiologists from the DC SE Asian are significantly different (P<0.05) from Australian radiologists.

counterparts (P= 0.0079).

With regards to specificity, readers from the DDC demonstrated significantly higher scores than those from the DC (P= 0.008).

Demographic data

Table 2 shows that radiologists from the DC have lower values for age (P <0.0001), hours of reading per week (P<= 0.0004) and years of mammography experience (P <0.0001) when compared with their counterparts from the DDC and Australia. In addition, the readers from the DC read fewer mammographic cases per week (P <0.0001) than the Australian readers.

Discussion

This work examined the performance of radiologists based in a developing country when asked to diagnose breast cancer using mammographic images. As a baseline we compared their diagnostic efficacy against two countries with mature breast imaging training programs, one is a typical westernised country (Australia), and the other is a developed country located in SE Asia.

The DC radiologists displayed a significantly lower location sensitivity than both their DDC and Australian counterparts. Location sensitivity refers to the ability to accurately locate lesions, which has previously been linked to radiologists' ability to detect smaller more difficult lesions (Mello-Thoms et al., 2014). Importantly, this large difference in location sensitivity could have harmful implications for subsequent biopsy location and outcome actions for patients in developing countries. It may be tempting to highlight that in this study we showed no difference between the radiologist groupings for case-based sensitivity and that this metric is a key feature of any screening program, this result must be considered in light of the specificity results. Specificity for the DC radiologists was lower than the other two groupings (although only significant when compared to the DDC radiologists), which would imply that DC radiologists are recalling more women than DDC radiologists. This means they are sliding up their ROC curve, thus potentially inflating the case-based sensitivity figures, at the expense of accurately recognising the normal cases. The unintended harmful effects of unnecessary recalls include long term impacts on the patient's psychological well-being when compared with those who weren't recalled (Brodersen and Siersma, 2013).

Differences in the location sensitivity and specificity of the DC radiologists and their counterparts and subsequent effects on JAFROC and ROC values can be attributed to their experience in reading mammograms. Our study shows that the DC radiologists were significantly younger than their counterparts in both the other countries in addition to having fewer years of experience reading mammograms. These results confirm the findings of other authors who have linked experience with performance: the most experienced radiologists had better location sensitivity scores than the least experienced radiologists (Rawashdeh et al., 2013; Suleiman et al., 2014). Experience is a determining factor in the risk of false positives, with younger and less experience radiologists more likely to have higher recall rates (Elmore et al., 1998; Reed et al., 2010; Alberdi et al., 2011; Hawley et al., 2016) with direct relationships between reader volume (Rawashdeh et al., 2013; Reed et al., 2010) and radiologist performance. Clearly therefore making sure that the experience of radiologists specifically in breast reading is critically important. However, this cannot be achieved simply by allowing lots of young radiologists' freedom to continually report on many different types of images. Some level of specialism is required to develop and fine tune the reader skills required to achieve good levels of diagnostic efficacy for this radiologic domain.

The significant differences identified in this experiment also highlight the need for effective training and reading strategies to minimise the variance in radiologic diagnosis between DC radiologists and their Australian and DDC counterparts. In the situation where readers have low levels of experience and less dedicated time devoted to reading mammograms, one approach would be the development of innovative, interactive training programs that imposes little expense and inconvenience. For example, the BREAST training programs (Suleiman et al., 2014) available to radiologists in several countries including Australia, New Zealand, Singapore and the Middle East would allow clinicians wherever they are located to login in a confidential way and in their own environment and test their ability to diagnose mammograms using several available test sets. Immediate feedback is available with details on performance levels as well as clear, localised information provided on reader-specific errors for each image diagnosed. As shown to be the case elsewhere (Suleiman et al., 2014), this approach should increase the efficacy of the DC radiologists in demonstrated areas of weakness and has been shown to be a promising method in training radiologists and residents both by the current authors and others (Suleiman et al., 2014; Poot and Chetlen, 2016). Another potential solution would be to have radiologists performing double readings with consensus decisions on cases, as is the case in most non-US westernised countries. This approach has shown to increase detection rates, whilst minimising recall rates (Anttinen et al., 1993), however the authors acknowledge that there are resource implications associated with this approach as well as a potential delay in diagnoses, so the feasibility of such an approach in a developing country would need to be fully evaluated. Alternatively, utilizing computer assisted diagnosis (CAD) as a first reading may be more feasible in DC's. However, the use of CAD even with experienced radiologists has shown to increase recall rates (Bargalló et al. 2014). Finally, establishing National Accreditation Standards around minimum levels of readings per year to be proficient at diagnosing breast cancer may also be a useful strategy. In countries such as Australia, it is clearly defined that breast reading radiologists should read a minimum number of 2,000 cases per year (Reed et al., 2010) with larger numbers being stated elsewhere. To achieve such numbers however, it may be necessary to reduce the numbers of individuals in DC countries reading mammograms so that minimum numbers of readings can be achieved per radiologist.

There are a number of limitations of this study that need to be acknowledged. It is noted for example that this study's results rely on a test set methodology to describe performance rather than clinical audit data, however work elsewhere has shown a strong agreement in performance across these two environments (Soh et al., 2013). Also, reader performance may be attributed to DC radiologists interpreting mammograms from unfamiliar populations, since all the images came from Australian clinics where for example the mammographic density may be lower than that seen typically in their own country. It is important therefore that future studies should develop test-sets based on populations that radiologists most commonly work in. Furthermore, for statistical robustness, the BREAST test-set has a much higher rate of cancer than typically found in clinical situations and such higher prevalence may affect performance, although the impact of such prevalence has previously been shown to be minimal (Reed et al., 2010).

In summary this work has shown that mammographic diagnostic efficacy in a DC may not be as high as levels demonstrated within developed countries. This discrepancy may be linked to experience levels and some corrective strategies have been suggested. The solution however requires a collaborative approach to embrace educational, professional and regulatory components.

Acknowledgements

The authors would like to acknowledge the participating radiologists for their contribution to this study. This study would not have been possible without the generous funding of the National Breast Cancer Foundation (NBCF) and the Sydney South East Asia Centre. The authors decided to exclude the names of the countries so that specific health systems that were not performing as well as others would not be identified. The aim is not to identify where low performance is occurring, but rather that it is occurring and the breast imaging community must come together to establish effective educational strategies.

References

Alberdi R, Llanes A, Ortega R et al (2011). Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. *Eur Radiol*, **21**, 2083-90

Anttinen I, Pamilo M, Soiva M, Roiha M (1993). Double reading of mammography screening films-one radiologist or two?. *Clin Radiol*, **48**, 414-21.

Bargalló X, Santamaría G, del Amo M, et al (2014). Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. *Eur J Radiol*, **83**, 2019-2023.

Bray F, McCarron P, Parkin D (2004). The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Res*, **6**, 229-39.

Brodersen J, Siersma V (2013). Long-term psychosocial consequences of false-positive screening mammography. *Ann Fam Med*, **11**, 106-15.

Chakraborty DP, Yoon H-J (2009). JAFROC analysis revisited: figure-of-merit considerations for human observer studies. *Proc SPIE*, **7263**, 72630T1-12.

Elmore J, Wells C, Howard D (1998). Does diagnostic accuracy in mammography depend on radiologists' experience?. *J Womens Health*, **7**, 443-9.

Ferlay J, Soerjomataram I, Dikshit R, et al (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, **136**, 359-86.

Hawley J, Taylor C, Cubbison A, et al (2016). Influences of radiology trainees on screening mammography interpretation. *J Am Coll Radiol*, **13**, 554-61.

Lalkhen AG, McCluskey A (2008). Clinical tests: sensitivity and specificity. *BJA Educ*, **8**, 221-3.

Mello-Thoms C, Trieu PD, Rawashdeh MA, et al (2014) Understanding the role of correct lesion assessment in radiologists' reporting of breast cancer. In: Fujita H., Hara T., Muramatsu C. (eds) Breast Imaging. IWDM 2014. *Lect Notes in Comput Sc*, **8539**, 341-7.

Nguyen B (2009). Breast cancer situation in women in some provinces/cities from 2001 to 2007. *Vietnamese J Oncol*, **1**, 5-11.

Poot J, Chetlen A (2016). A simulation screening mammography module created for instruction and assessment. *Acad Radiol*, *23*, 1454-62.

Rawashdeh M, Lee W, Bourne R, et al (2013). Markers of good performance in mammography depend on number of annual readings. *Radiology*, **269**, 61-7.

Reed W, Lee W, Cawson J, Brennan P (2010). Malignancy detection in digital mammograms. *Acad Radiol*, **17**, 1409-13.

Soh B, Lee W, McEntee M, et al (2013). Screening mammography:

test set data can reasonably describe actual clinical reporting. *Radiology*, **268**, 46-53.

Soh B, Lee W, Wong J, et al (2016). Varying performance in mammographic interpretation across two countries: Do results indicate reader or population variances? Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment. 97870X

Suleiman W, Lewis S, Georgian-Smith D, Evanoff M, McEntee M (2014). Number of mammography cases read per year is a strong predictor of sensitivity. *J Med Imag*, **1**, 015503.

Trieu P, Mello-Thoms C, Brennan P (2105). Female breast cancer in Vietnam: a comparison across Asian specific regions. *Cancer Biol Med*, **12**, 238–45.

Vuong D, Velasco-Garrido M, Lai T, Busse R (2010). Temporal trends of cancer incidence in Vietnam, 1993-2007. *Asian Pac J Cancer Prev*, **11**, 739-45.

World Health Organization (2017). Global Health Estimates. Available at: http://www.who.int/healthinfo/global_burden_disease/en/. Accessed June 25, 2017.