Journal of Medical Radiation Sciences

Open Access

ORIGINAL ARTICLE

Evaluating radiographers' diagnostic accuracy in screen-reading mammograms: what constitutes a quality study?

Josephine C. Debono, AssocDip(DR), MAppSc[1] & Ann E. Poulos, PhD, BAHons, DipEd, DipRad[2]

[1]Westmead Breast Cancer Institute, Westmead, New South Wales, Australia
[2]Discipline of Medical Radiation Sciences, Faculty of Health Sciences, University of Sydney, Lidcombe, New South Wales, Australia

Keywords
Accuracy, evaluation tools, mammogram, quality, radiographers, screen-readers

Correspondence
Josephine C. Debono, Westmead Breast Cancer Institute, PO Box 143, Westmead, New South Wales 2145, Australia.
Tel: +61 402 232 511; Fax: + 61 2 9845 8491;
E-mails: josephine.debono@bci.org.au; jodebono61@gmail.com

Funding Information
No funding information provided.

Received: 17 December 2013; Revised: 30 June 2014; Accepted: 1 July 2014

J Med Radiat Sci **62** (2015) 23–31

doi: 10.1002/jmrs.68

Abstract

Introduction: The aim of this study was to first evaluate the quality of studies investigating the diagnostic accuracy of radiographers as mammogram screen-readers and then to develop an adapted tool for determining the quality of screen-reading studies. **Methods:** A literature search was used to identify relevant studies and a quality evaluation tool constructed by combining the criteria for quality of Whiting, Rutjes, Dinnes et al. and Brealey and Westwood. This constructed tool was then applied to the studies and subsequently adapted specifically for use in evaluating quality in studies investigating diagnostic accuracy of screen-readers. **Results:** Eleven studies were identified and the constructed tool applied to evaluate quality. This evaluation resulted in the identification of quality issues with the studies such as potential for bias, applicability of results, study conduct, reporting of the study and observer characteristics. An assessment of the applicability and relevance of the tool for this area of research resulted in adaptations to the criteria and the development of a tool specifically for evaluating diagnostic accuracy in screen-reading. **Conclusions:** This tool, with further refinement and rigorous validation can make a significant contribution to promoting well-designed studies in this important area of research and practice.

Introduction

Diagnostic accuracy in medical imaging is essential for appropriate patient management and treatment.[1] Accurate screen-reading of mammogram images is critical for the early detection of breast cancer, the goal of population screening programs.[1] Screen-readers of mammogram images are predominantly, but not exclusively, radiologists.[2] Currently, there are workforce issues in radiology which impact on their availability for screen-reading.[3] In the United Kingdom, this shortage has been addressed by the training and employment of radiographers as screen-readers.[2,4,5] A range of studies have investigated the diagnostic accuracy of radiographers in this role,[2,6–15] and provide evidence that radiographers have comparable accuracy to radiologists.[2,6–15] More recent studies provide evidence of the ability of radiographers to contribute to improvements in the efficiency of the screening process, and most importantly that combining radiologist and radiographer screen-reading has been found to improve cancer detection rates.[7–9]

Since diagnostic accuracy in screen-reading underpins the goal of breast screening to detect breast cancer early and reduce mortality, the quality of these studies is paramount. A systematic review published in 2008 by van den Biggelaar et al.[16] excluded articles without evidence of sensitivity and specificity and an appropriate gold standard, resulting in a total of six. This systematic review raised questions of what constitutes a well-designed study and how quality is defined in studies investigating screen-reading accuracy by radiographers. More specifically, the authors emphasised the necessity of determining the key components of a well-designed study in this area of

research to increase the rigour and applicability of the outcomes to the clinical environment.

Quality evaluation tools for studies of diagnostic accuracy

A number of tools for evaluating the quality of diagnostic accuracy studies have been identified in the literature.[17] The Standards for the Reporting of Diagnostic accuracy studies (STARD), was developed from an initiative to improve the accuracy and completeness of reporting studies of diagnostic accuracy.[18] The Quality of Diagnostic Accuracy Studies (QUADAS) tool was later developed and validated by Whiting et al.[19] to determine the quality of primary studies in systematic reviews of diagnostic accuracy.

Subsequently, Whiting, et al.[17] conducted a systematic review of existing quality assessment tools to examine both the extent and type of quality assessment being incorporated in diagnostic accuracy systematic reviews. Aspects of quality considered in their review were classified as: potential for bias; conduct of the study; applicability of the results; and quality of reporting. Following data extraction, the data were synthesised according to purpose and summarised as items.[19]

This classification is useful since it is all-inclusive and includes items of quality drawn from an extensive review of systematic reviews. As well as determining the individual items related to quality in diagnostic accuracy studies, the classification also synthesises these items into aspects of quality. Importantly, this classification includes quality items relating to the reporting of studies.[20] The comprehensive nature of this classification facilitates the adaption of the quality items or criteria to a specific area of diagnostic accuracy research such as screen-reading.

Importance of observer characteristics and variability

The importance of observer characteristics and variability on diagnostic accuracy in medical imaging have been emphasised by Brealey and Westwood,[21] who claim that observers are frequently ignored in diagnostic accuracy studies in medical imaging in spite of their ability to affect the study outcomes. The number of observers, for example, influences the internal and external validity of research studies, while the profession and experience of observers affect estimates of accuracy.[22,23] Brealey and Westwood[21] strongly recommend the inclusion of observer assessment criteria in a quality assessment tool evaluating diagnostic accuracy in medical imaging.

The aim of this study was firstly to evaluate the quality of the studies investigating the diagnostic accuracy of radiographers as screen-readers using a quality evaluation tool constructed by combining the criteria for quality of Whiting et al.[17] and Brealey and Westwood.[21] Secondly, the applicability and appropriateness of the criteria were determined and an adapted quality evaluation tool was developed specifically for use in evaluating diagnostic the accuracy in screen-reading studies.

Method

Stage 1: Quality evaluation tool for studies using imaging

To construct this quality evaluation tool the classifications and items of Whiting et al.[17] were combined with the observer characteristics recommended by Brealey and Westwood[21] to provide a comprehensive all-inclusive quality assessment tool for diagnostic accuracy studies using imaging. An Ethics statement is not applicable to this study.

Stage 2: Literature search

A literature search was undertaken within the Medline, PubMed, Web of Science and Cinahl databases, using combinations of the terms: mammogram, radiographer, technologist, screen-reading, accuracy and interpretation. This search was undertaken in 2010, and therefore limited to articles published at that time. An initial review of titles and abstracts enabled the exclusion of papers that were clearly not relevant to the subject of interest. Studies investigating the diagnostic accuracy of radiographers reading mammograms were selected. Further studies were located using the reference lists. As only a small number ($n = 11$) of papers were located, no further inclusion/exclusion criteria were applied.

Stage 3: Quality evaluation of reviewed studies

The quality evaluation of studies was carried out by two experienced researchers. The role of these researchers was firstly to adapt the 'generic' diagnostic accuracy study quality items in Table 1 to specific criteria of quality in radiographer mammography screen-reading studies. This required knowledge and experience in mammography and the diagnostic process of screen-reading. Secondly, these researchers required research skills and experience in critical analysis of the reviewed studies to determine the extent to which they complied with the quality criteria. Finally, knowledge and familiarity with current relevant literature was required for stage 4. Any variation between the researchers was dealt with by discussion and consensus.

Table 1. Classification of items included in quality assessment tools (Source, with permission: Whiting et al.[17] p. 3, © 2005, Elsevier) plus observer characteristics (Source, with permission: Brealey and Westwood[21] p. 676, © 2006, the British Institute of Radiology).

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

Stage 4: Development of quality assessment tool for mammography screen-reading

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

relevant literature was carried out for evidence of specific quality criteria.

Results and Discussion

Stage 1: Development of quality evaluation tool

The developed tool is presented in Table 1.

Stage 2: Literature search

Eleven studies were identified in the literature relating to the diagnostic accuracy of radiographers reading screening mammograms and are presented in Table 2. No studies were excluded from the review.

Stage 3: Quality evaluation of reviewed studies

Study quality of each of the 11 studies was evaluated using the developed tool (stage 2 of method); the results of these evaluations are presented in Table 3.

The 'Total' row under each category A–E indicates the numbers of negative responses to the criteria for each study, while the numbers in the final 'Total' column indicate the number of negative responses to each of the 34 criteria. If a partial negative response was indicated then 0.5 was allocated. This relates positively to the scoring method used by Whiting et al.[24]

The 'scoring' of quality is fraught with difficulties in interpretation. Whiting et al.[25] emphasised the need to investigate individual quality items and their association with estimates of diagnostic accuracy rather than produce scores. So while identification of negative responses to criteria is a simplistic method of scoring quality, for the purposes of this study this method has provided detail of the categories demonstrating a large number of negative responses.

The largest number of negative responses is identified in section A: *Potential for bias*. Potential bias can severely compromise outcomes and must be minimised wherever possible. Bias can be minimised by ensuring the research design is similar to the screen-reading process in practice, using criteria A1–A8 (Reference Standard; Disease Progression bias; Verification bias; Incorporation bias; Review bias; Clinical review bias; Instrument variation. A5 was not applicable). The highest number of negative responses for bias potential were A4 (Incorporation bias: 7.5), A8 (Observer/instrument variation: 11) and A9 (Handling of uninterpretable results: 11). Incorporation bias (A4) did occur in the studies since it is an immutable aspect of the screen-reading process. Potential confounders, which affect test performance and relate to the varying classification systems used (A8), can be reduced by using a validated reporting instrument such as the BIRADS® (Reston, Virginia) classification lexicon.[26] Uninterpretable results (A9) were not included in the reviewed studies since the results were known prior to the test.

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

Table 2. Screen-reading studies, in chronological order.

Authors	Title
Haiart and Henderson[10]	A comparison of interpretation of screening mammograms by a radiographer, a doctor and a radiologist
Bassett et al.[6]	Effects of a program to train radiologic technologists to identify abnormalities on mammograms
Pauli et al.[2]	Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening
Pauli et al.[12]	Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions
Tonita et al.[14]	Medical radiologic technologist review: effects on a population-based breast cancer screening program
Wivell et al.[15]	Can radiographers read screening mammograms?
Sumkin et al.[13]	Prescreening mammography by technologists: a preliminary assessment
Holt[11]	Evaluating radiological technologists' ability to detect abnormalities in film-screen mammographic images: A decision analysis pilot project
Duijm et al.[7]	Additional double reading of screening mammograms by radiologic technologists: impact on screening performance parameters
Duijm et al.[8]	Introduction of additional double reading of mammograms by radiographers: effects on a biennial screening programme outcome
Duijm et al.[9]	Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome

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

Study	Haiart and Henderson[10]	Bassett et al.[6]	Pauli et al.[2]	Pauli et al.[12]	Tonita et al.[14]	Wivell et al.[15]	Sumkin et al.[13]	Holt[11]	Duijm et al.[7]	Duijm et al.[8]	Duijm et al.[9]	Total
A. Potential for bias												
A1	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	1
A2	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	1
A3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
A4	–	✓	–	Partial	–	✓	–	✓	–	–	–	7.5
A5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
A6	✓	✓	✓	✓	Partial	✓	✓	✓	✓	✓	Partial	0
A7	N/S	–	✓	✓	N/S	✓	✓	–	✓	✓	✓	2
A8	–	–	–	–	–	–	–	–	–	–	–	11
A9	–	–	–	–	–	–	–	–	–	–	–	11
A10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
Total	3	2	3	2.5	3.5	2	5	3	3	3	3.5	33.5
B. Applicability of results												
B1	✓	N/S	✓	✓	✓	✓	✓	–	✓	✓	✓	1
B2	✓	N/S	✓	✓	✓	✓	–	–	✓	✓	✓	2
B3	✓	–	✓	✓	✓	✓	–	–	✓	✓	✓	3
B4	–	–	–	–	–	–	–	–	–	–	–	11
Total	1	2	1	1	1	1	3	4	1	1	1	17
C. Conduct of the study												
C1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
C2	–	✓	✓	✓	–	–	✓	–	✓	✓	✓	4
C3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
C4	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	0
Total	1	0	0	0	1	1	0	1	0	0	0	4
D. Reporting of the study												
D1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D2	Partial	✓	✓	✓	Partial	✓	✓	✓	✓	✓	✓	1
D3	✓	✓	✓	–	✓	✓	✓	✓	✓	✓	✓	1
D4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
D5	✓	✓	✓	✓	–	–	–	✓	✓	–	✓	4
D6	–	–	–	–	✓	–	–	–	✓	–	✓	8
D7	–	–	✓	–	–	–	–	–	–	–	–	10
D8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
D9	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
Total	2.5	2	1	3	2.5	3	3	2	1	3	1	24
E. Observer characteristics												
E1	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	1
E2	✓	✓	Partial	✓	✓	✓	✓	✓	✓	✓	✓	5
E3	–	–	✓	✓	–	–	✓	✓	✓	✓	✓	4
E4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2
E5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
E6	✓	✓	–	–	–	–	–	✓	–	–	✓	7
E7	–	–	–	–	–	✓	–	–	–	–	–	10
Total	2	2	2.5	2	3	2	4	2	2	2	1	24.5

N/S, not stated; N/A, not applicable.

Negative responses in section B: *Applicability of results*, were the highest in B4 (Change in technology of index test). This is explained by the introduction of digital technology since the reviewed articles were published. It is possible that this new technology may provide increased diagnostic accuracy in screen-reading, and so results of the reviewed studies may not be generalisable to facilities using the digital equipment. This change, however, did not influence the applicability of the results at the time of publication.

Section C: *Conduct of the study* demonstrated low but significant negative responses to C2 (Sample size). Appropriate sample size is a critical component of a research study and in the field of research covered by the reviewed studies, sample refers to both number of images read and number of observers reading the images. This criterion, therefore, requires clarification.

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

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

In summary, these evaluation results emphasise the need for a specific evaluation tool for diagnostic accuracy in screen-reading. The specific screen-reading processes which minimise bias can be clearly enunciated, appropriate

Table 4. Developed tool named DASQUART (*Diagnostic Accuracy Study QUality And Reporting Tool)* for determining quality in studies investigating diagnostic accuracy in screen-reading.

	Criteria	Description of criteria
A1	Reference standard	An appropriate reference standard of pathology and at least 1 year follow-up used to determine the presence or absence of breast cancer
A2	Disease Progression bias	An interval cancer could not occur between the initial mammogram and the reference standard
A3	Verification bias	Same reference standard applied across the study
A4	Incorporation bias	The reading of the screening mammogram does not form part of the reference standard
A6	Review bias	Mammograms read blinded to knowledge of reference standard and interpretation by other readers
A7	Clinical review bias	Previous image rounds available for comparison
A8	Instrument variation	No reporting instrument variation which will affect estimates of test performance, e.g., use of BIRADS® lexicon[26]
A9	Handling of uninterpretable results	Uninterpretable results included in the analysis
A10	Arbitrary choice of threshold value	Threshold value of normal chosen independently of results
B1	Spectrum composition	Image sample similar to one of interest (test sets, e.g., PERFORMS,[31] BREAST[32] and consecutive screening)
B2	Population recruitment	Image sample selected adequate to include appropriate spectrum (test sets, e.g., PERFORMS,[31] BREAST[32])
B3	Disease prevalence/severity	Spectrum of breast cancer prevalence similar to one of interest (test sets, e.g., PERFORMS,[31] BREAST[32] and consecutive screening)
B4	Change in technology of index test	No change in mammography technology which will affect applicability of results
C1	Subgroup analysis	Subgroup analyses were appropriate and specified
C2	Sample size	Appropriate number of images included in study
C3	Objectives	Study objectives relevant to study question
C4	Study design	The purpose, method, results and conclusions demonstrate logical coherence and consistency
D1	Inclusion criteria	Included in systematic reviews
D2	Test execution (a) images	Sufficient details of mammogram reading reported to permit its replication. Details include number of images read in total and at one sitting, how images were selected (test sets), degree of difficulty (test sets), types of breast cancers included (test sets).
D2	Test execution (b) environment	Time taken to read, background lighting and type of monitors
D3	Reference execution	Sufficient details provided of reference standard used to permit its replication
D4	Normal defined	Authors clearly reported what was considered a normal reading result
D5	Appropriate results	Appropriate results of accuracy presented, e.g., sensitivity, specificity, ROC and JAFROC analysis
D6	Precision of results	Estimate of precision of results presented as appropriate
D7	Drop-outs	All images and observers accounted for
D8	Data table	Test performance reported in a data table
D9	Utility of test	Clinical relevance of the test emphasised
E1	Image allocation to observers	Image allocation to observers described
E2	Number of observers	Number of observers presented
E3	Observer experience	Experience of observers described
E4	Observer training	Training of observers described
E5	Observer profession	Profession of observers presented
E6	Analysis of observer variability	Observer variability in analysis, e.g., Kappa statistic

sample sizes of images and observers identified and criteria relating to study reporting increased in relevance.

Stage 4: Development of quality assessment tool for screen-reading studies

The development of topic-specific quality evaluation tools for diagnostic accuracy studies has been supported by Whiting et al.[24] The quality tool used to evaluate the reviewed studies (Table 1) was adapted to provide a specific tool for diagnostic accuracy studies in screen-reading and for ease of identification has been named the DASQUART (Diagnostic Accuracy Study Quality And Reporting Tool). The DASQUART is presented in Table 4.

The quality evaluation criteria of Whiting et al.[17] and additional criteria related to medical imaging of Brealey and Westwood[21] have been adapted to enhance relevance, clarity and precision and to contribute to the development of a user-friendly quality assessment tool. These adaptations are described below.

Changes to existing criteria

To maintain consistency in the structure of the tool, definitive statements rather than questions are presented throughout as descriptions of criteria. A positive response to these statements indicates an aspect of quality. Criteria for which a negative response indicates quality have been changed (A4: Incorporation bias, A8: Observer/instrument variation and B4: Change in technology of index test). One criterion not relevant to this area of study has been removed (A5: Treatment paradox) since treatment does not typically begin until verification has been made through pathology results. Criterion A8 of observer variation is similar to criteria E1–E7 of observer characteristics and has been removed. Only instrument variation, specifically the reporting form used to interpret the images, now comprises A8. For criterion C2 (Sample size), participants are changed to images while number of observers (screen-readers) comprises E2. The inclusion in D5 (Appropriate results) of receiver operating characteristic (ROC) and Jackknife Free-response Receiver Operating Characteristic (JAFROC) analysis rigorously assesses observer accuracy. This method allows quantitative analysis of observers interpreting images which could contain more than one lesion. For D7 (Drop-outs), patients are replaced by images and observers.

Additional criteria

Criterion D2 (Test execution) now provides further detail to allow replicability as well as identify variables which influence diagnostic accuracy to further adapt the tool to this area of study. Details included are related to the screen-reading process and include: number of images read at one sitting; how images were selected; degree of difficulty of interpretation; details of types of breast cancer; time taken to read; and environmental conditions such as lighting and type of monitors. Observer variability among radiologists has been found to be related to years of experience and numbers of images read[22,23] and so these criteria have been added to E: Observer characteristics.

Evidence for criteria

This adaptation has been carried out using evidence from the literature: van den Biggelaar et al.[16] (A1: Reference Standard, D5: Appropriate results), Reed et al.[27] (D2: Test execution) and Brealey and Westwood[21–23] (E1–E7: Observer Characteristics). As well, details of the breast screening process contained within the BreastScreen Australia National Accreditation Standards (NAS)[1] were also used. The NAS is not only based on rigorous international evidence relating to best practice[1] but also encourages the research design in these studies to mimic the real-life environment of screen-reading and consequently provide the most clinically useful outcomes.

One aspect of the screen-reading process which is typically impractical for research purposes is screen-reading consecutive populations. This has led to the use of test sets in research studies. However, for these studies to be clinically useful a correlation between test set results and real-life clinical results is essential.

Test sets and clinical practice

Much debate surrounds the testing of diagnostic accuracy using test sets which have artificially inflated breast cancer prevalence versus consecutive screening images which mimic the real-life clinical situation. Minimal or no correlation between test set outcomes and clinical outcomes has been identified by Scott et al.,[28] Rutter and Taplin.[29] Gur et al. reported a significant difference between performance in the clinic than completing test sets.[30] A study by Pauli et al. found a strong correlation between test set outcomes and consecutive screening outcomes when used together in the same research design.[12] These studies, however, used varying numbers of breast cancers, images and types of breast cancer to comprise the test set.

This variation can be overcome by the use of a validated test set such as PERFORMS (Scott and Gale)[31] and BREAST (Brennan, Lee and Tapia)[32] which increases the rigour of the study and provides consistency in the

important aspects of study as *spectrum composition, spectrum of images and spectrum of disease* (B1–B3). The degree of difficulty in terms of types of cancers, proportions of breast density and numbers of images read would be consistent and so comparisons between study outcomes could be more readily applied.

Incorporating validated test sets into the quality evaluation tool specifically developed to evaluate screen-reading accuracy, may well lead to an identification and understanding of the specific causal agents for any lack of correlation between clinical audits and screen-reading test sets, which as Soh et al. state is needed to facilitate the process of evaluating the diagnostic accuracy of screen-readers in practice.[20]

Validation of DASQUART

Since this reported analysis was carried out, an updated version of the QUADAS tool has been developed by Whiting et al. (*plus other authors*) and named Quadas-2 which includes a number of additional and improved features particularly focusing on bias.[33] A review in 2013 by the same authors provides an updated classification and overview of the sources of bias and variation in test accuracy studies.[34] As part of the validation process of the DASQUART careful analysis of these two Whiting et al. (*plus other authors*) studies should be undertaken to identify any further classifications and adaptations required to improve the tool.

Conclusion

This reported study has developed a quality assessment tool specifically for evaluating the quality of studies investigating the diagnostic accuracy of screen-readers. This tool now needs further refinement and rigorous validation processes including critical evaluation by a panel of clinical experts in the area of screen-reading. A limitation of this study is the focus on evaluation of studies only involving radiographers as screen-readers; future research should include a quality evaluation of studies conducted by radiologists as screen-readers to provide evidence for further refinement of the tool.

This tool, with further refinement and validation, can make a significant contribution to promoting well-designed quality studies in this important area of research and practice. Most importantly it can facilitate consistency in study design which can increase the rigour and applicability of the outcomes to the clinical environment.

Conflict of Interest

The authors declare no conflict of interest.

References

1. BreastScreen Australia. BreastScreen Australia National Accreditation Guidelines. BreastScreen Australia Quality Improvement Program [Internet]. 2008 [p. 43]. Available from: http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/br-accreditation/$File/standards.pdf. (accessed 2013 December 2).
2. Pauli R, Hammond S, Cooke J, Ansell J. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. *J Med Screen* 1996; **3**: 18–22.
3. RANZCR. 2010 RANZCR Radiology Workforce Report. Australia, 2010.
4. Price R, Miller L, Mellor F. Longitudinal changes in extended roles in radiography. *Radiography* 2002; **8**: 223–34.
5. Bennett R, Sellars S, Blanks R, Moss S. An observational study to evaluate the performance of units using two radiographers to read screening mammograms. *Clin Radiol* 2011; **67**: 114–21.
6. Bassett L, Hollatz-Brown A, Bastani R, Pearce J, Hirji K, Chen L. Effects of a program to train radiologic technologists to identify abnormalities on mammograms. *Radiology* 1995; **194**: 189–92.
7. Duijm L, Groenewoud J, Fracheboud J, de Koning H. Additional double reading of screening mammograms by radiologic technologists: impact on screening performance parameters. *J Natl Cancer Inst* 2007; **99**: 1162–70.
8. Duijm L, Groenewoud J, Fracheboud J, van Ineveld B, Roumen R, de Koning H. Introduction of additional double reading of mammograms by radiographers: effects on a biennial screening programme outcome. *Eur J Cancer* 2008; **44**: 1223–8.
9. Duijm L, Louwman M, Groenewoud J, van de Poll-Franse L, Fracheboud J, Coebergh J. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. *Br J Cancer* 2009; **100**: 901–7.
10. Haiart DC, Henderson J. A comparison of interpretation of screening mammograms by a radiographer, a doctor and a radiologist. *Br J Clin Pract* 1991; **45**: 43–5.
11. Holt JJ. Evaluating radiological technologist's ability to detect abnormalities in film-screen mammographic images: a decision analysis pilot project. *Can J Med Radiat Technol* 2006; **37**: 24–9.
12. Pauli R, Hammond S, Cooke J, Ansell J. Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions. *Br J Radiol* 1996b; **69**: 10–4.
13. Sumkin JH, Klaman HM, Graham M, et al. Prescreening mammography by technologists: a preliminary assessment. *AJR Am J Roentgenol* 2003; **180**: 253–6.

14. Tonita J, Hillis J, Lim C. Medical radiologic technologist review: effects on a population-based breast cancer screening program. *Radiology* 1999; **211**: 529–33.
15. Wivell G, Denton E, Eve C, Inglis J, Harvey I. Can radiographers read screening mammograms? *Clin Radiol* 2003; **58**: 63–7.
16. van den Biggelaar F, Nelemans P, Flobbe K. Performance of radiographers in mammogram interpretation: a systematic review. *Breast* 2008; **17**: 87–92.
17. Whiting P, Rutjes A, Dinnes J, Reitsma J, Bossuyt P, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005a; **58**: 1–12.
18. Bossuyt P, Reitsma J, Bruns D, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Radiol* 2003; **58**: 575–80.
19. Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BCM Res Methodol* 2003; **3**: 25.
20. Soh B, Lee W, Kench P, et al. Assessing reader performance in radiology, an imperfect science: lessons from breast screening. *Clin Radiol* 2012; **67**: 623–8.
21. Brealey S, Westwood M. Are you reading what we are reading? The effect of who interprets medical images on estimates of diagnostic test accuracy in systematic reviews. *Br J Radiol* 2007; **80**: 674–7.
22. Rawashdeh M, Lee W, Bourne R, et al. Markers of good performance in mammography depend on number of annual readings. *Radiology* 2013; **269**: 61–7.
23. Reed W, Lee W, Cawson J, Brennan P. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. *Acad Radiol* 2010; **17**: 1409–13.
24. Whiting P, Rutjes A, Dinnes J, Reitsma J, Bossuyt P, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004; **8**: 1–234.
25. Whiting P, Harbord R, Kleijen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BCM Res Methodol* 2005; **5**: 19.
26. American College of Radiology. BI-RADS-Mammography, 4th edn. 2003. Available from: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/BIRADSAtlas/BIRADSAtlasexcerptedtext/BIRADSMammographyFourthEdition.aspx. (accessed 2010 February 4).
27. Reed W, Poulos A, Rickard M, Brennan P. Reader practice in mammography screen reporting in Australia. *J Med Imaging Radiat Oncol* 2009; **53**: 530–7.
28. Scott H, Evans A, Gale A, Murphy A, Reed J. The relationship between real life breast screening and an annual self assessment scheme. *Proc SPIE* 2009; **7263**: 72631E-1.
29. Rutter C, Taplin S. Assessing mammographers' accuracy: a comparison of clinical and test performance. *J Clin Epidemiol* 2000; **53**: 443–50.
30. Gur D, Bandos A, Cohen C, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008; **249**: 47–53.
31. Scott H, Gale A. Breast screening: PERFORMS identifies key mammographic training needs. *Br J Radiol* 2006; **79**: S127–33.
32. Brennan P, Lee W, Tapia K. BreastScreen Reader Assessment Strategy – BREAST. University of Sydney. 2012. Available from: http://www.sydney.edu.au/health_sciences/breastaustralia. (accessed 2012 September 16).
33. Whiting P, Rutjes A, Westwood M, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155**: 529–36.
34. Whiting P, Rutjes A, Westwood M, Mallett S, Group Q-S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* 2013; **66**: 1093–104.