*Title Page (with author details)

Mammographic image quality in relation to positioning of the breast: A multicentre international evaluation of the assessment systems currently used, ~~for assessing mammographic image quality~~ to provide an evidence base for establishing a standardised method of assessement.

Kathryn Taylor[a] Dr Deepak Parashar[b] George Bouverat[c] Dr Ann Poulos[d] Randi Gullien[e] Elizabeth Stewart[f] Rand Aarre[g] Dr Pavel Crystal[h] Dr Matthew Wallis[a]

[a]Cambridge Breast Unit, Addenbrookes Hospital, Hills Road, Cambridge, CB2 0QQ, ENGLAND

[b]Cancer Research Unit, Division of Health Sciences, Warwick Medical School, The University of Warwick, Coventry, CV4 7AL, England

[c]Statistics and Epidemiology Unit, Division of Health Sciences, Warwick Medical School University of Warwick, Coventry CV4 7AL, England

[d]Faculty of Health Sciences, The University of Sydney, Camperdown NSW 2006, Australia

[e]Breast Imaging Centre, Oslo University Hospital, Oslo, NORWAY

[f]Breast screen Victoria, 78 Gheringhap Street, Geelong 3220 **VIC** Australia

[g]Hvidovre Hospital, Mammografiklinikken, afsnit 034, Pavillon 1, Østvej, 2650 Hvidovre, Denmark

[g]Mount Sinai Hospital, Women's College Hospital, Room 574, Mount Sinai Hospital, 600 University Avenue, Toronto, ON M5G 1X5

kathryn.taylor@addenbrookes.nhs.uk +44 01223 217627

D.parashar@warwick.ac.uk

g.bouverat@warwick.ac.uk

ann.poulos@sydney.edu.au

uxraul@ous-hf.no

estewart@breastscreen.org.au

Randi.Aarre@regionh.dk

Pavel.crystal@sinaihealthsystem.ca

Matthew.wallis@addenbrookes.nhs.uk

Corresponding author: Kathryn Taylor

Word count: 3213

Abstract: 244

*Abstract

ABSTRACT

Introduction

Optimum mammography positioning technique is necessary to maximise cancer detection. Current criteria for mammography appraisal lack reliability and validity with a need to develop a more objective system.

We aimed to establish current international practice in assessing image quality (IQ), of screening mammograms then develop and validate a reproducible assessment tool.

Methods

A questionnaire sent to centres in countries undertaking population screening identified practice, participants for an expert panel (EP) of radiologists/radiographers and a testing panel (TP) of radiographers. The EP developed category criteria and descriptors using a modified Delphi process to agree definitions.

The EP scored 12 screening mammograms to test agreement then a main set of 178 cases. Weighted scores were derived for each descriptor enabling calculation of numerical parameters for each new category. The TP then scored the main set. Statistical analysis included ANOVA, t-tests and Kendalls coefficient.

Results

11 centres in 8 countries responded forming an EP of 7 members and TP of 44 members.

The EP showed moderate agreement when the scoring the mini test set $W=0.50$ $p<0.001$ and the main set $W=0.55$ $p<0.001$, 'posterior nipple line' being the most difficult descriptor.

The weighted total scores differentiated the 4 new categories Perfect, Good, Adequate and Inadequate ($p<0.001$).

Conclusion

We have developed an assessment tool by Delphi consensus and weighted consensus criteria. We have successfully tabulated a range of numerical scores for each new category providing the first validated and reproducible mammography IQ scoring system.

Highlights

- establish current international practice in assessing Image quality of screening mammograms
- develop and validate a new assessment tool
- weighted consensus list of criteria
- potentially the closest evidence to date of a quantitative assessment.

Introduction.

In order to achieve a high quality diagnostic mammogram, a number of factors need to be considered, not least the expertise of the mammographer in producing optimally positioned breasts (Fig 1). ~~Indeed~~ P~~-~~ositioning has been cited as the single most important factor in optimising mammographic image quality (IQ). Without all breast tissue included on a mammogram or not optimally visualised, all other aspects of IQ are not relevant. R~~r~~esearch has shown ~~has shown~~ a direct link between mammographic image quality ~~mammographic image quality (~~IQ~~)~~ and cancer detection[1]. O~~; o~~ptimal IQ leads~~in~~g to earlier detection, higher detection rates, fewer interval cancers and reduced dose[2,3,4] ~~(1,2,3)~~

Current UK categories for mammography evaluation are PGMI (Perfect, Good, Moderate and Inadequate)~~-~~. Use of PGMI was established by the National Health Service breast screening programme in 2006. ~~-~~and there is evidence this has been adopted by other countries[5,6,7] ~~(4, 5, 6)~~. Some parts of Europe and the United States use evaluation tools provided by the Commission of the European Communities (CEC) and the American College of Radiologists (ACR) respectively[8,9,10] ~~(7, 8, 9~~). The common theme is a list of categories and associated criteria which largely relate to positioning of the breast. These are used to ~~for~~ determine~~ing~~ ~~image quality (~~IQ~~)~~ that informs the assignment of the image as excellent, acceptable or inadequate quality. There is evidence that these systems currently lack reliability and validity; guidelines for their implementation have always been subjective and have also not been reviewed commensurate with altered imaging practice such as the move from analogue to digital image acquisition[11,12] ~~(10, 11~~).

Difficulties involved in developing and validating any image assessment tool are ~~two fold~~twofold. First deciding ~~combining~~ which anatomical structures should be ~~are~~ included in the image then ~~and~~ the level of importance given by the observer to the inclusion of each

structure ~~their inclusion~~. There is a documented need to develop a visual grading scale for consistency in evaluating image quality[7,13] ~~(6, 12~~) combining both aspects of the assessment. In this study we will first establish current international practice in assessing the IQ of screening mammograms. Then develop and validate an assessment tool, incorporating a weighted consensus list of criteria derived from currents systems and deemed most relevant when assessing mammographic IQ.

Methods

The study comprised several phases. Participants completed an on line questionnaire to establish current practice in assessing the IQ of screening mammograms. Using PGMI as a starting point, a sub group (expert panel) employed a Delphi process and test set of mammograms to develop and test revised, weighted criteria and numerical ranges for overall scoring categories.

Phase 1

A questionnaire containing both closed and open questions aimed at assessing current practice in appraising mammographic IQ was initially sent to 2 UK breast screening units to test for content validity. No subsequent changes were made and it was then sent via on line Survey Monkey to centres in countries with a national mammography screening programme. In addition to establishing current practice, responses highlighted which centres had 4 radiographers meeting the inclusion criteria i.e. with a minimum of 4 years' experience in performing (not reporting) screening mammograms that could be taken forward for participation in phase 4 of the study.

Any information from the questionnaire requiring further elucidation was followed up by skype/telephone interview then all data transferred to MS Excel (Microsoft Corp, Redmond WA).

An expert panel was assembled from the respondents, inclusion criteria being a breast radiologist, or radiographer/other professional who either trains radiographers in undertaking mammography or has published work in peer reviewed journals investigating assessment of mammographic IQ and the associated criteria used.

Phase 2

The first panel task was to develop a list of criteria and their definitions to be used in the new assessment tool, then for each criterion, its level of importance in the assessment process. Members were individually sent a preliminary list of suggested criteria largely derived from UK PGMI guidelines. The panel was asked to consider several aspects of a revised system, first the *inclusion* and *wording* of the criteria using a Likert scale of 1-5 where 5 is complete agreement and 1 is no agreement. They were also invited to add any further criteria they felt should be included. A modified Delphi like process was used repeatedly to adjust responses until consensus was reached. The Delphi process was repeated to score the level of *importance* of each newly agreed criterion. In any cases of poor 'importance' agreement the criterion was either dropped from the list or sent back to the panel for re wording. The mean importance score for each criterion was calculated. Finally the panel was asked to consider categories and whether or not the PGMI categories should be replaced or altered.

Phase 3

To test agreement within the expert panel in interpreting the new criteria a mini test set of digital mammograms from 12 consecutively screened women aged 50 to 70 years of age was compiled by the principal investigator independent of the expert panel. Each case comprised four images (2x Cranial Caudal, 2x Medio Lateral Oblique views). The images were anonymised and numbered and then enriched with mammograms demonstrating a range of

image quality flaws. Women were excluded if they had previously undergone breast cancer surgery, had implants, only one breast or a pacemaker.

To enable all participants to view and score the study images a web based image collection and annotation software developed as part of the OPTIMAM project was used[14] ~~(13~~). The test set was uploaded onto the OPTIMAM server only accessible initially to the expert panel. On the advice of our institutions Research and Development department, w~~W~~e did not require Ethical or Trust approval to use NHS staff in this research. As~~ and as~~ all images used were anonymised and retrospectively viewed, we did not require local ethics committee approval, or patient consent.

The expert panel initially comprised 7 members; 3 breast radiologists with > 20 years' experience, 1 associate professor with > 20 years' experience in mammography image evaluation research and practice and 3 radiographers, one who has published papers previously and 2 involved in training radiographers in mammography technique. The panel individually scored each criterion against each image 1-5 where 1 was the poorest representation and 5 was optimal representation. They were also invited to comment on how effectively the new criteria worked in practice. Finally the panel members assigned each image an overall score of Perfect, Good, Adequate or Inadequate, these newly agreed categories being broadly based on the most currently used system PGMI.

Dedicated mammography workstations with 5 megapixel monitors were used to view the DICOM images under the same conditions as in the clinical setting i.e. in a reporting room with low ambient lighting and no windows thus excluding the risk of background light. Other computers/monitors in the rooms were positioned so that no glare from them compromised the viewing conditions[7] ~~(6~~). Participants could optimise viewing by locally manipulating the workstation settings.

Phase 4

A main test set of digital mammograms from 178 consecutively screened women was compiled. Each case comprised four images (2x Cranial Caudal, 2x Medio Lateral Oblique views). As there would be fewer mammograms which were *moderately good* or *inadequate* in the screening population from which the test set was drawn, the study was designed to have sets enriched with images from each of the PGMI categories. To enable this, i~~I~~mages were collected from consecutively screened women until the ~~enriched~~ set included ~~contained~~ 30 moderately good and 30 inadequate images and 59 Perfect and 59 Good images. This process was undertaken by the principal investigator using current PGMI criteria independent of all participants who would review the mammograms for the study.

The expert panel scored each descriptor against each image 1-5 as before and gave an overall category score to each image as before.

The intention was then for the rest of the study participants (breast radiographers identified in phase 1) to independently repeat the process just undertaken by the expert panel i.e. score each descriptor against each image 1-5 and give an overall category score to each image.

Statistical analysis

Phases 2, 3 and 4

Kendall's coefficient of concordance (Kendall's W) is a non-parametric measure used to indicate the degree of association or agreement of ordinal (ratings made on a scale) measurements made by multiple raters on the same samples. It tests the null hypothesis that the raters have produced ranking of the objects that are independent of one another. In other words, it assesses the degree to which a group of raters provides a common ranking

for a set of objects. In the present study, Kendall's W was used to establish the agreement between the expert panel members regarding the importance scores for the descriptors. At this point there was a possibility the study would not to go ahead if agreement within the exert panel in assigning importance scores for the descriptors was not statistically significant. The concordance was then applied to the scores for mini and main test set.

~~Kendall's coefficient of concordance was used, based on a nonparametric two-way analysis of variance to establish the agreement between the expert panel members regarding the importance scores for the descriptors.~~

~~At this point there was a possibility the study would not to go ahead if agreement within the exert panel in assigning importance scores for the descriptors was not statistically significant. The same test was again used to establish agreement between expert panel members~~

~~scoringscores for the mini and main test set~~

Phase 4

We now describe a new measure, the Weighted Total Score, as a potential predictor of the PGAI classification. Each descriptor score for individual expert panel members was multiplied by the mean importance score for that descriptor (assigned in phase 2) giving a weighted score. For each expert, the sum of scores given for each image was calculated. Collating these for all 5 panel members gave a range of scores for each overall image category such that an image categorised 'Perfect' overall may have a range of scores within the expert panel of say 80-100 and an image categorised as 'Good' overall may have a range of scores within the expert panel of 65 to 85 (there may be some overlap). The intention was to achieve numerical parameters for each P, G, A and I category which would form the

reference standard against which radiographer participants were assessed in the latter part of phase 4

Statistical tests at 5% significance level were carried out as follows:

1. ANOVA, to detect any significant difference in the weighted total scores across the 4 categories – PGAI.
2. t-tests, to detect any significant difference in the weighted totals between P & G, between G & A, and between A & I.
3. Ordered Logistic Regression model fitted with PGAI Classification as the outcome and weighted total score as the predictor, to explore the relationship and quantify the likelihood of an image being assigned a higher PGAI category given a unit increase in the weighted total.

   These tests were also repeated for each of the views – RMLO, LMLO, RCC and LCC.

Results

Phase 1

13 centres in 10 countries responded to the questionnaire. 3 centres did not have national guidelines for assessing the technical image quality of screening mammograms advocating a particular method e.g. PGMI, ACR, CEC so were excluded from further evaluation. All remaining centres confirmed they had some form of national mammography IQ assessment programmes and used unmodified national guidelines (table 1). All centres undertook digital mammography and had at least 4 radiographers meeting the inclusion criteria i.e. a minimum of 4 years' experience in performing (not reporting) screening mammograms that could be taken forward for participation phase 4 of the study

All centres said their radiographers performed mammography IQ assessment on an ongoing basis when mammography examinations are undertaken. 2 centres undertook regular

image review as part of an organised Continuous Professional Development programme with radiographers reviewing both their own and colleagues work, 1 centre did this but only reviewing colleagues work, not their own. In 2 centres radiographers work was assessed by senior colleagues periodically and an additional 2 centres undertook image review but on an ad hoc basis only. 1 centre relied on radiologists highlighting any recurring image flaws.

Phase 2

The preliminary list of suggested criteria largely derived from UK PGMI guidelines sent to the expert panel is shown in table 2. 4 rounds of Delphi process ensued until the panel agreed on the inclusion and wording of criteria -table 3

The importance scoring of the expert panel (phase 2) is shown in table 4

Using Kendalls coefficient of concordance on the importance scoring (ranges 0-1), W=0.35, $p<0.001$ demonstrating weak to moderate agreement.

Once the criteria and categories were agreed, 2 expert panel members (one consultant radiologist, one radiographer) left the study due to other commitments. This left a panel of 5 members but with diversity in expert contribution similar to the original panel of 7 members.

Phase 3

Using Kendalls coefficient on the mini test set of 12 images, (48 observations, 4 views per case), W=0.50, $p<0.001$ demonstrating moderate agreement in scoring the mini test set. The panel members reported satisfactory practical application of the descriptors. Those relating to the posterior nipple line created most comment and discussion ultimately necessitating 3 associated criteria 1a, 1c and 2f.

Phase 4 – expert panel

Using Kendalls coefficient on the main test set of 178 images (712 observations, 4 views per case), W=0.55, p<0.001 demonstrating moderate agreement.

Expert panel weighted scores tabulated against the overall PGAI classification are shown in table 5 and the distribution of the weighted total scores for each overall PGAI category is shown in fig 1.

Both ANOVA and t-tests demonstrated a p-value < 0.001which is statistically significant in demonstrating both differences in the weighted total scores across the 4 categories – PGAI and differences in the weighted totals between P and G, between G and A, and between A and I.

The ordered logistic regression model yields p<0.001 meaning that weighted total score is a significant predictor of the PGAI Classification. Further, an Odds Ratio of 1.23 [95%CI: 1.20-1.26] shows that a one unit increase in the weighted total increases the odds of being in a higher PGAI category (from I to A/G/P or from A to G/P etc.) by 23%.

The above statistical tests have also been repeated for each of the views – RMLO, LMLO, RCC and LCC yielding significant results similar to the above.

Phase 4- radiographers

Using Kendalls coefficient on the main test set of 178 images: W=0.46, p<0.001 demonstrating moderate agreement.

Then using Kendalls coefficient for the expert panel main test set merged (on Case ID and view) with the radiographers main test (N=178, 712 observations): W=0.66, p<0.001 demonstrating fair agreement.

Discussion

It is generally agreed there is a need for more valid and reliable mammography evaluation criteria[15,16] ~~(14, 15~~). Multiple systems are currently being used[17] ~~(16~~). PGMI is the most

common globally with 7 out of 10 countries in this study using unmodified national guidelines. However, previous work by the authors has demonstrated variation in interpretation between centres even when using the same system[7] ~~(6~~). There has also been discussion regarding which anatomical structures should be included, how they should be described to avoid ambiguity and how much importance is placed on the inclusion of each structure[18] ~~(17~~). Criticism of PGMI and other systems has predominantly focused on the use of subjective descriptors in their guidelines including the word 'appropriate' and 'general amount' in relation to the inclusion of the pectoral muscle in the medio lateral oblique (MLO) view[19] ~~(18~~)

The pectoralis muscle and its relation to the posterior nipple line (PNL) on the MLO view is one of the areas which caused most discussion for our expert panel. As described by authors previously this may be due to this criteria being particularly multifactorial involving the nipple position, the width, length and angle of the pectoral muscle[20] ~~(19~~).

One of the advantages of using a Delphi process is maximizing the use of anonymity which can be problematic in collective decision making[21] ~~(20~~). The iterative process enabled the panel members to alter or develop their opinions and from this it became apparent that a definition was needed for the PNL. Also 2 descriptors were necessary to incorporate the superior and inferior portions of pectoral muscle and their relation to the PNL (1a and 1b). It evolved during Delphi that these descriptors could not be mutually exclusive. This became a developing theme for several of the descriptors which were thought to be interrelated including the length of the PNL in each view (1c and 2f). During scoring of the mini test set, feedback raised the question of how accurately this should be assessed, by actual measurement or just visually checking. After further discussion the latter was agreed.

Debate in this study surrounding how the pectoralis muscle and PNL should be included in

mammography assessment is consistent with previous papers, which have failed to reach an evidence based conclusion[16,22] ~~(15,22~~). Our weighted scores and PGAI classification demonstrated in fig 1 is potentially the most valid and reliable evidence to date regarding inclusions and wording of descriptors

Another area discussed was training. The panel wanted to incorporate reasons /aims for some of the MLO descriptors to be used as a teaching aid for trainee mammographers and to reduce ambiguity, also to remind radiographers generally why they are using the criteria. There was a highly significant p value ($p<.001$) in the importance scoring of descriptors across the panel demonstrating enough concordance to continue with the study. This is a small set and there was weak agreement ($w=0.35$) although good agreement amongst the panel using a level of consensus established as a mean rating of ≥3.5 for all descriptors except for 3e 'No transparent skin folds or creases' scored 2.6 and 3f 'Symmetrical 'mirror' L and R images' scored 2.7. Also just below the 3.5 cut off was 2a' Breast centrally placed, nipple in midline subject to anatomical presentation' scored 3.4 and 3g 'No artefacts' scored 3.3. Interestingly the lower scores for these were assigned by radiographer panel members who are perhaps professionally more influenced than radiologists by an aesthetically pleasing and technically optimised image rather than looking purely at whether the necessary clinical information is contained within the image[6,23] ~~(5~~). These professional differences have been noted before and are areas for future training in establishing common ground when appraising mammograms[12] ~~(11~~).

There was overall moderate agreement in the expert panel scoring both the mini and then the main test set ($W=0.5$ and $W=0.55$ respectively). The improvement as the test set gets bigger provides a more significant result.

The most important finding of this study is the relation between weighted total scores and PGAI classification by expert raters. To our knowledge this has not been published before in relation to mammographic IQ and is potentially the closest evidence to date of a quantitative assessment. For the expert panel main test, each image view descriptor is assigned a score of 1 to 5 for image quality. The mean of these scores for the 5 raters was multiplied by the mean importance scores (obtained earlier by the expert raters) for the corresponding descriptor leading to a weighted score. The weighted scores were then added up for all the descriptors per image view and the resulting weighted total score assigned to each of the 4 views – RMLO, LMLO, RCC, LCC such that each P, G, A and I category gained numerical parameters. Ordinal logistic regression suggests that the weighted total score~~WTS~~ is a significant predictor of the PGAI classification The box plot demonstrates that each category P, G, A and I is significantly different ~~almost separate~~ with minimal~~only some~~ overlap. ~~which is inevitable.~~ The majority of scores are grouped separately for each category, the biggest range being within the category 'inadequate' (scores 141-214). This is perhaps due to local screening culture and suggests an area for training to promote uniformity of practice.

The intention was that this would then form the reference standard against which radiographer participants were assessed in the latter part of phase 4. Unforeseen IT difficulties preventing ease of access of the larger number of radiographer participants to the web based image collection and annotation software, also local issues preventing the radiographers having time to score each descriptor against each image 1-5 as the expert panel had done before meant that radiographers could only give an overall perfect, good, adequate or inadequate score to each image as before. It is encouraging that there was moderate agreement between radiographers scoring (W=0.46), and fair agreement between

expert panel and radiographers in main test set scoring (W=0.66). However this must be taken in context as it is unknown whether an overall category e.g. 'good' scored by the expert panel contains the same descriptors and signs as radiographers giving the same overall score. This is a limitation of the study and we suggest further work where radiographers use the new descriptors to assess the main test set. The level of concordance between them and the expert panel scores will inform credibility and validation, also training needs in using the new system.

A further limitation includes some inherent subjectivity associated with this type of image assessment but the large numbers of images for analysis help minimise intra observer variability.

Conclusion.

This study has established current international practice in assessing the IQ of screening mammograms. We have developed a mammography image evaluation tool that incorporates a weighted consensus list of criteria derived from currents systems in use. A range of numerical scores for each category P,G,A and I has been identified which is reliable and valid thus far although this needs further evaluation and validation by radiographers which will also inform future training .

References

1. Bassett L, Dione M. Farria, MD, MPH et al. Reasons for Failure of a Mammography Unit at Clinical Image Review in the American College of Radiology Mammography Accreditation Program1

~~1.~~2. Feig S. Image Quality of Screening Mammography: Effect on Clinical Outcome. Commentary American Journal of Roentgenology 2002;178:805–807

~~2.~~3. Taplin SH, Rutter CM, Finder C, Mandelson MT, Houn F, White E. Screening mammography: Clinical image quality and the risk of interval breast cancer. American Journal of Roentgenology. 2002; 178(4):797-803.

~~3.~~4. O'Leary D, Rainford L Can radiation dose in mammography be further reduced by increasing the image quality? Breast Cancer Research 2011

~~4.~~5. National Health Service Breast Cancer Screening Programme, National Quality Assurance Coordinating Group for Radiography. Quality assurance guidelines for mammography including radiographic quality control. NHSBSP Publication No. 63; ISBN 1 84463 028 5; 2006. Sheffield

~~5.~~6. Moreira C, Svoboda K, Poulos A, Taylor R, Page A, Rickard M, (2005), Comparison Of The Validity And Reliability Of Two Image Classification Systems For The Assessment Of Mammogram Quality, Journal Of Medical Screening, 12 (1), p38-42

~~6.~~7. M. Boyce, R. Gullien, D. Parashar, K. Taylor. Comparing the use and interpretation of PGMI scoring to assess the technical quality of screening mammograms in the UK and Norway. Radiography 2015 21(4) 342–347

~~7.~~8. Breast Imaging Clinical Education Review Committee, BreastScreen NSW. Research Proposal: Evaluation of Mammogram Quality. A Comparison of 2 Image Classification Systems. Sydney: State Co-ordination Unit, BreastScreen NSW, 2001

~~8.~~9. European Commission. European Guidelines on Quality Criteria for Diagnostic Radiographic Images. EUR 16260; 1996CEC (1996)

~~9.~~10. Committee ACR. Mammography quality control manual. American College of Radiology; 1999. [3]

~~10.~~11. Yankaskas B International comparison of performance measures for screening mammography: can it be done? Journal of Medical Screening 2004;11:187–193

~~11.~~12. Hofvind S et al. Quality assurance of mammograms in the Norwegian Breast Cancer Screening Program European Journal of Radiography (2009) 1, 22e29

~~12.~~13. Hill C, Robinson, L. Mammography image assessment; validity and reliability of current Scheme Radiography 21 (2015) 304e307

~~13.~~14. Looney P, Young K, Mackenzie A and Halling-Brown M MedXViewer: an extensible web-enabled software package for medical imaging Proc. SPIE 9037, Medical Imaging 2014: Image Perception, Observer Performance, and Technology Assessment, 90371K (March 11, 2014)

~~14.~~15. The framework for higher education qualifications in England, Wales and Northern Ireland. August 2008.

~~15.~~16. Naylor S, York J. An evaluation of the use of pectoral muscle to nipple level as a component to assess the quality of the medio-lateral oblique mammogram. Radiography 1999;5:107e10.

~~16.~~17. http://radiopaedia.org/articles/posterior-nipple-line [Accessed June 2016]

~~17.~~18. Li Yanpeng et al. A review of methods of clinical image quality evaluation in mammography. European Journal of Radiology 74 2010 e122–e

~~18.~~19. Bentley KM, Poulos A, Rickard M. Mammography image quality: analysis of evaluation criteria using pectoral muscle presentation. Radiography 2008;14(3):189e94.

~~19.~~20. Spuur K, Poulos A. Evaluation of the pectoral muscle in mammography images: The Australian experience. European Journal of Radiography Volume 1, Issue 1, March 2009, Pages 12–21

~~20.~~21. Hasson F, Keeney S, J McKenna H. Research guidelines for the Delphi Survey Technique. Journal of Advanced Nursing, 2000, 32(4), 1008-1015

~~21.~~22. Spuur K et al Mammography: Correlation of pectoral muscle width and the length in the mediolateral oblique view of the breast. Radiography (2010) 16, 286e291

~~22.~~23. Martin CJ, Sutton DG, Sharp FP. Balancing patient dose and image quality. Applied
Radiation and Isotopes 1999;50(1):1e19.

CENTRE	COUNTRY	NATIONAL GUIDELINES
Cambridge	UK	PGMI
Leuven	Belgium	EUROPEAN
Hvidovre	Denmark	PGMI adopted in 1995
Odense C	Denmark	PGMI adopted in 1995
Oslo	Norway	Norwegian Breast Cancer Screening Program (NBCSP) criteria based on PGMI
Vienna	Austria	EUROPEAN
Toronto	Canada	Ontario Breast Screening Program (OBSP) image assessment criteria
Charleston	USA	ACR
New South Wales	Australia	PGMI adopted from UK
Victoria	Australia	PGMI adopted from UK

Table 1. Participating centres with a national mammography IQ assessment programmes using unmodified national guidelines

DESCRIPTION OF ASSESSMENT CRITERIA	SCORE 1-5
CRITERIA SPECIFIC TO MLO IMAGE	
The lower edge of the pectoral muscle shadow should reach nipple level whenever possible, to ensure that the posterior aspect of the breast is satisfactorily included on the image.	
Pectoralis muscle at an appropriate angle in accordance with good practice. This angle varies according to the variations in physical constitution of the individual. It should be at an appropriate angle to enable the axillary tail of the breast to be demonstrated clear of the muscle shadow on the mammogram.	
Inframammary angle clearly demonstrated. The inframammary angle should be clearly shown without overlying or underlying tissue. This indicates that the breast has been lifted and that the inferior-posterior part of the breast has been correctly imaged.	
CRITERIA SPECIFIC TO CC IMAGE	
Medial border of the breast included on the image	
Some of axillary tail of the breast included on the image	
Pectoral muscle shadow may be shown on the posterior edge of the breast on some CC views depending on anatomical characteristics.	
CRITERIA RELEVANT TO BOTH CC AND MLO IMAGES	
Nipple in profile with no breast tissue obscured	
Appropriate compression – no blur/ movement	
Appropriate exposure	
No skin folds which obscure breast tissue	
No transparent skin folds	
Symmetrical ‘mirror’ L and R images	
No artefacts	

Table 2. The preliminary list of suggested criteria largely derived from UK PGMI guidelines sent to the expert panel

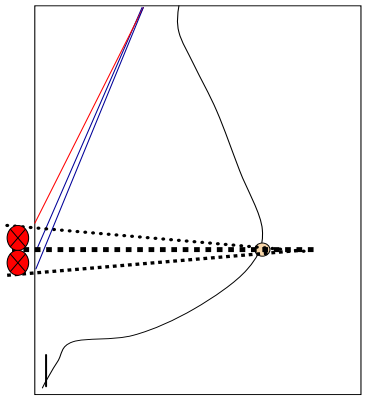
DESCRIPTION OF ASSESSMENT CRITERIA
1. CRITERIA SPECIFIC TO MLO IMAGE
a. The inferior edge of the pectoral muscle is no more than 2cm short of a straight horizontal line drawn from the nipple to transect the posterior edge of the image such that the transection creates a right angle. This line is known as the posterior nipple line (PNL) *Ensures the breast has been lifted up and out and all posterior breast tissue included*  X ≤ 2from PNL
b. A minimum of 3cm width of pectoral muscle is visible at the upper image border and the axillary tail demonstrated clear of the edge of the muscle. Signs a. and b. are inextricably linked *Ensures the axillary tail is optimally visualised*
c. The length of the posterior nipple line is within 1cm of the posterior nipple line on the CC view (see 2f) *Indicates all posterior breast tissue included*
d. Inframammary angle is clearly demonstrated. It should be clearly shown without any overlying or underlying tissue. *Indicates that the breast has been lifted up and pulled out to ensure the inferior-posterior part of the breast is included and optimally visualised*
e. The nipple is in profile or transected by skin as long as the PNL meets criteria in 1a.
2. CRITERIA SPECIFIC TO CC IMAGE
a. Breast centrally placed, nipple in midline subject to anatomical presentation
b. The nipple is in profile or transected by skin.
c. Medial border of the breast included on the image
d. Lateral border of the breast includes a portion of the axillary tail on the image
e. Breast tissue from the nipple to the edge of the pectoral muscle is included. Individual anatomical variables determine whether the pectoral muscle will be included on the image.3
f. The length of the posterior nipple line is within 1cm of the posterior nipple line on the MLO view (see 1c)
3. CRITERIA RELEVANT TO BOTH CC AND MLO IMAGES
a. Appropriate image processing algorithm used.
b. Appropriate compression – no evidence of inadequate sharpness of breast tissue structures
c. Appropriate exposure for modality – • Sufficient to differentiate between fatty and fibro glandular breast tissue • Sufficient to differentiate between breast tissue and subtle abnormalities such as small masses and architectural distortion
d. No skin folds which obscure breast tissue
e. No transparent skin folds or creases
f. Symmetrical 'mirror' L and R images
g. No artefacts
h. Correct image identification (patient ID, image projection etc)

Table 3. The inclusion and wording of criteria agreed by the expert panel

DESCRIPTOR	EXPERT PANEL MEMBER							MEAN SCORE
	1	2	3	4	5	6	7	
1a	5	5	4	5	5	5	5	4.9
1b	4	5	4	5	4	4	4	4.3
1c	4	4	5	5	3	5	3	4.1
1d	3	4	4	5	5	5	5	4.4
1e	3	4	4	3	4	4	5	3.9
2a	4	5	4	3	3	1	4	3.4
2b	4	5	4	3	3	2	5	3.7
2c	3	5	4	5	4	5	5	4.4
2d	3	4	4	3	4	4	4	3.7
2e	5	5	5	3	3	5	5	4.4
2f	3	4	5	5	3	5	3	4.0
3a	3	4	5	5	4	5	5	4.4
3b	4	5	5	5	4	4	5	4.6
3c	4	4	5	5	4	5	5	4.6
3d	3	5	5	5	3	4	5	4.3
3e	2	4	4	1	3	2	2	2.6
3f	1	4	4	3	3	3	1	2.7
3g	2	4	4	4	4	2	3	3.3
3h	5	5	5	5	5	5	5	5.0

Table 4. The importance scoring of the expert panel for each newly agreed criterion

	CATEGORY			
SCORE	Perfect	Good	Adequate	Inadequate
Minimum	229	204	180	141
Maximum	246	243	232	214
Mean	237	227	210	185

Table 5. Expert panel weighted scores tabulated against the overall PGAI classification

Acknowledgements

The authors would like to acknowledge the help of Mark Halling-Brown with all aspects of using the OPTIMAM annotation software

Figure(s)

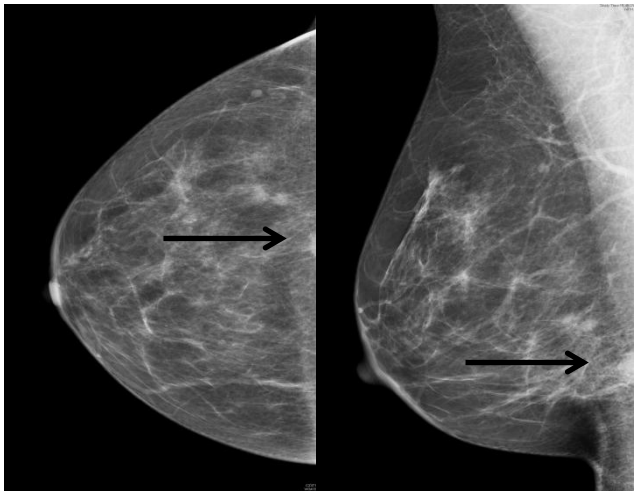


Fig 1. 15mm indistinct mass partially seen at the very back of the right breast on both MLO and CC views. Core biopsy histology = Invasive ductal carcinoma Grade 3 ER/PR +ve HER 2-ve

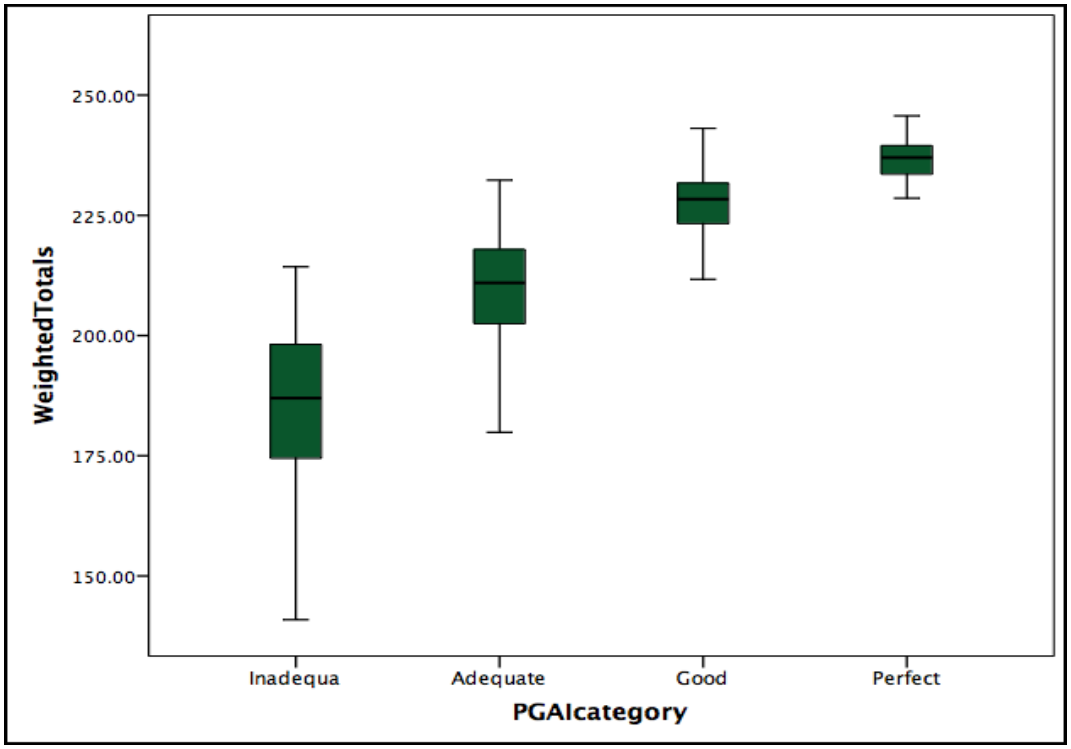


Fig 2. The distribution of the weighted total scores for each overall PGAI category