European Journal of Radiology 139 (2021) 109686

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Validation of a candidate instrument to assess image quality in digital mammography using ROC analysis

Joana Boita [a,b], Ruben E. van Engen [b], Alistair Mackenzie [c], Anders Tingberg [d], Hilde Bosmans [e,f], Anetta Bolejko [g], Sophia Zackrisson [g], Matthew G. Wallis [h], Debra M. Ikeda [i], Chantal van Ongeval [f], Ruud Pijnappel [b,j], Mireille Broeders [b,k], Ioannis Sechopoulos [a,b,*]

[a] Department of Medical Imaging, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, the Netherlands
[b] Dutch Expert Centre for Screening (LRCB), Wijchenseweg 101, 6538 SW, Nijmegen, the Netherlands
[c] National Coordinating Centre for the Physics of Mammography, Royal Surrey NHS Foundation Trust, Guildford, GU2 7XX, UK
[d] Department of Medical Radiation Physics, Translational Medicine Malmö, Lund University, Skåne University Hospital, Carl Bertil Laurells gata 9, SE-20502 Malmö, Sweden
[e] Department of Imaging and Pathology, Radiology, KUL, Herestraat 49, Leuven B-3000, Belgium
[f] Department of Radiology, Radiology, UZ Gasthuisberg, Herestraat 49, Leuven B-3000, Belgium
[g] Department of Medical Imaging and Physiology, Translational Medicine Malmö, Lund University, Skåne University Hospital, Carl Bertil Laurells gata 9, SE-20502 Malmö, Sweden
[h] Cambridge Breast Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge & NIHR Cambridge Biomedical Research Centre, Cambridge, CB2 0QQ, UK
[i] Department of Radiology, Stanford University School of Medicine, 875 Blake Wilbur Dr, Stanford, CA, 94305, USA
[j] Department of Radiology, University Medical Center Utrecht, PO Box 85500, 3508 GA, Utrecht, Utrecht University, the Netherlands
[k] Department for Health Evidence, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Digital mammography
Image quality evaluation
Type testing
Visual grading analysis
Receiver operating characteristic
Validation study

ABSTRACT

Purpose: To validate a candidate instrument, to be used by different professionals to assess image quality in digital mammography (DM), against detection performance results.
Methods: A receiver operating characteristics (ROC) study was conducted to assess the detection performance in DM images with four different image quality levels due to different quality issues. Fourteen expert breast radiologists from five countries assessed a set of 80 DM cases, containing 60 lesions (40 cancers, 20 benign findings) and 20 normal cases. A visual grading analysis (VGA) study using a previously-described candidate instrument was conducted to evaluate a subset of 25 of the images used in the ROC study. Eight radiologists that had participated in the ROC study, and seven expert breast-imaging physicists, evaluated this subset. The VGA score (VGAS) and the ROC and visual grading characteristics (VGC) areas under the curve (AUC_{ROC} and AUC_{VGC}) were compared.
Results: No large differences in image quality among the four levels were detected by either ROC or VGA studies. However, the ranking of the four levels was consistent: level 1 (partial AUC_{ROC}: 0.070, VGAS: 6.77) performed better than levels 2 (0.066, 6.15), 3 (0.061, 5.82), and 4 (0.062, 5.37). Similarity between radiologists' and physicists' assessments was found (average VGAS difference of 10 %).
Conclusions: The results from the candidate instrument were found to correlate with those from ROC analysis, when used by either observer group. Therefore, it may be used by different professionals, such as radiologists, radiographers, and physicists, to assess clinically-relevant image quality variations in DM.

1. Introduction

Image quality plays an important role in the early detection of breast cancer during mammographic screening. The quality of a digital mammogram can be affected by the parameters with which the image was acquired and by the image processing used to change its

Abbreviations: AUC, Area Under the Curve; CI, Confidence Interval; DM, Digital Mammography; MRMC, Multi-Reader Multi-Case; pAUC, partial Area Under the Curve; ROC, Receiver Operating Characteristics; VGA, Visual Grading Analysis; VGAS, Visual Grading Analysis Score; VGC, Visual Grading Characteristics.

* Corresponding author at: Radboud University Medical Center, Department of Medical Imaging, P.O. Box 9101 (766), 6500 HB Nijmegen, the Netherlands.
E-mail address: ioannis.sechopoulos@radboudumc.nl (I. Sechopoulos).

https://doi.org/10.1016/j.ejrad.2021.109686
Received 26 January 2021; Received in revised form 23 March 2021; Accepted 26 March 2021
Available online 30 March 2021

appearance. There are other non-system related issues, like patient positioning-related issues, that can also affect mammographic image quality [1–3]. Therefore, it is crucial to guarantee an optimal performance of digital mammography (DM) systems, including their image processing, so that they produce images that allow the interpreting radiologists to separate suspicious structures from the background [1, 3–10].

Guidelines for evaluation of DM systems have been developed to assess, optimise, and approve these devices before they are introduced into clinical practice and after major changes are made. An example of such evaluation procedures of DM systems is type testing, which is included in the Supplement to the European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis [11]. In type testing, both physical and clinical assessments are performed, to identify issues in the DM systems that may negatively affect lesion detection and interpretation.

Ideally, the clinical assessment of type testing should include an estimate of diagnostic performance by using task-based methods, such as Receiver Operating Characteristics (ROC) analysis or one of its variants [12–14]. However, ROC analysis is a very laborious and time-consuming procedure, involving a large number of observers and a considerable number of both negative and positive cases in order to reach sufficient statistical power. These limitations make it impractical to use this type of analysis repeatedly to evaluate image quality in type testing.

In current practice, a number of clinical observers are asked to grade the images using a set of criteria regarding physical parameters (such as resolution, noise, and contrast) and relevant normal anatomical structures reproduced in the digital mammogram [6,11]. This assessment is based on Visual Grading Analysis (VGA) where each evaluated feature, stated as a criterion, can be visually graded with an absolute score. An alternative is to compare the test image to a reference and score whether the quality of the former is worse or better than the reference image [15–17].

In comparison to ROC, VGA is a much faster and simpler method to evaluate image quality, making it appropriate for use in type testing [15–17]. However, the criteria that are currently used during VGA were developed based on the assumption that the possibility to clinically detect a lesion is associated with the reproduction and visibility of normal anatomic structures [18]. This is not necessarily applicable to calcification depiction in DM, for instance [19,20]. In general radiography, there is evidence that the outcome of a VGA correlates with clinical performance [21,22]. However, in digital mammography, Zanca et al. did not find any correlation between VGA and ROC analysis when assessing different image-processing algorithms [20].

A new set of criteria that does correlate with clinical performance is therefore needed for clinical assessment in type testing. According to Kundel et al., an image is of good quality if it allows for the distinction and representation of clinically relevant structures and features [23]. Therefore, a candidate instrument was previously developed by identifying the structures and features related to breast tissue and breast lesions that are important for radiologists to interpret a mammogram. This resulted in 18 items related to features of normal tissue, calcifications, and soft tissue lesions [24]. Even though the instrument does not evaluate physical parameters directly, the depiction of the breast tissue and lesion features is impacted by image acquisition-related and post-acquisition image processing-related issues. However, the clinical relevance of this instrument was not yet investigated [24]. In addition, it also remains to be determined if and how well the items in the instrument all assess and provide coherent information on the DM image quality.

Currently, radiologists who are highly experienced in mammography screening and image quality evaluation perform the clinical assessment for type testing [11]. However, these are in short supply. Therefore, it is of interest to know if other experts in breast imaging evaluation, like physicists and radiographers, could potentially also perform this assessment, making it easier to implement in regular practice.

Therefore, the aim of this study was to validate this previously-developed candidate instrument to assess image quality in digital mammography (DM), by investigating the correlation of its results with predictions of clinical performance resulting from receiver operating characteristics (ROC) analysis. In addition, the reliability and item correlation of this candidate instrument, and the possibility of this instrument being used by physicists experienced in breast imaging evaluation was investigated.

2. Methods

Validation of the candidate instrument was performed in terms of criterion validity, i.e., assessing the candidate instrument against the standard method used for clinical performance assessment, i.e., ROC analysis [25,26]. For this purpose, fourteen expert breast screening radiologists performed an ROC study for the task of lesion detection on 80 DM cases degraded to three levels of quality, in addition to evaluating the original non-degraded images. The outcomes of this ROC study were compared to the results obtained with eight expert breast screening radiologists performing a VGA study with the candidate instrument to assess the quality of a subset of 25 of the images used in the ROC study degraded to the same levels of quality. To determine whether it is feasible for physicists to use the candidate instrument, seven breast-imaging expert physicists also performed the VGA-based evaluation of the same 25 DM cases.

The candidate instrument was also evaluated in terms of construct validity and reliability, i.e., the extent to which the instrument can successfully and reliably, with an acceptable measurement error, assess image quality [25–27]. To this end, scaling assumptions were tested for each of the instrument items, and internal consistency reliability was evaluated for the entire instrument [25–29]. In this way, clues on which items in the instrument do not work satisfactory and should probably be removed to improve validity and reliability of the assessment were obtained. The ROC-VGA comparison analysis results were used to investigate the impact of removing those items. Details of the construct validity evaluation and reliability test can be found in the online supplement.

2.1. Digital mammography images

In the ROC study, a total of 80 DM cases with breast density varying from fatty to dense, with and without lesions, acquired using Lorad Selenia (Hologic, Inc., Bedford, MA, USA) DM systems, were selected and retrieved, under license, from the OMI-DB anonymised database of mammograms, which is part of the OPTIMAM project [30]. Each case consisted of two views (craniocaudal and mediolateral oblique) of each breast. This set contained a total of 60 lesions (40 cancers and 20 benign findings) and 20 normal cases. The positive cases contained different types of malignant and benign lesions, as described in Table 1, in order to include most of the types of lesions found in the clinical setting. The number of cases (n = 80) was estimated by using the sample size table from Obuchowski et al. [31] that uses the area under the ROC curve

Table 1
Breakdown of the 80 and 25 DM cases included in the ROC- and VGA-image set, respectively.

	ROC		VGA	
	Benign (n = 20)	Malignant (n = 40)	Benign (n = 7)	Malignant (n = 15)
Ill-defined mass	6	6	2	3
Well-defined mass	6	5	2	3
Spiculated mass	–	10	–	3
Architectural distortion	–	10	–	3
Calcification	8	9	3	3
Normal	20		3	

(AUC) as a measure of accuracy of the levels of quality. For this, the following parameter values were assumed: the number of cancer cases should be equal to the number of non-cancer cases, the number of observers that agreed to participate was high, the accuracy would be moderate (with AUC = 0.75), the differences in accuracy between the levels of quality would be small (of 0.05), and the observer variability would be between small to moderate [31]. All cases had either pathological (in case of the lesion-containing cases) or follow-up confirmation. The cases with lesions, as well as the respective lesion locations, were reviewed and annotated by an expert breast radiologist. Asymmetries were not included because these lesions involve normal tissue rather than specific lesion features, which is already evaluated in normal cases, and therefore including such cases would not be of added value. In the VGA study, a total of 25 DM cases were selected from the image set used in the ROC study, with varying breast density, lesion presence, and types of lesions, as described in Table 1.

2.2. Image quality modification

In total, images with four levels of quality were used. The first level, representing the highest quality, was the original images with no modification. The other three quality levels were generated by corrupting the original images using algorithms previously developed [19, 32–34] and validated [35]. These algorithms simulate previously observed image quality issues in DM systems undergoing type testing. Specifically, resolution was reduced, contrast was increased or decreased, texture of the structures was modified by increasing the correlation of the pixel values in the image, or noise was added by simulating lower dose acquisitions. Each image was corrupted to reflect a single image quality issue at a time, degraded into three levels. Examples of the effects of the quality levels in the image quality and lesion visibility can be found in the online supplement.

All reference images, i.e., the original images without any change in quality, were determined to be adequate for interpretation, but not necessarily perfect, as is the case with the majority of acquired mammographic images. The levels of quality were chosen according to what the radiologists selected as the level that they still felt was acceptable for interpretation for each image quality issue, as reported previously [19]. The first level corresponded to the reference level, i.e. the original uncorrupted images; the second level correponded to the first acceptable image selected by radiologists for each of the five simulated image quality issues; the third level corresponded to the first selected unacceptable image; and the fourth level corresponded to the second selected unacceptable image.

2.3. Observer studies

2.3.1. ROC study

A fully-crossed multi-reader multi-case (MRMC) ROC study was conducted over four sessions to assess the lesion detection performance when using the images from the four levels of quality. The observer was fully blinded to diagnostic reports and prior imaging. The sessions were separated by at least three weeks. In each session, images from the four levels of quality were displayed in a random order.

Fourteen expert breast screening radiologists from five countries (UK, USA, Sweden, Belgium, and the Netherlands) with a median of 22.5 years (range: 6–32 years) of experience in breast imaging and being involved in regular reading of screening mammography participated in the ROC study. On a per case basis, the observer was asked to score the images according to the probability of malignancy present (10-point scale, from definitely no malignancy present (1) to definitely malignancy present (10)), and to indicate the case-recall decision (yes or no). A training set composed of 10 DM cases was shown first to familiarise the observers with the workstation and the workflow of the study. The DM cases were viewed on either 5- or 12-megapixel calibrated liquid crystal diagnostic mammography displays (Barco, Kortrijk, Belgium). The images were displayed according to the DICOM standard for presentations using ViewDEX, a software specifically developed for observer studies [36]. The ambient light level was low and fulfilled conditions for reading rooms. The software allowed the radiologists to zoom, pan, and scroll over the images of each case.

2.3.2. VGA study

A fully-crossed MRMC VGA study was conducted to assess the quality of four levels of quality seven months after the ROC study, using the candidate instrument. Like the ROC study, the VGA study was conducted over four sessions. There was no time restriction in terms of a waiting period between sessions. Since this was not a detection-task study, it would be expected that the possibility of remembering the case would not induce bias in answering the VGA questions. In each session, images from the four levels of quality were displayed in a random order, as was done in the ROC study.

Eight radiologists that participated in the ROC study also participated in the VGA study. They again represented five countries (UK, USA, Sweden, Belgium, and the Netherlands) and had a median of 27.5 years (range: 10–32 years) of experience in breast imaging. Of these eight radiologists, three have English as their native language. In addition, seven physicists from five countries (UK, Germany, Ireland, Belgium, and the Netherlands) with a median of 20 years (range: 11–24 years) of experience in breast imaging and image quality evaluations participated in the VGA study. One of these physicists' native language is English. On a per case basis, the observer was asked to complete the candidate instrument, whose development was previously described [24]. Briefly, the instrument involves a set of 18 items on the visibility of structures present in normal tissue (items 1–8), calcifications (9–14), and soft tissue lesions (14–18) by selecting a score from 1 (completely disagree) to 8 (completely agree) [24]. High scores correspond to high image quality, since they correspond to complete agreement with the statement that was phrased using positive framing. In the lesion-containing cases, the lesion was annotated in the views in which it was visible. In case an item could not be scored because it was about a structure that was not present in the image, the observer answered not-applicable (N/A). The observers were instructed to answer based on their evaluation of the whole case and reflecting the worst rating given to any of the four views. A training set with 3 DM cases was shown first to familiarise the observers with the workstation and the workflow of the study, and to verify if all questions were understandable to the observer. In case of the physicists, examples of each image quality issue were shown to familiarise them with the levels and types of degradation because, in contrast to the radiologists, they did not participate in the ROC studies where the same levels and types of degradation were used. The displayed conditions and study software used in the VGA study were the same as those used in the ROC study.

2.4. Analyses

2.4.1. ROC study

The ROC curves for the average observer for each level and the respective areas under the curve (AUC_{ROC}) and partial AUCs ($pAUC_{ROC}$) were estimated from the probability of malignancy present score in order to compare the four levels of quality. The cut-off value of the $pAUC_{ROC}$ was selected as the sensitivity value that limited the region of interest to the portion of the curve most relevant to real world image evaluation. Standard random-reader random-case analysis of variance was performed to compare the AUC_{ROC} and $pAUC_{ROC}$ calculated for the average observer for each level using the OR-DBM (Obuchowski-Rockette and Dofman-Berbaum-Metz) and MRMC program (version 2.52, University of Iowa, Iowa City, IA). The respective 95 % confidence interval (CI) was also calculated. This analysis was also performed stratified by type of lesion, i.e., separated by calcifications (and normal) and soft tissue lesion (and normal) cases. Sensitivity and specificity were estimated using the case-recall decision score. The average of these

parameters across observers for each level was estimated using the iMRMC software (version 4.0.3., Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, MD). In this statistical analysis, the Bonferroni correction method was used to adjust the cut-off for statistical significance due to the six multiple comparisons performed, being $p < 0.008$ used for this situation.

2.4.2. VGA study

The VGA results were analysed using the visual grading characteristics (VGC) method [37]. The VGC analysis is a non-parametric rank-invariant method for comparing data from two conditions, a reference condition (in this case, the reference level, level 1) and a test condition (separately, levels 2, 3, and 4) using the observer scores for each item [37,38]. In VGC analysis, the distributions of ratings for the two conditions are used to produce a VGC curve in a similar way as ROC curves are constructed with the distributions of scores for positive and negative cases [37]. The average area under the VGC curve (AUC_{VGC}) was estimated for each group of items ('Normal Tissue', 'Calcifications', and 'Soft Tissue Lesions' group) and each comparison (level 1 vs level 2, level 1 vs level 3, and level 1 vs level 4), by averaging the AUC_{VGC} across observers and items and using the VGC analyser software [38]. The respective 95 % CIs were also calculated. An AUC_{VGC} of > 0.5 indicates higher image quality for the test condition, and an AUC_{VGC} of < 0.5 indicates higher image quality for the reference condition.

The VGA score (VGAS), which corresponds to the average of the ratings across observers and across applicable items for each case, considering the structures found in that case, and the respective 95 % CIs were also calculated for each level. As mentioned, the analysis was performed on a per case basis, and the items did not only represent structures, but also features of breast tissue and lesions. Therefore, the standard expression for the VGAS was adapted as follows:

$$VGAS = \frac{\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{o=0}^{O} G_{c,i,o}}{C \times I \times O}$$

where $G_{c,i,o}$ corresponds to the score for a case c, item i, and observer o, C is the number of cases, I the number of items, and O the number of observers [15–17].

In the event of any missing scores, due to observer error, for a certain item, the missing score was replaced by the mode of the scores for that item and case from the other observers. If the items addressing calcifications were answered for non-calcification cases, those scores were ignored [39].

2.4.3. Criterion validity

To investigate the clinical relevance of the instrument, the

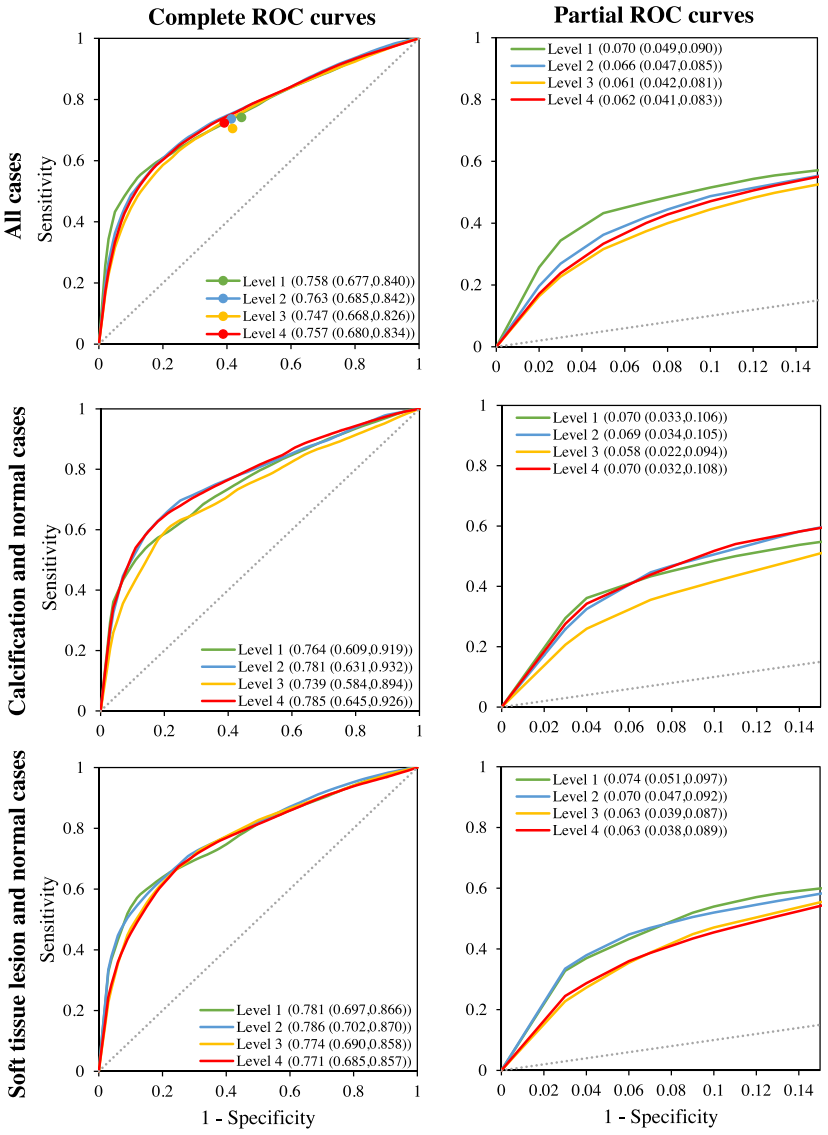

Fig. 1. The (left) complete and (right) partial ROC curves estimated to compare the performance of the radiologists when assessing the reference level (level 1) and test levels (levels 2, 3, and 4) for all cases, and for the calcification (and normal) and soft tissue lesion (and normal) cases separately. The partial ROC curves show the operating region that is most relevant to real world image evaluation. The points in the top left image reflect the operating points of the average reader obtained from the recall decision scores. The AUC_{ROC} and $pAUC_{ROC}$ values and respective 95 % CIs corresponding to each ROC and partial ROC curve are shown in parentheses.

correlation between the VGA and ROC results was calculated by comparing the VGAS and the $pAUC_{ROC}$ estimated from radiologists' scores for each level.

3. Results

3.1. ROC study

No large and statistically significant differences ($p > 0.008$) in radiologists' performance were observed in the lesion detection performance among the four levels of quality when comparing the average AUC_{ROC} estimated for each level, both for all cases and for each type of case separately (Fig. 1 and Table 2). The performance of the radiologists decreased with decreasing image quality, with, as expected, level 1 resulting in the highest $pAUC_{ROC}$ ($pAUC_{ROC}$: 0.070) compared to the levels with lower image quality (level 2, $pAUC_{ROC}$: 0.066; level 3, $pAUC_{ROC}$: 0.061; and level 4, $pAUC_{ROC}$: 0.062). The same pattern was observed when assessing the soft tissue lesions (and normal) cases, but not observed when assessing the calcification (and normal) cases.

On average, sensitivity decreased with decreasing image quality while specificity increased with decreasing image quality (Table 3). However, the observed differences were small to moderate, and not significant ($p > 0.008$).

3.2. VGA study

Fig. 2 shows the AUC_{VGC} corresponding to each comparison calculated for each group of items and for both radiologists and physicists. The AUC_{VGC} was lower than 0.5 for all conditions, which indicates a higher image quality for the reference condition (level 1). For both radiologists and physicists and for all groups of items, the ranking of the comparisons matched the expected. Similar to the ROC results, for the 'Calcifications' group of items, the differences among levels are smaller than for the 'Soft tissue lesions' group of items.

The same ranking of the levels was observed in the VGAS scores for radiologists and physicists (Table 4, and Fig. 3), i.e., a high VGAS corresponded to the level with higher image quality (level 1, VGAS: 6.77 and 6.20 for radiologists and physicists, respectively), while a low VGAS corresponded to the level with lower image quality (level 4, VGAS: 5.67 and 4.75 for radiologists and physicists, respectively). The observed differences in image quality between the four levels were not substantial but were statistically significant. Table 4, and Fig. 3 also show that there were significant differences between the radiologists and the physicists, with the latter being consistently stricter in assessing the image quality of the four levels using the candidate instrument. The average difference between the VGAS calculated for radiologists and physicists was of 10 %.

Table 2
AUC_{ROC} and $pAUC_{ROC}$ calculated for each level of quality and respective difference with 95 % CI corresponding to all cases, and the calcification (and normal) and soft tissue lesion (and normal) cases separately.

	Level 1	Difference		
		Level 1 - Level 2	Level 1 - Level 3	Level 1 - Level 4
All cases (n = 80)				
AUC_{ROC} (95% CI)	0.758 (0.677, 0.840)	−0.005 (-0.036, 0.027)	0.012 (-0.020, 0.043)	0.002 (-0.030, 0.033)
$pAUC_{ROC}$ (95% CI)	0.070 (0.049, 0.090)	0.004 (-0.004,0.011)	0.009 (0.001, 0.016)	0.008 (0.001, 0.015)
Calcification and normal cases (n = 47)				
AUC_{ROC} (95% CI)	0.764 (0.609,0.919)	−0.017 (-0.054,0.020)	0.026 (-0.012,0.063)	−0.021 (-0.058,0.016)
$pAUC_{ROC}$ (95% CI)	0.070 (0.033,0.106)	0.001 (-0.010, 0.012)	0.012 (-0.001,0.023)	−0.000 (-0.011,0.011)
Soft tissue lesion and normal cases (n = 63)				
AUC_{ROC} (95% CI)	0.781 (0.697,0.866)	−0.005 (-0.039,0.031)	0.007 (-0.028,0.043)	0.010 (-0.025,0.046)
$pAUC_{ROC}$ (95% CI)	0.074 (0.051,0.097)	0.005 (-0.003,0.013)	0.011 (0.003,0.019)	0.011 (0.003,0.019)

Table 3
Average sensitivity and specificity (in %) across radiologists of lesion detection calculated for each level of quality.

	Level 1	Level 2	Level 3	Level 4
Sensitivity	74.1 (59/80)	73.6 (58/80)	70.5 (56/80)	72.3 (57/80)
Specificity	55.4 (44/80)	58.6 (46/80)	58.2 (46/80)	61.8 (49/80)

3.3. Criterion validity

The VGAS versus $pAUC_{ROC}$ calculated for each level is shown in Fig. 4. The ranking of the four levels of quality was generally the same for the two methods and a distinction between the levels with lower quality (levels 2, 3, and 4) and the level with higher quality (level 1) is observed. Only levels 3 and 4 were in reverse order in the ROC results when compared to the VGA results. However, the differences between these two levels were very small.

4. Discussion

A candidate instrument to assess image quality in DM was previously developed based on clinically relevant features from breast tissue and breast lesions. In this study, this instrument was evaluated by investigating the correlation between the instrument performance and clinical performance using ROC analysis. Additionally, the relationship of the items to the instrument, and the possibility of the instrument being used by physicists were also investigated.

Overall, no large differences in image quality were observed between the four levels of quality in the ROC and VGA studies. However, the ranking of the four levels was, generally, consistent across the two methods (Fig. 4). These results demonstrated that there is a correlation between the visual grading instrument and the ROC analysis to assess mammographic image quality (R-squared: 0.852). Also supporting this correlation is the finding that, in both ROC and VGC results, the impact of the different levels of quality was less clear in the calcification cases than in the soft tissue lesion cases (Figs. 1 and 2). As a proof of concept, these findings showed that the candidate instrument is potentially able to characterise image quality variations that are clinically relevant in DM. This means that clinical image assessment can be performed in an easier and less laborious way than if performed using ROC methods. This instrument requires the collection and inclusion of cases with lesions with different appearance in the set, to be representative of the type of lesions found in the clinical setting. It is, however, not necessary to include a large number of them, as shown in this study (in the ROC study, 60 lesion-present cases were included, while in the VGA study, only 22 lesion-present cases were included). Thus, with the VGA it is possible to provide clinically and statistically meaningful results without the need for collecting a large number of cases as needed for an ROC study.

The agreement between the VGA and ROC methods has been previously investigated [20–22,40]. In a study by Zanca et al., the VGA and ROC analysis were compared in the assessment of different image-processing algorithms in DM [20]. The VGA was performed using the criteria adapted for the European Guidelines [6,11], and compared to location-specific variant free-response ROC (FROC) for the specific task of microcalcification detection. No correlation between the two methods was found. The authors reported the lack of correlation might at least partly be attributed to investigating only one type of lesion,

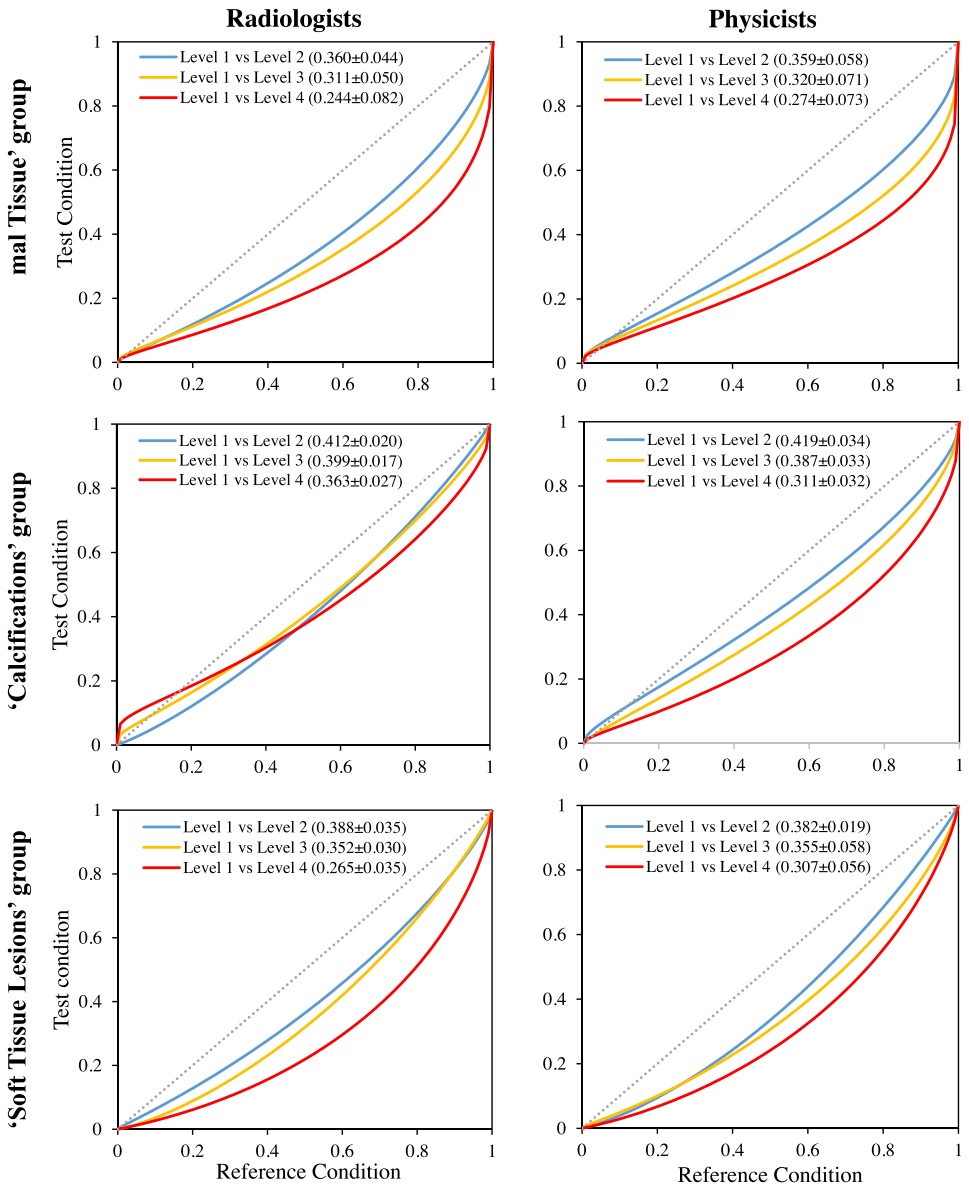

Fig. 2. Average VGC curves calculated across radiologists and items for each group of items and each test condition (levels 2, 3, and 4) in relation to the reference condition (level 1) and corresponding to (left) radiologists and (right) physicists. The AUC_{VGC} values and respective standard deviations corresponding to each VGC curve are shown in parentheses. An AUC_{VGC} of < 0.5 indicates higher image quality for the reference condition.

Table 4
VGAS with 95 % CI calculated for radiologists' or physicists' ratings and for each level of quality and respective differences.

	Level 1	Level 2	Level 3	Level 4
Radiologists	6.77 (6.64,6.90)	6.15 (6.00,6.30)	5.82 (5.65,5.98)	5.37 (5.18,5.55)
Physicists	6.20 (6.08,6.32)	5.55 (5.43,5.68)	5.23 (5.11,5.36)	4.75 (4.60,4.89)
Difference (Radiologists - Physicists)	0.57	0.60	0.58	0.62
Percentage difference (%)	8	10	10	12

microcalcifications, via a FROC study. In contrast, this study performed the VGA using the candidate instrument, which involves items of features of calcifications and soft tissue lesions, and the ROC study was conducted for the lesion detection task, including calcification and soft tissue lesion detection. Therefore, the inclusion of the same type of lesions in both studies might have contributed to the increased agreement between the VGA and ROC methods. Some of the differences between the VGA and ROC methods may, however, be related to the fact that with the VGA it is possible to detect issues that affect the reading task, even if not the detection task. That is because with the VGA, the whole image is assessed in detail when rating specific features about texture, density, margins, etc., of the breast tissue. Therefore, good visibility of normal tissue structures may not be the crucial prerequisite to improve clinical performance, and only assessing the visibility of normal tissue structures in the image might not be sufficient to correlate image quality assessment with clinical performance. However, normal tissue depiction still needs to be taken into account in order to assess possible issues in the image that affect the reading task. Although an increased agreement between the VGA and ROC methods was observed in this study, it is important to mention that a different study design, including the selected cases and the changes in image quality, might have led to a different result. Therefore, it is of interest to continue with the validation of the candidate instrument and investigate it in other VGA studies.

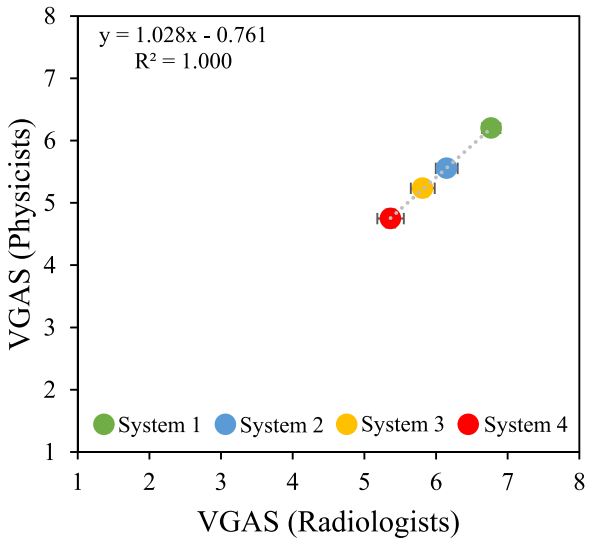

Fig. 3. Comparison of the VGAS and respective 95 % CIs calculated from the physicists' and radiologists' ratings.

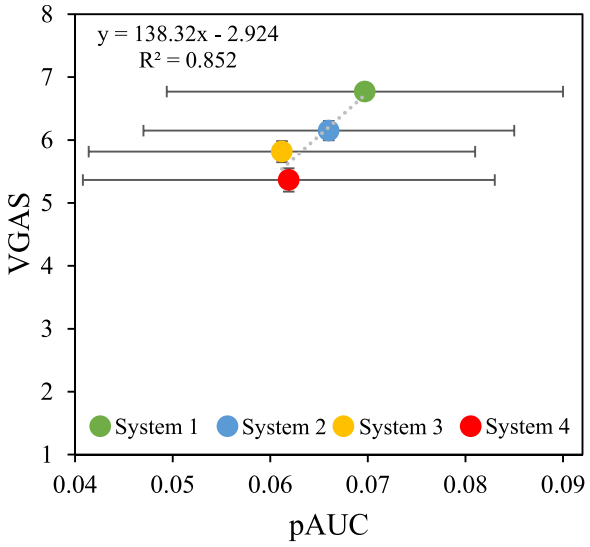

Fig. 4. VGAS versus $pAUC_{ROC}$ and respective 95 % CI calculated for each level of quality.

Currently, radiologists with extensive experience in mammography interpretation and image quality assessment perform the clinical assessment in type testing. As in some countries, these are in short supply, it is of interest to investigate whether professionals from other groups expert in breast imaging evaluation, like physicists and radiographers, could also perform this assessment. This does not mean that the radiologists would be replaced by these other professionals. Instead, a multidisciplinary team could be involved in this assessment and the different professionals could complement each other in assessing image quality from different perspectives. In this study, the difference in outcomes between radiologists and physicists when using the candidate instrument to assess image quality was investigated, and, although the differences between radiologists and physicists in performing this assessment were statistically significant, they were not large and there was a correlation between the two groups (average percent difference of 10 %, R-squared: 1.000). Moreover, the VGAS values showed that the physicists were consistently stricter in using the scale, i.e., their scores were consistently lower than the radiologists' scores. This could mean that the physicists were more cautious in assessing the quality of an image as acceptable for their clinical colleagues to have to work with. It is perhaps not surprising that a group of people from a different profession, even if experts in the field, would be more cautious in accepting the images that a different professional group has to work with. Also, as opposed to radiologists, physicists are not trained for the clinical task, and therefore, they may not know exactly what is the quality needed for interpreting a mammogram. In the future, structured training on the use of the candidate instrument, including the scale and numerous examples of images of varying quality could be provided to expert observers in breast imaging evaluation, like radiographers and physicists. Moreover, as previously mentioned, this study should be extended to investigate the performance of radiographers when using the instrument to assess image quality. Finally, in this case the instrument demonstrated to work independently of the professionals that used it, radiologists and physicists. However, it would still be valuable to verify this result among other experts in breast imaging evaluation.

A validation in terms of construct validity and reliability was also performed and included in the online supplement. Internal consistency reliability of the instrument was supported by high Cronbach's alpha coefficient (Cronbach's α) values (> 0.80). However, there were 4 out of 18 items that showed lower item-total correlations or deviated item-total correlation values in comparison to the other items, or that also resulted in increased Cronbach's α if item removed. Although these items demonstrated rather poor construct validity, there was not enough evidence to remove them from the instrument: the correlation between the VGA and the ROC methods did not improve if these items were removed, and in the previous study, most of these items demonstrated excellent content validity [24]. Validation studies represent a process where no simple test is either necessary or sufficient, and in this perspective no analysis is more important than another [27]. The final interpretation is the collective one of a range of evaluation tests.

The findings of this study showed that a lower VGAS was associated with a negative impact on clinical performance caused by low image quality. However, the impact of the levels with lowest image quality (levels 3 and 4) and the difference between these two levels was not so obvious. Due to time and evaluation capacity restrictions, decisions in the design were made that may have contributed to the relatively small differences between the four levels of quality, for instance the simulated levels represented levels of quality identified as plausible to be encountered during type testing in a previous study [19]. For this reason, and also considering that these images would be part of an observer study that should be pragmatically doable, extreme cases of low image quality were not included.

5. Conclusions

A candidate instrument previously developed to assess image quality in digital mammography was evaluated and it showed potential to correlate to clinical performance. This instrument can be used in testing procedures and studies to assess clinically-relevant image quality variations in digital mammography, which may be performed by professionals from different groups, such as radiologists, radiographers, and physicists. In this way, testing procedures can be optimised to account for clinically-relevant image quality issues in digital mammography that can now be indirectly assessed with the candidate instrument.

Statement of ethics

The authors have no ethical conflicts to disclose.

Funding source

Alistair Mackenzie was funded as part of the OPTIMAM2 project and is supported by Cancer Research UK (grant, number: C30682/A17321).

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Joana Boita: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Ruben E. van Engen:** Conceptualization, Methodology, Writing - review & editing. **Alistair Mackenzie:** Conceptualization, Methodology, Writing - review & editing. **Anders Tingberg:** Conceptualization, Methodology, Writing - review & editing. **Hilde Bosmans:** Conceptualization, Methodology, Writing - review & editing. **Anetta Bolejko:** Conceptualization, Methodology, Writing - review & editing. **Sophia Zackrisson:** Conceptualization, Methodology, Writing - review & editing. **Matthew G. Wallis:** Conceptualization, Methodology, Writing - review & editing. **Debra M. Ikeda:** Conceptualization, Methodology, Writing - review & editing. **Chantal van Ongeval:** Conceptualization, Methodology, Writing - review & editing. **Ruud Pijnappel:** Conceptualization, Methodology, Writing - review & editing. **Mireille Broeders:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing. **Ioannis Sechopoulos:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

Some of the authors of this manuscript declare relationships with companies: Ioannis Sechopoulos: research agreements, Siemens Healthcare, Canon Medical Systems, ScreenPoint Medical, Sectra Benelux, and Volpara Health Technologies, and speaker agreements, Siemens Healthcare; Mireille Broeders: speaker agreements, Siemens Healthcare and Hologic; Sophia Zackrisson Speaker agreements Siemens Healthcare, research agreement ScreenPoint Medical; Anders Tingberg: Research agreements, Siemens Healthcare; Matthew Wallis' institution has received grants from Philips; Chantal Van Ongeval: speaker agreement Siemens Healthcare; Hilde Bosmans: research agreements, Siemens Healthcare and GE Healthcare; Ruud Pijnappel: speaker agreement Hologic; and Debra Ikeda: consultant to Hologic.

Acknowledgments

The VISUAL group, which is the group that includes all the observers that participated in the project from which this study is part of: F. Jansen, L. Duijm, H. de Bruin, A. Bluekens, I. Andersson, C. Behmer, E. Johansson, K. Rönnow, K. Taylor, F. Kilburn-Toppin, P. Moyle, M. van Goethem, R. Prevos, A. van Steen, N. Salem, S. Pal, E. Rosen, H. Lelivelt, K. Michielsen, L. Cockmartin, N. Phelan, P. Baldelli, S. Schopphoven. The authors thank the Medical Physics Department, Royal Surrey NHS Foundation Trust for the use of mammograms from the OPTIMAM Mammography Image Database funded by Cancer Research UK (C30682/A28396), Sander van Woudenberg for all the help with the image processing, and Gustav Hellgren for all the help with the setup of the observer studies performed by the Swedish readers.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ejrad.2021.109686.

References

[1] S.H. Taplin, C.M. Rutter, C. Finder, M.T. Mandelson, F. Houn, E. White, Screening mammography: clinical image quality and the risk of interval breast cancer, Am. J. Roentgenol. 178 (2002) 797–803.
[2] A.K. Abdullah, J. Kelly, J.D. Thompson, C.E. Mercer, R. Aspin, P. Hogg, The impact of simulated motion blur on lesion detection performance in full-field digital mammography, Br. J. Radiol. 90 (2017), 20160871.
[3] L.W. Bassett, D.M. Farria, S. Bansal, M.A. Farquhar, P.A. Wilcox, S.A. Feig, Reasons for failure of a mammography unit at clinical image review in the american college of radiology mammography accreditation program, Radiology 215 (2000) 698–702.
[4] S.A. Feig, Image quality of screening mammography: effect on clinical outcome, Am. J. Roentgenol. 178 (2002) 805–807.
[5] G.W. Eklund, G. Cardenosa, W. Parsons, Assessing adequacy of mammographic image quality, Radiology 190 (1994) 297–307.
[6] C. Van Ongeval, H. Bosmans, A. Van Steen, Current challenges of full field digital mammography, Radiat. Prot. Dosimetry 117 (2005) 148–153.
[7] E.D. Pisano, E.B. Cole, B.M. Hemminger, M.J. Yaffe, S.R. Aylward, A.D. Maidment, R.E. Johnston, M.B. Williams, L.T. Niklason, E.F. Conant, L.L. Fajardo, D.B. Kopans, M.E. Brown, S.M. Pizer, Image processing algorithms for digital mammography: a pictorial essay, Radiographics 20 (2000) 1479–1491.
[8] F. Zanca, J. Jacobs, C. Van Ongeval, F. Claus, V. Celis, C. Geniets, V. Provost, H. Pauwels, G. Marchal, H. Bosmans, Evaluation of clinical image processing algorithms used in digital mammography, Med. Phys. 36 (2009) 765–775.
[9] E.B. Cole, E.D. Pisano, D. Zeng, K. Muller, S.R. Aylward, S. Park, C. Kuzmiak, M. Koomen, D. Pavic, R. Walsh, J. Baker, E.I. Gimenez, R. Freimanis, The effects of gray scale image processing on digital mammography interpretation performance, Acad. Radiol. 12 (2005) 585–595.
[10] L.M. Warren, R.M. Given-Wilson, M.G. Wallis, J. Cooke, M.D. Halling-Brown, A. Mackenzie, D.P. Chakraborty, H. Bosmans, D.R. Dance, K.C. Young, The effect of image processing on the detection of cancers in digital mammography, Am. J. Roentgenol. 203 (2014) 387–393.
[11] R. van Engen, H. Bosmans, P. Heid, B. Lazzari, S. Schopphoven, M. Thijssen, K. Young, Digital mammography update. European protocol for the quality control of the physical and technical aspects of mammography screening. S1, part 2: European type testing, in: N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, L. von Karsa (Eds.), Eur. Guidel. Qual. Assur. Breast Cancer Screen. Diagnosis. Fourth Ed. Suppl., European Commission, Office for Official Publications of the European Union, Luxembourg, 2013, pp. 55–71.
[12] C.E. Metz, Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems, J. Am. Coll. Radiol. 3 (2006) 413–422.
[13] D.P. Chakraborty, K.S. Berbaum, Observer studies involving detection and localization: modeling, analysis, and validation. Med. Phys., 2004, pp. 2313–2330.
[14] N.A. Obuchowski, M.L. Lieber, K.A. Powell, Data analysis for detection and localization of multiple abnormalities with application to mammography, Acad. Radiol. 7 (2000) 516–525.
[15] A. Tingberg, Quantifying the Quality of Medical x-ray Images. An Evaluation Based on Normal Anatomy for Lumbar Spine and Chest Radiography, Lund University, 2000.
[16] L.G. Månsson, Methods for the evaluation of image quality: a review, Radiat. Prot. Dosimetry 90 (2000) 89–99, https://doi.org/10.1093/oxfordjournals.rpd.a033149.
[17] M. Båth, Evaluating imaging systems: practical applications, Radiat. Prot. Dosimetry 139 (2010) 26–36.
[18] P. Sund, M. Båth, S. Kheddache, L.G. Månsson, Comparison of visual grading analysis and determination of detective quantum efficiency for evaluating system performance in digital chest radiography, Eur. Radiol. 14 (2004) 48–58.
[19] J. Boita, A. Bolejko, S. Zackrisson, M.G. Wallis, D.M. Ikeda, C. Van Ongeval, R. E. van Engen, A. Mackenzie, A. Tingberg, H. Bosmans, R. Pijnappel, M. Broeders, I. Sechopoulos, How does image quality affect radiologists' perceived ability for image interpretation and lesion detection in digital mammography? Eur. Radiol. (2021) 1–9.
[20] F. Zanca, C. Van Ongeval, F. Claus, J. Jacobs, R. Oyen, H. Bosmans, Comparison of visual grading and free-response ROC analyses for assessment of image-processing algorithms in digital mammography, Br. J. Radiol. 85 (2012) e1233–e1241.
[21] A. Tingberg, C. Herrmann, B. Lanhede, A. Almen, J. Besjakov, S. Mattsson, P. Sund, S. Kheddache, L.G. Mansson, Comparison of two methods for evaluation of the image quality of lumbar spine radiographs, Radiat. Prot. Dosimetry 90 (2000) 165–168.
[22] P. Sund, C. Herrmann, A. Tingberg, S. Keddache, L.G. Mansson, A. Almen, S. Mattsson, Comparison of two methods for evaluating image quality of chest radiographs, Med. Imaging 2000 Image Percept. Perform., SPIE (2000) 251–258.
[23] H. Kundel, Images, image quality and observer performance: new horizons in radiology lecture, Radiology 132 (1979) 265–271.
[24] J. Boita, A. Bolejko, S. Zackrisson, M.G. Wallis, D.M. Ikeda, C. Van Ongeval, R. E. van Engen, A. Mackenzie, A. Tingberg, H. Bosmans, R. Pijnappel, I. Sechopoulos, M. Broeders, Development and content validity evaluation of a candidate instrument to assess image quality in digital mammography: a mixed-method study, Eur. J. Radiol. (2020), 109464.
[25] D.L. Streiner, G.R. Norman, J. Cairney, Health Measurement Scales: a Practical Guide to Their Development and Use, 5th ed., Oxford University Press, New York, NY, 2015.
[26] P.M. Fayers, D. Machin, Quality of Life: Assessment, Analysis, and Interpretation, John Wiley & Sons, Ltd, West Sussex, 2000.
[27] J. Hobart, S. Cano, Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods, Health Technol. Assess. (Rockv) 13 (2009) 1–200.
[28] L.J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika 16 (1951) 297–334.
[29] M. Tavakol, R. Dennick, Making sense of Cronbach's alpha, Int. J. Med. Educ. 2 (2011) 53–55.
[30] M.D. Halling-Brown, L.M. Warren, D. Ward, E. Lewis, A. Mackenzie, M.G. Wallis, L. S. Wilkinson, R.M. Given-Wilson, R. McAvinchey, K.C. Young, OPTIMAM

mammography image database: a large-scale resource of mammography images and clinical data, Radiol. Artif. Intell. 3 (2021), e200103.
[31] N.A. Obuchowski, Sample size tables for receiver operating characteristic studies, Am. J. Roentgenol. 175 (2000) 603–608.
[32] A. Mackenzie, D.R. Dance, A. Workman, M. Yip, K. Wells, K.C. Young, Conversion of mammographic images to appear with the noise and sharpness characteristics of a different detector and x-ray system, Med. Phys. 39 (2012) 2721–2734.
[33] A. Mackenzie, D.R. Dance, O. Diaz, K.C. Young, Image simulation and a model of noise power spectra across a range of mammographic beam qualities, Med. Phys. 41 (2014) 1–14, 121901.
[34] A. Mackenzie, H.L. Dunn, J. Boita, D.R. Dance, K.C. Young, A method to modify mammography images to a appear as if acquired using different radiographic factors, Proc. SPIE (2019), 10948, 109482F.
[35] J. Boita, A. Mackenzie, I. Sechopoulos, Validation of a method to simulate the acquisition of mammographic images with different techniques, Proc. SPIE (2019), 10948, 109481K.
[36] M. Hakansson, S. Svensson, S. Zachrisson, A. Svalkvist, M. Bath, L.G. Mansson, VIEWDEX: an efficient and easy-to-use software for observer performance studies, Radiat. Prot. Dosimetry 139 (2010) 42–51.
[37] M. Båth, L. Månsson, Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation, Br. J. Radiol. 80 (2007) 169–176.
[38] M. Båth, J. Hansson, VGC Analyzer: A software for statistical analysis of fully crossed multiple-reader multople-case visual grading characteristics studies, Radiat. Prot. Dosimetry 169 (2016) 46–53.
[39] S. van Buuren, Flexible Imputation of Missing Data, CRC, 2018.
[40] A. Tingberg, M. Båth, M. Håkansson, J. Medin, J. Besjakov, M. Sandborg, G. Alm-Carlsson, S. Mattsson, L.G. Månsson, Evaluation of image quality of lumbar spine images: a comparison between FFE and VGA, Radiat. Prot. Dosimetry 114 (2005) 53–61.