Comparison of visual grading and free-response ROC analyses for assessment of image-processing algorithms in digital mammography

¹F ZANCA, PhD, ²C VAN ONGEVAL, MD, PhD, ²F CLAUS, MD, PhD, ¹J JACOBS, MSc, ²R OYEN, MD, PhD and ¹H BOSMANS, PhD

¹Leuven University Center of Medical Physics in Radiology, University Hospitals Leuven, Leuven, Belgium, and

²Department of Radiology, University Hospitals Leuven, Leuven, Belgium

Objective: To compare two methods for assessment of image-processing algorithms in digital mammography: free-response receiver operating characteristic (FROC) for the specific task of microcalcification detection and visual grading analysis (VGA).

Methods: The FROC study was conducted prior to the VGA study reported here. 200 raw data files of low breast density (Breast Imaging–Reporting and Data System I–II) mammograms (Novation DR, Siemens, Germany)—100 of which abnormal—were processed by four image-processing algorithms: Raffaello (IMS, Bologna, Italy), Sigmoid (Sectra, Linköping, Sweden), and OpView v. 2 and v. 1 (Siemens, Erlangen, Germany). Four radiologists assessed the mammograms for the detection of microcalcifications. 8 months after the FROC study, a subset (200) of the 800 images was reinterpreted by the same radiologists, using the VGA methodology in a side-by-side approach. The VGA grading was based on noise, saturation, contrast, sharpness and confidence with the image in terms of normal structures. Ordinal logistic regression was applied; OpView v. 1 was the reference processing algorithm.

Results: In the FROC study all algorithms performed better than OpView v. 1. From the current VGA study and for confidence with the image, Sigmoid and Raffaello were significantly worse ($p < 0.001$) than OpView v. 1; OpView v. 2 was significantly better ($p = 0.01$). For the image quality criteria, results were mixed; Raffaello and Sigmoid for example were better than OpView v. 1 for sharpness and contrast (although not always significantly).

Conclusion: VGA and FROC discordant results should be attributed to the different clinical task addressed.

Advances in knowledge: The method to use for image-processing assessment depends on the clinical task tested.

Received 26 January 2012

Revised 14 May 2012

Accepted 28 May 2012

DOI: 10.1259/bjr/22608279

© 2012 The British Institute of Radiology

Image processing applied in two-dimensional digital mammography has been suggested as a means of improving image quality, especially for patients with dense breast tissue [1–6]. In practice, edge enhancement, histogram equalisation and other greyscale adjustments [3, 4] are applied that alter the pixel values of the images in order to enhance the conspicuity of relevant findings like masses and microcalcifications. The algorithms may vary substantially among different vendors and there is no consensus on what processing algorithm is considered optimal among radiologists [4]. The process by which image-processing algorithms and settings are optimised often remains undocumented, and testing of algorithms is neglected in any European acceptance protocol for digital mammography systems [7]. Whether image processing provides improved diagnostic accuracy or simply improves the subjective appearance of images deserves further study.

With the advent of new technologies such as digital breast tomosynthesis, manufacturers have been and are introducing even more complex image processing and (iterative) reconstruction. Many researchers are studying the effects of such algorithms on clinical performance [8–13]. In most of these studies image quality has been quantified using physical measurements [8, 9, 13] (such as noise, signal-to-noise ratio, detective quantum efficiency, modulation transfer function) or psychophysical measurements [10, 11] (using, for example, contrast-detail phantoms). Although physical methods are of fundamental importance in describing the image quality, establishing the link between physical image quality measures and clinical utility has been pursued for decades and their relationship is not yet fully understood [14, 15].

Receiver operating characteristic (ROC) analysis [16] and its location-specific variants [17–19] are currently considered the best methods to quantify and report diagnostic performance: they measure the ability of an observer to detect and correctly interpret pathological structures, such as microcalcifications and masses in a

Address correspondence to: Mrs Federica Zanca, Department of Radiology, UZ Gasthuisberg, Herestraat 49, 3000 Leuven, Belgium. E-mail: federica.zanca@uzleuven.be

mammogram. Unfortunately, these methods require a large number of patient images and a large number of pathological lesions to reach a sufficient statistical power; moreover, the true health state of the patient must be known to classify an image as normal or abnormal, requiring a follow-up of typically 1 year to confirm benign and normal cases. Further, the cases need to be subtle so that false positives occur during the image interpretation. These requirements make the measurement of clinical performance difficult in practice and very time consuming. ROC analysis is typically performed at the introduction of new imaging modalities; for the evaluation of more subtle updates in a technology, the required effort may be considered excessive.

A simpler method has been published by the European Commission (CEC) [20] and is based upon the scoring of image quality by inspecting normal anatomical structures; good visibility of these (normal) structures on a radiograph is considered to define appropriate image quality and accurate diagnosis. The set of anatomical criteria is specific for the given type of examination [20–22], and visual grading analysis (VGA) is typically applied to evaluate them. Images have to be graded with an absolute score by a number of observers or alternatively one can use the side-by-side approach, which involves simultaneous viewing of two images, and the score is meant to express the comparison of the two images. The number of studies using this approach has increased in the past few years [15, 23–27].

Correlation between VGA experiments and more objective ROC studies has previously been investigated, but the results of the few published studies are contradictory [28–30]. Pilgram et al [28] found a strong correlation between diagnostic performance and the subjective evaluation of image quality on the diagnosis of craniosynostosis. They studied in detail the responses of one observer, who read radiographs having different image quality (obtained with three different methods) for a set of patients with this diagnosis and found that the increase in diagnostic performance resulted primarily from increased specificity, suggesting that, when image quality is being evaluated, specificity and the evaluation of normal structures should be the focus of attention. In the studies of Tingberg et al [29, 30], the influence of different characteristic curves on the diagnostic quality of radiographs of the lumbar spine and the chest evaluated with both VGA and free-response forced error (FFE) (an ROC variant) experiments showed mixed results. In the first experiment [29], lumbar spine images were manipulated by adding artificial lesions, and image processing was applied to simulate image appearance from three different screen–film combinations; VGA evaluation and FFE results were in good agreement. In the later paper [30], the authors found that images with greater contrast than the original had significantly improved image quality in the VGA study although this did not affect detectability in the FFE experiment, indicating that VGA has a stronger discriminative power than FFE. These contradictory results suggest that the influence of image processing on image interpretation justifies further research.

The aim of this study was to investigate the role of a VGA method in the assessment of image-processing algorithms for digital mammography relative to an

earlier applied free-response ROC (FROC) study that investigated the same algorithms but for the specific task of microcalcification detection.

Methods and materials

Collection of mammographic images

200 raw [Digital Imaging and Communications in Medicine *i.e.* (DICOM) “for processing”] craniocaudal mammograms were randomly selected from the clinical images routinely acquired in our mammography department. Only images without lesions and rated as Breast Imaging–Reporting and Data System (BI-RADS) I and II were selected. All images had been acquired under automatic exposure conditions on a Siemens Mammomat Novation DR (Erlangen, Germany) at 25–32 kVp (average of 28.5 kVp). Compressed breast thickness varied from 18 to 71 mm (average of 46 mm); breast densities, according to BI-RADS [31], varied from fatty to extremely dense: 13.5% of the 200 mammograms depicted an almost entirely fatty breast, 41% depicted scattered fibroglandular densities, 19% were heterogeneously dense and 26.5% were extremely dense. Patient age varied from 35 to 81 years (average of 51 years). Mean glandular dose ranged from 0.52 to 1.85 mGy (average of 0.99 mGy).

In order to create abnormal mammograms, clusters of realistic microcalcifications were simulated into half (100) of these raw images using a methodology previously described and validated [32]; the number of clusters per image varied between 1 and 3 (average 1.4). All morphology types as described by LeGal et al [33] were equally represented. The background at the location of the simulated lesions was selected as being dense, mixed or fatty areas of the breast, in equal proportion.

Image processing

All 200 raw mammograms were processed using the following four algorithms: 1=IMS Raffaello 1.2 (IMS, Bologna, Italy); 2=Sigmoid (Sectra Mamea AB, Linköping, Sweden); and 3=OpView v. 2 and 4=OpView v. 1 (Siemens, Erlangen, Germany). For each algorithm, the default parameter settings were applied. More details on these algorithms are given elsewhere [34]. These processed images were then used for the FROC and for VGA observer performance experiments.

Observer performance experiments

Evaluation of image quality using the FROC paradigm

The FROC study had been conducted 8 months prior to the VGA study reported here; for details of the FROC study see Zanca et al [34]. 800 images (200 × 4 image-processing algorithms) were objectively assessed for the detection of clusters of microcalcifications by 4 radiologists: 3 with more than 15 years' experience and 1 resident in the final year of radiology training. All

images were interpreted using the FROC method. In this paradigm, the observer is free to mark and rate as many suspicious regions as are considered reportable; the mark is the location of the suspicious region and the rating is the confidence level that the region harbours a lesion. The rating scale in the study was a five-point scale (1=highly unlikely to be a cluster, 5=very confident it is a cluster) and each mark was classified as lesion localisation if it was within the smallest rectangle containing the cluster, otherwise it was classified as non-lesion localisation. The Simulating and Scoring Artefacts for Radiological Applications (SARA) software package [35] was used to collect the localisation and rating information. All images were viewed on a five-megapixel liquid crystal display monitor specified for mammography applications (Barco LCD 5 MP MFGD5621HD; Barco, Kortrijk, Belgium) and calibrated to the DICOM greyscale standard display function [36]. The ambient light level in the reading room was between 5 and 10 lux, in accordance with recommended limits [7].

Four reading sessions were organised as follows: each observer interpreted 200 images in each session with a minimum interval of 1 week between two consecutive sessions, in order to reduce the influence of learning effects on the results. Within each session, images processed with each of the four algorithms were presented randomly, with a given mammogram shown only once per session. A typical reading session lasted about 5 h including pauses. The images were initially displayed at manufacturer-recommended window/level settings for the algorithms, but the observers were allowed to alter these settings as well as to zoom and to pan. Rating data were analysed using jackknife AFROC (JAFROC) software (www.devchakraborty.com/downloads.html; accessed March 2009) and differences in figures of merit (FoMs) between all pairs of processing algorithms were calculated, along with 95% confidence intervals (CIs).

Evaluation of image quality using visual grading analysis

8 months after the conclusion of the four-modality FROC study [34], a subset of the 800 images was interpreted by the same radiologists, this time using the VGA methodology in a side-by-side approach. The images were shown on the same display as used for the FROC study and the readers were allowed to window level, zoom and pan. The subset encompassed all original 200 raw mammograms (and included therefore also abnormal cases), each one processed with 2 out of the 4 image-processing algorithms, with the pair of processing algorithms being randomly sampled. It was not possible to show every combination to every reader as this would have required a very large number of images to be read. The processing pairs were randomly selected among the four algorithms: Raffaello, Sigmoid, OpView v. 2 or v. 1. The same image, processed using two different algorithms, was then displayed with the first processing algorithm on the left and the other processing algorithm on the right monitor, for a side-by-side ranking. In this way, approximately equal numbers of processing pairings were compared.

The questions asked in the side-by-side evaluation were related to specific characteristics of the image as

described by Van Ongeval et al [37]. The investigated parameters were noise, saturation in fatty regions, saturation in dense regions, contrast and sharpness. A typical question was "Processing A has noise compared to processing B". To provide meaningful numerical descriptors, an 11-point rating scale running from -5 to 5 was used to indicate the quality of the images. Finally, we asked the reader to score his or her confidence with the image for breast cancer detection examinations. As for the FROC study, the SARA software [35] was used to collect the rating information. The average time needed by each reader to interpret the images was also recorded.

For each processing pair, the observed relative frequencies of all scores, obtained as the fraction of times an answer occurred and expressed as a percentage, was calculated for the image confidence criteria. In order to account for the position of the image with a particular processing algorithm on either the left or the right side, the signs of the scores were inverted when necessary. Inspection of the bar plots revealed a skewed distribution of the data, indicating that the non-parametric Kruskal-Wallis analysis of variance test between the medians was appropriate for comparisons of the different processing algorithms. A non-parametric multiple comparison test was used as a *post hoc* test. Statistica v. 10 (StatSoft Inc., Tulsa, OK) was used for statistical analysis.

Ordinal logistic regression was also applied to study the influence of the image quality criteria. An appropriate method has been proposed by Smedby and Fredrikson [38] and uses the following regression equation:

$$\text{Logit}[p(y \leq n)] = a_1 \text{PP1} + a_2 \text{PP2} + a_3 \text{PP3} + D_p + E_o - C_n \quad (1)$$

where:

- $\text{Logit}[p(y \leq n)]$ is the probability of obtaining a value of the dependent variable y (one of the image quality criteria in our case) not greater than n , with n being the score given to the reference processing algorithm (OpView v. 1 in this case, see below) (n varies in the current study from -5 to 5).
- PP1-PP3 are dummy-variable-like (they can only assume a value of 0, 1 or -1 for each observation) for the image-processing algorithms.
- a_1 - a_3 are the parameters to be estimated in the logistic model; a negative estimate of the model parameters is associated with a poorer score of the reference processing algorithm than of the predictor processing algorithm.
- D_p , E_o and C_n have a value specific for each patient, observer and image quality level, respectively.

A reference category against which the other processing algorithms are compared is necessary as the logistic model (Equation 1) requires the independent variables not to be linearly dependent on each other. In Equation 1, the number of independent processing algorithms variables is three and not four as OpView v. 1 is used as the reference category. Similarly, a reference observer (Observer 1) has been chosen from the four observers; this was the most experienced radiologist.

Table 1. Difference in figures of merit (FoMs) between all pairings of image processing along with the 95% confidence intervals (CIs). A CI that does not include 0 indicates that the corresponding *p*-value is <0.05, implying a statistically significant difference in FoM

Compared image-processing algorithms	Difference in FoM	95% CI
1–2	0.0409	0.0172, 0.0646 ^a
1–3	–0.0105	–0.0342, 0.0132
1–4	0.0443	0.0206, 0.0680 ^a
2–3	–0.0514	–0.0751, –0.0277 ^a
2–4	–0.0033	–0.0204, 0.0271
3–4	0.0548	0.0311, 0.0785 ^a

1, IMS Raffaello; 2, Sectra Sigmoid; 3, Siemens OpView v. 2; 4, Siemens OpView v. 1.

^aSignificant differences.

Statistica v. 10 was used for statistical analysis. Results of the analysis were used to assess the model by means of the ratio of Pearson's χ^2 test to the respective degrees of freedom: a value close to 1 indicates a good fit. To ascertain whether each independent processing algorithm is a significant predictor of the outcome *y* (rating) when compared with the reference processing algorithm, the Wald statistic was used and a *p*-value calculated; *p*<0.05 was considered significant. The odds ratio (OR) is also reported; an OR of 1 for a specific image quality criterion implies that the probability of getting the same rating score is equally likely for both processing algorithms. An OR>1 implies that the event of getting a higher score is more likely for the reference processing algorithm; an OR<1 indicates the opposite. The larger the value of the OR, the larger the difference is between the two processing algorithms compared.

For a graphical representation of the results, the relative frequencies as predicted by the ordinal logistic model were compared with the observed relative frequencies for the same processing algorithm.

Results

FROC study

Results have shown that a statistically significant difference was found among the different algorithms in terms of the numbers of clusters of microcalcifications detected. The results are reported in Table 1: a CI that does not include zero indicates that the corresponding

p-value is <0.05, implying a statistically significant difference in FoMs (indicated by ^a). A graphical representation of the calculated FoMs per processing algorithm along with its standard deviation, for each observer as well as for the average observer, is given in Figure 1. It can be appreciated that the rankings of the individual observers of the different algorithms were quite consistent.

VGA study

The average time per reader needed to interpret the 200 images of the VGA study was 3.8 ± 0.8 h (range 2.99–4.89 h). The observed relative frequency for the image confidence and for the paired comparisons among all processing algorithms is reported in Figure 2, along with the calculated skew value *S*, for the average reader. The score on the *x*-axis reflects the evaluation of the processing algorithm on the left monitor (first image-processing algorithm listed in the legend of the graph) with respect to the processing algorithm on the right monitor. From Figure 2b–e, it is evident that when one of the two Siemens processing algorithms is compared with the Sectra or IMS processing algorithms, the distribution is skewed to the right with a positive *S* value, indicating that the perception of confidence with the image for the Siemens algorithms was in general associated with a higher score than for the other algorithms. When the two Siemens processing algorithms are compared (Figure 2f), the distribution of the observed score frequency is more gaussian, with an *S* value around 0 (0.06). The Kruskal–Wallis analysis for the image confidence showed statistically significant

Figure 1. Calculated figures of merit with the jackknife AFROC (JAFROC) method, for each and for the average observer for the different image-processing algorithms. Error is reported as $\pm$ standard deviation.

Figure 2. Graphs (a–f) reporting the observed relative frequency of the scores –5 to 5 of all image-processing comparisons during the VGA experiment. S is the skew value.

differences ($p < 0.01$) between the algorithms; the *post hoc* tests showed significant difference among all the processing algorithms ($p < 0.01$); only OpView v. 1 vs v. 2 was not significant ($p = 1$).

The logistic regression model described the data reasonably well, with a ratio between the Pearson χ^2 test and the degrees of freedom close to 1 for all image quality criteria (range 1.0–1.2). No statistically significant difference was found between each of the three observers and the reference observer (p -value ranged from 0.1 to 0.9 for the three observers). Table 2 reports the model estimates a_1 – a_3 as from Equation 1, the p -value to determine whether a processing algorithm was a significant predictor of the outcome, and the OR for each processing algorithm and for each physical criterion. It can be seen that, for the visualisation of noise, the Raffaello processing algorithm performed better than OpView v. 1 (negative model estimate -0.28) and the difference is statistically significant. Similarly, for the visualisation of structures in fatty areas, Raffaello and Sigmoid performed worse than

the reference processing algorithm while OpView v. 2 was better, and all three differences were statistically significant. For the visualisation of structures in dense areas, no statistically significant difference was found between the two Siemens processing algorithms, while Raffaello and Sectra were statistically significantly worse than Siemens OpView v. 1. Sharpness was significantly better for the Sectra processing algorithm, but no difference was found for Raffaello and Siemens OpView v. 2.

In terms of confidence with the image, OpView v. 2 was significantly better than OpView v. 1, whereas the IMS and Sectra processing algorithms were considered significantly worse. For the overall image quality, the ORs for OpView v. 1 vs Raffaello, Sigmoid and OpView v. 2 were, respectively, 14.07, 7.68 and 0.76, showing that the largest difference in image quality existed between the reference image-processing algorithm OpView v. 1 and Raffaello and the smallest between the two Siemens processing algorithms. The same type of reasoning in terms of ORs holds for the other image quality parameters investigated.

Table 2. Results of the logistic regression analysis applied to the data obtained from the visual grading analysis experiment vs OpView v. 1, for each image quality criterion investigated

Image quality criteria	Estimate	p-value	OR	95% CI for OR
Noise				
IMS Raffaello Mammo	-0.28	0.01	0.75	0.60, 0.95
Siemens OpView v. 2	-0.04	0.73	0.96	0.77, 1.21
Sectra Sigmoid	0.05	0.65	1.05	0.84, 1.32
Structures in fatty areas				
IMS Raffaello Mammo	0.92	<0.001	2.52	2.02, 3.13
Siemens OpView v. 2	-0.31	0.005	0.74	0.59, 0.91
Sectra Sigmoid	1.47	<0.001	4.34	3.45, 5.45
Structures in dense areas				
IMS Raffaello Mammo	1.03	<0.001	2.79	2.23, 3.48
Siemens OpView v. 2	0.13	0.23	1.14	0.92, 1.41
Sectra Sigmoid	0.51	<0.001	1.66	1.34, 2.05
Contrast				
IMS Raffaello Mammo	-0.01	0.90	0.99	0.80, 1.22
Siemens OpView v. 2	0.13	0.24	1.13	0.92, 1.40
Sectra Sigmoid	-0.04	0.73	0.96	0.78, 1.19
Sharpness				
IMS Raffaello Mammo	-0.13	0.24	0.88	0.71, 1.09
Siemens OpView v. 2	0.01	0.94	1.01	0.81, 1.25
Sectra Sigmoid	-0.64	<0.001	0.53	0.43, 0.66
Overall image quality				
IMS Raffaello Mammo	2.64	<0.001	14.07	10.69, 18.53
Siemens OpView v. 2	-0.27	0.01	0.76	0.62, 0.95
Sectra Sigmoid	2.04	<0.001	7.68	5.98, 9.86

CI, confidence interval; OR, odds ratio.

Figure 3 represents the graph of the relative frequencies of the VGA scores as predicted by the ordinal logistic model and compared with the observed relative frequencies for the same processing algorithm, for the criteria of confidence with the image. Differences in Figure 3 are small, indicating that the fitted model is realistic.

Discussion

This study compared two approaches (VGA and FROC) for the assessment of image-processing algorithms in digital mammography. The earlier FROC study [34] that had quantified the effect of image processing on the clinical task of detection of microcalcification clusters is the implicit gold standard.

The results of the VGA experiment for the confidence with the image (which showed that only OpView v. 2 was significantly better than OpView v. 1) do not follow the ranking obtained from the earlier FROC experiment [34] (which showed that all three algorithms performed significantly better than OpView v. 1). From this, it is doubtful that the fulfilment of the investigated quality criteria for the visualisation of normal structures in a VGA experiment is sufficient for accurate detection of microcalcifications. Similar results, but for a different clinical task, were found in the Tingberg et al study [29]: they investigated the correlation between a VGA experiment and the objective ROC study for the detection of lumbar lesions, and also found no agreement between the two methods.

The validity of using VGA (and the CEC image criteria) as a measure of clinical image quality or performance has been questioned before, especially because of the subjectivity of the evaluation [39] and of

the high intra- and interobserver variability [37, 39]. In this specific study, observer variability did not seem to limit the use of VGA (no statistical significance was found between each observer and the reference observer; $p > 0.05$), perhaps because the processing comparisons were carried out within one radiology unit, where the radiologists follow a similar method of interpreting and discussing image quality criteria. Incidentally, this underlines some of the benefits of discussing image quality regularly in a team of radiologists and at teaching sessions. In our opinion, the subjectivity or the low validity of the VGA approach is not the reason for the discrepancy between the two methods: we attribute it to the different clinical tasks addressed. In the VGA experiment the task was the evaluation of the visibility of normal anatomy in terms of specific image quality criteria, when different processing algorithms are used; the FROC study instead aimed to define how these image-processing algorithms affect diagnosis in terms of microcalcification detection. Good visibility of normal anatomy seems a needed, but not a sufficient, prerequisite of a performance indicator.

At this point it seems appropriate to consider a question that may define future test methodologies for image-processing algorithms: is it enough that the radiologist feels confident with the global appearance of an image or should it be tested whether a processing algorithm also guarantees better lesion detection? Both aspects are very relevant for the daily use of an imaging system, and therefore both testing methods, VGA and a task-based evaluation, are useful.

A test of diagnostic detection accuracy by means of objective ROC experiments remains necessary and fundamental when testing image processing, as diagnostic accuracy has a direct impact on the patient. The

Figure 3. Graphs (a–c) reporting the comparison between predicted and observed score frequency for each image-processing comparison with respect to the reference algorithm Siemens OpView v. 1, for the overall image confidence criteria.

drawback of this approach is that the task of conducting such studies for every possible diagnosis is daunting and very time-consuming. Still, at least for new algorithms, this is the approach which should be pursued. On the other hand, a radiologist should not be forced to work with images he or she does not feel confident with. The favourable role of VGA experiments in assessing such algorithms is at a different level from absolute patient

outcome. First, they can indicate the clinical acceptability of the presentation of the images that only radiologists can give and that lately has been very critical to device utilisation. For example, a processing algorithm with a high noise perception in the images by the radiologists will most likely not be widely adopted in clinical practice, even if good diagnostic accuracy had been proved. Second, VGA experiments provide an overall assessment of visibility of normal structures in terms of contrast, noise and sharpness, which ideally allow pathology to be excluded more reliably. This is especially true in a breast-screening environment, where only about 5 in 1000 images will contain pathology and a prompt decision is required. Third, visual grading studies are also relatively easy to conduct, especially in comparison with ROC studies: the average reading time per observer in the current investigation was about 3.5 h for the VGA method and 15 h for the FROC experiment. Note also that the number of images read in the FROC experiment was 4 times (800) the number of images read in the VGA experiment (200). The longest time however was for the preparation of sets of images with subtle lesions. Whether these lesions are retrieved from large clinical data sets or simulated with templates, this part of the test should not be underestimated. This is especially appealing if a whole spectrum of clinical tasks has to be tested.

Therefore, the image quality evaluation method for image-processing assessment that should be used is clearly dependent on the purpose and the resources that can be used for accomplishing it. The use of the VGA methodology might be more appropriate for the fine tuning of processing parameters of an algorithm which has already been validated quantitatively for specific tasks. Ideally, regulations like the pre-market notification or pre-market approval procedures of the US Food and Drug Administration (unfortunately, the European Guidelines [7] neglect any processing algorithms in the acceptance protocol for digital mammography systems) should aim for such an approach. Virtual clinical trials using, in part, simulated data may be a realistic scenario to avoid excessive costs. Hence, ROC studies should remain the gold standard against which other assessment methods, like quantitative subjective studies, should be validated.

The VGA approach therefore seems more useful at a local level, especially in a breast-screening programme, and more specific in decentralised centres, where radiologists are confronted with images that look very different (as algorithms may vary substantially among digital mammography vendors); VGA may help to quantify their occasional concerns about specific image-processing algorithms. In Belgium, which has a decentralised screening programme, the stability of all types of processing algorithms and the global appearance of the images are tested using image quality criteria [37] of 25 examinations (100 images, 4 views per examination) per system, before an image-processing algorithm can be used in the screening setting. This type of test procedure has to be repeated in cases of substantial updates to the processing algorithms or their parameter settings. We have also observed that parameters like noise, saturation and contrast can be different even within the same commercial processing algorithm, as a manufacturer might tune the algorithm parameters differently in

different hospitals in order to accommodate the radiologist's requirements.

Our study has a few limitations:

1. The FROC study was conducted for the specific task of microcalcification detection: it is possible that for a different clinical task, like mass detection, a better agreement between VGA and FROC data would be obtained.
2. In our study, not all readers were familiar with IMS and Sectra processing algorithms. The prejudice against the unfamiliar look of a particular processing algorithm could have been a source of bias in the VGA evaluation; however, the radiologists were blinded to the processing algorithm.
3. All images were acquired on a Siemens device and general conclusions which advocate one processing method as being the best in an absolute sense were therefore outside the scope of this paper. Instead, we aimed at comparing image-processing assessment methods.

Acknowledgments

The authors are grateful to Dr Orjan Smedby for his suggestions regarding the statistical analysis and Remy Klausz for his useful comments on the manuscript.

References

1. Cole EB, Pisano ED, Kistner EO, Muller KE, Brown ME, Feig SA, et al. Diagnostic accuracy of digital mammography in patients with dense breasts who underwent problem-solving mammography: effects of image processing and lesion type. *Radiology* 2003;226:153–60.
2. Cole EB, Pisano ED, Zeng D, Muller K, Aylward SR, Park S, et al. The effects of gray scale image processing on digital mammography interpretation performance. *Acad Radiol* 2005;12:585–95.
3. Pisano ED, Chandramouli J, Hemminger BM, Glueck D, Johnston RE, Muller K, et al. The effect of intensity windowing on the detection of simulated masses embedded in dense portions of digitized mammograms in a laboratory setting. *J Digit Imaging* 1997;10:174–82.
4. Pisano ED, Cole EB, Hemminger BM, Yaffe MJ, Aylward SR, Maidment AD, et al. Image processing algorithms for digital mammography: a pictorial essay. *Radiographics* 2000;20:1479–91.
5. Pisano ED, Gatsonis CA, Yaffe MJ, Hendrick RE, Tosteson AN, Fryback DG, et al. American College of Radiology Imaging Network digital mammographic imaging screening trial: objectives and methodology. *Radiology* 2005;236:404–12.
6. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005;353:1773–83.
7. van Engen R, Young K, Bosmans H, Thijssen M. The European protocol for the quality control of the physical and technical aspects of mammography screening. Part B: digital mammography. In: The fourth edition of the European guidelines for breast cancer screening. Luxembourg: Office for Official Publications of the European Communities; 2006.
8. Zhang Y, Chan HP, Sahiner B, Wei J, Goodsitt MM, Hadjiiski LM, et al. A comparative study of limited-angle cone-beam reconstruction methods for breast tomosynthesis. *Med Phys* 2006;33:3781–95.
9. Wu T, Moore RH, Rafferty EA, Kopans DB. A comparison of reconstruction algorithms for breast tomosynthesis. *Med Phys* 2004;31:2636–47.
10. Rakowski JT, Dennis MJ. A comparison of reconstruction algorithms for C-arm mammography tomosynthesis. *Med Phys* 2006;33:3018–32.
11. Mertelmeier T, Ludwig J, Zhao B, Zhao B. Optimization of tomosynthesis acquisition parameters: Angular range and number of projections. *Digital Mammography Lecture Notes in Computer Science* 2008;5116:220–7.
12. Das M, Gifford HC, O'Connor JM, Glick SJ. Penalized maximum likelihood reconstruction for improved microcalcification detection in breast tomosynthesis. *IEEE Trans Med Imaging* 2011;30:904–14.
13. Bliznakova K, Kolitsi Z, Speller RD, Horrocks JA, Tromba G, Pallikarakis N. Evaluation of digital breast tomosynthesis reconstruction algorithms using synchrotron radiation in standard geometry. *Med Phys* 2010;37:1893–903.
14. Sund P, Bath M, Kheddache S, Mansson LG. Comparison of visual grading analysis and determination of detective quantum efficiency for evaluating system performance in digital chest radiography. *Eur Radiol* 2004;14:48–58.
15. Sandborg M, Tingberg A, Ullman G, Dance DR, Alm CG. Comparison of clinical and physical measures of image quality in chest and pelvis computed radiography at different tube voltages. *Med Phys* 2006;33:4169–75.
16. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol* 2006;3:413–22.
17. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004;31:2313–30.
18. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad Radiol* 2000;7:516–25.
19. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. *Radiology* 1975;116:533–8.
20. European Commission. European guidelines on quality criteria for diagnostic radiographic images. EUR 16260. Luxembourg: Office for Official Publications of the European Communities; 1996.
21. European Commission. European guidelines on quality criteria for computed tomography. EUR 16262. Luxembourg: Office for Official Publications of the European Communities; 1996.
22. European Commission. European guidelines on quality criteria for diagnostic radiographic images in paediatrics. EUR 16261. Luxembourg: Office for Official Publications of the European Communities; 1996.
23. Carlander A, Hansson J, Soderberg J, Steneryd K, Bath M. Clinical evaluation of a dual-side readout technique computed radiography system in chest radiography of premature neonates. *Acta Radiol* 2008;49:468–74.
24. Geijer H, Geijer M, Forsberg L, Kheddache S, Sund P. Comparison of color LCD and medical-grade monochrome LCD displays in diagnostic radiology. *J Digit Imaging* 2007;20:114–21.
25. Geijer H, Norrman E, Persliden J. Optimizing the tube potential for lumbar spine radiography with a flat-panel digital detector. *Br J Radiol* 2009;82:62–8.
26. Park J, Larson AC, Zhang Q, Simonetti O, Li D. High-resolution steady-state free precession coronary magnetic resonance angiography within a breath-hold: parallel imaging with extended cardiac data acquisition. *Magn Reson Med* 2005;54:1100–6.
27. Wiltz HJ, Petersen U, Axelsson B. Reduction of absorbed dose in storage phosphor urography by significant lowering of tube voltage and adjustment of image display parameters. *Acta Radiol* 2005;46:391–5.
28. Pilgram TK, Vannier MW, Hildebolt CF, Marsh JL, McAlister WH, Shackelford GD, et al. Craniosynostosis:

- image quality, confidence, and correctness in diagnosis. *Radiology* 1989;173:675–9.
29. Tingberg A, Herrmann C, Lanhede B, Almen A, Sandborg M, McVey G, et al. Influence of the characteristic curve on the clinical image quality of lumbar spine and chest radiographs. *Br J Radiol* 2004;77:204–15.
 30. Tingberg A, Bath M, Hakansson M, Medin J, Besjakov J, Sandborg M, et al. Evaluation of image quality of lumbar spine images: a comparison between FFE and VGA. *Radiat Prot Dosimetry* 2005;114:53–61.
 31. American College of Radiology. Breast Imaging Reporting and Data System Atlas (BI-RADS® Atlas). Reston, VA: American College of Radiology; 2003.
 32. Zanca F, Chakraborty DP, Van OC, Jacobs J, Claus F, Marchal G, et al. An improved method for simulating microcalcifications in digital mammograms. *Med Phys* 2008;35:4012–18.
 33. LeGal M, Chavanne G, Pellier D. Diagnostic value of clustered microcalcifications discovered by mammography (apropos of 227 cases with histological verification and without a palpable breast tumor). [In French.] *Bull Cancer* 1984;71:57–64.
 34. Zanca F, Jacobs J, Van Ongeval C, Claus F, Celis V, Geniets C, et al. Evaluation of clinical image processing algorithms used in digital in digital mammography. *Med Phys* 2009;36:765–75.
 35. Jacobs J, Zanca F, Marchal G, Bosmans H. Implementation of a novel software framework for increased efficiency in observer performance studies in digital radiology. Proceedings of the Scientific Assembly and Annual Meeting of the RSNA, Chicago, IL, 2008.
 36. Digital Imaging and Communications in Medicine (DICOM) part 14: Gray scale standard display function. NEMA PS 3, National Electrical Manufacturers Association: Rosslyn, VA; 2000.
 37. Van Ongeval C, Van Steen A, Geniets C, Dekeyzer F, Bosmans H, Marchal G. Clinical image quality criteria for full field digital mammography: a first practical application. *Radiat Prot Dosimetry* 2008;129:265–70.
 38. Smedby O, Fredrikson M. Visual grading regression: analysing data from visual grading experiments with regression models. *Br J Radiol* 2010;83:767–75.
 39. Krupinski EA. Practical applications of perceptual research. In: Beutel J, Kundel H, Van Matter R, eds. Handbook of medical imaging, Vol 1: Medical physics and psychophysics. Bellingham, WA: SPIE; 2000.