Received:
9 July 2014

Revised:
26 November 2014

Accepted:
10 December 2014

doi: 10.1259/bjr.20140482

Cite this article as:

Wolstenhulme S, Davies AG, Keeble C, Moore S, Evans JA. Agreement between objective and subjective assessment of image quality in ultrasound abdominal aortic aneurysm screening. *Br J Radiol* 2015;88:20140482.

FULL PAPER

Agreement between objective and subjective assessment of image quality in ultrasound abdominal aortic aneurysm screening

¹S WOLSTENHULME, DCR, MHSc, ²A G DAVIES, BSc, MSc, ³C KEEBLE, BSc, MSc, ⁴S MOORE, HND, MSc and
²J A EVANS, PhD, FIPEM

¹School of Healthcare, University of Leeds, Leeds, UK

²Division of Medical Physics, University of Leeds, Leeds, UK

³Division of Epidemiology and Biostatistics, University of Leeds, Leeds, UK

⁴Department of Medical Physics, Leeds Teaching Hospitals, Leeds, UK

Address correspondence to: Mr Andrew Graham Davies

E-mail: a.g.davies@leeds.ac.uk

Objective: To investigate agreement between objective and subjective assessment of image quality of ultrasound scanners used for abdominal aortic aneurysm (AAA) screening.

Methods: Nine ultrasound scanners were used to acquire longitudinal and transverse images of the abdominal aorta. 100 images were acquired per scanner from which 5 longitudinal and 5 transverse images were randomly selected. 33 practitioners scored 90 images blinded to the scanner type and subject characteristics and were required to state whether or not the images were of adequate diagnostic quality. Odds ratios were used to rank the subjective image quality of the scanners. For objective testing, three standard test objects were used to assess penetration and resolution and used to rank the scanners.

Results: The subjective diagnostic image quality was ten times greater for the highest ranked scanner than for the lowest ranked scanner. It was greater at depths of <5.0 cm (odds ratio, 6.69; 95% confidence interval, 3.56, 12.57) than at depths of 15.1–20.0 cm. There was a larger range of odds ratios for transverse images than for longitudinal images. No relationship was seen between subjective scanner rankings and test object scores.

Conclusion: Large variation was seen in the image quality when evaluated both subjectively and objectively. Objective scores did not predict subjective scanner rankings. Further work is needed to investigate the utility of both subjective and objective image quality measurements.

Advances in knowledge: Ratings of clinical image quality and image quality measured using test objects did not agree, even in the limited scenario of AAA screening.

The quality of images produced by a medical imaging device is an important consideration when gauging its suitability for a specific clinical task—it is essential that the system produces images that are of sufficient fidelity for the clinical user. As such, image quality will form an important consideration in the selection of equipment and in the ongoing quality assurance procedures following installation.

The assessment of medical image quality can be performed in a number of ways, both subjectively (for example, using visual grading^{1,2}) and objectively using test phantoms specifically designed for that purpose.^{3,4} Even for a specific imaging modality such as ultrasound, the level of agreement between these methods has not been thoroughly investigated, although there is some evidence of poor agreement between ratings of quality scores from test

objects with those of clinical users when asked to rate clinical images from the same scanner.⁵

The need to provide more objective image quality assessment is highlighted when there are national programmes requiring common standards. The breast cancer, foetal abnormalities and abdominal aortic aneurysm (AAA) detection programmes are good examples requiring ultrasound imaging of a uniform quality. It is critical that there is good agreement between clinical users as to what constitutes an acceptable image for these purposes. This will form the basis of a gold standard of performance against which the utility of any objective testing can be evaluated.

In this study, we have used the ultrasound-based aortic aneurysm screening programme as an exemplar. In the UK,

the National Abdominal Aortic Aneurysm Screening Programme (NAAASP) was implemented in 2013.⁶ This programme is primarily community based, necessitating the use of portable ultrasound scanners to allow transportation to screening centres. Measurements of the anteroposterior (A-P) inner to inner (ITI) abdominal aortic diameter in longitudinal section (LS) and transverse section (TS) planes are taken.

The quality of images depends upon the skill of the practitioner, the habitus of the patient and the performance of the scanner. Together they may influence the reliability and accuracy of measurements.^{7,8} Small errors in measurements may impact on clinical decision making, for example, resulting in inappropriate enrolment into the surveillance programme, at the 30-mm threshold, or delayed referral for a vascular surgical opinion, at the 55-mm threshold.

Selection of the ultrasound scanner to carry out national screening is the responsibility of the service provider, although in the UK, some guidance on specification is available from the National Screening Committee. It is less clear what method providers should use to make their choice of scanner and whether this choice has any impact on the diagnostic image adequacy and the service provided. When faced with similar procurement decisions, providers have invited competing manufacturers to supply equipment for evaluation over a short time. The service providers commonly use subjective assessment of the image quality to make a decision, while recognizing on a small sample, differences between subjects, *e.g.* body habitus, may affect differences between scanners.^{5,9} An alternative approach is to use one or more test objects to objectively assess image adequacy thus removing intersubject variation. Such objective measures also have the potential advantages that they are quick to perform, can be reproduced exactly at different centres and are ought to be less affected by the subjective opinion of the operator. A variety of test objects have been described for evaluation of ultrasound image quality, and each of these can be used to measure a range of different parameters.⁴ However, there is a paucity of evidence as to how results from such tests relate to subjective assessment. We are not aware of any specific advice or publication aimed at evaluating portable AAA scanners.

The aim of this study was to investigate the level of agreement between the subjective assessment of the aortic images from portable ultrasound scanners and objective assessments obtained using test objects. If the agreement is good, then the implication is that test objects could be used with confidence in the assessment of image quality both for purposes of scanner selection and in monitoring ongoing performance. If the agreement is poor, then either the use of test objects as objective evaluators of performance should be seriously questioned or the assumption that clinical subjective performance is useful is called into question.

METHODS AND MATERIALS

This was a prospective study in which selected ultrasound scanners were used by the same operator in a routine screening environment with later viewing by blinded observers.

Equipment

The following ultrasound scanners, nominated by their manufacturer as being suitable for aortic aneurysm screening, were made available for evaluation:

- CX50 (Philips Healthcare, Bothell, WA)
- LOGIQ® book XP and LOGIQ e (GE Healthcare, Chalfont St Giles, UK)
- Micromax, M-Turbo® and Nanomax (SonoSite Inc., Bothell, WA)
- SIUI CTS-900 (MIS Healthcare, London, UK)
- Viamo (Toshiba Medical Systems, Tochigi, Japan)
- z-One (Zonare Medical Systems Inc., Mountain View, CA).

These scanners are referred to in no particular order as being scanners A–J. The rotation of the scanners through one local screening programme of the NAAASP was arranged by the Purchase and Supply Agency in negotiation with the manufacturers. Each scanner was evaluated for 1 week within the local screening programme and was taken to at least two general practitioner practices. The transducers used were curvilinear arrays recommended by scanner manufacturer for this application. For each scanner, the same transducer was used for both clinical image acquisition and objective testing.

Subjective evaluation of image quality

Acquisition of images

On the first day of each week, one screening technician and the scanner manufacturer's clinical application specialist worked together to achieve familiarization with the portable ultrasound scanner. The screening technician, with 5 years' post-certification experience of carrying out abdominal aorta ultrasound examinations, acquired all images for aortic diameter assessment. For each examination, the screening technician varied the operator's scanning position (sitting/standing) and the degree of tilt of the monitor. This variation depended on the height of both the examination couch and the scanner's monitor. The room lighting was dimmed when carrying out the examination. Scanner controls such as gain, compound and tissue harmonic imaging and depth of field were changed, as required, to obtain the perceived optimal ultrasound image. Each patient was examined using only one scanner. For each patient, four images of the abdominal aorta were acquired, one LS image and one TS image with measurements of the ITI diameter for NAAASP, and one LS image and one TS image without callipers. These images were stored in digital imaging and communications in medicine (DICOM) format on the scanner's hard drive and transferred to a secure hospital information technology server.

The subject's informed consent to have an ultrasound examination was obtained as per NAAASP Standard Operating Procedures.⁶ Ethical approval was not required, as the images were routinely acquired and anonymized and the practitioners, who rated the images in the study, were National Health Service employees.

The DICOM images without callipers were exported, without any image adjustment or enhancement, to a computer. They were then cropped to remove subject name, hospital and ultrasound scanner manufacturer identity, but retained the vertical

measurement scale data. A unique identification number was added to each image. The anonymized images allowed blinded scanner ranking. At the end of the clinical data collection phase, 900 anonymized images were stored in a database.

Image selection and scoring

Five LS and five TS images were randomly selected from each scanner, subject to the constraint that one of each LS and TS image set contained an image of an aorta with an A-P diameter subjectively >40 mm. This was to ensure that each set contained one aneurysmal aorta. 90 images (45 LS and 45 TS) were used for analysis. 90 images permitted each observer to complete the study in a realistic time scale. The reason for choosing the same 90 images rather than providing a random set of 90 images from the 900 total images was to enable analysis of the same images to determine the variation in the scores. The ultrasound scanner's control settings likely to affect image quality (depth of field, compound imaging and tissue harmonic imaging) that were used for the 90 images were recorded. Readers unfamiliar with these ultrasound control settings are referred elsewhere.¹⁰

33 practitioners completed a demographics questionnaire and undertook scoring of images using a web-based tool. The practitioners were from radiology or vascular departments in the UK and the six NAAASP early implementer sites. Each practitioner was given a unique identifier. The demographics requested were the practitioners' profession and the level of experience (number of years they have been in their profession). The practitioners included a variety of professions: medical physicists (1), screening technicians (1), radiologists (1), ultrasound practitioners (12), vascular surgeons (3) and vascular technologists (15). Their mean (range) level of experience was 11.2 years (1–30 years).

All 33 observers were blinded to the scanner type and subject characteristics. To achieve this, the alphanumeric text and logos were removed from the images prior to viewing. Since the operator acquiring the images was not involved in the image viewing, all of the observers were blinded to any patient data.

The web-based tool allowed the observers to view the 90 images in 1 session or to pause the session and complete it in stages and at their own pace. This was performed on their own personal computer accessing the custom written web-based survey software. The observers were advised to score in dimmed lighting. At the beginning of each scoring session, the observers were presented with a challenge response test to confirm the monitor and viewing conditions offered sufficient viewing quality to make meaningful judgments for the study. The test involved reading low contrast letters against differing background intensities.¹¹

The observers viewed one image at a time and were required to answer "yes/no" to the question: "Is this image of adequate diagnostic quality?" Each observer viewed the images in a random and different order. Images were resized for display purposes using a bilinear interpolation, such that all images were

displayed at the same size. No images were minified (*i.e.* had their resolution reduced).

Objective evaluation of scanner performance

In the absence of clear guidelines for the objective evaluation of this type of scanner, a judgment was needed to decide which parameter(s) to evaluate. Given that the aorta is a relatively large organ, it was deemed to be unlikely that imaging it normally would be a challenge for any modern scanner. Consequently, traditional spatial resolution assessment was not carried out. However, the ability of the system to image the aorta at depth in large patients was regarded as critical and therefore penetration-type measurements using tissue-equivalent test objects were adopted. Three such test objects were selected and used on all scanners. The scanners were delivered in turn to the Medical Physics Department of the Leeds Teaching Hospitals Trust, Leeds, UK, and all measurements were undertaken by the same operator, experienced in ultrasound quality assurance (QA). The screen used in each case was that supplied with the scanner. It was not possible to blind the operator to the scanner's identity, but this was regarded as unimportant owing to the objective nature of the test. In each case, the preset recommended for AAA scanning by the manufacturer was selected with tissue harmonic imaging turned off. The gain was set to maximum, unless that led to saturation, and the time gain compensation was adjusted to give a speckle display at the greatest possible depth.

The Cardiff resolution test object (RTO) is a rather old device that has been used extensively by many workers. Its primary purpose is to assess spatial resolution, but in our case, we used only sections that were free of resolution targets. The penetration value that was recorded was defined as the depth at which the speckle was judged to change into noise or base dark level.

The Edinburgh pipe test object (EPipe) was kindly supplied by the Department of Medical Physics, Edinburgh Royal Infirmary, Edinburgh, UK. It has a tissue mimicking background but contains a number of small diameter pipes that are scanned along their lengths. In this case, however, the pipes were ignored and only the penetration in pipe-free regions was considered. Two different measurements were made with this object. The penetration [EPipe(pen)] was recorded using a region of the test object that was free of pipes. The second measurement was the maximum depth at which the 6-mm pipe could be seen [EPipe(vis)]. The rationale for this is that the quality of the image is likely to relate to the ability to image a small object at depth.

The Gammex® 408LE spherical lesion phantom was used (Gammex-RMI, Nottingham, UK). This device has a number of simulated spherical lesions at a range of depths. It was thought that the ability of the scanner to detect these lesions at depth would be similar to that found with the EPipe(vis) test. The protocol used was the same as for the penetration measurement. This time, the maximum depth at which spherical lesions could be clearly seen was recorded.

The attenuation in the test objects was 0.86, 0.50 and 0.70 dB cm⁻¹ MHz⁻¹ for the RTO, EPipe and Gammex test objects, respectively.

Scanners were then ranked by object visibility (in millimetres), and the rankings were compared with the subjective scanner rankings using Spearman's rank correlation coefficient.

Statistical analysis

Summary statistics and logistic regression were used to generate odds ratios, with 95% confidence intervals (CIs), to rank the scanners in order of their odds of producing an image with diagnostic image adequacy compared with the lowest ranked scanner, that is, how many more times likely an adequate diagnostic image would be from a given scanner compared with the least successful scanner. Three logistic regression models were used: one with LS images, one with TS images and one with all images. Analysis was carried out using Microsoft Excel® (Microsoft, Redmond, WA) and the statistical software R.¹² The independent variables included in the logistic regression were the nine scanner types; the 33 practitioners; the depth categorized into four ranges (<5.0, 5.1–10.0, 10.1–15.0, 15.1–20.0 cm); compound imaging (on/off); and tissue harmonic imaging (on/off).

RESULTS

Scanner control settings

The scanner settings used, and the depths at which the aortas were located are summarized in Table 1. The median depth of field was 10 cm (range, <5–20 cm), with the majority of images being obtained with the aorta at a depth in the 10 to 15-cm range. Eight of the nine scanners had compound imaging available, and it was used at least once in seven (77.8%). The use of compound imaging in these seven scanners ranged from 20% (scanner D) to 100% (scanner A and B). For five scanners (55.6%), tissue harmonic imaging was selected at least once, with usage ranging from 20% (scanner D) to 100% (scanners C, E and H).

Subjective assessment

Overall, 70.9% of images were ranked as adequate. The ordering of scanner types, overall and for LS and TS separately, when ranked using the odds of producing an image of diagnostic

image adequacy compared with the least successful scanner, is shown in Table 2 and Figure 1. The combined LS and TS scores show that the highest ranked scanner (A) was 10.71 (95% CI, 6.48, 17.69) times more likely to have diagnostic image adequacy than did the least successful scanner (J). Less variation was shown when rating LS images (greatest odds ratio, 5.14) as adequate compared with the TS images (greatest odds ratio, 34.28). Two images from the study, both from the TS set, are shown in Figure 2. Neither image contains an aneurismal aorta.

The images where compound imaging was used had, statistically significant, lower odds of producing an adequate image (0.38; 95% CI, 0.27, 0.53), whereas those using tissue harmonic imaging had higher odds of producing an adequate image (1.77; 95% CI, 1.00, 3.11). These odds ratios were calculated allowing for the scanner type, observer and depth. The relationship between the depth of field and the odds (with 95% CI) of scoring an abdominal aorta ultrasound image as adequate for the nine portable ultrasound scanners, rated by all observers, is shown in Table 3. For all 90 images, as the depth of field increased, the odds of producing an image of diagnostic image adequacy decreased.

Objective assessment

A summary of the test object measurements is shown in Table 4 and summarized in Figure 3. This shows variation in the measurements when using different test objects. Little agreement was seen between the order of the overall subjective ranking of the scanners and the objective test object rankings (Table 5). Spearman's rank correlation coefficient, ρ , was 0.00, 0.27, 0.10 and -0.27 between the combined subjective rank and the RTO, EPipe(pen), EPipe(vis) and Gammex test objects, respectively, indicating no strong correlations. No significant or strong correlations were found when the LS and TS subjective ranks were similarly compared.

DISCUSSION

Our findings show the observers regarded 70.9% of the images to be of diagnostic image adequacy, which is in disagreement with

Table 1. Variation in the depth, compound imaging (CoI) and tissue harmonic imaging (THI) control settings used by one screening technician when the nine portables ultrasound scanners were used to examine the longitudinal and transverse sections of the abdominal aorta

Scanner	Depth (cm)					CoI on	THI on
	Median (minimum, maximum)	(≤5)	(>5 to ≤10)	(>10 to ≤15)	(>15 to ≤20)		
A	11 (6.6, 13)	0	2	8	0	10	0
B	7.8 (1, 13)	1	5	4	0	10	0
C	9 (5, 18)	1	7	1	1	9	10
D	10 (6, 14)	0	6	4	0	2	2
E	9 (7, 15)	0	7	3	0	7	10
F	12 (8, 19)	0	3	4	3	3	0
G	14 (8, 17)	0	2	6	2	0	9
H	13 (7, 20)	0	3	3	4	8	10
J	12 (5, 15)	1	3	6	0	0	0

Table 2. Odds ratios (and 95% confidence intervals) of diagnostic image adequacy ratings

Scanner	Overall	Longitudinal section	Transverse section
A	10.71 (6.48, 17.69)	5.14 (2.55, 10.36)	34.28 (14.81, 79.34)
B	10.19 (6.14, 16.90)	3.89 (1.93, 7.86)	26.16 (11.53, 59.35)
C	4.30 (2.24, 8.26)	2.02 (0.81, 5.01)	3.01 (0.72, 12.54)
D	3.79 (2.56, 5.60)	1.43 (0.79, 2.61)	6.01 (3.48, 10.36)
E	2.88 (1.53, 5.43)	1.42 (0.60, 3.37)	1.91 (0.46, 7.98)
F	2.55 (1.77, 3.68)	0.64 (0.37, 1.09)	9.73 (5.31, 17.82)
G	2.04 (1.09, 3.83)	0.89 (0.40, 2.02)	1.51 (0.32, 7.12)
H	1.87 (0.98, 3.57)	1.32 (0.48, 3.68)	0.48 (12.09, 1.98)
J	1.00	1.00	1.00

the screening technician, who regarded all abdominal aorta images as optimal for screening purposes, when acquired in real time. The screening technician would have considered the subject characteristics and the degree of difficulty in identifying the anatomical relationships and landmarks to measure the A-P abdominal aorta ITI diameter. We can only speculate the reasons for the observers rating the 90 images differently to the screening technician. The diagnostic image adequacy may have been affected by the observers viewing the images on different computers at different light levels, although the bias associated with this was reduced by undertaking the challenge response test.¹¹ The observers may also have been assessing different aspects of the image when scoring the images. The data collection was performed at the time when observers may have been using either NAAASP standard operating procedures to determine diagnostic image adequacy for control settings, anatomical relationships and landmarks to measure the A-P ITI diameter⁶ or local guidelines to determine the position of the callipers on the aortic wall.^{7,8} It has also been demonstrated that guidelines alone are not sufficient for agreement on what comprises an acceptable image.¹³

The strengths of this study included the acquisition of the clinical images by one experienced screening technician, and all objective testing by a single experienced technologist, from all nine scanners, reducing variation in image acquisition. By using

a web-based system, we were able to include assessment from a wide range of expert practitioners.

The effect of depth on image adequacy for all 90 images (Table 3) was likely to be owing to increased ultrasound beam divergence and attenuation. This leads to decreased spatial and contrast resolution, which could impact on the identification of anatomical structures and landmarks to measure the A-P abdominal aorta ITI diameter. The analysis of the subjective preference scores controlled for the influence of depth on a scanner's subjective image quality performance. When comparing the subjective quality of ultrasound scanners for AAA applications, it is important that a suitable range of depths are included in the images sets.

The use of compound imaging decreased the odds of diagnostic image adequacy, which is contrary to the predicted use of this control in practice.^{4,14} This may be owing to the blurring of lateral borders. The use of tissue harmonic imaging, which might have been predicted to improve contrast resolution,^{4,15} resulted in an increase in diagnostic image adequacy, but this was not statistically significant.

The data show that the variation in odds ratio from the lowest scoring scanner (J) to the highest scoring scanner (A) was wider for TS than LS images. The scanner rankings show the 95% confidence intervals for the LS and TS sections to be wider than the overall rankings as their analysis uses smaller data sets. This suggests observers may be more confident when determining adequate image quality with LS images. This may be owing to observers only assessing if they could identify the aorta and determine the landmarks to measure the A-P diameter. When scoring the TS images, the observer was assessing the anatomical relationships of the aorta, such as the inferior vena cava, lumbar spine and bowel and the landmarks, to measure its A-P diameter. This may explain the differences in repeatability and reproducibility between LS and TS abdominal aorta A-P diameter measurements.^{7,8,16–19}

Objective tests

All of the objective measurements were performed by a single person. This would have ensured consistency, although the testing was not blinded, which may have introduced bias. The operator

Figure 1. Subjective image quality scores—odds ratio of each scanner producing an image of acceptable quality.

Figure 2. Two clinical transverse images from the subjective image comparison, showing (a) a highly rated image and (b) a poorly rated image.

was highly experienced in ultrasound QA and familiar with a range of equipment of differing manufacturers. Analysis of previous repeated measurements of penetration indicate that a variation of the larger of 2 mm or 5% would be expected for such measurements.

Given that the speed of sound and attenuation is claimed to be the same in all three test objects, it would be predicted that there would be a high level of agreement in the ranking of penetration values obtained in all three tests. This was clearly not the case. One possible explanation is that the relative contribution to the attenuation from scatter may differ between the three test objects. Data on these values were not available. This is important because it is the scattering component that is being measured in the image as a surrogate for real penetration. Furthermore, it is not known whether the scatter from normal tissue is similar to any of the three test objects, although all three test objects claim to mimic liver parenchyma.

An additional problem is that the rank order of the scanners was different for the three test objects. This suggests that some factor other than scatter is involved since scattering differences alone would have been expected to change the magnitude of penetration values but not the rank order of the scanners. It can be speculated that the discrepancy lies in the different greyscale transfer curves and other image processing algorithms used by different manufacturers.

Comparison of objective and subjective image quality measures

Our selection of objective test to perform was based on the conjecture that penetration would be an important factor in predicting the quality of the clinical image given the nature of the examination, and spatial resolution would be less critical, since the abdominal aorta is relatively large. Conversely, the clarity with

which the landmarks of the abdominal aorta A-P diameter are displayed is presumably important, and this should be related to the greyscale transfer curve and/or dynamic range of the scanner. Our choice was to measure penetration and detection of spherical cystic targets with test objects. The variation between our subjective and objective image quality rankings may be owing to the subjective assessment of abdominal ultrasound images with varying depths compared with the detection and resolution of the small (diameter, 4 mm) spherical cystic targets in the test objects at pre-defined depths. It is unknown whether either the observers' subjective or the test object objective rankings have any bearing on the precision and/or reproducibility of the abdominal aortic diameter measurement or, more importantly, patient outcome.

Limitations

At the patient image acquisition stage, the screening technician had greater familiarity with scanner E than with the other eight scanners. This may have affected the screening technician's confidence in manipulating the scanner control settings to obtain an optimal image. For this reason, and as we neither wished to identify the best and worst scanners nor infringe national procurement confidentiality, we anonymized the scanners in the findings. As the observers were blinded to the scanner on which the image was acquired, we believe the bias was reduced. The general practitioner (GP) examination rooms had different background light levels. In a room with excessive background lighting, the illumination of the screen would be increased.²⁰ To compensate for this, the screening technician may have increased, in dynamic scanning, the gain to visualize the abdominal anatomy. This could potentially impact on the diagnostic image adequacy of the static image.²¹ The portable ultrasound scanners were used without a dedicated scanner stand. This meant, between the GP practices, the scanners were placed on dressing trolleys of different heights, leading to discrepancies of the height of the scanner's monitor. The screening technician needed to change scanning positions, the degree of tilt of the monitor and the resulting viewing angle to allow better visualization of the abdominal anatomy. This may have impacted on the diagnostic image adequacy by causing image distortion or anisotropy, leading to change in contrast resolution.¹⁷ The scanners were not used randomly in the different rooms, and this may increase the risk of bias in the scanner rankings.

The scanner presets were used as a starting point for both objective and subjective tests, and the operators were free to alter the settings as they felt appropriate. This may have resulted in

Table 3. Odds ratios (95% confidence intervals) describing the relationship between depth of field and diagnostic image adequacy

Depth (cm)	Odds ratio
<5.0	6.69 (3.56, 12.57)
5.1–10.0	4.24 (3.03, 5.93)
10.1–15.0	3.13 (2.26, 4.34)
15.1–20.0	1.00 (NA)

Table 4. Summary of the objective measurements of the nine portable ultrasound scanners

Scanner	Resolution test object (mm)	EPIPE(pen) (mm)	EPIPE(vis) (mm)	Gammex® (mm)
A	130	190	140	52
B	145	180	110	45
C	115	155	117	42
D	140	200	145	36
E	135	180	133	52
F	155	200	146	50
G	135	170	129	76
H	125	180	120	40
J	135	158	115	61

EPIPE(pen), Edinburgh pipe test object (penetration); EPIPE(vis), Edinburgh pipe test object (visibility).
Gammex; Gammex-RMI, Nottingham, UK.

different settings being used on the two image quality measures. However, this is likely to be the situation when such tests are carried out in hospital environments.

It is possible that the images from patients used in this study may not be representative of the nine portable ultrasound scanners, on which they were acquired, but the random selection of images should have reduced the bias, as the ten images per scanner are likely to represent a variety of patients. Since sample size calculations are not suitable for categorical predictors and binary outcomes (the scanner type being categorical and the outcome being yes/no), the required sample size cannot be calculated to achieve a target power. We enrolled 33 observers each analysing the same 90 images producing 2970 responses in total. We believe this number of observers and the large image data set is sufficient to draw conclusions.

Figure 3. Objective image quality scores: test object measurements for each scanner. Gammex®; Gammex-RMI, Nottingham, UK. EPIPE(pen), Edinburgh pipe test object (penetration); EPIPE(vis), Edinburgh pipe test object (visibility); RTO, resolution test object.

There were a number of professions and range of experience in the observers in the subjective study. It is entirely possible that both of these factors could affect the responses of the observer. More specifically, staff dealing with AAA screening may score differently to other users. Broadly stated profession and experience, as recorded in this study, are not likely to be good predictors of how often an individual routinely uses AAA ultrasound images. Future work to investigate the effect of profession and experience in AAA ultrasound screening on quality score responses would be useful in establishing how prescriptive future study design should be in terms of the background of participating observers.

Implications for image quality assessment of ultrasound scanners

Even in a screening setting, where there is a well-defined clinical task, with specific criteria for the features that are required for an adequate image, we failed to select an objective test that could predict the subjective assessment of image quality by users. This does not mean that such objective testing is not useful, however, as such tests are likely to be sensitive to changes in scanner performance over time, and therefore should play a role in quality assurance programmes. Objective tests are also minimally affected by differences in the subject variability.

Care must also be taken in drawing the conclusion that any objective test is not useful if it does not predict subjective image quality. It is not established that observer rating of image quality is able to predict diagnostic accuracy. In particular, in the context of AAA screening, it is the diameter measurement that is the purpose of the imaging and not for instance the detection of a lesion. For other clinical applications, better agreement between objective and subjective tests may be found.

It is not clear what criteria, if any, should be used when assessing the image quality performance of scanners in this AAA screening context. To answer such a question, it would be necessary to study the effect of scanner selection on patient outcomes, and such a study would be long and expensive. It is likely that by the time that the results of such a study were available, the scanners in the study would no longer be available on the market. In lieu of

Table 5. The ranking of the nine ultrasound scanner scores for the subjective scores, compared with the objective test object scores

Scanner	Subjective	Objective			
	Rated by practitioners	Resolution test object	EPipe(pen)	EPipe(vis)	Gammex®
A	1	7	3	3	3
B	2	2	4	9	6
C	3	9	9	7	7
D	4	3	1	2	9
E	5	4	4	4	3
F	6	1	1	1	5
G	7	4	7	5	1
H	8	8	4	6	8
J	9	4	8	8	2
Spearman rank order correlation with subjective rank		0.00	0.27	0.1	−0.27

EPipe(pen), Edinburgh pipe test object (penetration); EPipe(vis), Edinburgh pipe test object (visibility).

This shows none of the test objects scores helps to predict the subjective study results.

Gammex®; Gammex-RMI, Nottingham, UK.

such a study, we would encourage the development of task-specific test phantoms for image quality assessment, especially for common tasks such as those in screening programmes such as NAAASP. It might be possible that given a phantom with anthropomorphic characteristics, where the observer task is aortic diameter measurement can also be combined with a subjective opinion on quality. For subjective ratings on clinical images, image selection should contain a number of challenging cases, with the aorta at greater depths within the patient. Care must be taken with the viewing conditions, although it is unlikely to be practical to allow all of the images to be viewed on the scanner's own monitor by all observers, therefore, the viewing system must be controlled via methods such as the monitor quality check employed in this study. Careful selection of observers, so that the observers are selected from the specific staff group likely to use the equipment would be good practice, although given differences between observers, it may be difficult to recruit sufficient numbers of very tightly selected observers.

CONCLUSION

The study shows large variation in the performance of the nine portable ultrasound scanners evaluated, for use in the primarily community-based NAAASP, when assessed both subjectively and objectively. Test object measures of image quality do not predict

subjective scanner image quality rankings, and it is not clear which of these methods of assessment is better linked to clinical outcomes. Further development of task-specific test objects could be of great benefit in future quality assessments and in the understanding of the relationship between subjective and objective measurements of image quality.

FUNDING

The work was funded by the Department of Health National Service AAA Screening Programme. AGD receives a research grant from Philips Healthcare.

ACKNOWLEDGMENTS

We would like to thank the following: the ultrasound machine manufacturers for allowing their machines to be evaluated at the Leicester National Abdominal Aortic Aneurysm Screening Programme centre; National Health Service Purchasing and Supply Agency for organizing for the ultrasound machines to be taken to Leicester for evaluation; Gillian Hussey for acquiring the abdominal aorta ultrasound images; Kari Dempsey for performing the *in vitro* analysis; the practitioners who rated the ultrasound images; and Professor David Brett and Medipex for the use of the login verification tool on the web-based software.

REFERENCES

1. Båth M, Månsson LG. Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. *Br J Radiol* 2007; **80**: 169–76.
2. Smedby O, Fredrikson M. Visual grading regression: analysing data from visual grading experiments with regression models. *Br J Radiol* 2010; **83**: 767–75. doi: [10.1259/bjr/35254923](https://doi.org/10.1259/bjr/35254923)
3. Launders JH, McArdle S, Workman A, Cowen AR. Update on the recommended viewing protocol for FAXIL threshold contrast detail detectability test objects used in television fluoroscopy. *Br J Radiol* 1995; **68**: 70–7.
4. Browne JE, Watson AJ, Gibson NM, Dudley NJ, Elliott AT. Objective measurements of image quality. *Ultrasound Med Biol* 2004; **30**: 229–37.
5. Metcalfe SC, Evans JA. A study of the relationship between routine ultrasound quality assurance parameters and subjective operator image assessment. *Br J Radiol* 1992; **65**: 570–5.
6. National Screening Programme Standard Operating Procedures and Workbook. [Cited 26 November 2014.] Available from: <http://www.aaa.screening.nhs.uk>

7. Beales L, Wolstenhulme S, Evans JA, West R, Scott DJ. Reproducibility of ultrasound measurement of the abdominal aorta. *Br J Surg* 2011; **98**: 1517–25. doi: [10.1002/bjs.7628](https://doi.org/10.1002/bjs.7628)
8. Long A, Rouet L, Lindholt JS, Allaire E. Measuring the maximum diameter of native abdominal aortic aneurysms: review and critical analysis. *Eur J Vasc Endovasc Surg* 2012; **43**: 515–24. doi: [10.1016/j.ejvs.2012.01.018](https://doi.org/10.1016/j.ejvs.2012.01.018)
9. Tapiovaara MJ. Review of relationships between physical measurements and user evaluation of image quality. *Radiat Prot Dosimetry* 2008; **129**: 244–8. doi: [10.1093/rpd/ncn009](https://doi.org/10.1093/rpd/ncn009)
10. Hoskins PR, Martin K, Thrush A, eds. *Diagnostic ultrasound: physics and equipment*. Cambridge, NY: Cambridge University Press; 2010.
11. Brettle DS, Bacon SE. Short communication: a method for verified access when using soft copy display. *Br J Radiol* 2005; **78**: 749–51.
12. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for statistical computing. 2014. Available from: <http://www.r-project.org>
13. Keeble C, Wolstenhulme S, Davies AG, Evans JA. Is there agreement on what makes a good ultrasound image? *Ultrasound* 2013; **21**: 118–23.
14. Elliott ST. A user guide to compound imaging. *Ultrasound* 2005; **13**: 112–17.
15. Shapiro RS, Wagreich J, Parsons RB, Stancato-Pasik A, Yeh HC, Lao R. Tissue harmonic imaging sonography: evaluation of image quality compared with conventional sonography. *AJR Am J Roentgenol* 1998; **171**: 1203–6.
16. Stather PW, Dattani N, Bown MJ, Earnshaw JJ, Lees TA. International variations in AAA screening. *Eur J Vasc Endovasc Surg* 2013; **45**: 231–4. doi: [10.1016/j.ejvs.2012.12.013](https://doi.org/10.1016/j.ejvs.2012.12.013)
17. Hartshorne TC, McCollum CN, Earnshaw JJ, Morris J, Nasim A. Ultrasound measurement of aortic diameter in a national screening programme. *Eur J Vasc Endovasc Surg* 2011; **42**: 195–9. doi: [10.1016/j.ejvs.2011.02.030](https://doi.org/10.1016/j.ejvs.2011.02.030)
18. Thapar A, Cheal D, Hopkins T, Ward S, Shalhoub J, Yusuf SW. Internal or external wall diameter for abdominal aortic aneurysm screening? *Ann R Coll Surg Engl* 2010; **92**: 503–5. doi: [10.1308/003588410X12699663903430](https://doi.org/10.1308/003588410X12699663903430)
19. Bredahl K, Eldrup N, Meyer C, Eiberg JE, Sillesen H. Reproducibility of ECG-gated ultrasound diameter assessment of small abdominal aortic aneurysms. *Eur J Vasc Endovasc Surg* 2013; **45**: 235–40. doi: [10.1016/j.ejvs.2012.12.010](https://doi.org/10.1016/j.ejvs.2012.12.010)
20. Moore SC, Munnings CR, Brettle DS, Evans JA. Assessment of ultrasound monitor image display performance. *Ultrasound Med Biol* 2011; **37**: 971–9. doi: [10.1016/j.ultrasmedbio.2011.02.018](https://doi.org/10.1016/j.ultrasmedbio.2011.02.018)
21. Oetjen S, Ziefle M. A visual ergonomic evaluation of different screen types and screen technologies with respect to discrimination performance. *Appl Ergon* 2009; **40**: 69–81. doi: [10.1016/j.apergo.2008.01.008](https://doi.org/10.1016/j.apergo.2008.01.008)