Early experience with automated B-mode quality assurance tests

NJ Dudley¹ and NM Gibson²

¹Medical Physics Department, United Lincolnshire Hospitals NHS Trust, Lincoln LN2 5QY

²Medical Physics Department, Nottingham University Hospitals NHS Trust, Nottingham NG5 1PB

Corresponding author: Nick Dudley. Email: nick.dudley@ulh.nhs.uk

Abstract

Although quality assurance programmes have been recommended for many years, there is limited evidence of their efficacy. This study aimed to assess whether an automated image quality analysis method could demonstrate changes in scanner performance in a quality assurance programme. Test object images were analysed, measuring lateral resolution, low contrast penetration, slice thickness, anechoic target visibility and grey-scale target contrast and visibility. Known and suspected scanner faults were investigated and routine results were reviewed. At least one variable changed in response to each known or suspected scanner fault. Resolution and grey-scale target visibility changed due to image shadowing. Slice thickness, lateral resolution and grey-scale target contrast were affected where users reported deterioration in image quality. A single probe fell out of tolerance on routine testing, due to an unrecorded change to the default preset by the supplier's representative. Interpretation of individual results is not always intuitive, observed changes depending on the shape of the grey-scale transfer curve and on target and background echo levels. Our results have provided evidence for the efficacy of this method of performance testing. Further experience is required to evaluate this method for prospective detection of faults and further work is required to determine optimum scanner settings and test object properties to maximise fault detection and to reduce the dependence of results on confounding factors.

Keywords: Quality assurance, image processing, test object

Ultrasound 2014; **22**: 15–20. DOI: 10.1177/1742271X13516896

Introduction

A number of professional bodies have recommended performance testing of ultrasound scanners as part of a quality assurance (QA) programme. Professional guidelines and the necessary test objects have been available for many years. In our experience, manual measurement and visual assessment of test object images are of little value.¹ These tests are prone to subjectivity and are labour-intensive. Grey-scale and anechoic target image visibility is recorded in a coarse fashion as either 'seen' or 'not seen' (or possibly 'seen, but unclear'). There is a need for objective, verifiable and repeatable testing devices and analysis techniques, i.e. evidence-based QA.² Hangiandreou et al.³ recently recommended quarterly assessments of mechanical integrity (careful inspection of scanner, transducers and cables) and image uniformity (using subjective acceptance criteria), having found that these tests detected 91.4% of equipment failures (1.6% and 7% being detected by depth of penetration measurement and clinical users, respectively). There is, however, no way to determine how many faults were not detected by any of their methods.

We have developed an automated, objective method for measuring recommended imaging performance parameters

using conventional test objects (the Nottingham USQC system).⁴ Initial work on the system showed that the speed and reproducibility of the measurements are superior to current manual and visual methods. The results provided for grey-scale and anechoic target image visibility are no longer quantised and can be used for serial comparisons.

The aim of this study was to assess the ability of the system to demonstrate changes in the performance of individual scanners using (1) known scanner or probe faults; (2) suspected scanner or probe faults and (3) analysing routine QA results.

Method

The test objects used in our QA programme were an ATS 539 (ATS Laboratories, Bridgeport, CT, USA) tissue-mimicking test object and, for small parts probes, a Gammex RMI 404GS (Gammex RMI, Middleton, WI, USA) test object. Each test object contains a column of nylon filaments for spatial resolution measurement, anechoic targets for assessing high contrast performance and variable backscatter targets for assessing overall grey-scale performance. The RMI 404GS is gel-based, with a speed of sound of 1540 m s^{-1} and specified attenuation of $0.5 \text{ dB cm}^{-1} \text{ MHz}^{-1}$. The ATS 539 is

urethane-based, with a speed of sound of 1450 m s^{-1} and specified attenuation of $0.5 \text{ dB cm}^{-1} \text{ MHz}^{-1}$, although attenuation is not linearly related to frequency.⁵ Although the effect of this low speed of sound on beam profile has been well documented,^{6–9} such test objects remain suitable for serial testing of individual scanners, where relative, rather than absolute, measurements are important. There is an advantage in the defocusing effect of urethane, in that changes in resolution are more likely to be demonstrated.¹⁰

Software was written, in house, to analyse the test object images using Matlab® (Mathworks Inc, Natick, MA, USA).⁴ This processing software was interfaced with Microsoft Access® (Microsoft, Redmond, WA, USA) to record results from routine QA testing.

For each test, images of each set of targets were analysed to provide the following measurements: (1) lateral resolution, expressed as the full-width-half-maximum (FWHM) of the nylon filament images above the mean speckle background (mean speckle measured adjacent to each nylon filament, typically over about 20 pixels for a lower frequency probe in the ATS 539); (2) slice thickness, expressed as the FWHM of the filaments imaged with the probe angled at 45° to the side of the test object¹¹; (3) low contrast penetration (LCP, the maximum depth of speckle visibility) by calculating profiles of standard deviations of speckle and of noise from (summing and differencing) a successive pair of images captured with the probe stationary; (4) anechoic target visibility, using the correlation of an 'ideal' image (black circle on white background) with the actual image; (5) grey-scale target contrast, expressed as the ratio of mean pixel values on circles inside and outside the target image (0.7 and 1.35 times the radius of the target, respectively); (6) grey-scale target visibility, calculating the standard error of the difference between the mean pixel values on circles inside and outside the target image and expressing visibility as the ratio of difference between the means and the standard error.

These measurements were chosen, as several are widely used in testing imaging performance, e.g. resolution and contrast are used in all modalities, they make use of currently available test objects and the large uncertainty associated with visual/manual measurement methods can be reduced by image analysis.⁴ In order to achieve reproducible settings, an appropriate scanner preset is used, swept gain is set (usually) to mid-range settings and all scanner settings are recorded at initial baseline testing. At baseline testing, tolerance levels for the results were set based on two standard deviations of the results from five sets of images for each scanner.

Gorny et al.¹² tested three methods for determining LCP (or DOP – maximum depth of penetration – in their terminology), finding that image-pair methods such as ours are subject to problems of hand movement. We currently use the cine-loop function to capture two sequential images from a visually stable image sequence, with failure to compute LCP being very rare. Our image processing, filtering and profile analysis method was refined by careful experimentation.⁴ Gorny et al.¹² used different methods to produce and analyse their signal and noise profiles, which may

be a reason for their high failure rate due to hand motion. Their alternative 'single phantom image' method has merits and deserves further evaluation.

The automated system has been gradually introduced into our routine QA programme. Baseline tests with the automated system have been performed on 50 scanners. This study analysed the first set of routine (biannual) tests carried out on 48 scanners, the second set on 15 scanners and the third set on 5 scanners, i.e. 68 routine tests in total, comparing a single set of results with the baseline tolerances. In addition to this, any user-reported fault or suspected fault, whether on scanners within or outside this scanner cohort, was followed up by further investigation using the automated test method as felt appropriate. If a fault was suspected on a scanner outside the cohort, then baseline test results were not available, but analysis was achieved, for example, by comparing test measurements taken with an alternative or replacement transducer.

Where possible, images were transferred from on-board storage using an uncompressed format (DICOM images are our currently preferred format). In some cases it was necessary to digitise the video signal from the scanner; this was achieved using a Video Capture Essentials VCE-B5A01 video capture system (Imperx Incorporated, Boca Raton, FL, USA) capable of capturing images of 640×480 pixels with a grey level range of 0–255.

Results and discussion

No faults were found on small parts probes tested with the Gammex RMI 404GS test object, or on phased array probes, and no faults were reported or suspected by users of these probes. The following results are from curvilinear array probes used for general abdominal or obstetric scanning and are summarised in Table 1.

Several isolated results fell slightly out of tolerance, e.g. for a single target within a series; these were regarded as being due to chance, with follow-up at the next scheduled QA visit (no results available at the time of analysis). One routine test was significantly out of tolerance. In addition to this, two scanners with known faults and two with user-suspected faults were analysed and further investigated.

Routine tests out of tolerance

Results for a single curvilinear probe fell considerably outside tolerance; LCP was reduced and grey-scale contrast was increased, results for all grey-scale targets falling out of tolerance. On investigation, it was found that the supplier's clinical applications specialist had visited and altered the grey-scale transfer curve for the default obstetric setting for this probe, resulting in a generally darker image. Results for a repeat test, using the original grey-scale transfer curve, were within tolerance.

It is reassuring that the automated system demonstrated a gross change in grey-scale transfer characteristics following this deliberate, but unrecorded, change to a pre-processing function.

Hangiandreou et al.³ demonstrated an annual scanner component failure rate of 10.5% and an annual transducer failure rate of 13.9% (both averaged over 4 years) using

Table 1 Summary showing parameters where results under fault or routine test conditions were significantly different from reference results ($p < 0.05$ unless otherwise stated)

Fault/test	Significant results	Action
Shadowing in air only	None	None
Shadowing in test object	Resolution degraded; 3 dB grey-scale target visibility reduced	Probe replaced
User reported image deterioration (1)	Slice thickness degraded; grey-scale contrast increased; LCP marginally reduced ($0.1 > p > 0.05$)	Probe replaced
User reported image deterioration (2)	Grey-scale target contrast reduced; resolution degraded	Probe replaced
Routine testing	LCP reduced; grey-scale contrast increased	Reverted to original grey-scale setting

LCP: low contrast penetration.

routine tests of mechanical integrity, image uniformity (presence of artefacts in a test object image and in-air), distance accuracy (no failures) and DOP. These results highlight the importance of a QA programme, but further data are required on failure rates for image quality parameters other than uniformity and DOP.

As image quality faults are less prevalent with modern scanners, extensive experience will be required in order to evaluate the routine use of these methods in the prospective detection of faults. Further work is required to determine the optimum scanner settings to produce maximum change in testing results in response to changes in performance and to facilitate interpretation of results.

Work is also required to determine the ideal test object properties to demonstrate and allow understanding of changes in scanner performance. Test objects for routine QA should be designed to maximize the deviation in results with changes in performance, rather than to meet ill-defined criteria of tissue equivalence. The purpose of routine performance testing is to detect changes, rather than to predict clinical performance.

Known faults

At baseline testing, two probes presented with shadowing in the image, presumably due to an electrical connection or mechanical coupling fault; this is sometimes described as 'crystal dropout.' In one case, shadowing was evident only in the reverberation pattern with the probe operating in air (no gel) and not in the image of a patient or test object. In the other case, shadowing was also seen in the test object image. Images were captured with the ATS 539 targets in two positions within the image, aligned with the probe defect and distant from the probe defect.

Where shadowing was seen only in the reverberation pattern with the probe operating in air and not in the test object, no significant differences in test results were found between the two sets of images (shadowed and non-shadowed areas). Where shadowing was also seen in the test object, resolution under the shadowing area was significantly degraded in the distal part of the image (Figure 1). Visibility of the 3 dB target, imaged at a depth of 4 cm directly under the shadowing area, was significantly

Figure 1 Lateral resolution expressed as the full-width-half-maximum (FWHM) of the imaged target profile in a shadowing area (diamonds) and a non-shadowing area (squares; ± 2 standard deviations)

reduced (from 9.2 to 5.6, $p < 0.05$). No other significant differences were demonstrated.

The degradation in resolution in the distal image (Figure 1) may be due to broadening and flattening of the beam profile from the inefficient area of the probe. The reduced visibility of the 3 dB grey-scale target can be explained by slightly reduced signal amplitude in the shadowing region increasing the standard error of the difference between pixel values inside and outside the target; this reduction will cause a proportionally larger change in visibility of lower contrast targets.

We are not convinced of the need for detailed tests of image quality in cases of drop-out. If there is shadowing in the in-air reverberation pattern there is evidence to suggest that any Doppler beam in that location will be adversely affected¹³; the probe could be used clinically with care in positioning pulsed wave Doppler beams. If shadowing is evident in clinical images, the probe should be replaced; if the probe is kept in use, areas of clinical interest should be imaged using other parts of the array.

User suspected faults (1)

Users reported gradual deterioration in image quality with a 3.75 MHz curvilinear probe used to guide renal biopsies. The antimicrobial fluid used for skin antisepsis prior to

Figure 2 Slice thickness expressed as the full-width-half-maximum (FWHM) of the imaged target profile for the faulty probe (diamonds) and the reference probe (squares; ± 2 standard deviations)

Figure 3 Gray-scale target contrast for the faulty probe (diamonds) and the reference probe (squares; ± 2 standard deviations)

biopsy had discoloured the probe face. Routine tests of sensitivity and noise¹ had shown no change over several years. As the scanner was outside the cohort for automated QA baseline tests, the probe was tested on a recently decommissioned scanner (within the cohort) of the same model and compared with that scanner's probe (the 'reference' probe) whose automated QA baselines had been established and performance had remained within tolerance.

Several tests showed a significant difference in performance between the probe with a discoloured face and the reference probe. Figure 2 shows a difference in slice thickness, which is significant ($p < 0.05$) for 12 of the 17 targets. LCP was marginally less ($0.1 > p > 0.05$) for the faulty probe (131 mm) relative to the reference probe (141 mm). Figure 3 shows a significant difference ($p < 0.05$) in contrast for five of the six targets for the discoloured probe compared to the reference probe; no other significant performance differences were demonstrated. The users were supplied with the reference probe as a replacement for the discoloured probe and were satisfied that the image quality showed an improvement.

The increased contrast in images from the discoloured renal biopsy probe (Figure 3) was unexpected. Close inspection of the grey-scale histograms revealed the reasons for this. For example, the modal grey-scale within the 15 dB

Figure 4 Gray-scale target contrast for the preferred scanner (circles), the scanner where users perceived deterioration in performance (diamonds) and the latter scanner at baseline testing (squares; ± 2 standard deviations)

target was 20% less for the discoloured probe, whilst the modal grey-scale in the background material was almost 50% less, resulting in increased contrast. The clinically perceived deterioration in image quality was due to the reduction in overall signal level, an observation supported by the reduction in LCP. The concomitant degradation in slice thickness (Figure 2) suggests that this was due to a change in the acoustic properties of the lens material and it is expected that degradation in slice thickness will affect clinical image quality.¹⁴ In the context of routine testing, the results from the discoloured probe would be unequivocally out of tolerance, giving confidence in the future interpretation of routine results.

User suspected faults (2)

Six months after the commissioning of two identical scanner/probe combinations (only one within the automated QA cohort), users reported a preference for one scanner over the other. There were no differences in routine sensitivity and noise tests. Reactive measurements were made on both scanners and compared with the automated QA baseline measurements made on one of the scanners (the least preferred).

Figure 4 shows measured grey-scale target contrast versus specified contrast for the two identical scanners, where one was preferred over the other. There was a significant difference ($p < 0.05$) in contrast between baseline and reactive measurements on the least preferred scanner at +3 dB and +15 dB. Further analysis of absolute differences between paired contrast results¹⁵ showed a significant difference between baseline and reactive measurements on the least preferred scanner (mean absolute difference: 0.046; 95% confidence interval: 0.011–0.081; $p < 0.05$); this corresponds to our visual assessment of differences in shape between the contrast curves (shallower slope for the least preferred scanner). A significant deterioration in resolution ($p < 0.05$) was also demonstrated over the range 20 to 70 mm (Figure 5).

The probe from the deteriorating scanner was retested on the preferred scanner and produced similar results to those on its host scanner, suggesting that the suspected fault was associated with the probe and not the scanner.

Figure 5 The improvement of resolution with a replacement probe (circles), compared to the original probe after user perceived deterioration in performance (diamonds). Baseline results for the original probe are also shown (squares; ± 2 standard deviations). A single transmit focus was set at 90 mm

The probe was replaced; the grey-scale contrast then matched that of the preferred scanner, and resolution improved, as shown in Figure 5, confirming a fault on the original probe.

We were able to support user perceptions of image deterioration on one of a pair of identical scanners. Resolution had deteriorated over a short range (Figure 5) and contrast had fallen (Figure 4); replacing the probe improved both resolution and contrast to earlier levels. It is of note that the grey-scale targets are at a similar depth to the region of deterioration in resolution; the loss of contrast may be due to a change in the speckle pattern and increased side lobe noise.

In the context of routine testing, the contrast results (Figure 4) would not appear to be unequivocally out of tolerance. Further work is required to facilitate the interpretation of small changes in contrast curves; interpretation may be simplified by choosing a linear grey-scale transfer curve¹⁶ if this is available and can be determined from the scanner controls and display.

The resolution results (Figure 5) were difficult to interpret but the fault was isolated to the probe by testing on another scanner, hence a probe replacement was appropriate (in the absence of another scanner we could have obtained a loan probe from the supplier to isolate the fault).

The objective confirmation of users' subjective judgement of deterioration in image quality indicates that such a QA system has value, at least for reactive testing (provided that baseline results have been obtained at commissioning or following upgrades that will affect image quality). The large differences between the performance of faulty and reference probes in Figures 2 and 3 suggest that scheduled testing may detect progressive deterioration earlier.

Conclusions

We have demonstrated changes in one or more parameters in cases of known or suspected faults, providing evidence for the efficacy of this method of performance testing.

Clinical users may believe that they can detect faults without a formal QA programme. Their vigilance, especially for acute faults, is a vital component of any QA programme, but it is likely that by the time they detect an image quality fault it will be clinically significant; our reactive testing has confirmed such faults and, if performed as scheduled testing, may have detected progressive faults earlier.

DECLARATIONS

Competing interests: None

Funding: None

Ethical approval: Not required

Guarantor: NJD

Contributorship: NJD and NMG researched and conceived the study. NMG developed the Nottingham US QA software. NJD analysed the data and wrote the first draft of the manuscript. Both authors reviewed and edited the manuscript and approved the final version.

REFERENCES

1. Dudley NJ, Griffith K, Houldsworth G, et al. A review of two alternative ultrasound quality assurance programmes. *Eur J Ultrasound* 2001;12:233–45
2. Moore GW, Gessert A, Schafer M. The need for evidence based quality assurance in the modern ultrasound clinical laboratory. *Ultrasound* 2005;13:158–62
3. Hangiandreou NJ, Stelkel SF, Tradup DJ, et al. Four-year experience with a clinical ultrasound Quality Control program. *Ultrasound Med Biol* 2011;37:1350–7
4. Gibson NM, Dudley NJ, Griffith K. A computerised ultrasound quality control testing system. *Ultrasound Med Biol* 2001;27:1697–711
5. Browne JE, Ramnarine KV, Watson AJ, et al. Assessment of the acoustic properties of common tissue-mimicking test phantoms. *Ultrasound Med Biol* 2003;29:1053–60
6. Goldstein A. The effect of acoustic velocity on phantom measurements. *Ultrasound Med Biol* 2000;26:1133–43
7. Dudley NJ, Gibson NM, Fleckney MF, et al. The effect of speed of sound in ultrasound test objects on lateral resolution. *Ultrasound Med Biol* 2002;28:1561–4

8. Goldstein A. Beam width measurements in low acoustic velocity phantoms. *Ultrasound Med Biol* 2004;**30**:413–6
9. Chen Q, Zagzebski JA. Simulation study of effects of speed of sound and attenuation on ultrasound lateral resolution. *Ultrasound Med Biol* 2004;**30**:1297–306
10. Green HL, Dudley NJ, Gibson NM. A comparison between urethane and gel test objects for ultrasound imaging performance testing. *Ultrasound* 2007;**15**:105–8
11. Skolnick ML. Estimation of ultrasound beam width in the elevation (slice thickness) plane. *Radiology* 1991;**180**:286–8
12. Gorny KR, Tradup DJ, Hangiandreou NJ. Implementation and validation of three automated methods for measuring ultrasound maximum depth of penetration: application to ultrasound quality control. *Med Phys* 2005;**32**:2615–28
13. Weigang B, Moore GW, Gessert J, et al. The methods and effects of transducer degradation on image quality and the clinical efficacy of diagnostic sonography. *J Diagnos Med Sonogr* 2003;**19**:3–13
14. Browne JE, Watson AJ, Muir C, et al. An investigation of the relationship between in vitro and in vivo ultrasound image quality parameters. *Ultrasound* 2004;**12**:202–10
15. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991:189–9115
16. Thijssen JM, Weijers G, de Korte CL. Objective performance testing and quality assurance of medical ultrasound equipment. *Ultrasound Med Biol* 2007;**33**:460–471