2004-01-01

An Investigation of the Relationship Between In-vitro and In-vivo Ultrasound Image Quality Parameters

Jacinta Browne
Dublin Institute of Technology, jacinta.browne@dit.ie

Amanda Watson
Western Infirmary Glasgow

Cathy Muir
Western Infirmary Glasgow

Peter Hoskins
University of Edinburgh

Alex Elliott
University of Edinburgh

Recommended Citation

Browne, J., Watson, A., Muir, C., Hoskins, P., Elliott, A.: An investigation of the relationship between in-vitro and in-vivo ultrasound image quality parameters. Ultrasound, Vol. 12 (4), pp.202-210. 2004. doi:10.1179/174227104X5016

AN INVESTIGATION OF THE RELATIONSHIP BETWEEN IN-VITRO AND IN-VIVO ULTRASOUND IMAGE QUALITY PARAMETERS

JACINTA E BROWNE[1], AMANADA J WATSON[1], CATHY MUIR[2], PETER R HOSKINS[3] AND ALEX T ELLIOTT[1]

Address: Dept. of Clinical Physics, Western Infirmary Glasgow[1], Dept. of Radiology, Western Infirmary Glasgow[2], and Medical Physics Dept., Edinburgh University[3].

Dr Jacinta Browne, Nuclear Cardiology, Queen Elizabeth Building, Glasgow Royal Infirmary, 16 Alexandra Parade, Glasgow, G31 2ER.

Tel: 0141 - 2110501

Fax: 0141 - 2111252

Email: Jacinta.Browne@northglasgow.scot.nhs.uk

Key Words: B-mode ultrasound, Colour Doppler, Performance testing, quality assurance and subjective perception.

The aim of this paper is to investigate the relationship between B-mode and colour Doppler technical test methods with the clinical perception of B-mode and Doppler in-vivo test parameters. It was found that technical and clinical comparisons between the B-mode test parameters: lateral resolution versus clinical resolution; anechoic target detection versus clinical noise; and penetration depth versus clinically useful penetration depth, demonstrated moderate correlations, ($r= -0.69$, $p<0.003$; $r= 0.5$, $p=0.14$; and $r= 0.56$, $p<0.03$ respectively). However, axial resolution versus tissue texture variation; slice thickness versus overall clinical image quality; and contrast resolution versus clinically useful dynamic range demonstrated poor correlations. The majority of the colour Doppler performance parameters were found to demonstrate moderate correlations: sensitivity performance index and clinical Doppler sensitivity ($r= 0.52$, $p<0.07$); axial / lateral resolution and clinical colour Doppler resolution ($r= -0.64$, $p=0.02$ / -0.55, $p=0.05$); and temporal resolution and clinical temporal resolution ($r= -0.59$, although not statistically significant $p=0.4$). The poor correlations for axial resolution slice thickness and contrast resolution suggest that some revision of the test protocols may be required or that these quantities are not as important as previously thought in image quality, whereas the tests protocols for anechoic target detection, penetration depth and in particular lateral resolution appear promising in their prediction of clinical perception of image quality. The newly developed colour Doppler test protocols and test objects also appear promising in their prediction of clinical perception and merit further investigation.

INTRODUCTION

There are currently in existence a number of technical performance protocols recommended by various professional bodies [1-3]. However, it has become evident that these performance tests fall short of adequately testing state of the art ultrasound scanners as they have not been able to detect deterioration in the scanners' performance, as reported by the operators [4-6]. Furthermore, there has been mounting evidence that current performance tests and test object results do not reflect perceived clinical performance of ultrasound scanners [7; 8]. The study conducted by the Common Services Agency evaluated cardiac scanners by scanning patients with known heart defects and did not compare individual parameters of image quality [7]. The study conducted by Metcalfe et al evaluated abdominal scanners by scanning one healthy volunteer and compared the individual image quality parameters [8]. A correlation was found for the former study, whereas none was found for the latter study. For test procedures and test devices to successfully evaluate the performance of B-mode and colour / power Doppler imaging, they should be able to differentiate scanners of varying complexity, and the results should be reflective of the clinical perception of the scanners' performance. In this paper, the efficacy of current and newly developed B-mode and colour Doppler test protocols and phantoms was evaluated by investigating the correlation between these in-vitro test results and the results obtained from a clinical study evaluating the B-mode and Doppler in-vivo test parameters on a range of ultrasound scanners of varying complexity.

METHODS

Subjective Image Assessment – Clinical Study

All B-mode images and colour Doppler cine-loops were acquired by the same experienced operator from one healthy female volunteer with an average amount of body fat. The abdomen was chosen as the anatomical site of interest for this clinical study for two main reasons. Firstly, most of the technological advancements have been applied to abdominal imaging (e.g. tissue harmonic imaging, compound imaging and broadband Doppler imaging). Secondly, abdominal imaging presents the majority of challenges and problems associated with ultrasonic imaging in general, all concentrated into one imaging procedure. Beam distortion, phase aberrations, clutter from slow moving tissue around deep, small vessels with slow flow, and vessels situated close together are problems typically encountered in abdominal imaging. When scanning the liver or kidney for example, the ultrasound pulse must first pass through a number of layers of fat and tissue, thereby undergoing distortion and attenuation. Furthermore, the liver and kidney have deep veins and arteries, while the kidney has small closely spaced vessels. The study was given ethics permission from the North Glasgow University Hospitals NHS Trust. The volunteer read and signed an informed consent form after the methods of the study had been verbally explained. The scanning protocol used in this study was a standard imaging protocol normally used to obtain ultrasound images of the right lobe of the liver and the right kidney within the same image plane [9; 10]. The B-mode and colour Doppler acquisition parameters were optimised in order to obtain the best images and cine loops from the different modes on each of the ultrasound scanners. The first image and cine loop obtained was used as a reference and the operator was asked to replicate that anatomical site as closely as possible on all subsequent scans. The different B-mode techniques available for each ultrasound scanner were also assessed. The colour Doppler cine loops were between 30 and 60 seconds in duration. The healthy female volunteer was scanned using all the scanners assessed. The B-mode images were captured using a video frame grabber (VideoPort^{TM}, MRT, Germany) whereas the colour and power Doppler cine loops were recorded using an S-VHS video recorder (SVO-9500MDP, Sony, USA). The images and the cine loops were graded by the operator during acquisition using the image quality parameters discussed above. The operator also graded the B-mode images and Doppler cine loops two weeks after all the clinical data had been collected in order to determine how viewing the images off-line affected the grades given. The frame grabber was calibrated using the SMPTE [11] calibration procedure. The video-recorder was used in routine clinical service to record ultrasound scans and was not calibrated. A total of nine ultrasound scanners of varying age and complexity were used in the clinical study (Table 1).

Ultrasound Scanners

Nine ultrasound scanners were assessed in terms of B-mode techniques, while only seven of the ultrasound scanners were assessed in terms of colour and power Doppler. The different B-mode techniques available on each ultrasound scanner were also assessed. A summary of the probes used and the modes tested for each of the ultrasound scanners is presented in Table 1. The ultrasound scanners were chosen to represent a cross-section of scanners, ranging from about 10 years to just a few months in age, as well as varying greatly in complexity (Table 1).

A total of 17 static B-mode images and 14 cine loops were presented to each of the 23 participants (5 sonographers and 18 radiologists), the range of experience of the participants was between 2 and 26 years. Along with a questionnaire in which the participants were asked to rate the images in terms of the B-mode and colour / power Doppler image quality parameters as outlined below. The participants were asked to grade the images and cine loops on a scale of poor (1) to very good (5) for each of the image quality parameters. The questionnaire consisted of four sections requesting information regarding: (i) the number of years experience of ultrasound imaging; (ii) the model of ultrasound scanner they most frequently use; (iii) their assessment of the images presented; and (iv) other comments. The participants were randomly divided into two groups: one group was presented with the images and cine loops in one random sequence, while the second group was presented with them in a different random sequence. This was done in order to remove any bias deriving from the order in which images and cine loops were seen. A duplicate of one of the images was included in the selection of images presented to the participants to determine the intra-observer variability and also as a control measure. Furthermore, one participant graded the images on two different occasions, two weeks apart, in order to determine the intra-observer variability over time. The B-mode images were all the same size (700x500x256) and were presented in a PowerPoint (Microsoft, USA) presentation slide show, while the cine loops were all displayed using the same monitor and video-recorder, and thus all of the B-mode images and cine loops were presented under the same conditions. The viewing conditions were similar to normal viewing room conditions and were kept very similar for each of the participants.

The in-vivo image quality parameters listed below were selected following discussion with a number of experienced ultrasound operators and with reference to the literature [2; 7; 8]. No attempt was made to define these clinical parameters to each of the participants, as it was important to determine whether the same definitions for these image quality parameters are universally accepted among ultrasound operators. The scanners and images were assessed in terms of the following categories for B-mode Image Quality:

(i) Overall image quality

Overall image quality was assessed to determine the participants' general impression of the image. The participants were asked to grade the images on a scale of poor (1) to very good (5) for overall image quality. This was used as the baseline category to determine if the image was considered "very good" (score of 5) or "poor" (score of 1); the subsequent parameters were used to explore the basis for this judgement. The range of grades given to the ultrasound scanners for each of the image quality parameters were found to be consistent, less than one standard deviation (std = 0.7), on a scoring scale ranging from one to five, indicating that the scoring scale chosen was appropriate.

(ii) Visualisation of variances in tissue texture

This image quality parameter is representative of the quality of fine detail between the tissue texture within the liver and the participants were asked to grade the images for visualisation of variances in tissue texture.

(iii) Clinical resolution

Clinical resolution is a measure of the ability to visualise fine detail and small structures and the participants were asked to grade the images for clinical resolution.

(iv) Clinical dynamic range

Clinical dynamic range is a measure of the range of signal magnitudes or signal-to-noise ratio, S/N (low signal magnitude-limited by the noise of the system and high signal magnitude-limited by the system saturation point) that can be handled by the ultrasound scanners and is important in differentiating structures of variable contrast and thus be important for diagnosis of tissue pathology. The participants were asked to grade the images for clinical dynamic range.

(v) Clinical penetration depth

Clinical penetration depth is the deepest point in the ultrasound image from which meaningful information could be obtained and the participants were asked to grade the images for clinical penetration depth.

(vi) Clinical noise

Clinical noise is the electronic noise and the speckle pattern of the image from the ultrasound scanner and the participants were asked to grade the images for clinical noise.

The scanners and images were assessed in terms of the following categories for Colour Doppler Image Quality:

(i) Overall Image quality

As with B-mode, the overall image quality was used to determine the participants' general impression of the image. This was used as the baseline category to determine if the image was considered "very good" (score of 5) or "poor" (score of 1); the subsequent parameters were used to explore the basis for this judgement. The range of grades given to the ultrasound scanners for each of the image quality parameters were found to be consistent, less than one standard deviation (std = 0.73), on a scoring scale ranging from one to five, therefore, indicating that the scoring scale chosen was appropriate.

(ii) Clinical temporal resolution

Clinical temporal resolution is a measure of the ultrasound scanners' ability to respond to changes in flow patterns and to display the vessel filling at a physiological rate. It was determined by the participants observing the rate at which the ultrasound scanner displayed the kidneys filling and the participants were asked to grade the cine loops for colour Doppler temporal resolution.

(iii) Clinical Doppler spatial resolution

Clinical spatial resolution is the ability to visualise small vessels and the participants were asked to grade the cine loops for colour Doppler spatial resolution.

(v) Clinical sensitivity / Clinical penetration depth

Sensitivity or clinical penetration depth is the deepest point in the colour / power Doppler image from which a signal could be obtained and the participants were asked to grade the cine loops for colour Doppler penetration depth.

(vi) Clutter suppression performance

Clutter can be caused by vessel wall, tissue movement, patient breathing or probe movement and is suppressed by the clutter filters. The amount of clutter present in the cine-loop is representative of the clutter filter's performance and the tissue-blood discriminator performance and the participants were asked to grade the cine loops for colour Doppler clutter suppression performance.

Objective Image Quality Assessment

The B-mode in-vitro test parameters evaluated using the Model 403 general purpose test object (Gammex RMI, Nottingham) and an automated image analysis program [12] were: axial resolution, lateral resolution, slice thickness, anechoic target detection, contrast resolution and penetration depth. Of the in-vitro parameters measured, it was hypothesised following discussions with a number of experienced ultrasound operators and with reference to the literature [7; 8], that the following correlations might be expected with the in-vivo parameters:

Axial resolution versus tissue texture variation – axial resolution may be defined as the image of a point source in the axial direction and a full characterisation of this parameter can be given as either the point spread function (PSF) or the spatial frequency response (modulation transfer function MTF) and it describes the scanner's ability to detect and clearly display closely spaced objects that lie along the beam's axis.

Lateral resolution versus clinical resolution – lateral resolution may be defined as the image of a point source in the lateral direction and a full characterisation of this parameter can be given as either the point spread function (PSF) or the spatial frequency response (modulation transfer function MTF). Lateral resolution describes the scanner's ability to distinguish structures that are closely positioned within the image plane along a line perpendicular to the beam's major axis.

Slice thickness versus overall image quality – Slice thickness may be defined as the image of a point source in the elevation direction and a full characterisation of this parameter can be given as either the point spread function (PSF) or the spatial frequency response (modulation transfer function MTF). Slice thickness or elevation focus describes the scanner's out-of-plane focus.

Anechoic target detection (Sensitivity) versus clinical noise – the anechoic target imaging test examines the scanner's ability to detect and accurately display round, negative contrast objects of various sizes at different depths within the test object.

Contrast range versus clinical dynamic range – Contrast resolution is the scanner's contrast dynamic range and gives an indication of the low and high contrast resolution achievable by the system. It was evaluated by determining the contrast resolution of four targets of varying backscatter levels (-12 dB, -6 dB, 6 dB and 12 dB)[6].

Penetration depth versus clinical penetration depth – Penetration depth is the point at which the weakest echo signal level can be detected and clearly displayed.

The axial resolution, lateral resolution and slice thickness results at a depth of 60 mm (elevation focus) for each scanner and imaging mode were measured and correlated with their respective in-vivo scores (tissue texture variation, resolution and overall image quality respectively). The in-vitro contrast range was determined from the difference between the

+12 dB and the -12dB target backscatter ratios, and then correlated with its corresponding in-vivo test parameter.

The in-vitro colour Doppler test parameters evaluated using the newly developed test procedures and test objects [13] were: axial resolution, lateral resolution, sensitivity, temporal resolution and tissue movement suppression performance. Of the in-vitro parameters measured, it was suggested that the following correlations might be expected with the in-vivo parameters:

Axial resolution versus clinical Doppler spatial resolution – Axial resolution is the minimum separation in space in the axial direction for which two separate point or line targets can be resolved. It describes the scanner's ability to distinguish vessels that are closely positioned along the beam's axis.

Lateral resolution versus clinical Doppler spatial resolution – Lateral resolution is the minimum separation in space in the lateral direction for which two separate point or line targets can be resolved. It describes the scanner's ability to distinguish vessels that are closely positioned within the image plane along a line perpendicular to the beam's major axis.

Sensitivity versus Doppler sensitivity / overall image quality – Sensitivity is the minimum signal strength (from different diameter vessels and from different depths) that the lowest detectable velocity can be detected unambiguously. It describes the scanner's ability to distinguish low velocity flow (1 – 10 cm s^{-1}) in small diameter vessels (5 – 1 mm) at depth.

Temporal resolution versus clinical temporal resolution – temporal resolution is the minimum separation in time for which two separate events can be identified.

Tissue movement suppression performance versus clutter filter performance – tissue movement suppression performance is the ability of the tissue suppression algorithm and the clutter filters to remove strong signals from slowly moving tissue, while still preserving the low velocity content of the colour flow signal.

The axial and lateral resolution, sensitivity performance and temporal resolution for each scanner and imaging mode under test were determined and the results were correlated against the in-vivo score for spatial resolution, sensitivity (overall image quality) and temporal resolution respectively. The tissue movement performance score was correlated against the in-vivo score for clutter filter performance (noise).

RESULTS

B-mode In-Vivo Image Quality Performance

The average differences between the grades given to the B-mode real-time images and the off-line images of the image quality parameters by the operator had an average standard deviation of less than ± 0.66 of a grade level.

The influence of the number of years experience on the grades given to the image quality parameters for each of the ultrasound scanners was investigated and it was found that there was no relationship between the two ($r=0.004$, $p=0.6$).

The grades given to the duplicate images within the same grading session by each of the participants were analysed using the analysis of variance (ANOVA) test and it was found that for a significance level of $p < 0.001$, the sets of grades were not significantly different ($p = 0.57$). The intra-observer variability of one of the participants between two separate grading sessions over two weeks was found to be good, with a standard deviation of ± 0.4 of a grade level; in all cases, the images were given the same grade or within one grade of the previous grading session. These results indicate a high level of grading consistency for ultrasound operators within a grading session and over time.

The inter-observer variability was found to be relatively consistent, demonstrating a standard deviation of ± 0.7 of a grade level (grading standard error = ± 0.14). This standard error (s.e. = ± 0.14) was very similar to that found by Metcalfe and Evans (s.e. = ± 0.12) in a similar study of the relationship between quality assurance image quality parameters and subjective operator image assessment [8].

It was found that the resolution had the largest contribution to the overall image quality, with a correlation coefficient r = 0.79. The tissue texture (r = 0.64), the noise level (r = 0.64) and the dynamic range (r = 0.6) were all found to have a strong correlation with the overall image quality, while the penetration depth (r = 0.49) only had a moderate correlation with the overall image quality.

The correlation coefficient found between the commonly used in-vitro B-mode test parameters and the corresponding in-vivo test parameters are presented in Table 2. It was found that the scores for the B-mode and the THI in-vivo image quality test parameters were not statistically different, which was an unexpected result and may have been due to the average body habitus of the healthy volunteer. Therefore, the correlation between the different B-mode in-vivo and in-vitro parameters was investigated without the scores of the THI modes of the different scanners. The correlation between the in-vitro results and the in-vivo scores improved marginally when the THI scores were removed.

Colour / Power Doppler In-Vivo Image Quality Performance

The average differences between the grades given to the colour / power Doppler real-time images and the off-line cine-loops of the image quality parameters by the operator had an average standard deviation of less than ± 0.76 of a grade level. This demonstrated that the colour / power cine loops presented to the participants gave a good representation of the scanners' image quality. However, the grades for these cine loops demonstrated more variability between the real-time images and the off-line images than for the B-mode images, which may be due to the limited length of the cine-loops (30 – 60 seconds).

As with B-mode, the number of years of ultrasound experience of the participants had no effect on the grades awarded ($r = 0.003$).

The grades given to the duplicate cine clips within the same grading session by each of the participants were analysed using the analysis of variance (ANOVA) test and it was found that for a significance level of $p < 0.001$, the sets of grades were not significantly different ($p = 0.49$). It was found that the intra-observer variability for two separate grading sessions over two weeks demonstrated very little variance with a standard deviation of less than ± 0.4 of a grade level. These results indicate an acceptable grading consistency for ultrasound operators within a grading session and over time.

The inter-observer variability for colour / power Doppler was also found to demonstrate good consistency of responses, with a standard deviation of less than ± 0.73 of a grade level (standard error (s.e) = ± 0.15).

To determine the most important contribution to the overall image quality of colour and power Doppler imaging of an ultrasound scanner, scatter plots of the grades for overall image quality were plotted against the individual image quality parameters. It was found that the temporal and spatial resolutions had the largest contribution to the overall image quality, with coefficients of correlation $r = 0.78$ and $r = 0.79$ respectively. The clutter filter performance ($r = 0.75$) and the velocity resolution ($r = 0.56$) were both also found to have a strong correlation with overall image quality, while the sensitivity performance ($r = 0.54$) only had a moderate correlation with overall image quality.

The effect of body habitus on the colour / power Doppler images quality was investigated by scanning three healthy subjects of different fat mass (below average, average and excess fat levels). It was found that the scores given by the participants for the image quality parameters of a mid-range scanner (SonoSite 180) were not significantly different ($p = 0.17$) as determined by ANOVA at a significance level of $p < 0.01$ for the sensitivity performance, the velocity resolution and the clutter filter performance. However, it was found that habitus did affect the following image quality parameters: overall image quality ($p = 1.7 \times 10^{-6}$), temporal resolution ($p = 0.0039$) and spatial resolution ($p = 8.2 \times 10^{-6}$), as determined by ANOVA.

The correlation coefficient found between the commonly used in-vitro Colour Doppler test parameters and the corresponding in-vivo test parameters are presented in Table 3.

DISCUSSION

In this paper, comparisons were made between commonly used and newly developed in-vitro B-mode and Doppler test parameters and their corresponding in-vivo parameters assessed as part of the clinical evaluation. It was found that the in-vitro B-mode test parameters axial resolution, slice thickness and contrast resolution all had poor correlations with the corresponding in-vivo test parameters. It is possible that the interpretation of these corresponding B-mode in-vivo test parameters was not the same for all observers, although the relative similarity that was observed between observers, which had a standard error $\cong 0.14$, suggested that this is not the case. The scores for the in-vivo image quality test parameters found no difference between the image quality of B-mode and tissue harmonic imaging modes, which was an unexpected result and may have been due to the average body habitus of the healthy volunteer. It does appear from the literature and from communications with ultrasound operators that body habitus may affect the diagnosis of tumours and masses due to limited field of view and degraded image quality [14-16]. The degree of image degradation which occurs for different amounts of body fat is not known; furthermore, men and women deposit fat in different ways, usually around the organs in men whereas women have more fat deposited in layers next to the skin around the abdomen. Therefore, the correlation between the different B-mode in-vivo and in-vitro parameters was also investigated without the scores of the THI modes of the different scanners. The correlation between the in-vitro results and the in-vivo scores improved marginally when the THI scores were removed. It was expected that axial resolution would be strongly correlated with tissue texture variances, as this is often considered to be a measure of fine speckle pattern, which in turn tends to be associated with fine detail and good resolution [17]. However, no correlation was found between these two parameters with or without the THI results included. The in-vitro axial resolution results did not appear to be representative of the scanners' complexity (r=0.002), whereas the in-vivo scores for tissue texture were representative of the scanners' complexity. This suggests that the test method recommended by the different professional bodies may not have been suitable [1-3], or indeed that this test parameter may not be a useful indicator of scanner performance. In order to determine which may be the case, it is suggested that axial resolution should be determined using alternative test objects, and that in a future study the efficacy of the axial resolution determined using these alternative test objects should be investigated. Examples of alternative test objects which could be used to measure axial resolution are a point target such as a ball bearing in speed of sound corrected water or TMM [18], or an anechoic target TMM phantom [19]. The relationship between axial resolution and perception of speckle and resolution cell volume may also warrant investigation in this area as good resolution tends to be associated with a fine speckle pattern [17].

It was found that lateral resolution had a moderate negative correlation with clinical perception of resolution; however, it was expected that a stronger correlation than that measured would exist. When the in-vitro results and the in-vivo scores for the THI modes of the different scanners were removed, the correlation between lateral resolution and clinical perception of resolution was found to be very significant ($r = -0.93$). In the clinical study it was found that the resolution scores had a strong positive correlation ($r = 0.79$) with overall image quality. Therefore, lateral resolution appears to be an important technical parameter for predicting clinical image quality for abdominal imaging and appears to be a promising quality control (QC) test.

It was further expected that slice thickness would be strongly correlated with overall image quality, as it has an impact not only on the overall image quality but also on the detectability of cyst structures, particularly when they are located in a region of the beam at which the width of the elevation plane is greater than that of the cyst [20]. However, no correlation was found between these two parameters with and without the THI results included. The in-vitro slice thickness results did not appear to be representative of the scanners' complexity, suggesting that the method recommended by the different professional bodies may not have been suitable [1-3], or that again the test parameter may not be a useful indicator of scanner performance. However, slice thickness has a strong influence on the clinical image quality and the detectability of cysts and, therefore, the use of slice thickness measurement at the fixed depth of 60mm instead of the measurement of slice thickness as a function of depth or the method of evaluation are most likely to be the reasons for the poor correlation.

It was found that sensitivity had a moderate positive correlation with clinical perception of noise, with and without the THI results included. The trend demonstrated in the sensitivity results appeared to be representative of the ultrasound scanners' complexity. The clinical noise level performance demonstrated limited differentiation between low-range and mid-range levels of ultrasound scanner's, this was most likely due to the participants grading the noise level from a static image instead of a cine loop. The moderate correlation is suggestive that the test method used to determine B-mode sensitivity appears to be a promising QC test.

It was found that contrast resolution had a slight positive correlation with clinical perception of dynamic range, with and without the THI results included. The ability of a scanner to display a low contrast or high contrast lesion in a tissue background is limited by the noise or the saturation point of the scanner respectively and as such the contrast resolution of a scanner may be representative of the dynamic range gradient of the scanner. The trend found in the contrast resolution results did not appear to be very representative of the ultrasound scanners' complexity. The weak correlation found for this parameter is probably due to the fact that dynamic range is just one aspect of the scanners grey-level characteristics and furthermore, the test phantom used for determining this parameter lacked contrast targets of varying size and at varying depths. Therefore, the grey-level transfer curves of the scanner could not be determined, alternatively, the use of a contrast-detail test phantom may yield better correlation between the in-vitro and in-vivo results [21].

It was found that penetration depth had a moderate positive correlation with clinical perception of penetration depth, with and without the THI results included. The trend found in the penetration depth results did not appear to be very representative of the ultrasound scanners' complexity. Penetration depth is very dependent on the examination type and the probe frequency used, and has been found not to be a major contributing factor to operator assessment [4].

A range of scanners of varying cost and age were evaluated as part of this study and it was found that the cost of the scanner was moderately strongly correlated to clinically perceived B-mode image quality (r = 0.7, p>0.05) while no correlation was found between age of the scanner and clinically perceived B-mode image quality (r = 0.04). The lack of correlation between age of the scanner and clinically perceived B-mode image quality was unexpected however, it may have been influenced by the fact that the majority of the scanners included in the study were less than 1 year old and this had a disproportionate effect on the result.

Overall, it was found that the in-vitro Doppler test parameters showed a moderate correlation with the corresponding in-vivo test parameters. It was found that axial and lateral colour Doppler resolution both had a moderate negative correlation with clinical perception of spatial resolution which was expected as it is a commonly held belief that these objective and subjective parameters are linked, due to resolution being considered an important factor in the visualisation of small areas of flow. Furthermore, the trend found in the colour Doppler resolution results appeared to be representative of the ultrasound scanners' complexity, indicating that this test may be a promising QC test.

It was expected that temporal resolution would be strongly correlated with the clinical perception of temporal resolution which is of clinical importance as, flow events within the body change very rapidly, particularly for flow in the heart, therefore, a high frame rate is needed to follow these changes. However, no correlation was found between these parameters. The in-vitro temporal resolution results appeared for the most part to be representative of the scanners' complexity, suggesting that this method of evaluation is suitable. It was found that the temporal resolution of the Acuson 128 was very good although it had a poor temporal resolution clinical score and it is probable that the poor clinical score given for the temporal resolution was strongly influenced by its poor sensitivity and spatial resolution. When the temporal resolution result for the Acuson 128 is removed, there is moderate negative correlation (r = -0.59, p=0.4) between the two parameters. Consequently, given the limited number for scanners (five ultrasound scanners) tested, the Acuson 128 result invariably had a disproportionate effect on the overall correlation result.

As expected a moderate correlation was found between the in-vitro parameter sensitivity performance index and the clinical perception of sensitivity performance and the in-vitro scores for the sensitivity performance indices appeared for the most part to be representative of the scanners' complexity, suggesting that the method of evaluation is suitable. Doppler sensitivity is an important parameter for predicting an ultrasound scanners ability to distinguish between slow flow and no flow - if flow cannot be detected then no other aspect of Doppler performance matters.

It was expected that tissue movement suppression performance would be strongly correlated with the clinical perception of clutter filter performance or noise. However, only a weak correlation was found between these parameters. The in-vitro scores for tissue movements suppression performance appeared for the most part to be representative of the scanners' complexity, suggesting that the method of evaluation was suitable however, further work needs to be carried out to optimise the test parameter.

A range of scanners of varying cost and age were evaluated as part of this study and it was found that the cost of the scanner was moderately correlated to clinically perceived Doppler image quality (r = 0.4, p>0.05) and the age of the scanner was also moderately correlated to the clinically perceived Doppler image quality ($r = 0.5$, $p>0.05$). The reason for a moderately strong correlation existing between age of the scanner and clinically perceived Doppler image quality may be due to the vast technological improvements due to piezoelectric material design, matching layer design and developments in digital electronics which have had a significant effect on the image quality of colour Doppler imaging[13; 22; 23].

This study, which evaluated the efficacy of currently recommended B-mode in-vitro test parameters and TMM test phantoms, and newly developed colour / power Doppler in-vitro test parameters and test devices, was limited to the evaluation of ultrasound scanners within close proximity to the testing laboratory. Apart from two recently decommissioned scanners, the scanners which were evaluated were in regular clinical use at the time of this study, which limited the amount of clinical data which could be acquired during the study. Any future study should involve the co-operation of ultrasound departments at multiple sites. This would facilitate the use of a larger number of ultrasound scanners spanning a range of different manufacturers, as well as a wider group of observers, ideally with different experiences of ultrasound scanners across a range of manufacturers.

CONCLUSIONS

In this paper, it was found that for B-mode imaging, the axial resolution, slice thickness and contrast resolution did not reflect the clinical perception of corresponding in-vivo parameters, while lateral resolution, sensitivity and penetration depth demonstrated moderate or greater correlations with the corresponding in-vivo parameters, demonstrating that these in-vitro performance tests are efficacious and promising QC tests. In the clinical study it was found that the resolution scores had a strong positive correlation ($R^2 = 0.79$) with the overall image quality scores; therefore, the in-vitro parameter, lateral resolution may be an important technical parameter for predicting clinical image quality for abdominal imaging. The poor correlations for axial resolution, slice thickness and contrast resolution tests suggest that new test methods may be needed to evaluate these technical parameters, in addition, to some revision of the test protocols being required. It was found that all of the in-vitro colour / power Doppler test parameters demonstrated moderate to strong correlations with the corresponding in-vivo parameters, with the exception of tissue movement suppression test therefore, these new test protocols and test objects appear to be efficacious and particularly promising. These results help to provide a better understanding of the influence the individual in-vivo parameters have on overall image quality, which could help to define minimum diagnostic criteria for abdominal ultrasound, similar to those used in X-ray for the different clinical applications [24]. However, to gain a more complete impression of the clinical perception of ultrasound scanners' image quality performance would necessitate a more extensive clinical trial, ideally incorporating the following: (i) a larger number of observers, ideally with different experiences of ultrasound scanners across a range of manufacturers as a bias may exist towards ultrasound scanners manufactured by the manufacturer whose scanner is used within the department; (ii) a larger number of ultrasound scanners should be included from different manufacturers and of varying complexity; and (iii) a larger number of healthy volunteers with different types of body habitus.

The findings of this study have implications for quality control testing, since the purpose of QC testing is to monitor changes in performance of the different test parameters of the ultrasound scanner over time before they become noticeable to the user. Therefore, the test parameters which were found not to be reflective of clinical perception for a range of scanners of varying complexity, will most probably not be able to detect subtle changes in the scanners performance over time.

REFERENCES

Price R. Routine quality assurance of ultrasound imaging systems. *Institute of Physics and Engineering in Medicine, Report No 71*. 1995;

Hoskins PR, Sherriff SB, and Evans JA. Testing Doppler Ultrasound Equipment. *Institute of Physical Sciences in Medicine, Report No 70*. 1994;

AIUM. Performance criteria and measurements for Doppler ultrasound devices. *American Institute of Ultrasound in Medicine Standards Committee*. 1993;

Donofrio NM, Hanson JA, Hirsch JH, and Moore WE. Investigating the Efficacy of Current Quality Assurance Performance Tests in Diagnostic Ultrasound. *Journal of Clinical Ultrasound*. 1984;12:251-260.

Dudley NJ, Griffith K, Houldsworth G, Holloway M., and Dunn MA. A review of two alternative ultrasound quality assurance programmes. *European Journal of Ultrasound*. 2001;12:233-245.

Browne JE, Watson AJ, Gibson NM, Dudley NJ, and Elliott AT. Objective Measurements of Image Quality. *Ultrasound in Medicine and Biology*. 2004;30:229-237.

Common Services Agency. Comparative evaluation of imaging and Doppler ultrasound systems for examination of the heart. *EEV/91/3*. 1991;1-38.

Metcalfe SC and Evans JA. A Study of the Relationship Between Routine Ultrasound Quality Assurance Parameters and Subjective Operator Image Assessment. *British Journal of Radiology*. 1992;65:570-575.

Meire HB and Farrant P. Liver, pancreas and bilary system. 1995;108-109.

Allan PL, Dubbins PA, Pozniak MA, and McDicken WN. Clinical Doppler Ultrasound. 2000;123-190.

Society of Motion Picture and Televisions Engineers. SMPTE Recommended Practice: Specifications for medical diagnositc imaging test pattern for television monitors and hard-copy recording cameras. *RP 133-1991*. 1991;

Gibson NM, Dudley NJ, and Griffith K. A computerised quality control testing system for B-mode ultrasound. *Ultrasound in Medicine and Biology*. 2001;27:1697-1711.

Browne JE. Diagnostic ultrasound real-time and colour Doppler imaging assessment by in-vivo and in-vitro methods. *PhD Thesis, University of Glasgow*. 2002;

Carpenter DA, Kossoff G, and Griffiths KA. Correction of Distortion in Us Images Caused by Subcutaneous Tissues - Results in Tissue Phantoms and Human-Subjects. *Radiology*. 1995;195:563-567.

Shapiro RS, Wagreich J, Parsons RB, Stancato-Pasik A, Yeh HC, and Lao R. Tissue harmonic imaging sonography: Evaluation of image quality compared with conventional sonography. *American Journal of Roentgenology*. 1998;171:1203-1206.

Tranquart F, Grenier N, Eder V, and Pourcelot L. Clinical use of ultrasound tissue harmonic imaging. *Ultrasound in Medicine and Biology*. 1999;25:889-894.

Wagner RF, Insana M, and Smith SW. Fundamental correlation lengths of cohernet speckle in medical ultrasound images. *Ieee Transactions on Ultrasonics Ferroelectrics and Frequency Control*. 1988;35:34-44.

Whittingham TA. Towards a portable system for visualising and measuring the point spread function of an ultrasound scanner. *The Physics and Technology of Medical Ultrasound-Biennial Meeting*. 2001;

Madsen EL, Zagzebski JA, Macdonald MC, and Frank GR. Ultrasound Focal Lesion Detectability Phantoms. *Medical Physics*. 1991;18:1171-1180.

Skolnick ML. Estimation of Ultrasound Beam Width in the Elevation (Section Thickness) Plane. *Radiology*. 1991;180:286-288.

McCormack S, Evans JA, and Metcalfe SC. Assessing the improvement in contrast detail resolution using tissue harmonic imaging. *Ultrasound Quality Assurance 2002: B-mode, Doppler and New Modalities*. 2002;

Claudon M, Tranquart F, Evans DH, Lefevre F, and Correas JM. Advances in ultrasound. *European Radiology*. 2002;12:7-18.

Whittingham TA. New and future developments in ultrasonic imaging. *British Journal of Radiology*. 1997;70:S119-S132.

Commission of the European Communities. European guidelines and quality criteria for diagnostic radiographic images. *EUR 16260 EN*. 1996;

Table 1: Summary of probes used and modes tested in the clinical trial

Table 1: Summary of Ultrasound Scanners used with associated probes and modes tested

Ultrasound Scanner	Probe (Nominal Frequency)	Modes Tested	Approximate Cost- When Purchased	Age
Acuson 128	V4 (3.5 MHz)	B-mode Colour Doppler	£40,000	12 years
Acuson Aspen	4C1 (3 MHz)	B-mode, THI Colour Doppler, Power Doppler	£80,000	4 years
ATL HDI 5000	C5-2 (2–5 MHz)	B-mode, THI, SonoCT, HSonoCT Colour Doppler, Power Doppler	£100,000	3 years
Aloka SSD 5500	UST-9126 (3 MHz)	B-mode, THI	£90,000	1 month
Aloka SSD 5000	UST-9119 (5 MHz)	B-mode, THI Colour Doppler, Power Doppler	£80,000	2 months
Siemens Sienna	C5-2 (2-5 MHz)	B-mode, THI Colour Doppler, Power Doppler	£45,000	3 months
Hitachi EUB 420	C5-2 (3.5 MHz)	B-mode	£30,000	6 years
SonoSite 180	C5-2 (2–5 MHz)	B-mode Directional Power Doppler, Power Doppler	£25,000	1 year
SonoSite 180Plus	V4 (3.5 MHz)	B-mode, THI Directional Power Doppler, Power Doppler	£20,000	1 month

Table 2: Summary of the correlation coefficients between the B-mode In-Vitro Parameters and the In-vivo Parameters

In-Vitro Parameter	**In-vivo Parameter**	**Correlation Coefficient, r (p value)**	**Correlation Coefficient, r without THI values (p value)**
Axial Resolution (mm)	Tissue Texture Variance	r= 0.003 (not statistically significant, p = 0.99)	r = 0.002 (not statistically significant, p = 0.99)
Lateral Resolution (mm)	Resolution	r = -0.69 (statistically significant, p = 0.003)	r = -0.93 (statistically significant, p = 0.0001)
Slice Thickness (mm)	Overall Image Quality	r = -0.13 (not statistically significant, p = 0.62)	r = -0.21 (not statistically significant, p = 0.56)
Sensitivity	Noise	r = 0.66 (statistically significant, p = 0.005)	r = 0.5 (not statistically significant, p = 0.14)
Contrast Resolution	Dynamic Range	r = 0.45 (statistically significant, p = 0.008)	r = 0.69 (statistically significant, p = 0.003)
Penetration Depth (mm)	Penetration Depth	r = 0.56 (statistically significant, p = 0.003)	r = 0.53 (statistically significant, p = 0.09)

Table 3: Summary of the correlation coefficients between the Colour Doppler In-vitro Parameters and the In-vivo Parameters

In-vitro Parameter	**In-vivo Parameter**	**Correlation Coefficient, r (p value)**
Temporal Resolution (second)	Temporal Resolution	r= 0.16 (not statistically significant, p = 0.79)
Spatial Resolution (Lateral) (mm)	Resolution	r = -0.55 (statistically significant, p = 0.05)
Spatial Resolution (Axial) (mm)	Overall Image Quality	r = -0.64 (not statistically significant, p = 0.02)
Sensitivity	Clinical Sensitivity	r = 0.52 (statistically significant, p = 0.07)
Clutter	Clinical Clutter	r = -0.47 (not statistically significant, p = 0.15)

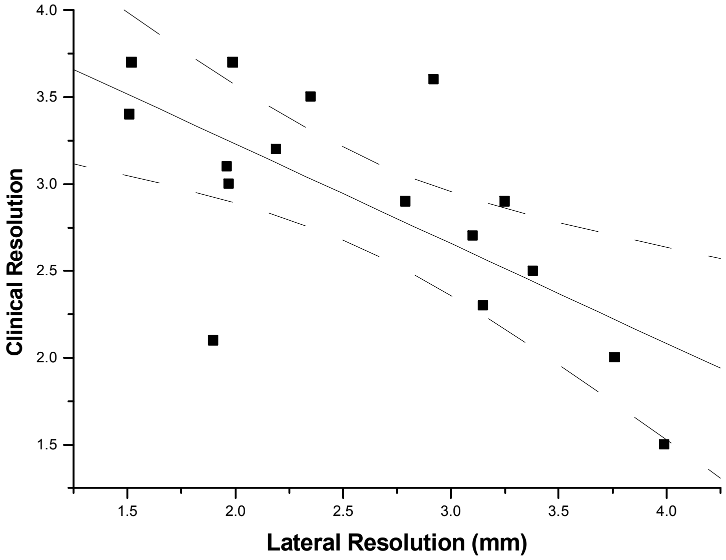

Figure 1a: ***Scatter Plot of Lateral Resolution versus Resolution (with 95 % confidence interval).***

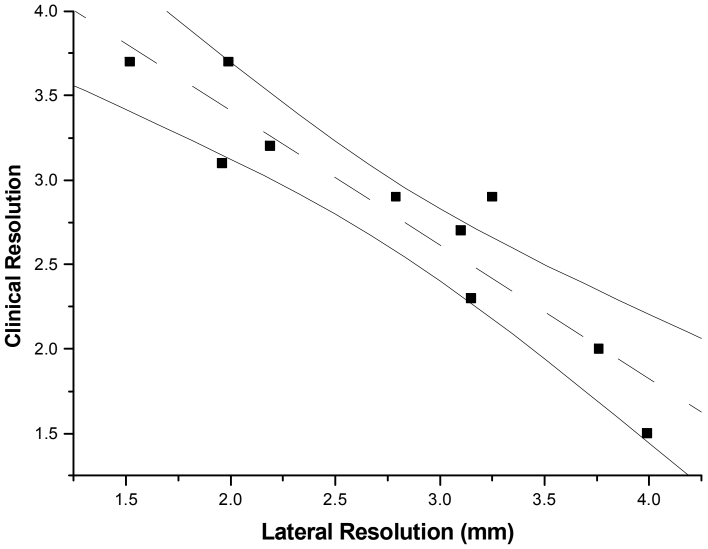

Figure 1b: ***Scatter Plot of Lateral Resolution versus Resolution, without the THI results (with 95 % confidence interval).***

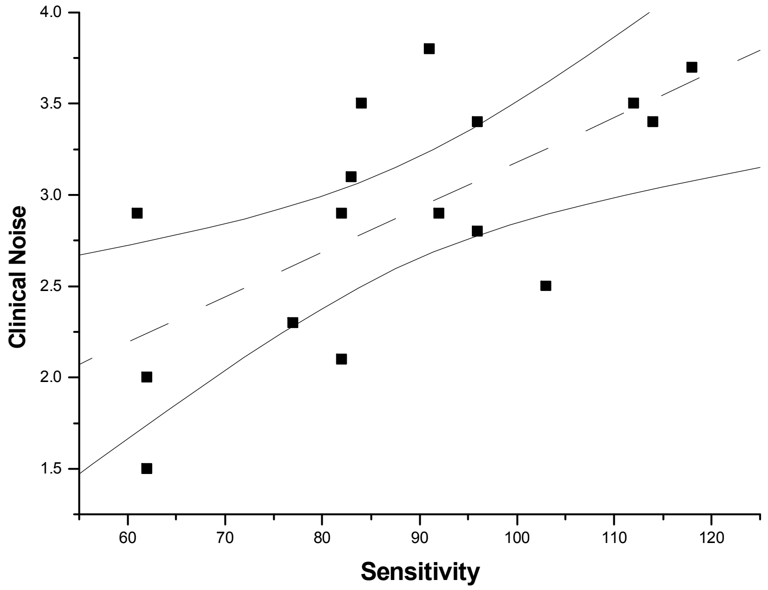

Figure 2: ***Scatter Plot of Sensitivity Index versus Noise (with 95 % confidence interval).***

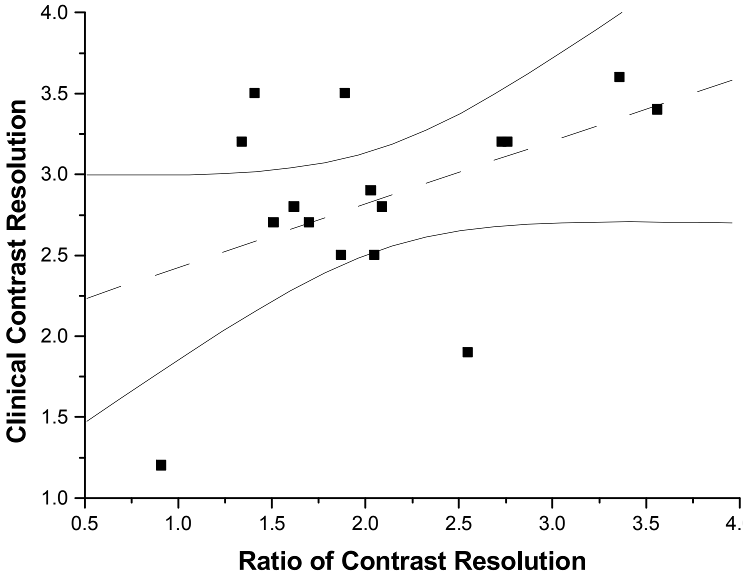

Figure 3: ***Scatter Plot of Contrast Range versus Dynamic Range (with 95 % confidence interval).***

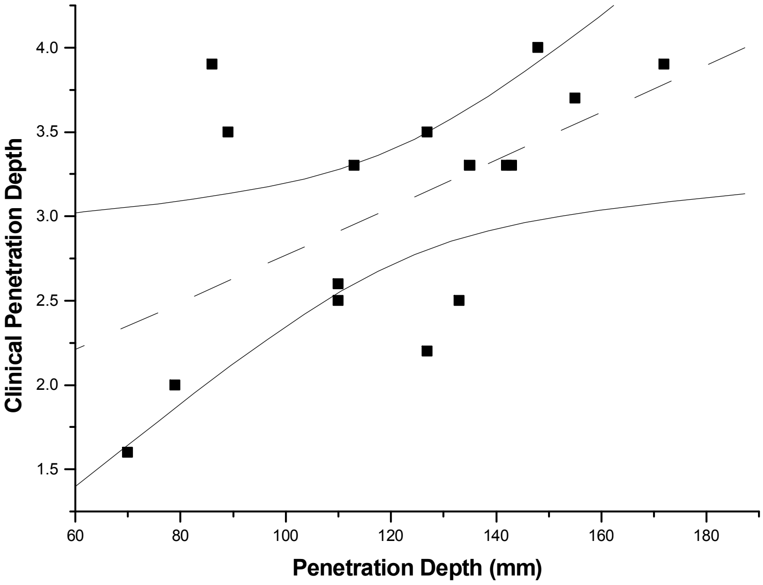

Figure 4: ***Scatter Plot of Penetration Depth versus Penetration Depth (with 95 % confidence interval).***

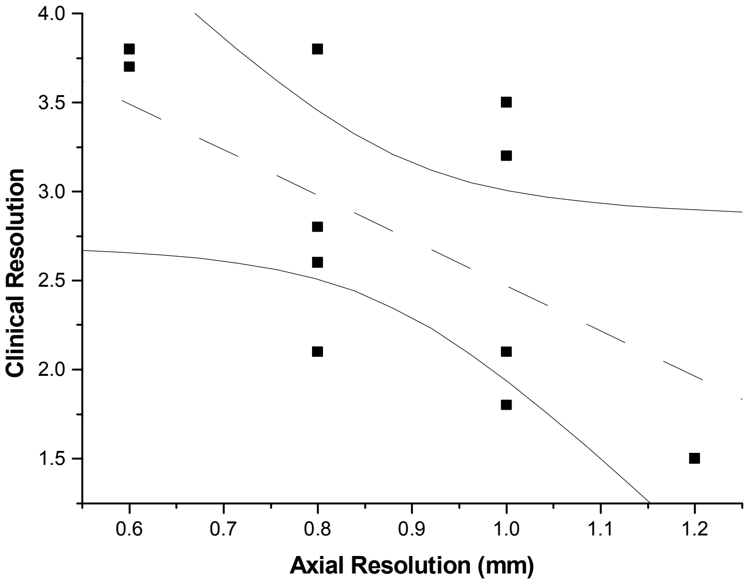

Figure 5: ***Scatter Plot of Colour Doppler Axial Resolution versus Resolution (with 95 % confidence interval).***

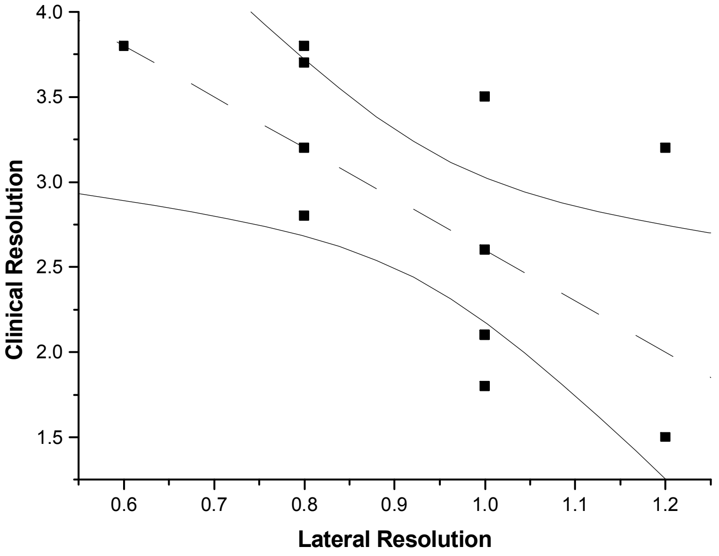

Figure 6: ***Scatter Plot of Lateral Resolution versus Resolution (with 95 % confidence interval).***

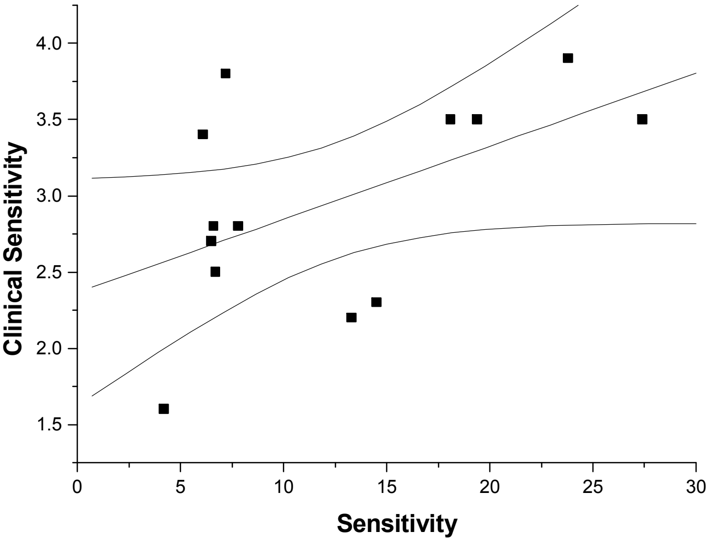

Figure 7: ***Scatter Plot of Sensitivity Performance Index versus Overall Image Quality (with 95 % confidence interval).***

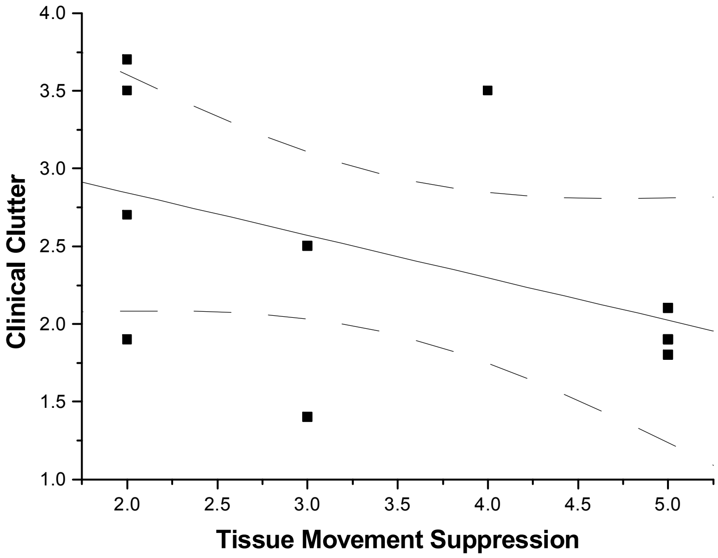

Figure 8: ***Scatter Plot of Tissue Movement Suppression Performance versus Clutter Suppression Performance (with 95 % confidence interval).***