NPL
National Physical Laboratory

Report

NPL REPORT AC 2

Standards to support performance evaluation for diagnostic ultrasound imaging equipment

Adam Shaw and
Rob Hekkenberg

NOT RESTRICTED

October 2007

National Physical Laboratory | Hampton Road | Teddington | Middlesex | United Kingdom | TW11 0LW

Switchboard 020 8977 3222 | **NPL Helpline** 020 8943 6880 | **Fax** 020 8943 6458 | **www.npl.co.uk**

Standards to support performance evaluation for diagnostic ultrasound imaging equipment

Adam Shaw and Rob Hekkenberg*
Quality of Life Division, NPL, Teddington, United Kingdom
* TNO-Quality of Life, Leiden, Netherlands

ABSTRACT

The National Physical Laboratory has been asked by the National Measurement System Policy Unit (NMSPU) of the UK Department for Innovation, Universities and Skills (DIUS) to explore whether there are ways in which the NMSPU Acoustics and Ionising Radiation programme can contribute to better or more useful performance evaluation methods for ultrasound. This report includes the findings of a survey of hospital departments and users, and makes recommendations for further research and development which can be followed up in the NMSPU programme either at NPL or elsewhere.

Four topics were identified as being of highest priority:

1. Traceability of measurements using phantoms (including the establishment of reference measurement capability for absorption coefficient, backscatter coefficient and nonlinearity parameter for tissue mimics used in imaging phantoms and their variation with frequency and temperature, to augment existing capabilities for measurement of speed of sound and attenuation coefficient); the development of a method for monitoring changes in the acoustic properties of intact phantoms (where the use of reference measurement methods would not be appropriate as they would require disassembly of the phantom); and the development of a system for measuring changes in dimensions and wire separations within intact phantoms.
2. The evaluation of more general imaging metrics such as the Point Spread Function, Modulation Transfer Function, noise power spectrum and contrast-to-noise ratio which are widely used in other imaging modalities.
3. The development of scatter-free phantom for the determination point spread function throughout the imaging plane.
4. A benchmarking exercise to establish in detail how much time and money is spent on ultrasound QA and performance assessment across the UK.

A further seven topics were identified as important but of lower priority either because they are under the control of manufacturers or because they naturally follow on from one of the highest priority topics.

ISSN 1754-2936

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

Approved on behalf of the Managing Director, NPL
by Martyn Sené, Director, Division of Quality of Life

CONTENTS

1 Introduction

1.1 Background to the project

The National Physical Laboratory has been asked by the National Measurement System Policy Unit (NMSPU) of the DIUS to explore whether there are ways in which the NMSPU Acoustics and Ionising Radiation programme can contribute to better or more useful performance evaluation methods. The aims of the project are to undertake a survey of hospital departments and users, produce a report regarding the requirements for standards for the performance of medical diagnostic ultrasound equipment, and to make recommendations for further work which could followed up within the NMS Acoustics and Ionising Radiation programme either at NPL or elsewhere. Since the Institute for Physics and Engineering in Medicine (IPEM) has already commissioned a replacement for the widely-used Report 71 (which gives practical guidelines for performance and Quality Assurance (QA) tests, there is no point in duplicating this effort, so our project will concentrate on looking ahead: defining requirements that would make a new generation of tests more clinically useful for modern scanners; and defining areas where basic metrics for performance are required, particularly where these can be assimilated into the International Standards..

QA and performance assessment of diagnostic ultrasound scanners has, in principle, a number of functions. It may be used to ensure that equipment performs to an agreed standard and is fit-for-purpose; to monitor changes in performance over time; to evaluate new imaging and processing techniques; to assist in the procurement and replacement process; and in the development and improvement of ultrasound transducers. However, objective technical evaluations of grey-scale ultrasound have always been limited by the fact that measurements of spatial and contrast resolution are made with test objects. These may be useful as consistency checks, but do not seem able to compare the image quality of different scanners, and seem to have very little correlation to perceived clinical performance. In the UK only breast screening has any formal requirements for the technical performance of scanning equipment [[1]]. The IPEM meeting in York (Quality Assurance of ultrasound scanners, 9 March 2005)[[2]] brought together an excellent group of participants and highlighted a general dissatisfaction with current practice.

Many different terms are used, often in conflicting or inconsistent ways. For clarity, the authors of this report consider three of the terms often used to have the meanings defined below and we have tried to use them consistently. However, where we quote other authors or replies given in response to the questionnaire described, these terms may be used in different ways.

Imaging performance evaluation or performance assessment: the quantitative evaluation of different aspects of the ability of an ultrasound system to generate an image. Performance evaluation does not mean the determination of a **single** 'figure of merit' which allows a series of systems to be ranked in order of overall performance. One system may perform better on some aspects (for instance lateral resolution) of image generation than another system, but may be less good on other aspects (for instance contrast resolution).

Image quality: this is a purely descriptive term. It has many aspects that may range from aesthetic and subjective factors to those that are capable of precise physical measurement. It may not be evaluated quantitatively.

Quality Assurance: the activity of checking, in a way which is demonstrable to customers or other stakeholders, that equipment meets needs, expectations, or other specified requirements. Imaging performance evaluation may be part of the QA process but QA contains many aspects not directly related to imaging, and QA may also take place without any reference to quantified evaluation of imaging performance (although, presumably, basic tests of imaging function will normally be carried out).

According to Dr. Stan Barnett, former president of the World Federation of Ultrasound in Medicine and Biology (WFUMB), in his lecture on QA during the conference of Advanced Metrology in

Ultrasound [3], QA is vital for the attainment of universally high standards of practice. Technical advances in equipment design and functionality have also created a notion about their ease of use to the stage where ultrasonographic imaging devices may be regarded as "diagnostic stethoscopes". It is often suggested that the greatest risk to the patient is that of misdiagnosis: this emphasises the need for standardised assessment of imaging ability and the use of QA procedures. The Medical Devices Directive [4] extends to the accuracy of indication of medical devices, such as obstetric scanners, used to determine quantities such as the femur length and bi-parietal diameter. In principle, the accuracy of these distance determinations may be assessed using test-objects. Within the NHS, there is an increasing requirement to demonstrate best practice and to accurately demonstrate the technical performance of a clinical system.

1.2 Scope and outline

The project was carried out in association with TNO in Leiden, The Netherlands, who acted as a subcontractor.

The scope of the project agreed with TNO was that it would:

- Concentrate on 'performance assessment' (including the establishment of a chain of 'traceability' from basic standards down to the end user) and not QA (which we consider to contain aspects not directly related to performance, and to be more concerned with identifying changes in the ultrasound system).
- Address the ability of systems to provide positional information (mostly associated with pulse-echo grey-scale imaging) and not velocity information (eg from spectral, colour-flow or power Doppler modes). However, the survey should include some questions related to the need for specific Doppler assessment methods.
- Include the assessment of 3-D volume imaging systems.
- Include the assessment of measurement functions for length, area and volume.
- Include the ability to image (the position of) moving objects.
- Both the 'quality' of the 'raw' image (eg by frame grabbing or DICOM) and of the image presented to the user (through a display unit in an illuminated room) are relevant. Ideally, assessment methods could be used on either of these.
- Address end-users of the assessment techniques – considered to be primarily hospital medical physicists and development labs of manufacturers (not the clinical end users). Different sets of questions may be used.
- Include non-UK participants but emphasis should be on UK.
- Include visits to a small number (3 to 5) of experts prior to finalising the questionnaire to discuss issues in detail and ensure the questions will draw out the information we seek.
- Involve the organisation of an invited workshop for further discussion prior to finalising our conclusions and recommendations.

The workplan was divided into a number of specific tasks:

- Design of questionnaire.
- Trial questionnaire on selected candidates and feed observations into a refinement of the questionnaire.
- Identify list of contacts to approach.
- Undertaken search of relevant literature, standards etc.
- Complete questionnaire process.
- Compile and analyse results.
- Complete draft of report, including recommendations.
- Organise and hold workshop to discuss draft report with selected experts.
- Prepare and publish final report.

1.3 Organisation of the survey

A survey about the needs and shortcomings concerning the image performance in the field of diagnostic ultrasound cannot be carried out without interviewing those who are actually performing these kinds of tests. A draft set of questions were produced and circulated to a group of experts at five UK hospitals. The group included senior members of the medical physics community (who might be considered also have an academic interest in image performance evaluation) and also less experienced members who were concerned more with the delivery of a service. They were:

Andy Coleman	Guy's and St Thomas's NHS Trust, London
Tony Evans	University of Leeds
Stephen Pye	Edinburgh Royal Infirmary
Stephen Russell	Christie Hospital NHS Trust, Manchester
Kate Wells and Paul Williams	University Hospital of Wales, Cardiff

We arranged to visit each of this group to gather feedback on the draft questions and to discuss image performance evaluation more generally. Following these visits, we decided to split the questionnaire into two separate parts which would be posted separately on a commercial survey website (www.surveymonkey.com). The reason for splitting the survey was to ensure maximum participation in the more straightforward set of questions relating to current practice. We considered that some of the questions in the second set related to possible improvements may be off-putting to some participants. Instead we decided to publicise the first part of the questionnaire widely to gain maximum participation and then invite only those people who had completed Part 1, to complete Part 2.

Part 1 of the survey was publicised by direct email to participants in the 2005 IPEM meeting in York and to people on the NPL Ultrasonics mailing list; it was also posted on the NPL Acoustics website and advertised in the NPL Health Matters newsletter. In addition, it was advertised to all members of the British Medical Ultrasound Society through a flyer placed in their Bulletin. All questions in Part 1 were mandatory and we received 104 completed replies, which was a very pleasing response and exceeded our expectations. The full responses to Part 1 can be found via links at www.npl.co.uk/acoustics.

A breakdown of the respondents to Part 1 is shown below. All of these respondents were invited to complete Part 2: these questions were mostly optional, so respondents could omit those where they were unsure or had no opinion. We received 48 completed replies to Part 2: this was also a very pleasing response rate. The full responses to Part 2 can also be found via links at www.npl.co.uk/acoustics.

1.4 Professional background of the respondents

Of the 104 responses to Part 1 of the questionnaire, 53,8 % were from people based in the UK and 46,2 % from persons abroad. The Table 1 show the distribution of respondents over several countries.

Table 1. Distribution of respondents over several countries

	%		%		%
- UK:	**53,8**	- Ireland:	2,9	- Greece:	1
- United States:	17,3	- Italy:	2,9	- India:	1
- Czech Republic:	5,8	- Netherlands:	1,9	- Japan:	1
- China:	4,8	- Korea, South:	1,9	- Mexico:	1
- Germany:	2,9	- Albania:	1	- Spain:	1

The majority of the respondents in the UK who completed the questionnaire were clinical scientist (hospital physicists), see Table 2, where as for the other countries a relative large number is

occupied in research and/or education. In the UK the response from industry was much lower compared to that for the other countries. This might be expected as there is a limited medical ultrasound industry in the UK.

Table 2. The profession of the respondents

	% In the UK	% In the other countries
Clinical Scientist (hospital physicist)	**62,5**	22,9
Clinical technologist (technician)	5,4	4,2
Manufacturer	7,1	22,9
Clinical user	12,5	2
Other	12,5	**47,9**

80 % of the respondents were already involved in the ultrasound area for more than 10 years, see Table 3. And 61% of the UK respondents spend 50% or more of their time on ultrasound, see Figure 1.

Table 3. Involvement in ultrasound

	In the UK	In the other countries
less than 2 years	1,8	10,4
2 to 5 years	5,4	4,2
5 to 10 years	10,7	10,4
more than 10 years	**82,1**	**75**

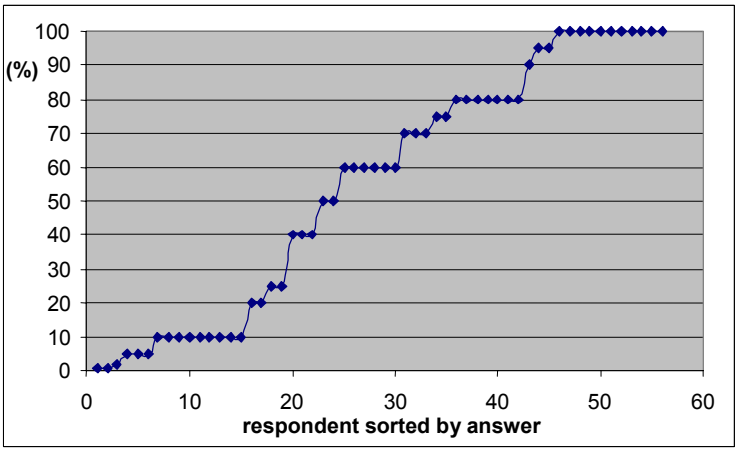

Figure 1. Fraction of time spend by UK respondents.

The result also shows a relative large involvement in standards development. That should imply that respondents are reasonably familiar with the "testing" possibilities. The involvement in standards development of the respondents outside the UK is much larger (in the UK: 46%, outside UK: 73%), but this may simply reflect the fact that the non-UK audience were in large part those people already in contact with NPL: often this contact arises through international committee activity and so the group is likely to be heavily biased towards those involved with Standardisation.

2 Imaging performance

2.1 Introduction

Not everyone defines performance in the same way. The WHO defines it as “The level of attainment of a goal in comparison to a given effort”. Here, goal is defined as: “A general objective towards which to strive. Unlike objectives and targets, goals are not constrained by time or existing resources, nor are they necessarily attainable but are rather an ultimate desired state towards which actions and resources are directed.” The dictionary of Webster defines performance in a number of ways, some of which are: “the execution of an action”, “the fulfilment of a claim, promise, or request”, “the manner in which a mechanism is able to carry out an action or pattern of behaviour”.

Often image performance is seen as a part of the whole system performance. E.g. in [5] equipment performance is divided into three related groups: (a) system sensitivity and dynamic range, (b) spatial resolution, and (c) geometric accuracy. In the IPEM report [6] QA is viewed as a procedure to establish that a scanner is set up correctly, and performs to agreed standards.

Literature [7] uses the term "quality assurance" to refer to rapidly accomplished routine monitoring of clinical diagnostic ultrasound equipment to assure that imaging performance of the unit is optimal. It has been divided into four categories: (1) daily maintenance, (2) detecting signs of degradation in human or phantom images, (3) periodic QA tests using phantoms and (4) documentation.

Literature [8] reads: “testing of the performance from the user’s point of view. In case of medical ultrasound, this may be summarized by imaging performance”.

In literature [9] one can find: “The efficiency for performing important clinical examinations is the ultimate assessment of image quality”.

Literature [10] is more descriptive: it lists: “The performance assessment and quality assurance of diagnostic ultrasound scanners have several roles: to ensure that equipment performs to an agreed standard and is fit-for-purpose; to assess new imaging modalities and signal processing techniques; to inform decision making in the procurement and replacement of equipment; to assist in the design of ultrasound transducers”.

The particular performance characteristics that are required for different imaging applications will in general be different. The ICRU report 54 [11] goes into more detail concerning specifically imaging. It takes the position that imaging system assessment depends on the task for which the system is intended and can, therefore, be cast in terms of task performance. This performance can be measured in two stages. First, task performance can be measured in terms of how well an ideal or Bayesian decision maker would perform the task using only the acquired data, i.e., before it is presented as an image to a human observer. The second stage involves measurement of the performance of the task by observers using displayed data.

Two key imaging metrics are the point spread function (PSF) which describes the response of an imaging system to a point source or point object. Another commonly used term for the PSF is a system's impulse response. A related parameter is the optical transfer function (OTF) which describes the spatial (angular) variation as a function of spatial (angular) frequency and is essentially the Fourier Transform of the PSF. OTF may be broken down into magnitude and phase components: the modulation transfer function (MTF) and the phase transfer function. The ICRU report [11] describes these metrics and also a number of different conceptual ‘observers’

Although the scope of this report is restricted to positional (grey-scale) imaging, a short comment on velocity (Doppler) imaging is worthwhile. Doppler ultrasound differs from ultrasound imaging in that the product of the examination is often a measurement, albeit an indirect one, of a physiological quantity. Whereas the "quality" of an image has many aspects that may range from aesthetic - and wholly subjective- factors to those that are capable of precise physical measurement, a clinical

Doppler examination is frequently directed toward a well defined question concerning blood flow [12]. However, good grey-scale imaging performance is also essential for all types of cardiovascular ultrasound and is considered, for instance, in IPSM report 70 [13]. Proposals for tests, however, were beyond its remit.

2.2 The survey

The survey contained a number of questions related to these important underlying questions about image quality and performance evaluation. The topics and answers are summarised in the subsections below; full answers are listed in Annex A. In each case, we provided a short introduction (shown in **bold**) to the topic and asked for opinions.

2.2.1 What defines imaging performance?

There are widespread opinions on what constitutes 'good' imaging performance for medical applications. Some people consider it should be assessed relative to technical issues such as resolution, geometric accuracy, and contrast; others consider that it is an image the clinician is 'comfortable' with; others again consider it to be the image that is most successful at diagnosing a particular condition.

From the responses we summarise that, although there seems not to be a clear correlation between images from test objects and the clinical situation, the majority of the respondents find that the image performance of a scanner best can be evaluated using rather technical indicators. There were dissenting voices: for example, "*the problem is that most clinicians seem to show a very subjective, qualitative, view on what kind of image is the best for them to work with*". Objectivity and reproducibility was an important consideration, especially for QA purposes. Interestingly, many replies proposed parameters related to visibility of objects against a background (like "lesion-signal-to-noise-ratio", contrast ratio) rather than to spatial resolution. Other views are the importance of SNR when it comes to the ability to analyse definable structures like cysts; the Edinburgh pipe phantom was preferred by some respondents.

All agree that good imaging performance is related to the success of diagnosing a particular condition and that work towards finding the relationship between the two would be very helpful. But as said earlier in general it was believed that system characteristics (i.e. technical parameters such as resolution and contrast) provide a bottom line for the image performance.

This issue will be discussed further in Section 6.

2.2.2 Who should define imaging performance?

It often seems that different categories of users (eg sonographers, physicists, manufacturers, Royal Colleges, regulators..) have very different views on how to define imaging performance. The question was raised whether different categories of user have different views and, if so, why?

A few of the respondents were of the opinion that the only category that is important are the clinical users of ultrasound; that physicists have become too much involved in the numbers game; and that the devices are developed to perform diagnosis. The majority of the respondents were, however, not of that opinion: "*As diagnostic accuracy is the ultimate goal, defining the image performance should be a team work of everyone involved, eg, engineers, physicists, sonographers, radiologists etc. Each group has its different perspective. That's why developing a reproducible and more quantitative Quality Control test approach is important to provide a bottom line*".

Some of the comments give the impression that a number of the measurements performed to evaluate the image performance are based too much on tradition and that the methods should be replaced by more "realistic" measures that take sidelobes and the 3-dimensional behaviour into account. One comment separates the performance criteria into two parts: "*first, a set of physical quantities to define quantitative criteria - this should be done by technical staff; in addition, more physiological tests would be helpful, for example structures of organs that should be recognized from medical staff or a computer program. So, both should be incorporated"*.

This issue will be discussed further in Section 6.

2.2.3 Should performance be defined as application specific?

There is a wide range of clinical applications for ultrasound imaging (obstetric, general abdominal, cardiovascular, peripheral vascular, skin, musculoskeletal, …) where the size and nature of structures to be identified or examined vary greatly. In the poll it was questioned whether it would be possible to incorporate most of these applications within a single general scheme for assessing imaging performance, or are the requirements for each application so different that they should be considered separately.

As a general response it was thought that the vast majority of soft tissue imaging applications could be handled within a single general scheme for assessing imaging performance. One research group claims to have demonstrated the feasibility of achieving this [[10]]. Albeit that for different applications and frequencies, different tolerances are to be allowed. For QA purposes it is suggested that a single scheme could be devised: "*For assessment of clinical performance relative to other scanners, it is actually possible to devise a scheme, it may well be different for some (but not all) applications, e.g. obstetrics. Also, for instance, Doppler performance and B-mode performance are not necessarily related"*. A number of respondents drew attention to the cost aspects of performance evaluation.

2.2.4 Computational or human observer?

Since ultrasound scanners are used by human beings, many people believe that evaluation of imaging performance should be carried out by a subjective human observer. Others believe that a 'computational observer' or a computer analysis programme will provide more objective and consistent measures of performance. In practice it may be that even an ideal 'computational observer' will require human input, for instance to select a region of interest for analysis. The questionnaire invited comment on the preference for human or computerised evaluation.

There was no clear preference for either one or the other type of observer. As could be expected the opinion is quite diverse, from "*The human observer is introducing bias and noise into the performance measurement and should NOT be part of it*" to "*Human evaluation of course. There are other things of importance to the clinical users than the quality of the image*". It was generally agreed that a computational observer (referred to as computerised evaluation system) will present more reproducible results. This is especially useful when it concerns QA. None of the respondents offered a proposal on how to remove the component of subjectivity so that results of human observers could be compared more easily.

2.2.5 Is imaging performance linked to safety?

The IEC series of 60601 standards deal with "basic safety and essential performance" of medical equipment. Imaging equipment is intended to provide images of the patient. The ability of the equipment to produce such images at a specified quality is considered to be performance. The decision whether such performance is "essential performance" may be strongly related to the question of how the images are used and whether a degradation or

absence of performance would cause an "unacceptable risk". It is useful to know whether adequate imaging performance should be considered a safety issue for ultrasound.

A large majority of respondents stated that image quality was linked to safety. Several different reasons were referred to in the responses. First, deteriorating image quality introduces risk of misdiagnosis. Secondly, it's been noticed that improved performance of the scanners is quite often achieved by increasing output power, which again is related to safety. If the image performance is inadequate a higher percentage of false-negative outcomes than would otherwise be the case will be the result. Also it may make the physician refer the patient on to a more risky study, now exposing the patient to a risk to which they would not otherwise have not been exposed. This can lead to more invasive tests which may result in "unacceptable risk". It was also remarked that the examiner performance is the important feature and the instrument performance is not so important.

2.2.6 What are the benefits of QA testing?

Some people argue that routine QA tests of ultrasound scanners identifies so few faults which are not obvious to the user that they are not worth doing. They may also feel that a better or more relevant set of tests would add value. Others claim that testing already significantly benefits the organisation or the patients either economically or in other ways. Now the question is: is QA testing for ultrasound imaging worthwhile, and what value is added in this area?

A large majority were of the opinion that image performance testing has a substantial benefit for the patient and financially for the hospital. Without identifying why, one respondent argued that all of us want regular maintenance on our car, but are questioning maintenance on a £100k ultrasound scanner. The responses in general feel that performance testing is beneficial to the lifetime and proper functioning of the device. Due to performance testing, problems were found with crystal dropout or decoupling of lens and damage to the cables where no action has been taken until reported. Also ghosting was found in one of the probes which had already been signed off by the supplier as the image looked ok clinically. Someone came across brand new scanners with faulty measurement callipers. A relatively large set of occurrences can be found in Section 5.1.2. Another respondent stated that "*QA testing of ultrasound systems was essential because it sets clear reference points as windows between which the machine is deemed to have been operating normally: it also allows for checking the results of up-grading and repair, as well as for deciding on the acquisition of new equipment*".

Some respondents argue that traditional manual/visual testing with phantoms has no benefit, and that the testing should be replaced by more adequate methods. Another respondent argues that "*if people doing ultrasound QA aren't finding problems with the scanners that they look after, they're not doing the measurements properly!*". In their opinion, it is very important that users have access to NHS-based Medical Physics Departments with ultrasound expertise.

3 Current practice in ultrasound

3.1 Interviews with selected UK experts

Before sending out the survey questionnaire to the departments and users, the views of a small group of UK experts in the field of QA were gathered. This was done by visiting and interviewing these experts at their place of work. Apart from the discussion on the appropriateness of the questions posed in the first draft of the questionnaire, the discussions were expanded to cover their views on current image performance testing for ultrasound and how this could be improved in future if needed.

Some of the points related to current practice raised in the interviews are listed briefly below. Points related to the role of the Medical Physics department in each centre are listed first.

First centre
- Act as firstline check for maintenance company.
- Provide general central support, but maintenance contracted out.
- Within region, different scanners on different level of control (no control/med phys/company/medphys+company/3rd party).
- Acceptance testing carried out against what would be expected subjectively from similar equipment.
- Departments are advised to leave one preset unaltered as a control.
- Weekly checks by sonographers with log kept.
- Equipment tested with static phantoms/ whereas clinically target is moving.
- Doppler tested with string phantoms.

Second centre
- Act as first line tester with maintenance carried out by Toshiba.
- Responsible for 4-5 hospitals within the Trust and others outside on a contracted basis.
- Processing functions often source of image problem so can't see problem with these turned off.
- Not following Report 71 strictly – eg use clinical settings.
- No routine Doppler testing.
- Harmonic imaging not tested.
- Mostly measurements taken off frozen image.

Third centre
- Medical Physics does not act as firstline contractor.
- Use KeyMed for maintenance.
- Don't do phantom work any more – experience showed that reports were not acted on.
- Only test electrical safety on purchase.
- Wide userbase who manage their own equipment – no universal QA scheme.
- In-house QA should target critical equipment and older equipment only.
- Equipment inventory is splitting into high-end and low-end.
- Range of cost of equipment is very wide compared to CT, MR etc which are all expensive.
- Ultrasound scanners are widely spread – but all x-ray equipment in one place.

Fourth centre
- Medical Physics is not generally firstline contractor.
- Responsible for any replacement and new purchases in the Region – this is a recent unified policy.
- Assess new equipment prior to purchase and most scanners acceptance tested during first year.
- Specification for all new equipment includes minimum acceptable performance requirement based on the Resolution Integral.
- Scanners mostly on preventive maintenance contract with manufacturers.

- No routine assessment – stopped due to insufficient benefit.
- Evaluation after software changes is important.
- Carry out 'twilight' testing prior to taking equipment out of service.

Fifth centre

- Medical Physics provide firstline maintenance contract.
- Provide service to city hospitals plus other major acute hospitals in the region.
- Also contracted to provide component of the breast screening programme over a wider area.
- Rarely carry out pre-purchase evaluation.
- Provide comprehensive test after acquisition (including output measurements) before scanner is used.
- Thereafter, try to check imaging every 6 months for each machine.
- Most equipment on manufacturer maintenance contract for 6-7 years.
- Recently, management of the maintenance contracts has passed to Medical Physics, but each department pays separately.
- Acoustic output measurements have identified some errors in Thermal Index display.
- Equipment is generally reliable and the evidence base for the economic benefit of QA testing is weak.
- Do not know how to relate QA measurements to clinical performance and the most detailed image does not necessarily give the best diagnosis.
- Different pulse-shaping affects ability to detect different structures.
- Properties of phantoms are important but not well known. For instance, incorrect speed of sound leads to poor focusing and lower resolution.
- The balance between attenuation and absorption in a phantom is important.
- Time spent on QA is worthwhile because it promotes interaction with departments, keeps manufacturers more alert and unearths problems; but the specific tests carried out are of questionable valuable.
- Users are required to carry out basic tests in the breast screening programme.

3.2 International publications on performance testing

In the last ten years, a relatively large number of papers concerning the performance testing of ultrasound diagnostic equipment have been published [8,14,]. Most of these papers show the ability of the methods preferred by the authors to provide information about the relative performance of the equipment. Some studies show the benefit of a computational observer with respect to objectivity and quick response time [9,15,16]. One study [17] observed how compromises in imaging system properties that affect overall image quality, affect low-contrast detectability most strongly. For this study it was important to characterize each system configuration in terms of contrast, resolution, and noise; however, no attempt was made to optimize the system for each scan head. The time in which a performance test of the image can be carried out seems very important. Some studies focus on a dominant parameter, some times a new one, that characterises the overall image performance [10,18]. During a workshop on ultrasound physics and quality control Lu presented information where QA had identified deficiencies [19] and they summarised the occurrence rate of the following problems: 20,5% probe; 17,9% image display and hard copy; 7,7% software; and 26,5% image quality. They concluded that a thorough visual inspection and a simple phantom scan could identify many deficiencies.

3.3 Standards and guidelines

Most ultrasound societies have prepared a set of guidelines concerning the determination of the image quality performance of diagnostic imaging equipment [5,6,7,20,21,22,23,24]: introductory extracts from the most important documents are given in the boxes below. Usually these guidelines are based on well established measurement methods. Basically, all listed standards use performance testing methods which originated about 10 years ago. Although the methods and procedures described are

still useful, they also do not distinguish very much between image modalities or the intended application of the device to be tested. There are also standards that just describe methods to evaluate the display performance [25,26,27]. Some investigators believe that the performance of the display should be tested independently of the ultrasound performance. The International Electrotechnical Commission (IEC), the organisation responsible for standardisation in ultrasound, has only published two methods for calibration specific parameters [28,29], one of them still being a draft. A new proposal concerns a very interesting, automated, testing method based on visualisation of voids of different sizes against a scattering background [30]. Its intention is to be restricted to the aspect of long-term reproducibility of testing results for essential imaging parameters, like noise level, echo contrast resolution, and ultrasonic beam characteristics. The method defined in the draft is based on Satrapa [31,32]. There exists a document [33] that covers essential image parameters for MRI equipment. It measures parameters like slice thickness, spatial resolution, ghosting artefacts, geometric distortion, uniformity and signal to noise ratio. Basically this standard does not define new ideas that can be used in ultrasound, but it is interesting that a set of simple 'engineering' parameters is used.

Extracts from the introduction in the original 1995 version of IPEM Report 71 [6].

Quality assurance is viewed here as a procedure to establish that a scanner is set up correctly, and performs to agreed standards. It does not include the correction of defects; this should be part of a separate procedure, carried out by the maintenance team. Some of the procedures described require the person carrying out the tests to make accurate measurements. This will require some familiarity with scanner controls, particularly the TGC settings and any software options which affect pre- and post-processing. Most scanners use tracker balls, joysticks or mice to position the screen cursors, and this exhibits a range of characteristics. Again, it will be necessary for the tester to attain same dexterity in the use of the pointing device of each individual scanner to be tested.

Three levels of testing are described in this report:
1. User tests - to be carried out at frequent intervals by the usual operator of the equipment. It is unlikely that most users would be prepared to carry out tests which were perceived as time-consuming, difficult or pointless. The activities at this level are therefore simple tests, requiring a minimum of equipment and consuming as little time as possible while relating directly to the aspects of scanner function on which users depend for clinically meaningful results.
2. Routine quality assurance tests - to be carried out by a third party, normally a member of the Medical Physics department but possibly a service engineer. They are designed to combine the requirement for a meaningful set of tests with the limitations on resources and scanner access time which the survey indicates is common experiences. Again, the aim has been to limit the tests to those which have relevance to normal use of the scanner and which are likely to detect deterioration in performance.
3. Baseline tests - also intended normally to be carried out by a member of the Medical Physics department. These include all the tests specified for routine quality assurance, for which they are intended to establish baseline readings. There are also some extra measurements which are intended to probe more deeply into the function of the scanner.
A series of tests for each level is recommended, but sufficient information is included in each Rationale section to enable quality assurance operators at each level to select their own set of tests if they so wish. If there are variations in the depth of the discussions in these sections, they are intended to reflect the variations in difficulty of understanding the essential background information.

1.2 Schedule
It is recommended that tests be carried out at the following intervals:
1. User testing - should be carried out at intervals of one to four weeks. For scanners with a variety of probes, successive tests should be carried out using different probes in rotation, so that each probe is tested at least every four weeks.
2. Routine quality assurance testing - should be carried out ideally every six months but at least every 12 months.
3. Baseline testing - should be done at acceptance, and whenever a new probe or major hardware or software upgrade is added (it may be argued that all upgrades are major, in that they are capable of affecting performance). These tests may also be useful in the selection of equipment.

Tables in the report give lists of tests and setting-up procedures for each level of testing, with references to the sections which give details of the testing procedure. Note that in several cases the order in which the tests are

performed is important (e.g. measurement accuracy should be checked before using the callipers to measure resolution). The order given in the report avoids such problems, but is obviously not a unique solution.

Extracts from the introduction in the Report of the American Association of Physics in Medicine (AAPM) on Ultrasound quality control [21], 1998.

WHY ULTRASOUND QUALITY CONTROL?

It is sometimes argued that there is no need for ultrasound (US) quality control (QC) testing because (1) the new machines are very reliable and rarely break down, and (2) the sonographer will detect image quality defects during normal scanning. Although both of these statements may be true, they do not necessarily negate the utility of US QC tests. A primary reason is that a set of periodic definitive measurements for each transducer and US unit can identify degradation in image quality before it affects patient scans. Another is that when equipment malfunction is suspected, QC tests can be employed to determine the source of the malfunction. Even equipment that is under warranty or service contract should be checked periodically. QC tests can verify that equipment is operating correctly and repairs are done properly.

A quality assurance (QA) program involves many activities including: quality control testing, preventive maintenance, equipment calibration, in-service education of sonographers, bid specification writing and bid response evaluation, acceptance testing of new equipment, and evaluation of new products. The purpose of the present document is to describe routine ultrasound QC tests to be performed by or under the supervision of a medical physicist. Descriptions of other QA activities, in particular acceptance testing of US equipment, are beyond the scope of this document. Further information on ultrasound QC tests can be found in other documents?

The report presents a detailed set of instructions for setting up and performing ultrasound QC tests. An abbreviated instruction set is also included in Appendix A for the operator's convenience. Examples of individual QC test farms are included in Appendix B, and examples of possible phantom designs are provided in Appendix C.

TEST SCHEDULE

There is a strong commitment to performing at least once a year comprehensive tests of x-ray imaging equipment such as mammographic and fluoroscopic units. Depending upon the complexity of the x-ray equipment, and the number and nature of the tests, the entire set of tests is completed in about 1-8 h. There are no factors in an US unit which would indicate the need for more frequent thorough QC evaluations than is needed for general radiography, so long as servicing is competent and the US technologists are well trained and attentive. However, in the interest of discovering problems before they become serious, it is recommended that certain tests of short duration be performed more frequently. These are termed the "quick scan" tests. They include display monitor fidelity, image uniformity, depth of visualization, hard copy fidelity, vertical distance accuracy, and horizontal distance accuracy. Only the most frequently employed transducer is evaluated in these tests. The quick scan tests plus a physical and mechanical inspection should be performed every three months for mobile and emergency room systems and every six months for others. The total time commitment for the quick scan tests plus the physical and mechanical inspection should be about 15 min per US unit. The more thorough set of tests, analogous to those for x-ray equipment, should be performed annually. Normally, it should take about 1-2 h to perform the more thorough set of tests on a single ultrasound unit with its associated transducers. The final record of an ultrasound exam is often in the form of images stored on transparent film. When this is true, it is imperative that film processor QC tests be performed.

Most of these processor tests are carried out on a daily basis. They are described in the ACR Mammographic Quality Control Manual. To make the present US manual independent and complete, descriptions of these tests are also included here. For the efficient implementation of a QA program, the authors advocate developing a QC test calendar which indicates the dates on which each unit and transducer are to be tested. This calendar should include an area to check off when the tests are completed.

PERFORMING THE BASELINE TESTS

The baseline represents the instrument's peak performance for a particular image quality indicator. Subtle changes in image quality can be detected by comparing the current value with the baseline value.

The baseline tests establish the instrument control settings to be used for the periodic image quality tests and determine the baseline values for each image quality indicator. For the best representation of an instrument's peak performance, the baseline tests should be performed immediately after the instrument has been installed and accepted. To ensure that existing systems are operating up to specification, it is best to perform the baseline tests immediately after preventive maintenance and service by a qualified engineer. If this is not possible, and a particular system is between service calls at the time of the baseline tests, one should immediately after the next service call measure each image quality indicator and adjust the original baseline values if the measurements improve (if the indicator values degrade, the system should be repaired). Remember, the baseline values are the landmarks for detecting changes in image quality.

A. Selecting instrument control settings
A good tissue mimicking phantom allows the use of normal control settings during QC tests. To select the control settings for the image quality tests, scan the phantom as if it were a patient and adjust the controls to produce the best possible clinical image, taking care not to emphasize or exaggerate a particular image attribute. Be sure to adjust and record the video monitor's brightness and contrast settings under "clinical" room lighting conditions. These same dim lighting conditions should be employed while performing all QC imaging tests. When the setup is deemed acceptable, record each of the control settings on the data sheet for the scanner-transducer pair in use. Examples of the settings that should be recorded include dynamic range, grey level map, body part menu selection, power level, gain level, and time gain compensation (TGC). Same of the image quality tests will require different settings for image and focal zone depth. Suggested initial settings for these tests are provided in the instructions, but the settings may need to be altered. Be sure to record the final settings on the data sheet and use them every time the tests are performed. It cannot be emphasized enough that the parameters that are being measured are highly dependent upon the machine and display monitor settings. If different settings are employed, the results may be meaningless. On same instruments, it may be possible to program the desired settings in a user-specified file. When this file is later invoked, the instrument will automatically adjust all of the imaging settings to the desired values. Use of such a file will greatly simplify machine setup for performing the tests, and should eliminate even minor differences in the settings which can be a cause of variability in the test results. Finally, during QC tests, one should be consistent when pairing a particular transducer with the ultrasound unit being tested (i.e., check the serial numbers) because ultrasound transducers sometimes "float" from unit to unit in departments that have more than one scanner from a single manufacturer.

Extracts from the introduction in the Report of the American Institute of Ultrasound in Medicine, AIUM [7], 1995.

Today's ultrasound scanners are generally highly reliable and consistent in their day-to-day performance; therefore, why dedicate time and effort to a quality assurance (QA) program? The reason is that no system is perfect, and, sooner or later, performance degeneration will occur in some scanner in a department with accompanying diagnostic shortcomings to patients; therefore, maintenance of a QA program that demands time dedication consistent with the likelihood of system deterioration is vital to assure that no group of patients is given inferior diagnostic tests.
This manual uses the term "quality assurance" to refer to rapidly accomplished routine monitoring of clinical diagnostic ultrasound equipment to assure that imaging performance of the unit is optimal. (This manual is not concerned with Doppler ultrasound.) It is assumed that most personnel using this manual are sonographers and sonologists. Quality assurance (QA) be divided into four components: (1) daily maintenance and the optimal operation and care of equipment, which is discussed in Chapter 2 along with tips for enhancing operator and instrument efficiency; (2) detecting signs of imaging degradation on patient images or on quickly obtained phantom images, discussed in Chapter 3; (3) periodic quality assurance tests using appropriate phantoms for equipment performance assessment, which is discussed in Chapter 4; and (4) documentation of problems and corrective actions taken, also discussed in Chapter 4. Chapter 1 describes the elements of image formation and physics parameters related to image quality and the appendices, which are highly recommended, contain a much more detailed description of scanners and peripheral equipment for image storage, etc.

The six imaging parameters to be determined are as follows: (a) horizontal distance measurement accuracy (perpendicular to the beam axis), (b) vertical distance measurement accuracy (parallel to the
beam axis), (c) maximum depth of visualization, (d) image uniformity, (e) resolution, and (f) grey-level range. Regarding resolution, all three conventionally designated resolutions, i.e., axial, lateral, and elevational, should be accounted for individually or collectively. Since resolution can vary with depth, this parameter should be determined at various depths over the depth of view or field of view (FOV) chosen. In addition, inspection for flaws in the transducer and its cable, monitor cleanliness, and filter cleanliness are suggested in the procedure below.

Resolution assessment:
Resolution assessed via Focal Lesion detectability: For a given echogenicity relative to its surroundings, the smallest spherical object that can be detected is a measure of collective resolution-including axial, lateral, and elevational. The minimum size detectable sphere will depend on the depth, i.e., distance from transducer to sphere. The technique recommended for rapidly assessing lesion detectability, as a function of depth, is to determine depth ranges over which low-echo spheres of different diameters can be detected in the image.
Resolution assessed via Fibres, Cylinders and Rough Inclined Planes

Extracts from the introduction as presented in the Report of the Royal College of Radiologists, [22], 2005.

STANDARDS FOR RADIOLOGICAL EQUIPMENT
1 The equipment should be capable of producing images of diagnostic quality for all clinical applications.
2 Changes in technology have resulted in considerable improvements in image quality and patient management High priority must, therefore, be given to the replacement of equipment which relies on obsolete or redundant technology, where this compromises image quality
3 An image archiving system should be in place, which allows rapid recall and review of images by those responsible for patient care and which fulfils the medico-legal requirements for the Trust. The system which properly fulfils all these requirements is full PACS (ie, Picture, Archive and Communication System)

ULTRASOUND
Ultrasound is non-invasive, readily available and inexpensive in comparison to other imaging modalities Changes in ultrasound technology and the increasing capabilities of the equipment are developing rapidly These technological changes have resulted in a widely used imaging modality with high diagnostic yield and cost effectiveness Today there is greater availability of equipment with a wide range of performance and cast Consequently, there is also greater complexity involved in the selection and replacement of ultrasound equipment to ensure fitness for purpose.

A number of professional and regulatory bodies have produced guidelines, standards and protocols for performance tests of ultrasound imaging equipment, including Doppler ultrasound machines (See Appendix A) These tests can aid assessment, but with new developments in technology, there is increasing need for the tests and test phantoms to be regularly updated. The mechanical and acoustic properties of materials in test phantoms should ideally be similar to those of typical biological tissues Performance tests of the equipment should be carried out in a clinical setting and be clinically relevant, produce objective results and be widely accepted. Newer phantoms, and updated protocols and assessment parameters that comply with requirements of existing guidelines and provide clinically relevant tests are likely to become available. Detailed discussion of the issues involved in new equipment testing is beyond the present scope of this document, which is intended to be a pragmatic guide to users of general clinical ultrasound Instead, the recommendations suggested here are intended to be used in conjunction with existing guidance, to provide additional assistance to purchasers of ultrasound equipment when drawing up specifications, and to provide assessors with a set of suggested requirements to gauge the likely fulfilment of clinical needs It is limited, at present, to suggested general requirements for all ultrasound equipment, and three widely used clinical applications: abdominal, small parts and vascular ultrasound These sections cover the majority of radiological clinical applications. Venous access for central lines is included in the vascular section.

Recommendations will suggest specifications that are currently assessable and achievable, but it is inevitable that rapid changes in technology will necessitate review of this guidance at regular intervals.

Extracts from the introduction as presented in the Report of the Japanese Industrial Standard, JIS T 1501 [23], 2005.

The standard specifies methods to determine the penetration depth, axial and lateral resolution and displays accuracies like distance accuracy, area accuracy. For this purpose it utilises a staircase low and high attenuation tissue phantom. For the axial resolution 2 wire rods, under a small angle, is used. For lateral resolution a wire rod array is used and the slice thickness is determined using a scattered sheet under a small angle. It also specifies a method for speed measurements used a thread phantom.

Extracts from the introduction of the IEC standard 61391-1, [28], 2006.

This standard describes test procedures that should be widely acceptable and valid for a wide range of types of equipment. Manufacturers should use the standard to prepare their specifications; the users should employ the standard to check specifications. The measurements can be carried out without interfering with the normal working conditions of the machine. Typical **test objects** are described in the annexes. The structures of the **test objects** have not been specified in detail, rather suitable types of overall and internal structures are described. The specific structure of a **test object** should be reported with the results obtained using it. Similar commercial versions of these **test objects** are available.
The performance parameters specified and the corresponding methods of measurement have been chosen to provide a basis for comparison with the manufacturer's specification and between similar types of apparatus of different makes, intended for the same kind of diagnostic application. The manufacturer's specification should

allow comparison with the results obtained from the tests in this standard. Furthermore, it is intended that the sets of results and values obtained from the use of the recommended methods will provide useful criteria for predicting the performance of equipment in appropriate diagnostic applications. This standard concentrates on measurements of images by digital techniques. Methods suitable for inspection by eye are covered here as well. Discussion of other visual techniques can be found in IEC 61390.

Extracts from the introduction of the draft IEC standard 61391-2, [[29]], 2007.

This standard describes test procedures for **measuring the maximum depth of visualization** and the **displayed dynamic range** of these imaging systems. Procedures should be widely acceptable and valid for a wide range of types of equipment. Manufacturers should use the standard to prepare their specifications; users should employ the standard to check performance against those specifications. The measurements can be carried out without interfering with the normal working conditions of the machine. Typical test objects are described in Annex A. The structures of the test objects have not been specified in detail; instead, suitable types of overall and internal structures for phantoms are described. Similar commercial versions of these test objects are available. The specific structure of a test object selected by the user should be reported with the results obtained when using it. The performance parameters described and the corresponding methods of measurement have been chosen to provide a basis for comparison between similar types of apparatus of different make, intended for the same kind of diagnostic application. The manufacturer's specification must allow comparison with the results obtained from the tests described in this standard. It is intended that the sets of results and values obtained from the use of the recommended methods will provide useful criteria for predicting the performance of equipment in appropriate diagnostic applications.

3.4 New developments in current practice

In 1999, Wear et al [[34]] from the FDA (USA) stated that, although the performance of a medical ultrasonic imaging system is to some extent determined by its spatial resolution properties (axial, lateral, and elevational), another important quantitative descriptor of imaging performance is focal lesion detectability. The lesion detectability refers to the ability of a system (in combination with a human or automated observer) to detect the presence of an object which differs somewhat in acoustic properties from the background. Lesion detectability is a complicated function of spatial resolution properties, statistical properties of the lesion and background signals, lesion contrast, and lesion size. Several investigators [[10,34,35,36,37]] have designed custom phantoms and evaluation methodologies for assessment of lesion detectability, which is now implemented as part of the assessment of image quality. There are a number of techniques that can help determining the contrast resolution, e.g.: Imaging fine wires or plastic filaments lying perpendicular to the scan plane, Imaging point scatterers (such as small metal spheres), Imaging a series of anechoic cylinders lying perpendicular to the scan plane, Imaging an array of anechoic spheres positioned in the scan plane, Imaging a series of anechoic pipes positioned in the scan plane.

In the early 1980s a test object was designed consisting of low-contrast conical speckle targets embedded within a speckle background [[35]]. A test object with totally anechoic targets (voids) was developed to get a better understanding of the influence of ultrasonic beams side lobes on the image [[37]]. Another test object consisting of randomly distributed spherical simulated lesions scattered throughout a tissue-mimicking material is described in [[36]]. In this test object the lesions were of the size of 3 or 4 mm in diameter and equal contrast, ranging from 0,5 to 15,5 cm in depth. Also a test object has been developed where lesions consisting of fully developed speckle were embedded within a background consisting of fully developed speckle [[34]]. They define fully developed speckle as to correspond to the situation where the scatterers are sufficiently numerous and positioned sufficiently random such that the signal intensity obeys a defined probability density function. Finally it is worth mentioning the development of a test object that employs a series of anechoic, wall less, pipe structures embedded under an angle in tissue mimicking material [[10]].

The goal of most of the work reported above is to investigate accuracy and usefulness of measurements of lesion detectability using phantoms for medical ultrasonic imaging systems.

Some institutions are not only carrying out testing for regular QA purposes but are developing methods that may improve performance evaluation. [[10, 16, , 37, 38, 39, 40, 41, 42]]. Using the "Edinburgh

pipe phantom", Pye *et al* [10] developed the use of the resolution integral, which provides a general figure for the image quality in the investigated region of a scanner. Hall [17] used the conical speckle targets test object to analyse contrast detail (CD) as a summary measure of image quality. They used five expert observers to observe images from five different contrast targets. It was found that differences in low-contrast detectability were due to differences in ultrasonic beam properties. Thijssen et al [14,16] developed a complete QA program called QA4US®; using commercial available test objects (ATS®). The main parameters they determined were: lesion signal-to noise ratio (SNR_L), point spread function (PSF), tissue to clutter ratio (TCR) and, later on, the geometric conformity. In the sense of QA, a point for discussion could be that for each chosen mode of operation they propose to fix the output power by using either the displayed thermal index (TI) or the mechanical index (MI). They recognised as a practical problem that to perform the measurements they need at the moment at least three different test objects. They also recognise that there is a need to distinguish between technical performance and clinical performance. Brown et al [41] also used in their survey commercially available test objects (Gammex-RMI), but used for the analyses a computerised QA system developed by Gybson in Nottingham. The goal of the study was to compare the image quality in THI, compound imaging and harmonic compound imaging with conventional B-mode imaging. Subtle differences between the different imaging modes were detected. This study was not directed to QA but shows a way to compare modes. Satrapa [15,38,39] and Coulthard [43] produced a series of papers describing the efficient use of the "Satrapa test object" [44]. The specific features of this test object has now been proposed as a Technical Specification document within IEC. Apart from the test object itself, the QA system consists of a computerised analysing system that analyses the SNR and presents also a curve of best fit that seems easy to understand. Finally, the measurement of the point spread function using small metallic ball targets or flat ended steel wires ahs bee developed by Dolezal [40] and Lubbers [42] potentially, this provides an objective measure of the imaging performance for discrete scatterers over the whole scan area.

These will further be discussed in Chapter 5.

3.5 The survey

Questions raised in the survey on the topic of currents practice were:

1. What is the main reason you carry out performance evaluation?
2. Do your QA procedures contain quantitative tests of imaging performance (for instance measurement of spatial or contrast resolution) or are they primarily qualitative (for instance "acceptable/unacceptable" assessment?
3. What is the main aim of your tests?
4. When do you carry out a performance test?
5. Who actually carries out the testing?
6. How often do you test?
7. What Standards or Guidelines do you generally follow?
8. On what sort of scanners do you carry out performance tests?
9. What test devices do you use?
10. Approximately how many scanners are subject to your tests?
11. Approximately how many transducers are subject to your tests?
12. Approximately what percentage of scanners do you test in the hospital or hospitals you are responsible for?
13. Approximately what are the replacement costs of the equipment you are responsible for?
14. How long is typically spent per transducer?
15. How long is typically spent per scanner?
16. Do you advise the end-user to carry out any routine equipment checks?
17. How do you decide which scanner pre-sets or settings to test?
18. How do you control environmental influences (such as lighting, screen condition etc)?
19. Do you consider your test procedures to be adequate?
20. Which of the following aspects do you measure?
21. On what type of image is the analysis carried out?

Answers to these questions are summarised below and full responses can be found through links at www.npl.co.uk/acoustics. As this survey was in principle directed to the UK users the responses on this part of the questionnaire are, unless identified otherwise, only given for the UK.

3.5.1 Why and How Often?

A number of questions were related to "*Why and how often?*. According to 47%, performance evaluation was carried out because it was a requirement of their QA system, see Table 4. An interesting result is that 38% of the respondents consider it of economic benefit but, surprisingly, only one respondent feels that safety is the main reason to perform these tests. When the respondents were asked on what occasion they perform the tests the majority identified that that is done on a routine schedule (77%), see Table 5. Tests are also relatively often carried out after a repair (57%) and as part of acceptance test after delivery (58%). One interesting suggestion was to perform the test prior to the end of any warranty on the equipment. If there is a high incidence of faults occurring shortly after the normal warranty period, and the tests were able to detect them, this timing could save money for the healthcare provider.

Table 4. Why carry out performance evaluation?

	%
Required as part of QA system	**47**
Economic benefit for the healthcare provider	38
Clinical benefit for the patient	26
To identify the best scanner	19
Other	21

Table 5. When are performance tests carried out?

	%
As part of a pre-purchase evaluation	21
As part of an acceptance test after delivery	58
Prior to disposal	9
On a routine schedule	**77**
When a fault is reported	57
Other	11

There was also a question on the QA procedure: did the respondent consider their own tests to be qualitative or quantitative? A large percentage (60%) of the respondents identified that the tests they carry out were mostly quantitative and only 36% considered their tests were mostly qualitative. It is generally easier to report the results of quantitative tests and potentially there could be correlation between the stated reasons for testing and whether the results are quantitative or qualitative. However, in this sample no obvious correlation was seen. As is shown in Table 6, the majority of the respondents aim to assess whether the scanner is performing as it did previously, and rather few are checking for performance against an agreed specification.

Table 6. What is the main purpose of the evaluation?

	%
Performing as it did previously	**43**
Performing to an agreed specification	17
Fit for purpose	32
Other	8

Annual testing was most common: 38% tested once per year, but a substantial number of respondents identified shorter periods, see Table 7. This implies that 68% tested their equipment at least once a year. Some respondents identified that the periodicity also depends on the application of the equipment. General purpose equipment usually was tested yearly. Although 80% of these tests are carried out by NHS staff, see Table 8, they do not perform the same tests or with the same period.

Table 7 How often is equipment tested?	
	%
Once per year	**38**
Once per 6 months	22
Once per 3 months	6
Once per month	2
Other	32

Table 8 Who actually carries out the testing?	
	%
NHS staff from own hospital	**62**
NHS staff from other hospital or regional medical physics department	18
Manufacturer	2
Other	18

It is of interest to know the percentage of scanners used in a hospital that are actually tested, see Figure 2**.** As discussed earlier, not all devices are tested: often it depends on the perceived importance of a high quality image, so general imaging scanners are less tested than special purpose scanners. Again there is a large spread in the results, with an average of 68% of the devices tested. It could be asked why at least 3 respondents did not test any of the hospital devices, although they identified that they spend time on transducer and scanner tests.

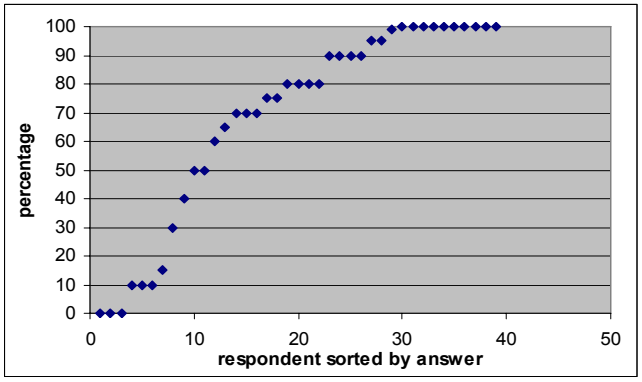

Figure 2. Percentage of scanners from the hospital(s) tested.

3.5.2 Number of devices tested?

The number of scanners available varied greatly among the respondents; ranging from 1 to 500 scanners in total, see Figure 3 (not all scanners are tested annually as indicated earlier). The same applies for the number of transducers tested, see Figure 4. According to the results of the questionnaire, a total of 3020 scanners are tested (although there could be some double counting due to multiple respondents from the same hospital), and they are equipped with at least 7080 transducers, with most being tested at least annually. Considering the results presented in Table 7 almost all of these are tested yearly. Considering the time spend to test an individual transducer, see Figure 5, or a complete scanner, see Figure 6, a total time spend on these tests can be estimated. The average time spend per transducer is about: 1,1 hours. The total time spend on testing the transducers will then be 8000 hours. The average time spend to test a scanner is about 4,4 hours: this results in a total testing time for scanners of 13300 hours, which is about 50% more than calculated on a transducer basis The difference between the total for the transducers compared to that for the scanners may simply be due to the different counting basis or it may be due to the fact that there are additional tests to be carried out on a scanner (such as electrical and mechanical safety or cleaning of air filters).

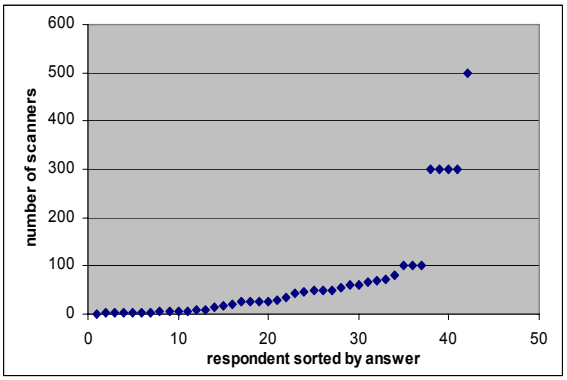

Figure 3. Number of scanners tested per respondent.

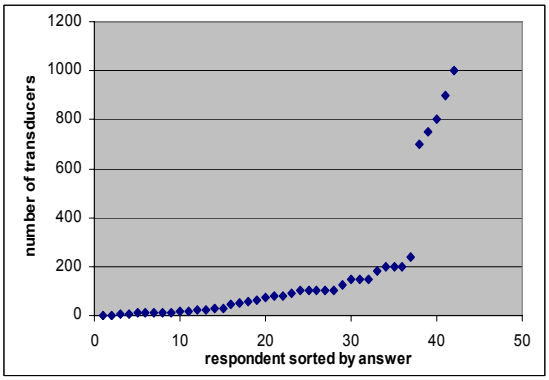

Figure 4. Number of transducers are tested.

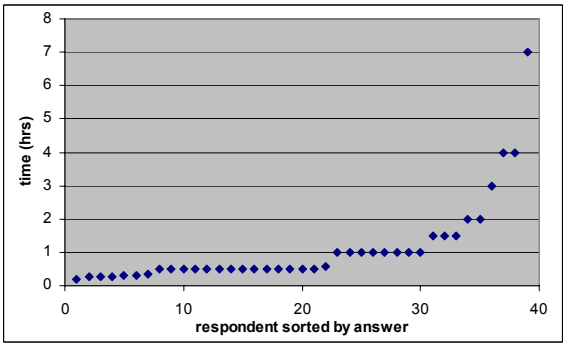

Figure 5. Time spent to test a transducer.

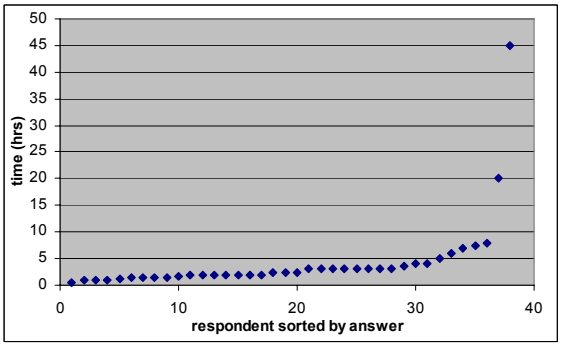

Figure 6. Time spent to test a scanner.

The respondents were asked what they expect the replacement costs for the scanners would be. Figure 7 gives a graphical overview. The maximum was £15 million for 300 scanners (£50,000 each); the minimum came to £30,000 for 30 scanners (only £1,000 each). Such a wide range probably reflects the fact that many respondents really don't know the replacement cost of the equipment they are testing. The average was £28,000 per scanner.

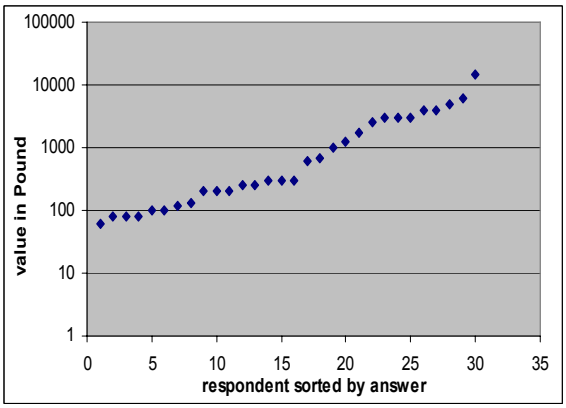

Figure 7. Replacement costs.

3.5.3 Which standards are followed?

It seems very clear that in the UK the IPEM Report 71 [6] is used as a reference for the testing of ultrasound image quality. From the results however it follows that a relatively large amount of respondents use other ways (36%), see Table 9. From their detailed comments, it seems that a number use their own guideline, others add parts of relatively new or experimental approaches in

addition to [6], and there is also a group that like to refer to the European guide on the Medical Device Directive MDA/98/52 combined with IEC documents.

Table 9. What standards or guidelines are followed?	
	%
IPEM Report 71	**54**
RCR ref No BFCR (05) 1	0
AIUM 1995	8
AAPM 1998	2
Other	36

Some specific responses in the "other" category were:
- In-house guidelines, based on various sources, particularly conferences, IPEM meetings, workshops and the literature.
- MDA/98/52 and IEC documents
- Criteria derived from Edinburgh Pipe Phantom performance tests, plus some elements of IPEM 71
- For SNR it is guidance from Satrapa for more subjective tests guidance from IPEM 71 but also guidance from Roy Preston on power measurement
- our own

3.5.4 Is the end-user advised to carry out any test?

An important link in the chain of QA, especially concerning a basic issue like image quality, is communication and interaction with the end-user. 66% of the respondents identified that they ask end-users to carry out at least some tests routinely. Table 10 gives an overview of the responses as to how the end-user is advised about participating in the QA process. At the moment there is no information about the effectiveness of these tests: for instance, are faults or incorrect functioning found earlier.

Table 10. Which tests are the end-users advised to carry out?
- Visual check of transducer surface, cables etc.
- Crystal dropout. E.g. by using a paper clip or wire across transducer face.
- Noise and sensitivity measurements for the reverberations in air image analysed using in house software
- Provide the end-user with Perspex blocks to scan and check, and tried to get them to count reverberations in air. But, no one ever did this.
- We are trying to build simple test objects for use by sonographers.
- Give advice on various aspects of equipment safety
- Weekly caliper and sensitivity tests
- Simple TMM phantoms recently introduced for radiographers use
- Fans and filters ensure are kept clean
- Set up of monitor brightness and contrast to a baseline.
- Ensure monitor and hardcopy settings correct.
- User checks as recommended by IPEM report no. 71
- The Users are instructed in doing a sequence of basic tests and a visual inspection. The tests are performed with a clean transducer in air and cover penetration, crystal drop out and noise. The visual inspection is a comprehensive visual check of the probe and machine condition. This is a recent change (last 12 months) previously advice was to use a tissue equivalent phantom for a wider range of tests. The new tests are advised more frequently (weekly) and take about 5 minutes, they can be split into smaller segments.
- Encouragement, support and training is provided to enable and promote local quality assurance, especially in breast assessment ultrasound. In future, we are considering creating new lower cost contracts whereby we have less direct involvement but contingent on local staff carrying out and reporting local quality work. We feel, overall, that this may promote better standards.
- All medical device users are required to perform pre use checks in accordance with manufacturer's and sometimes professional body recommendations.

3.5.5 What systems are tested?

There seems not to be a large difference between UK and other countries in testing when it concerns Grey-scale 2D systems, see Table 11**,** but when we compare tests related to Doppler and 3-D then it seems that other countries are more often testing. Another question was what image is used for analysing the image performance. From the replies we see that in the UK usually the image on the monitor is analysed, whereas in the other countries it is more common to use a frame grabber or a previously stored data file, see Table 12. Of the respondents who replied "other", the majority used a combination of the individual sources; a number of replies also involved the use of raw r.f. data from

the scanner. Of course this result is influenced by the specific type of respondents from the other countries, which, in general, were not the normal clinical physicists.

Table 11. What kind of systems are tested?

	% UK	% Other countries
Grey-scale 2D	**96**	**80**
Pulsed Doppler	26	50
Colour Doppler / Power Doppler	24	38
3-D volumetric imaging	6	28
Other	24	23

Table 12. What image is used for analysing?

	% UK	% Other countries
Directly on the monitor	**64**	**31**
From a photograph / screen dump	2	
From video frame grabber	3	16
From an image data file	3	22
No opinion	0	3
Other	20	28

Although fundamental to the process of identifiying changes in performance, it seems very difficult to decide at what scanner setting the test has to be performed. Table 13 shows the wide range of approaches adopted and highlights the need for some standardisation in this aspect.

Table 13. Which scanner setting is to be tested?

- Clinical application of interest - preset and then own preferred manual adjustments
- Settings for giving maximum acoustic power output for probe heating, acoustic power and pressure
- On breast scanners with phantom use the preset for breast
- Generally use typical clinical preset with uniform TGC slope, for ease of reproducibility. Generally test a central frequency for both fundamental and harmonic
- The ones in most frequent clinical use.
- select the factory default setting for the usual application of that type of probe
- Determining conditions that produce worst-case acoustic output\
- Try to use default settings as much as possible. There are all recorded to ensure that they do not change between surveys
- Have developed test protocols over time based on accepted parameters
- I use my judgment to optimize settings for imaging of our test object.
- follow IPEM report 71
- Power at max, focus near the transducer, gain is generally at max value, however this is being investigated at present due to saturation effects
- We test the preset that would most often be selected on a particular transducer
- Those which give best image of phantom plus maximum power
- Optimal settings by trial and error, discussion with manufacturer, discussion with user/clinicians

To be able to get reproducible test results it seems to be important that environmental conditions (such as lighting, screen setup and temperature) should also be controlled. The survey tried to get an answer on this issue and responses are given in the Table 14. From the responses it seems that in a number of cases the environment does affect the image performance tests. For some monitor systems the environmental situation is not that important, but for others it is. Also the situation in a test laboratory may not be the same as in the normal clinical surrounding.

Table 14. How is the environment controlled during the tests?
- Test in vivo - in the normal working environment
- Using phantom we would use low ambient light levels.
- We measure ambient room light and monitor luminance with a light meter. We have a comprehensive database which is used to store baseline conditions used in the test.
- As per QA manual and previous tests conditions
- Not very reproducibly. Tests are carried out in the same rooms with the same lighting sources on or off. Monitor assumed to have been left unchanged
- I clean the screen, lower the room lighting and check the brightness and contrast settings
- Use clinical settings and note them - number of lights on, blinds drawn etc.
- Depends on circumstance: for evaluation/acceptance we control ambient lighting to < 50 lux and allow the supplier to optimize screen settings
- Apart from our two micromaxx's all our systems have crt monitors which don't depend that much on external lighting
- Engineers during our tests set to the best available conditions, often there is inadequate control of the environment
- Try to ensure in same room as previously and use a lux meter
- Use a prescribed method
- Personal preference
- It is important to assess systems in the way they are used clinically to measure the effects of any poor environmental conditions. If necessary, thereafter, optimum or better conditions can be set up
- To demonstrate the effects of poor environmental control.
- At present I am developing a monitor assessment method based on the AAPM Tg18 luminance recommendations using softcopy test objects
- Have acquired photometer to improve QA procedures
- Our methods don't generally test the monitor. We use digitised video
- As for screen condition, this is one of our QA tests and a user test
- We have curtains in our lab, and a relatively constant air temperature.
- Air-conditioning in some rooms. Electric fan in one room.

3.5.6 Test devices and parameters measured

Some respondents reported using flow and vessel wall motion phantoms, “abnormal” patients, wires in alcohol mix, string phantoms and/or acoustic output measurements. However, the majority of respondents use normal imaging phantoms most of the time and, in some cases, human volunteers, see Table 15.

Table 15. What test devices are used?		
	% UK	% Other countries
Imaging phantom	**96**	**57**
Human volunteer	17	27
Electronic injection	6	14
Other	17	38

Using these test devices a number of aspects that could describe the image quality are determined. From Table 16 it can de seen that almost everyone is testing the low contrast sensitivity (penetration depth). Also some dimensional measures seems to be a regular measure (lateral and axial resolution, and lateral and axial calliper accuracy). Already earlier stated in the guidance to the end-user (Table 10), crystal dropout is almost always tested. It is interesting to see that aspects related to lesion detectability (e.g. lesion signal to noise, focal lesion detectability, contrast-detail, contrast to noise) are tested less often. The somewhat more sophisticated, relatively new, tests like tissue to clutter ratio, resolution integral, point-spread function analysis and information from scatterers are not yet generally incorporated into testing protocols Although these tests may eventually be shown to give quick insight into the imaging performance.

Table 16. Which aspects are measured?					
	% Almost always	% Frequently	% Occasionally	% Never	% Not sure
Low contrast sensitivity / penetration depth	**70**	7	16	2	5
High contrast sensitivity / penetration depth	**48**	9	23	14	7
Dynamic range	25	5	27	**36**	7
Depth of visualisation	**52**	11	18	11	7
Contrast to noise	16	5	18	**45**	16
Lesion signal to noise	20	5	9	**50**	16
Tissue to clutter ratio	2	0	9	**66**	23
Lateral resolution	**66**	9	20	2	2
Axial resolution	**61**	9	25	2	2
Elevational resolution (slice thickness)	**34**	14	27	14	11
Spatial resolution	**36**	11	18	11	23
Resolution integral	5	0	16	**61**	18
Point-spread function (PSF) analysis	5	7	9	**61**	18
Contrast-detail	11	7	16	**45**	20
Focal lesion detectability	23	9	20	**32**	16
Lateral calliper accuracy	**64**	5	25	5	2
Axial calliper accuracy	**64**	5	25	5	2
Area measurement accuracy	**39**	5	23	20	14
Information from scatterers	7	2	9	**57**	25
Image uniformity	**55**	7	20	14	5
Crystal dropout	**77**	16	5	2	0

4 Shortcomings of current practice in ultrasound

4.1 Introduction

Subjective evaluations by operators using phantoms, for instance, are time-consuming and operator dependent; but more objective evaluations by frame grab or digital data export and numerical image analysis techniques ignore the role of the operator and the significance of monitor/display aspects. Consequently, more objective methods may be more suited to standards but risk losing out on clinical usefulness. Perhaps good quality digital photo of the screen could be used instead – but these can introduce their own transfer problems.

Existing phantoms are all simple geometry and construction; they do not apparently predict which scanners will perform better clinically. Many users report that existing phantoms do not provide adequate discrimination between modern scanners. Hence one question is: what is the appropriate specification of a phantom or other test system which would provide adequate discrimination for basic imaging characteristics (resolution, contrast resolution etc.). Another is: "Can we construct an overall test method which could be used for specifying/acceptance testing and fault investigation?". Pye's resolution integral shows some promise in this [10], and it may also be possible to bring in approaches from CT, MRI and other imaging modalities to define and evaluate imaging performance.

4.2 Interviews with selected UK experts

During the initial interviews, we discussed the perceived weaknesses of current practice. Some of the points raised are listed below, not in any particular order. Obviously, the interviewees did not necessarily all share or express the same concerns.

- There is a very wide range of cost (from less than £10,000 to more than £200,000) and sophistication in ultrasound scanners. This is different to the situation with MRI and CT, where equipment is much more uniform. It is therefore difficult to have an ultrasound testing scheme which can be applied to all equipment.
- A careful weighting between critical and non-critical uses of ultrasound should be performed to limit the costs and improve effectiveness. Some critical applications are much more demanding in terms of the imaging performance required to be clinically effective than other non-critical applications.
- In ultrasound everyone wants to use their own procedure to evaluate performance. This is different to other medical fields. E.g. in radiology and nuclear imaging there exists a strict and organised regime of testing [11,45]
- Imaging phantoms used to identify image quality in the QA process should be traceably calibrated regularly. This seems to be of major importance as different test objects show great differences in results.
- There are no standard images to act as a reference. It would be easier if there were one or more images that could be used as a standard for determining the image quality.
- It is not possible to check that the scanner is using the correct look-up table for the transducer under test.
- Scanner software changes were seen as a problem: sometimes there were unexpected consequences such as a change in thermal index value. Also software changes often change the image performance, so after such a change a new QA baseline has to be drawn
- There was concern that image quality varies significantly within the image frame and should, in principle, be evaluated everywhere. To avoid numerous measurements the image quality should be defined at one position: probably the centre field of view.
- There seems to be a need for a moving target imaging test object – to test the ability of a scanner to image rapidly moving targets such as heart valves.
- Present methods and test objects are still not able to relate the image quality to the clinical performance. In that sense the influence of fat is largely underestimated.
- Test objects do not generally show the type of artefacts that appear normally in real tissue.

- Image ‘speckle’ and noise are two different characteristics. It could be important to measure the noise in an absolute way.
- In most places the performance of 3D images is not tested.

4.3 The survey

Respondents were asked to evaluate how strongly they agreed or disagreed with the following statements:

1. The results of the tests show good correlation with practical experience of the clinical performance of the scanner.
2. The tests I currently carry out give useful technical results.
3. The following parameters are valuable in the evaluation of imaging perfomance.
4. The test phantoms I use are fully adequate for the tests I perform.
5. There are commercially available test phantoms which would be better for the tests I perform.
6. There is no need for a physical test phantom: a good electronic injection test system would be fully adequate for my needs.
7. I am fully aware of the properties of the test-object I use
8. I find that the presence of text on the screen has an impact on the assessment of the quality of the image
9. The quality of captured images (e.g. as a photograph, image datafile or DICOM data) is often significantly different to the quality of real-time images on the scanner monitor.
10. It is important to test the monitor separately from the scanner.
11. The tests I make are cost effective and add significant value to my organisation and/or patients.

Answers to these questions are summarised below and full responses can be found at through links at www.npl.co.uk/acoustics. As this survey was in principle directed to the UK users the responses on this part of the questionnaire are, unless identified otherwise, only given for the UK.

4.3.1 Aspects that are worth evaluating

A number of the tests listed in Table 16 are performed for largely historical reasons. Some of the tests exist now for about 30 years: no large changes have been suggested, although modern ultrasound scanners are much more sophisticated than 30 years ago. On the question whether the tests that are performed are adequate, 61% of the UK respondents agreed but, amazingly, 59 % of the responses from other countries disagreed. It seems that a slight majority feels that the tests they perform show a good correlation with the practical experience, see Table 17. The technical value of the tests is rated more highly with more than 70% of the respondents feeling that the tests are cost effective, although they did not provide a measure for that statement. These results are not consistent with discussions at the IPEM 2005 meeting or interviews carried out as part of this survey: the general impression gained at these events was that many practitioners were dissatisfied with existing performance test methods and not convinced of their benefits.

An important issue is the adequacy of the test objects used to carry out tests for the image quality. From the results of the survey it followed that the majority find that their own test objects are not adequate; however, they do not believe that there are better phantoms available commercially. Also, 82% of the respondents consider that they are fully aware of the properties of their test object. This is an interesting result which again does not reflect the earlier discussions: perhaps the respondents mean that they are aware of what is contained in the information sheet provided with the phantom, whereas earlier discussion related more to data that was not provided. Finally, almost no-one believes that an electronic injection system could completely replace the task of the test phantom.

Table 17. User views on current practice					
	Strongly disagree	Generally disagree	Generally agree	Strongly agree	No opinion
The test results show good correlation with practical experience	7	28	**46**	7	12
The test currently carried out give useful technical results	2	23	**53**	17	2
The tests are cost effective and add significant value to the organisation	7	12	**49**	23	9
The test phantoms used are fully adequate	12	**44**	35	7	2
There are commercially test phantoms available which would be better	16	**53**	7	5	19
No need for a physical phantom, good electronic injection is fully adequate	**40**	**40**	2	0	19
I am fully aware of the properties of the test phantom	2	12	**47**	35	5
The presence of text on the screen has impact on the assessment of the image quality	7	**40**	16	2	35
The quality of the image captured is different to the quality of the real-time image	19	**51**	10	2	43
It is important to test the monitor separately	2	28	**40**	16	14

From the listing of possible aspects that are worth evaluating (Table 18**)**, those items that received a high score for the question "which aspects are measured" also received a good score on value. There was relatively little disagreement with any of the parameters: respondents tended to either agree they were valuable, or have no opinion (although it is not clear if this because they have no experience of them, or they have experience but haven't made up their minds about their merit. Crystal dropout was the most strongly supported parameter. Even aspects that got a low score for "actually measured" (dynamic range, contrast to noise ratio, elevational resolution and spatial resolution) were considered to be reasonable valuable. Again there were fewer opinions on subjects that give information about the lesion detectability (lesion signal to noise, focal lesion detectability, contrast-detail and contrast to noise ratio) but most of those who expressed an opinion considered them valuable. It seems logical that the somewhat more sophisticated, relative new, tests like tissue to clutter ratio, resolution integral, point-spread function analysis and information from scatterers are not yet widely used and they usually resulted in "no opinion".

4.3.2 Most important shortcomings

On the question: "What are the most important shortcomings of the tests you currently do?", the full responses are given in Annex A2 but can be summarised, into the following general areas:

- Results from IPEM report 71 are hard to interpret.
- Test objects do not cope with higher frequency probes (above 10 MHz).
- Standard tests not sensitive enough to detect subtle changes in performance.
- The mistaken belief amongst many that measurements in a phantom correlate with clinical performance.
- Contrast or Temporal resolution are not provided for in the current range of test objects.
- Limited traceability of the measurements made using tissue mimicking test objects.
- The reproducibility of test object characteristics.
- The test objects do not represent complex tissue structures.
- Test object-based tests are not "well-computerized", due to over design.
- The adjustment of the TGC setting.
- The time required to carry out testing.

Table 18. Which aspects are valuable?					
	Strongly disagree	Generally disagree	Generally agree	Strongly agree	No opinion
Low contrast sensitivity / penetration depth	7	2	**49**	33	9
High contrast sensitivity / penetration depth	16	9	**47**	16	12
Dynamic range	7	12	**44**	16	21
Depth of visualisation	7	2	**56**	19	16
Contrast to noise	2	0	**44**	9	44
Lesion signal to noise	2	5	**35**	7	51
Tissue to clutter ratio	2	7	**23**	5	63
Lateral resolution	5	5	**56**	30	5
Axial resolution	7	9	**49**	28	7
Elevational resolution (slice thickness)	5	5	**58**	16	16
Spatial resolution	5	2	**58**	14	21
Resolution integral	0	5	**23**	19	53
Point-spread function (PSF) analysis	0	5	**19**	9	67
Contrast-detail	0	2	**33**	14	51
Focal lesion detectability	2	2	**35**	19	42
Lateral calliper accuracy	9	7	**51**	28	5
Axial calliper accuracy	9	9	**51**	28	5
Area measurement accuracy	9	5	**49**	23	14
Information from scatterers	0	2	**30**	9	58
Image uniformity	7	2	**49**	30	12
Crystal dropout	5	9	33	**51**	2

5 Improvements to current practice

The possibilities for resolving perceived shortcomings depend on the cause. In some cases, definitions or techniques from other modalities could be adopted or adapted; or more demanding specifications and tighter tolerances for test phantoms may be required. Work may be required at an NMI level to establish fundamental definitions and basic measurement techniques, or at other points in the traceability chain. A series of questions in the survey concentrated on trying to identify the main causes of degradation and a specification for the perceived performance needed for an adequate test methods in the future.

5.1 The survey

Questions raised in the survey on the topic of currents practice were:

1. What are the most important shortcomings of the tests you currently do?
2. What are the main causes of image degradation?
3. What measures of imaging performance are most relevant to clinical use?
4. What resolution would be adequate for selected length-related parameters in future test methods?
5. What resolution would be adequate for selected dB-related parameters in future test methods?
6. What resolution would be adequate for selected area-related parameters in future test methods?
7. What resolution would be adequate for selected dimensionless parameters in future test methods?
8. What other medical imaging equipment do you test?.
9. Do you have any other comment on performance evaluation?

5.1.1 Routes to address perceived shortcomings in future

As might be expected, there was no clear preferred answer to this addressing the shortcomings. There was a desire for simpler methods and a single figure of the merit, but a route towards addressing these goals was not suggested. Respondents argue that first the goal of certain tests has to be defined: evaluation of clinical performance and testing for QA purposes are quite different tasks, requiring different approaches. For QA the task may be more easy. This requires the cooperation of the manufacturer and improved determination methods. Also, using computer analysis of digitally captured images to improve reproducibility, and the use of a photometer to routinely measure the luminance of scanner monitors were suggested. The complete set of comments is given in Annex A, but they can be distilled down to the following ideas:

- Provision of user-accessible self-test probe diagnostics within the scanner.
- Provision of QA mode within the scanner that remains constant despite software upgrades.

- Use computerized analysis of phantom images.
- Greater use of electronic probe testing.
- Measurements of crystal behaviour using in-air measurements with digital image capture and software analysis.
- Temporal resolution evaluated using cineloop capture.

- Methods of monitoring changes in phantom properties to ensure traceability in assessment.
- A test phantom with stable acoustic properties.
- Further development of the Edinburgh pipe phantom and resolution integral concept.
- Design of a suitable contrast resolution test object.
- Method to look at high echoic structures and to improve what we can offer cardiology.
- Use test-objects that are build in 3D in order to get a more realistic image.
- Design phantoms for specific imaging modes like harmonic imaging.

5.1.2 Main causes of image degradation

In the listing below the main causes of image degradation is given. As expected, the transducer is the main cause with the following reasons for failure given: element drop-out, lens/membrane problems, break in electrical screen, damaged pins in probe plug, air in mechanical probe head, cable or connector break due to mishandling or wrong construction. This indicates that putting effort into developing simple and rapid methods for probe testing would be one of the most cost effective steps. A relatively simple cause of image degradation seems to be the monitor itself: this part could generally be tested for its performance separately and is easy and relative cheap to replace. A point of note, is the training (or perhaps lack of training) of the user or operator which is indicated as one of the main causes of image degradation.

Table 19. Causes of image degradation sorted into related groups.

probes Probe deterioration piezoelectric element coupling in the transducer Transducer faults crystal drop-out lens de-lamination crystal degradation Probe Damage/Faults probe defects Transducer element drop-out transducer faults Transducer faults Probes probe degradation	monitors CRT monitor lifetime monitor failure Monitor degradation monitor degradation CRT monitor degradation degradation of monitors Pc Monitor ageing of screen ageing monitors Monitor performance display monitor drifts	Electronic faults Scanner port & channel faults Internal hardware faults Analog pulser-receiver degradation Power supply degradation	changes in operator user familiarity or knowledge Lack of user awareness operator training The perception of the user
Transducer elements failure Ageing of elements Acoustic lens damage Lens (contact) degradation defect of single elements element degradation/failure transducer degradation dead elements, bad cables PZT ceramic ageing element dropout Transducer degradation Deterioration of transducer probe defects by usage lens damage cable / connector damage	equipment age Age of machine equipment degradation Wear and tear on machine "wear and tear"	software faults / deterioration Software Errors/Bugs Scanner software (set-up)	non-uniformity of relative sensitivity among the channels (elements) of the array transducer and output powers from every elements of an array transducer

5.1.3 Clinically most relevant image performance parameters

Table 20 shows that contrast resolution was found to be the single most relevant parameter to evaluate the image performance, with various types of spatial resolution next, followed by penetration depth. Noise measurements were also often identified as being important. A few identified the more recent parameters like detection of cysts, characteristic resolution, 3D signal to noise as being important; image uniformity, image thickness (presumably slice thickness), side lobes, monitor performance, and acoustic output were listed as important.

Objective methods to evaluate the performance were also indicated, e.g. Thijssen‘s QA4US method, Satrapa’s 3D signal/noise integral method, UltraIQ program from RAMSOFT and transducer parameters evaluation with FirstCall 2000 from Sonora. Or the more simple daily routine check out system – like Kollmann’s Austrian test kit or more sophisticated and expensive Sonora – Nickel system, which is limited to check transducer and receiving module only.

Some of the comments listed in Section 2.2.2 gave the impression that a number of the measurements performed to evaluate the image performance are too much based on tradition and that the methods should be replaced by more “realistic” measures. From the results on the question “What are most relevant parameters”, there is no outstanding candidate for one which might be more “relevant”, but there is a sense that perhaps more effort should be directed at a scanner’s ability to distinguish contrast.

Table 20. Clinically most relevant image performance measures			
low contrast resolution at depth the high contrast resolution down through the image contrast resolution contrast resolution contrast resolution contrast sensitivity contrast performance contrast detail image contrast contrast-detail analysis contrast resolution contrast contrast ratio of the image low contrast contrast sensitivity	penetration penetration depth of penetration depth of penetration penetration penetration penetration penetration penetration sensitivity of far off wire displayed dynamic range	noise noise snr noise signal-to-noise signal-to-noise	cyst detection visualisation spherical cysts visualisation of low contrast objects cystic lesion detection spherical lesion detectability lesion and cyst detectability snr in anechoic cysts
resolution resolution resolution resolution resolution resolution	resolution integral and derived parameters (depth of field / characteristic resolution)	caliper accuracy caliper accuracy calliper accuracy	uniformity of crystal sensitivity
spatial resolution lateral resolution lateral/elevation resolution axial resolution lateral resolution axial and lateral resolution accuracy of size detection temporal resolution	image uniformity image uniformity	point spread function point spread function point spread function	acoustic output watching transmitted energy limits, ti and mi accuracy
-20dB pulse width and the fractional bandwidth of the array. Also the effective dynamic range of the ultrasound system itself	dead zone	monitor and ambient light (adjustment) quality monitor contrast response	electronic probe tests
image thickness	side lobes	transmit spectrum	receive spectrum
Periodical testing using Tissue mimicking phantoms combined with objective method of evaluation. E.g. Thijssen's QA4US method, Satrapa's 3D signal/noise integral method, UltraIQ program from RAMSOFT and transducer parameters evaluation with FirstCalll 2000 from Sonora.	Simple daily routine check out system – like Kollmann's AUStrian test kit or more sophisticated and expensive Sonora – Nickel system, which is limited to check transducer and receiving module only.	Uniformity of channels gain parameters, dynamic focussing stability and affectivity, lines density and lateral and transversal focusses position and resolution evaluation using	

5.1.4 Adequate resolutions for the performance parameters

The respondents were asked about the resolution of different parameters for future test objects. The responses are given in the next tables. About half of the respondents skipped the question, leaving between 17 and 26 replies for each parameter; and of those who replied, many had no opinion. In the tables the results are given in a percentage and for two frequency ranges: 4 to 8 MHz and 10 to 15 MHz. For length related parameters resolution values that were identified as adequate can be reasonably differentiated in the tables but, even here, there is a significant spread and the number of definite responses is low (typically, one reply is equivalent to between 4% and 7%). For the dB related parameters and for the resolution integral a relative large number of respondents did not have an opinion. But from those that had an opinion the result was reasonably focussed on specific resolutions. For the area measurement however the respondents did not come up with a clear

preferable adequate resolution. The results were about equally spread between < 1 mm^2 till 100 mm^2.

Table 21. Adequate resolutions for the length related parameters.							
Results (in %) for: 4-8 MHz / 10-15 MHz	<0,05 mm	0,05-0,1 mm	0,1-0,2 mm	0,2-0,5 mm	0,5-1,0 mm	>1,0 mm	no opinion
Low contrast sensitivity / penetration depth	0 / 0	0 / 9	4 / 4	13 / 9	13 / 13	**35 /** 30	35 / 35
High contrast sensitivity / penetration depth	0 / 10	5 / 0	0 / 5	10 / 10	14 / 10	**24 / 19**	48 / 48
Depth of visualisation	0 / 0	0 / 5	0 / 5	14 / 10	0 / 10	**48 / 35**	38 / 35
Lateral resolution	0 / 5	4 **/ 27**	22 / 14	**26 /** 23	13 / 14	13 / 0	22 / 18
Axial resolution	0 / 5	4 **/ 36**	30 / 27	**35 /** 5	4 / 9	4 / 0	22 / 18
Elevational resolution (slice thickness)	0 / 0	0 **/ 23**	13 / 9	17 / 9	17 / 18	**22 /** 14	30 / 27
Spatial resolution	0 / 5	4 **/ 23**	**22 /** 14	13 / 9	9 / 9	13 / 5	39 / 36
Lateral calliper accuracy	4 / 5	4 / 18	13 / 9	**26 / 32**	13 / 9	13 / 5	26 / 23
Axial calliper resolution	4 / 5	9 **/ 23**	17 / 18	**22 / 23**	17 / 5	4 / 5	26 / 23
Point spread function (PSF)	0 / 0	0 / **15**	**14** / 5	**14 / 15**	10 / 10	4 / 0	57 / 55

Table 22. Adequate resolutions for the dB related parameters								
Results (in %) for: 4-8 MHz / 10-15 MHz	0,5 dB	1 dB	2 dB	3 dB	4 dB	5 dB	6 dB	No opinion
Dynamic range	12 / **13**	6 / **13**	**18** / 6	6 / 6	0 / 0	12 / **13**	0 / 0	47 / 50
Contrast to noise	18 / 18	0 / 0	6 / 6	**24 / 24**	0 / 0	0 / 0	6 / 6	47 / 47
Lesion signal to noise	0 / 0	13 / 13	7 / 7	**20 / 20**	0 / 0	0 / 0	0 / 0	60 / 60
Tissue to clutter ratio	0 / 0	7 / 7	7 / 7	**27 / 27**	0 / 0	0 / 0	0 / 0	60 / 60
Contrast detail	7 / 7	**13 / 13**	7 / 7	**13 / 13**	7 / 7	0 / 0	7 / 7	47 / 47
Focal lesion detection	**13 / 13**	7 / 0	0 / 7	**13** / 7	0 / 7	0 / 0	7 / 7	60 / 60

Table 23. Adequate resolutions for the area related parameter								
Results (in %) for: 4-8 MHz / 10-15 MHz	<1 mm^2	1-2 mm^2	2-5 mm^2	5-10 mm^2	10-20 mm^2	20-50 mm^2	50-100 mm^2	No opinion
Area measurement accuracy	0 / **15**	**20** / **15**	10 / 10	15 / **15**	10 / 5	5 / 5	10 / 0	30 / 35

Table 24. Adequate resolutions for the dimensionless parameter						
Results (in %) for: 4-8 MHz / 10-15 MHz	1	1-2	2-5	5-10	>10	No opinion
Resolution integral	6 / 6	12 **/ 18**	**18 /** 12	0 / 0	6 / 6	59 / 59

5.1.5 Other imaging equipment

Imaging equipment tested in the departments that participated in the survey ranged from X-ray (majority) and MRI to Fluoroscopy. It was a minority of the respondents to the questionnaire that also test other medical imaging equipment. Respondents were asked if they consider that the

performance tests on other types of imaging equipment are adequate, and are there any lessons that ultrasound could learn from other modalities?

Whilst it may be true that many of the same problems exist as in the ultrasound world, testing for other imaging equipment, like X-ray, is more focussed on QA purposes and the evaluation of engineering parameters. The general opinion seems to be that these tests are better controlled and easier to carry out than for ultrasound, and they seem to be considered adequate. There is less concern expressed about subjective image "quality. An important question is: Are the system properties equal? Answer: Probably not. So specific test methods would not carry across to ultrasound. More test tools became available for X-ray equipment due to legislative requirements. So "enforcement" could be the key to get more resources to fund development of "better" test equipment. One advice for ultrasound is to move away from trying to mimic the patient and develop systems to measure "physics" variables like resolution in the absence of any specific tissue mimic. Further it is advised to pay more attention to the image quality of the monitor.

5.1.6 Other comment on performance evaluation

The respondents were invited to provide any other comments on aspects of QA or performance evaluation for ultrasound imaging. In response, some are argued on a preventive base: "*Despite the shortcomings many argued, ultrasound QA should be done and its efficacy in revealing "small" problems before they become "big" should be studied*". Others see a problem in convincing the other about the need and usefulness: "*The most important problem of ultrasound quality assurance is the political one: Nobody is willing to accept that it is really necessary. And the question for us is whether we should further try to convince the others or whether we should use our limited budget for other things*".

A very critical response, repeated from 5.1.1 but along the same lines as the previous one, is: "*Publish the evidence from the respondents to part 1 that backs up their responses about correlation with clinical performance, results being technically useful and being value for money! Further testing of the various systems in use in the field to assess their value in demonstrating degradation in performance*." This is augmented by: "*The lack of evidence base is a real problem; those who strongly believe in what they are doing should disseminate their evidence. Even professional bodies make statements about the efficacy of QA/QC without referencing evidence - very disappointing*."

An encouraging comment to improve and simplify and make the methods traceable is the following: "*I am a medical physicist specialising in medical ultrasound. However, when it comes to giving a professional opinion about the image quality and performance of scanners, I don't have any widely accepted way of doing this at my disposal. I can stand beside the scanner, watch a few patients getting scanned, and make encouraging noises. I can speak to the manufacturers and get to grips with new technologies and how they work. I can get hold of the probe myself and image various commercial test objects containing filaments and voids. What I haven't got is a standard or widely accepted way of objectively assessing imaging performance. My colleagues who deal with MRI, CT, diagnostic X-ray and nuclear medicine images can all do better than I when it comes to assessing imaging performance. That's a problem for me, and a problem for the health sector generally given that ultrasound scanning is the second most common imaging procedure after plain X-ray. Clinicians often rely on scanning just a few patients before deciding how to spend tens or hundreds of thousands of pounds on ultrasound equipment. That's not a good idea. As a scientist, I am looking for better evidence than this when it comes to choosing imaging equipment. My employer owns ultrasound equipment worth around £6M. Are they getting good value for money? Have they paid over the odds for bells-and-whistles that contribute little to better imaging performance and better patient outcomes? Could they purchase less expensive equipment and still achieve the same level of performance? These are pertinent questions for a health provider currently in a very challenging financial position. So, would better measurement techniques benefit providers and patients? My own view is that they would, because they would underpin the processes of evaluation, procurement and in-life assessment. New techniques need to be able to objectively summarise the imaging performance of a scanner in ways that allow it to be compared with other scanners, as well as with*

its own performance over time. The ability to do this would also benefit manufacturers, who, like users, have no objective way of assessing the imaging performance of their products. Our group has successfully employed the concept of a "Resolution Integral" measurement for the last 5 years. It has become apparent that the main source of uncertainty in making this measurement is the characterisation of individual test objects. The ability to obtain traceable measurements of attenuation, backscatter, B/A, and small-scale particle distribution within the TMM would add greatly to the robustness and reproducibility of the technique. We see no benefit at all to the user community if this, or any other performance assessment technique, is solely "artefact" based. Good science is underpinned by measurements that are objective and traceable. In ultrasound imaging, objective and traceable performance measurements would ensure reproducibility, and provide a depth of scientific understanding that is sadly lacking from the purely artefact-based ultrasound QA techniques which have been commonplace for many years.".

5.2 New methods in scientific literature

Apart from the opinions from the survey respondents, a selection of scientific literature was evaluated in a search for ideas that might be valuable for performance evaluation. Several investigators published work to improve methods for determining the quality of ultrasound imaging devices. The usually result in a figure related to the whole system. One of the improved approached is given in [42]. It uses a series of flat ended stainless steel wires which were experimentally evaluated as point targets giving a calibrated backscattering over a large range (up to 72 dB) for ultrasound frequencies in the range 2 to 10 MHz. It is proposed to apply the targets for calibration of the large dynamic range of a scanner and as point targets to establish resolution in three perpendicular directions, axial, lateral and elevational. One paper [16] describes a method that might be considered as a minimum set of objective QA measures. This paper is significant here because of the use of a complete test methodology called QA4US®, which includes among others the use of a computational observer that enables determination of contrast sensitivity from r.f. or DICOM images of test phantoms.

Several investigations show the effectiveness of using an automatic analysing system with a 3D artificial cysts test phantom, basically developed by Satrapa [15,37,38, 43]. Investigators report that it was possible to monitor the increase of the image quality after the repair of an ultrasound scanner and that the method is suitable to measure the quality of an ultrasound scanner [44].

A complex system based on the point reflector principle bas been developed by Dolezal et all [18] to analyze the Point Spread Function (PSF) and in this way measure a number of significant image quality parameters at any point in the area being imaged. Their system uses digital image analysis for accurate and objective measurement. They were able to plot the Lateral Resolution (LR) characteristics over the scanning plane. This can differentiate separate scanning lines and even multiple focal areas for dynamic focusing systems. The measuring system can detect malfunctions in dynamic focusing, size of aperture, time gain compensation function and/or transducer element failure.

The modulation transfer function (MTF) [11] is a widely used metric in a range of imaging techniques – especially optical ones. It is a measure of the transfer of modulation (or contrast) from the subject through to the image. In other words, it measures how faithfully the lens reproduces detail on different size scales from the object to the image produced by the lens. The test pattern and even lighting have to be defined, and there are many different patterns. MTF for optical lenses can be a very complicated function distance from the centre, spatial frequency, aperture sizes and target orientations, and such complexity would also be expected for ultrasound.

Several interesting points emerge from study of MTF [46]. First, it is quite possible for the MTF to be less than zero. In reality what this means is that black areas in a test pattern appear as white and white areas appear as black, so although a pattern may appear to be resolved (insofar as you may see black and white areas), in reality it isn't. Resolution above the point at which the MTF first reaches zero is known as spurious resolution. It seems possible that such spurious resolution could also occur in ultrasound imaging. Secondly, is the idea of Subjective Quality Factor (SQF). SQF takes the

quality of the lens into account, but also factors in the MTF of the human visual system. An important observation is that the human visual system has an MTF which peaks in the range of 10 to 20 line-pairs per mm on the retina. SQF factors this into the viewing equation, so that. for example, for a given lens at a given aperture, each print size (viewed from a constant distance) would have its own SQF value. SQF defines the subjective quality of an image rather than defining the quality of a lens. Extending this to ultrasound, we can imagine that a 'better' system which happens to produce structured noise on the monitor which coincides with the sensitive range of the human visual system, may not perform as well clinically as a 'worse' system.

We can also imagine another approach in which a complex but precisely known structure is imaged. The difference between the image and the known real structure can be evaluated numerically over the whole image and also over parts of the image. This would give information about the 'accuracy' of the image and the spatial-scale if the difference function may give information about the cause of the differences (eg speckle, limited resolution, low penetration…).

5.3 Traceable calibration of test objects

There seems to be a great need for a traceable calibration of test objects and the properties of their material. Stability and reproducibility assumed (or at least hoped for) by users, but one of the shortcomings observed is that the test objects used to determine the image quality are not regularly checked for their own performance. Some methods to measure the speed of sound, attenuation coefficient and backscatter coefficient are given by AIUM [47].Measurement methods for the determination of the actual position for reflecting targets, cists or voids in existing test objects are not available.

5.4 Other imaging devices

In a search for improvement for definitions on image quality, approaches from CT/MRI/photography are evaluated. In their document on digital testing methods [24] the AIUM states: "*For photon-based imaging systems (e.g., photography, radiography, nuclear medicine), the two-dimensional, high contrast spatial resolution can be described by the measurement of the spatially symmetric point spread function, since these systems are relatively isotropic. Ultrasound, however, possesses an asymmetric point spread function, which changes with depth in the image, and must be described by separate axial and lateral resolution measurements, representing the x and y dimensions of the point spread function. For three-dimensional spatial resolution, the slice thickness must also be measured, and represents the z -dimension of the point spread function.*".

To further complicate matters, ultrasound imaging systems also do not behave like photon imaging systems, due to the fact that ultrasound is a coherent radiation, which produces a phenomenon called "speckle". Speckle in an ultrasound image gives rise to the "texture," which is normally seen in a B-mode image of any structure that is made up of non-resolvable scatterers. Thus, we have two separate phenomena that describe ultrasound images: the production of speckle and the production of specular reflections. Specular reflections arise from the imaging of targets that are on the order of, or much larger than the wavelength of the ultrasound beam, and whose reflection coefficient is high, while speckle is produced when imaging a collection or ensemble of diffuse targets that are smaller than a wavelength and low in reflectivity. Also, the system resolution of the imaged specular reflectors is different from the system resolution of the imaged reflectors."

Already in 1995 the USA Food and Drug Administation felt the need to establish a coherent regulatory approach to computer-aided diagnosis (CADx), bringing together the relevant device experiences of various Office of Device (ODE) divisions [48]. A secondary goal was to advance the scientific basis for assessment of unconventional and artefact-limited imaging systems to the level previously achieved for conventional systems. Mammography, Fluoroscopy, Magnetic Resonance and Tissue Characterization using ultrasound were evaluated for a common regulation. The work, that continued after 1998 [49] was typically undertaken in research projects using automatic image or tissue recognition. No practical solutions for evaluation of image quality were included.

Image quality in chest radiography is usually considered in terms of the portrayal of normal anatomy or the depiction of potential pathology [50]. Radiographic display of normal anatomy provides examples of the compromises that arise when image quality is considered. As one example, technical factors that might improve the visibility of unobscured lung may tend to diminish the visibility of lung projecting behind the heart or other structures in the chest. Consideration of such compromises often dominates careful investigations of image quality for the examination.

The concepts of image contrast, image sharpness and image noise are the mainstays in the quantification of image quality in medical radiographic science. The concept of structured noise is very important in the understanding of image quality in chest radiography. Structured noise means the amount that the signals from anatomic structures interfere with the detection of significant pathology. Structured noise is also present in ultrasound images, usually referred to as 'artefacts' such as reverberation and shadowing, so this concept may be useful for ultrasound performance evaluation.

A new work has started in IEC to prepare a standard that concerns Medical Image Display Systems [51]. The goal of the intended standard is to describe evaluation methods for testing the performance of image display systems like CRT monitors, flat panel displays and projection systems.

6 Discussion

This report addresses the needs for standards and especially for standards relevant to NMSPU – *ie* those which are measurement based and have the eventual aim of establishing a traceability chain between appropriate basic definitions and practical tools for the end-user. This may require work at different levels for different aspects of performance assessment, so a strategy and sequence for this research or development should be identified.

The debate about ultrasound performance assessment seems to be confused by a failure to distinguish consistently between:
- the consequences of viewing an image (ie the accuracy of the clinical diagnosis)
- the subjective 'quality' of an image
- the performance of the equipment which produces the image.

Many replies to our survey assume implicitly that we should be searching for a single 'performance' parameter that predicts which scanner is best. The ultimate goal may indeed be to predict which equipment is most likely to allow diagnosis of a particular condition in a particular patient, but this is too big a problem to address in one step. If we take view that the clinical user is most important and that good performance is defined as the ability to diagnose disease, then the recommendations of this report seems simple: there is no need for NMS research because don't know how to address this 'complete' question. As one survey response stated: "*We physicists reduce image performance to an assemblage of individual measurements, but we may not know how to properly weight and combine these metrics into an overall "image performance*".

However, can we address smaller, more limited questions related to the performance of the equipment which produces the image. For example, not "which scanner has the best performance", but "how accurate is the image presented by this scanner?" or "what is the contrast in this image?". In our view, 'imaging performance assessment' is the assessment of different aspects of imaging performance – so more than one parameter is required. The availability of appropriate technical performance data should make it easier to purchase the ultrasound scanner or transducer which is most suited to the required task.

Arising form the survey responses, interviews and material in the literature, a number of topics for discussion were identified and a group of experts were invited to attend a workshop at NPL to discuss those topics and any other issues around future research requirements. The participants were:

Jacinta Browne	Dublin Institute of Technology,
Andy Coleman	Guy's and St Thomas's NHS Trust, London
Nick Dudley	Nottingham University Hospitals
Tony Evans	University of Leeds
Andrzej Jastrzebski	Barts and the London NHS Trust
Stephen Pye	Edinburgh Royal Infirmary
Stephen Russell	Christie Hospital NHS Trust, Manchester
Barry Ward	Newcastle General Hospital
Kevin Wells	Centre for Vision, Speech and Signal Processing, University of Surrey

Discussion at the workshop centred on six main topic areas and the key points in those areas are summarised below.

6.1 Benefits of QA and image performance assessment

To be really useful a QA test method should ideally be more sensitive to changes than a human observer. For instance, some studies have shown that a small number of damaged transducer elements can go unnoticed clinically, even though they presumably do have a small effect on

imaging performance. A good QA test should pick this (or other subtle changes) up before they become noticeable clinically and potentially compromises diagnosis.

The need for basic underpinning research was appreciated. It is natural for hospital medical physicists to concentrate on what they could apply personally in terms of carrying out evaluations, but this is only part of the story. If manufacturers can do better evaluations or have better quality control, this is also a benefit to the hospital user.

6.2 Traceability of measurements using phantoms

There is a need for reference facilities and methods for the measurement of speed of sound, attenuation coefficient, absorption coefficient, backscatter coefficient and nonlinearity parameter for tissue mimics used in imaging phantoms. Of these properties, only measurement of speed of sound and attenuation coefficient are fully established in the UK and, even for these, there are no International Standards. The determination of the variation of properties with frequency and with temperature should be included.

It has been observed that phantoms are often stored or transported at temperatures different to their intended temperature of use, and that it can take many hours for them to reach a uniform temperature. For proper traceable use, there should be some indication of the temperature inside the phantom – preferably close to the centre.

Some widely used commercial phantoms need regular ‘rejuvenating’ by the manufacturer. This now involves returning the phantom to the USA and the results of rejuvenating are uncertain. Some people felt that the rejuvenating fluid does not recombine with the phantom material and so does not reverse aging changes, but there is no way to check this properly.

Quality control by manufacturers is perceived to be poor, with experience showing that two examples of the same phantom type can give very different results, suggesting that the material properties are different: this could be improved by the publication of an IEC standard for acoustic material characterisation.

Phantoms can be expensive, especially when rejuvenation cost is included. There is an example in NMR of an expensive EC-developed phantom being shared amongst a group of hospitals and it was suggested that there could be some reference ultrasound phantoms held centrally (perhaps at NPL).

There was some support for the idea of a very simple and stable water/ethanol phantom to test consistency, but the general feeling was that this was not sufficient and that an element of tissue similarity was required.

The need to match acoustic properties over a wider frequency range (up to 20 MHz was suggested) was seen as important, with manufacturers usually specifying properties over a relatively small range.

The nonlinear behaviour of tissue mimics needs to be known and matched. This is mainly because of the much wider use of harmonic imaging (HI) nowadays, with some systems defaulting to HI on start up and probes often being designed to be optimal with HI.

The balance of scatter and absorption in a tissue mimic may be more important than is generally realised, with manufacturers often matching attenuation by adjusting scatter rather than absorption. Traceable measurements of backscatter would be useful here as would a comparison of results made using the same phantom design but with different scatter/absorption ratio.

6.3 Integrated scanner diagnostics

Integrated diagnostics accessible to the user or the medical physics department would be extremely useful, although this is not in itself a research topic for the NMS. In the past some ATL equipment did have diagnostics which were accessible to selected users but more recent models have not. The use of specific presets is helpful – some users record the presets either to disc or by making notes form the set-up menus – but it assumes that everything which is relevant is contained within the preset. Some aspects – notably TGC – usually have to be adjusted in addition.

6.4 Comparisons/evaluations

There was seen to be some merit in evaluating the repeatability of phantoms by circulating a small number to different centres. However, in general, there was little support for comparisons of phantoms of different types because of the difficulty in establishing a common reference and interpreting the results.

Investigation of the feasibility and benefits of using general imaging metrics such as MTF, PSF, noise power spectrum and contrast to noise ratio had widespread support. It was pointed out that CT imaging assessment (to which ultrasound is closely-related) generally assumes that the PSF is nearly invariant with position (which is not true with ultrasound). So the adoption may not be straightforward. The Satrapa phantom and Edinburgh pipe phantom also seem to be worthy of further investigation.

The use of computational or synthetic observers was also generally supported because of greater reproducibility and speed. However, it was noted that there may be differences between the data which is image data which is available for offline analysis and the live data which is presented on the scanner. These differences, together with issues such as compression, would have to be explored. ImageJ and Matlab are widely used environments for image analysis.

The monitor and video system should be tested separately if computer analysis of offline data is used. Ideally this should be done according to the principles of the DICOM standard although, in practice, it is not always possible to separate the acquisition part of an ultrasound scanner from the video system and monitor. It was also noted that there is a large ‘human’ element in the positioning of the transducer, which does not apply to MR and CT for instance.

There was a consensus that it is better to list relevant image performance parameters rather than try to combine them into a single overall figure of merit; although it may be that some subsets of parameters may combine naturally.

Electrical test systems like the Sonora FirstCall were seen to be valuable since the transducer is the component which is most likely to fail and the tests can be carried out quickly by a competent technician. The major perceived problem was the availability and cost of the connectors which interface to the transducer – which are different for different models and manufacturer of transducers. This difficulty could be overcome if the purchasing process specified that appropriate connectors must be provided by the scanner manufacturer. They are seen as a QA tool rather than an imaging performance tool; however they have high sensitivity for detecting faults/changes. There could be merit in trying to correlate changes in performance with electrical changes in the transducer.

6.5 New test objects

The development of scatter-free phantom for measuring PRF was seen as an important activity. However, it was noted that the PSF may not be independent of the reflectivity of the target and so a range of target strengths should be investigated.

Very simple phantoms (for instance wires in water/ethanol mixture) were not seen as particularly worthwhile because the in-air reverberation test seems to provide a useful basic test. Likewise a phantom to test geometric image accuracy was not favoured.

It was noted that existing phantoms do not cater for specialised transducers such as transvaginal, transoesophageal, intravascular, eye and skin scanners. It was also noted that there were no phantoms specifically for HI or compound imaging.

Another area of concern is the need to test the accuracy of measurement of dimensions, area and volume since these measurements are more and more being used for diagnosis.

6.6 Organisation

The questionnaire showed up the level of variability between departments as to the amount of ultrasound QA carried out. There was concern that, where the level of activity was too low, the level of skill and expertise may not be sufficient to identify subtle damage or changes to performance, and that there should be a minimum 'critical level of activity'. A great deal of the expertise is now leaving the NHS because of the retirement of those physicists who were involved in the development stages of ultrasound imaging. So it is important that proper training (and accreditation) be given to the people carrying out ultrasound QA and performance assessment and that they should be able to maintain their skills.

It was felt that a benchmarking exercise would be valuable to quantify more precisely how much time and effort is spent in different centres and what should be the critical minimum involvement to provide an effective and efficient service.

An evaluation of the economic benefit of QA or performance testing would also be carried out. As an example, the Foetal Anomaly Screening project, spent £12m on replacing ultrasound equipment based essentially on the age of the equipment, rather than on objective measures of its performance. As another example, providing a first line service in one centre reduced the maintenance contract cost from approximately 7% of the equipment value to approximately 5%, saving £35k p.a.

It was noted that the testing protocols vary greatly between hospitals – including factors such as range of parameters measured, monitor set up and room lighting conditions - and that this should really be more uniform. The replacement of IPEM report 71 will be significantly simplified which may result in greater uniformity.

6.7 Funding

This report is primarily written to identify ways in which the NMSPU Acoustics and Ionising Radiation programme can contribute to better or more useful performance evaluation methods. However, many of the important topics identified in this report are multi-disciplinary and would benefit from collaborative work which may be funded from multiple sources.

The NMS Acoustics and Ionising Radiation programme is most immediately relevant to evaluation of imaging metrics and the development of measurement methods and standards for characterising tissue mimics and monitoring changes in phantom properties. Apart from these, a number of other potential sources of research funding can be considered.

The National Institute for Health Research, which manages the NHS research budget: there could be possibilities under a number of schemes within NIHR – for instance Health Technology, or Research for Patient Benefit. One idea would be an evaluation of what level of performance is required for a number of specific clinical applications. This could study and model the sequence of events from patient arrival through to clinical outcome and patient/societal benefit in an attempt to optimise the net benefit with respect to equipment performance level and cost.

The Engineering and Physical Science Research Council, the Medical Research Council and perhaps the Economic and Social Research Council could be appropriate for more fundamental research into establishment of appropriate imaging metrics for ultrasound, their use with the NHS and the benefit to the patient population. There may be scope for a multi-disciplinary 'life-science interface' project jointly funded by the research councils and in collaboration with the NMS.

Development Agency or EC SME funding would be appropriate for companies working on imaging phantoms or the development of imaging systems. New methods and improved quality control would certainly offer the promise economic benefits to such companies.

There should also be interest from specific charities or professional organisations (for instance British Heart Foundation or Cancer Research UK) and indeed from the NHS Purchasing Authorities into methods for optimising the choice of equipment based on objective evaluation of performance characteristics.

7 Recommendations

This report has surveyed current practice relating to image performance assessment and QA for ultrasound imaging in the UK and has investigated the perceived needs for research to underpin improved standards for performance assessment. Although other aspects have been considered, the prime concern has been those needs which are relevant to the National Measurement System – *ie* those which are measurement based and which have the eventual aim of establishing a traceability chain between appropriate basic definitions and practical tools for the end-user.

Following a small number of initial interviews with selected experts, and a two-part survey of current practice, the responses to which are summarised in Sections 3, 4 and 5, a draft report was prepared. This report was circulated to nine experts who went on to participate in a workshop at NPL to discuss in more detail the research needs and priorities. These discussions are summarised in Section 6.

There was recognition that, although measurements and assessments made in hospitals are important, research should not be focused exclusively on this community. There is a need for more basic research and the establishment of a measurement infrastructure which will allow other stakeholders (for instance manufacturers of imaging equipment or test devices) to provide more reliable and relevant information which will benefit the hospital user and, ultimately, the patient. However, it was also recognized that testing which is carried out for the health services must be beneficial and cost effective and that there was also a need for research outside of the narrow technical considerations of pure performance assessment.

Four topics were identified as being of highest priority:

1. **Traceability of measurements using phantoms.**
   Three key components were identified for this topic
   1.1. *Reference measurement of phantom acoustic properties.*
   Establish reference facility and methods for the measurement of absorption coefficient, backscatter coefficient and nonlinearity parameter for tissue mimics used in imaging phantoms to augment existing capabilities for the measurement of speed of sound and attenuation coefficient. This should include determination of the variation of properties with frequency and with temperature. As a follow on to this, there is a need for internationally standards specifying how manufacturers should measure and specify material properties.

   1.2. *Monitoring changes in the acoustic properties of intact phantoms.*
   A device to monitor degradation or changes in certain key properties (probably speed of sound and attenuation coefficient) of intact phantoms over time to ensure that performance assessment is carried out using phantoms which are themselves within specification. Changes may occur gradually or as a result of a specific event such as being 'rejuvenated' by the addition of fluid or being allowed to become too hot or too cold. However, use of reference measurement methods is not appropriate as they would require disassembly of the phantom.

   1.3. *Development of a system for measuring changes in dimensions or separations in intact phantoms*.
   It has been observed that the tissue-mimicking part of many phantoms shrink over time, reducing their overall dimensions and also changing the separation of the wires or other internal components

2. **Evaluation of imaging metrics.**
   The applicability and potential benefits of using more general imaging metrics such as the Point Spread Function, Modulation Transfer Function and noise power spectrum and contrast-to-noise ratio which are widely used in other imaging modalities should be explored. It is recognized that these are not single-value quantities for ultrasound system but the same is true in other cases.

Mapping the variation of these quantities through the image plane should give basic performance data which can be compared with other test results (for instance the Resolution Integral and Satrapa method) and with clinical experience.

3. **Development of scatter-free PSF phantom**
   In determining the basic imaging capability, the presence of a scattering medium adds a level of noise which can obscure the underlying performance. To provide the maximum sensitivity to changes in performance (either over time or between equipment), the Point Spread Function should be determined in the absence of scatter, but with the correct speed of sound and attenuation to otherwise simulate tissue. The phantom should allow determination of PSF at different points in the image plane, and the variation of PSF with target strength should be investigated.

4. **Benchmarking**
   The survey has shown up the wide variation between hospitals in the amount of effort going into ultrasound QA and performance assessment. Whilst it is widely felt that there are significant benefits from this testing, there is little hard evidence on which to base a proper cost-benefit study, which would allow effort to be concentrated on those aspects of testing which are cost-effective. A benchmarking exercise should be carried out to establish in detail how much time/money is spent on ultrasound QA and performance assessment across the UK.

A further seven topics were identified as important but of lower priority either because they are under the control of manufacturers or because they naturally follow on from one of the highest priority topics. These topics were:

Integrated scanner diagnostics.
Since the most common cause of scanner degradation in the early years are attributed to mechanical or electrical damage to the transducers, or changes to the monitor and video system, built in diagnostic tests accessible to the user or the physics support would enable the vast majority of these faults to be picked up and repairs/replacements to be made. The user community should explore with industry and regulatory bodies the possibilities of establishing a set of integrated diagnostic tests to reduce the burden of QA. These could include: Reference or comparison images; U/S relevant test cards for monitors/video systems; low-level transducer diagnostic tests; and inclusion of specific QA pre-sets.

Evaluation of the repeatability of measurements with phantoms.
The idea would be to evaluate the repeatability and reproducibility of results when using the same type of phantom type in different clinical centres on a range of scanners. This is an important topic but it should wait until methods for monitoring changes in phantom properties are in place, Otherwise, it is difficult to distinguish differences between phantoms from differences between scanners or testing protocols.

Evaluation of electrical testing methods
Ideally, QA testing should identify faults before they become obvious to the user. Existing phantom test methods do net seem able to do this, but electrical test systems may be able to. The ability of these test systems to identify faults or incipient changes to transducer characteristics should be evaluated.

New test objects for specialised transducers/modes
The vast majority of imaging test phantoms are designed for use with general imaging transducers operating in normal grey-scale mode. Specialised transducers (such as intraluminal, intravascular, skin or eye scanners) have different requirements in terms of physical size and shape, frequency range, interfacial layers and so on. Similarly, harmonic imaging or tomographic imaging have somewhat different requirements for phantoms. More research is required on these designs but this is a lower priority than improving the basic test methodologies.

Quality systems in hospitals

There would be benefit in more unified testing protocols and/or quality systems in hospitals. Although many follow IPEM Report 71 in part, many do not and there is a wide variation in which parts are implemented amongst those that do (the selection of monitor settings and lighting conditions during testing was one aspect which was picked out). It may be that the simplified protocols expected in the updated IPEM guidelines will encourage more uniformity. The recommended benchmarking and cost-benefit exercise should also help in this.

Training and accreditation

Ultrasound QA and performance evaluation is not straightforward and it is often not a large part of an individual physicist's or technician's job. This makes it essential that organisational steps are taken within the hospital quality system to ensure that staff are properly trained and maintain their skills at an appropriate level. A training and accreditation scheme for ultrasound QA and performance evaluation should be established.

Research into these topics should be collaborative: different skills are required for each topic but together should involve measurement scientists, the health service, academia, industry and social scientists.

8 References

[1] Quality assurance guidelines for Medical physics services National breast screening quality assurance Coordinating group for physics NHSBSP publication no 33 Second edition June 2005.

[2] IPEM, Quality Assurance of Ultrasound Scanners Meeting Abstracts, York, UK, 03-2006.

[3] S.B. Barnett, Q.A. and the accreditation of ultrasound practitioners: is it really necessary, J.of Physics: Conference series 1, 2004, 11-12.

[4] Medical Device Directive, Directive 93/42/EEC, Council of the European Communities, 1993.

[5] AIUM, Standard Methods for Measuring Performance of Pulse-echo Ultrasound Imaging Equipment, AIUM standard, AIUM 1990

[6] IPEM, Routine quality assurance of ultrasound imaging systems, R. Price, Institute for Physics and Engineering in Medicine, Report 71, 2002 (originally published by Institute for Physical Sciences in Medicine, 1995).

[7] AIUM Quality Assurance Manual for Gray-Scale Ultrasound Scanners, Stage 2, AIUM, 1995.

[8] M.C. van Wijk, J.M. Thijssen, Performance testing of medical ultrasound equipment: fundamental vs. harmonic mode, Ultrasonics, 40 (2002), 585-591.

[9] M.F. Insana, T.J. Hall, Visual detection efficiency in ultrasonic imaging: A framework for objective assessment of image quality, J. Acoust. Soc. Am. 95(4), 1994.

[10] S.D. Pye, W. Ellis, T MacGillivray, Medical Ultrasound: a new metric of performance for gray-scale imaging, J.of Physics: Conference series 1, 2004, 187-192.

[11] ICRU, Medical imaging, The assessment of image quality, ICRU, Report 54, 1996.

[12] AIUM, Performance Criteria and Measurements for Doppler Ultrasound Devices, Technical Discussion 2nd Edition, AIUM Technical Standards Committee, 2002.

[13] IPSM, Testing of Doppler Ultrasound Equipment, Institute of Physical Science in Medicine (IPSM), Ed. P. Hoskins, S. Sherriff, J. Evans, Report no. 70, 1994.

[14] J.M. Thijssen, M.C. van Wijk, M.H.M. Cuypers, Performance testing of medical echo/Doppler equipment, Eur J. Ultrasound, 15 (2002), 151-164.

[15] H-J. Schultz, J. Satrapa, G. Doblhoff, Automatisierte Qualtitaetskontrolle von Ultraschalldiagnostikgeraete und Ordinationen …. Oder warum fuhren vorhandene messmethoden zur falschen einschaetzung von qualitaetsparametern, DGBMT Kongress, TU Ilmenau, 2004.

[16] Johan M. Thijssen, Gert Weijers, and Chris L. de Korte, Objective Performance Testing and Quality Assurance of Medical Ultrasound Equipment, UMB, Vol. 33, No. 3, pp.460-471, 2007.

[17] T.J. Hall, M.F. Insana, L.A. Harrison, N.M. Soller, Ultrasound contrast-detail analysis: A comparison of low-contrast detectability among scanhead designs, Med. Phys. 22(7) 1995.

[18] L. Dolezal, J. Hálek, Ch. Kollmann, R. Wiecek, An automated system for ultrasound scanner evaluation using PSF analysis of received signal, J of Physics: Conference series 1, 2004, 205-208.

[19] Z.F. Lu, R.L. Kruger, Hands-On Ultrasound Physics and Quality Control Workshop, 2006 ?? http://aapm.org/meetings/02AM/pdf/8394-19995.pdf.

[20] AAPM, Pulse echo system specification, acceptance testing and QC, P.L. Carson, M.M. Goodsitt, in AAPM Monograph 24, CT and Ultrasound in Medicine, 1995.

[21] AAPM, Real-time *B*-mode ultrasound quality control test procedures, Report of AAPM Ultrasound Task Group No.1, M.M. Goodsitt, P.L. Carson, D.L. Hykes, J.M. Kofler, Med. Phys, 25 (8) 08-1998.

[22] Standards for ultrasound equipment, Royal College of Radiologists, RCR refNo BFCR(05)1, 2005, web: www.rcr.ac.uk

[23] JIS 1501, (JEITA) General methods of measuring the performance of ultrasonic pulse-echo diagnostic equipment, 2005.

[24] AIUM, Methods for Measuring Performance of Pulse-echo Ultrasound Imaging Equipment, Part II: Digital Methods, Stage 1, AIUM, 1995.

[25] Ultrasound Quality Control, Basic tests, J.M. Koffler, D.S. Groth, Mayo Clinic and Foundation, Rochester, USA, Nuclear Associates, 1996.

[26] Assessment of display performance for medical imaging systems, AAPM on-line report no 3, 04-2005,

[27] DICOM, Digital Imaging and Communications in Medicine, 2003: http://medical.nema.org/dicom/2003/03_10PU.pdf.

[28] IEC 61391-1: Ultrasonics – Pulse-echo scanners – Part 1: Techniques for calibrating spatial measurement systems and measurement of system point spread function response, 2006.

[29] IEC 61391-2: Ultrasonics – Pulse-echo scanners – Part 2: Techniques for measurement of maximum depth of visualization and the displayed dynamic range, draft 02-2007.

[30] 87/375/NP, Ultrasonics – Real-time pulse-echo scanners – Phantom and methods for automated evaluation and periodic testing of 3-D distributions of signal-to-noise ratio voids.

[31] Ultraschall Med. 2006 Jun;27(3):262-72. Automated quality control of ultrasonic B-mode scanners by applying an TMM 3D cyst phantom. Satrapa J, Schultz HJ, Doblhoff G. (in German).

[32] www.northernphysics.co.uk/tcc.html.

[33] IEC 62464-1 Ed.1: Magnetic resonance equipment for medical imaging – Part 1: Determination of essential image quality parameters, 2007.

[34] K.A.Wear, R.M. Gagne, R.F. Wagner, Uncertainties in estimates of lesion detectability in diagnostic ultrasound, J.Acoust.Soc.Am. 106(2), 1999.

[35] S.W. Smith, R.F. Wagner, J.M. Sandrik and H. Lopez, Low contrast detectability and contrast/detail analysis in medical ultrasound, IEEE Trans. Sonics Ultrason, 30, 164-173, 1983.

[36] J.J. Rownd, E.L. Madsen, J.A. Zagzebski, G.R. Frank and F. Dong, Phantoms and automated system for testing the resolution of ultrasound scanners, Ultrasound Med. Biol, 23, 245-260, 1997.

[37] H.-J. Schultz, J Satrapa, G.Doblhoff, Deutsche Gesellschaft für Ultraschall in der Medizin e.V, Arbeitskreis Ultraschalsysteme Geschäftsstelle Frahmrodder 1160 D. 22000 Hamburg, Advanced Ultrasound Imaging and Imager Control by applying artificial 3D Cylinder Cyst Phantom.

[38] J. Satrapa, New strategy in automated measurement of imager quality parameters (Discussions basis for IEC meeting in Stuttgart on March, 20. 2007).

[39] Jaroslav Satrapa and Ivan Zuna, Differences of Ultrasound Propagation in Tissue and Tissue Mimicking Materials, privat communication.

[40] L. Dolezal, J. Mazura, J. Tesafik, J. Hálek, H. Kolárvá, A New Approach to an Ultrasound Imaging System Evaluation Using the Point Spread Function, EMBEC 2005, ISSN: 1727-1983.

[41] J.E. Browne, A.J. Watson, N.M. Gibson, N.J. Dudley, A.T. Elliott, Objective measurements of image quality, Ultrasound in med. & biol., vol. 30, no. 2, pp. 229–237, 2004.

[42] Jaap Lubbers and Reindert Graaff, Flat Ended Steel Wires, Backscattering Targets for Calibrating over a Large Dynamic Range, UMB, Vol. 32, No. 10, pp.1585-1599, 2006.

[43] P. Coulthard, Results from using the TCC 3D QC tool, for Routine QA Assessment of 21 Ultrasound Scanners over two years, IPEM, Biennial Ultrasound, Birmingham, UK, 02-2007.

[44] B. Blichenberg, Ch.Hamann, Fr. Ueberle, Image quality of ultrasound scanners, 2006 http://www.rzbd.haw-hamburg.de/~m4100105/Projekte-Start/abstract_blichenberg_hamann_bmt2006_english.pdf

[45] ICRU Report 70, *Image Quality in Chest Radiography,* Journal of the ICRU Volume 3, No 2, 2003.

[46] http://photo.net/learn/optics/mtf/.

[47] AIUM, Methods for Specifying Acoustic Properties of Tissue Mimicking Phantoms and Objects, Stage 1, AIUM, 1995.

[48] FDA, Diagnostic Imaging, Advanced Performance Assessment Techniques, Office of Science and Technology - Annual Report - Fiscal Year 1995, 1996 http://www.fda.gov/cdrh/ost/reports/fy95/diagnostic_imaging.html.

[49] FDA, Imaging, Advanced Problems in Statistical Classification and Estimation, Office of Science and Technology - Annual Report - Fiscal Year 1998, 1999, http://www.fda.gov/cdrh/ost/reports/fy98/Imaging.HTM

[50] ICRU, Image Quality in Chest Radiography, ICRU Report 70, An Executive Summary, 2003 http://www.icru.org/n_03_4.pdf.

[51] IEC 62B/664/NP, Medical Image Display Systems – Part 1: Evaluation methods, 2007.

Documents worth reading but not referenced from the present report:

IPEM, Guidelines for the Testing and Calibration of Physiotherapy Ultrasound Machines, IPEM, Ed. S. Pye and B. Zeqiri, Report no. 84, 2001

Th.L. Szabo, Diagnostic Ultrasound Imaging, Inside Out, Elsevier Academic Press, 2004

E.D. Angelini, et al, Comparison of ventricular geometry for two real-time 3D ultrasound machines with three-dimensional level set, Heffner Biomedical Imaging Lab, New York, USA.

Quantitative Ultrasound Image Quality Assurance package, UltraIQ Workstations, Ramsoft Inc., 1995

J. Satrapa, H-J. Schultz, G. Doblhoff, Conception progress of ultrasound propagation in living tissue and imager quality control, (DRAFT), Privat communication. 2006

J. Satrapa, H-J. Schultz, Advantages of quality control automation for ultrasound imager, 2006

H. Huynen, (in Dutch) Quality control, needed but then from A till Z, Workshop on image quality, Acc hosp. Maastricht, Netherlands, 2006

N.J. Dudley, M. Kirkland, Lovett, A R Watson, Clinical agreement between automated and calculated ultrasound measurements of bladder volume, The British Journal of Radiology, 76 (2003), 832–834

IEC 61223-3-2 Ed.2: Evaluation and routine testing in medical imaging departments - Part 3-2: Acceptance tests - Imaging performance of mammographic X-ray equipment, FDIS, 2007

IEC 61223-3-6 Ed.1: Evaluation and routine testing in medical imaging departments - Part 3-6 Acceptance Tests - Image Display Devices, CDV, 2006

Annex A

Survey, responses to the open-ended questions

A.1 Questions on current practice

A.1.1 Concerning: What defines imaging performance?

Responses from UK:

- I would say: technical issues and what clinicians like.
- As a physicist I feel more comfortable with an objective assessment although current technical results do not correlate at all well with clinicians visual assessment.
- the best indicators of imaging performance are high and low contrast penetration from IPEM report 71. This is advanced nicely by the Edinburgh resolution integral.
- from Physics & Eng point of view: physical image parameters i.e resolution, caliper accuracy , etc. as these can give an idea about the beam geometry. From clinical point of view: image clarity (contrast) and image details.
- Good imaging performance is that which allows accurate diagnosis or measurement. Clearly the technical parameters are important to clinical performance, but these parameters measured in TMM do not correlate well with perceived clinical performance. I believe this lack of correlation is real, although the assessment of clinical performance in order to assess the correlation is very subjective and often an image is assessed on aesthetic rather than diagnostic grounds.
- Good imaging performance related to success at diagnosing a particular condition. This however should relate to technical parameters such as resolution and contrast. Work towards finding the relationship between the two would be very helpful.
- measurable parameters linked to clinical outcomes that can be audited, but separable from training, experience and knowledge of operator
- The image performance needs to be set against measurable criteria such that standards for performance can be put in place. There needs to be a technical standard. Diagnostic tasks could then be set against this criteria so that minimum standards are maintained. These can be increased as equipment improves. This would clearly help with redeployment of ultrasound systems to less demanding tasks. If the criteria were to be diagnostic success then there would need to be a constant monitoring to ensure equipment maintains this level. This is difficult to achieve.
- I would agree with "the image that is most successful at diagnosing a particular condition" I do believe that SNR testing does have an important part to play in informing the user about cyst detection which is not intuitive from the image presented or by instruction in the user manuals.
- The single most important thing is the ability of a scanner to visualise small, preferably spherical, cystic objects in a TM material. Unfortunately, all of the commercially available phantoms with cystic objects inside them have problems of one kind or another. In my opinion, the Edinburgh pipe phantom (minus the resolution integral) is the best available option at the moment.
- When I began work in medical ultrasound, 25 years ago, I was thought by both experienced physicists and radiologists that the best probe to use for a particular procedure was the probe with the highest centre frequency that had the necessary tissue penetration. Although technology has progressed a long way since then, I think this still encapsulates the compromise between tissue penetration and image quality in ultrasound scanning. And it illustrates the point that scientists/engineers and clinicians can hold the same opinion about what constitutes good performance. The issue of image quality can of course be discussed from either a technical or a clinical perspective, and using technical or clinical vocabularies. This can make it difficult to distil the core opinion being expressed from the vocabulary and exemplars used to articulate it.

Responses from outside the UK:

- As ultrasound is used in more and more quantitative ways, objective measures of what constitutes imaging performance need to be identified. There are no current objective standards on what constitues "good" Image quality, and that is part of the problem - the term "good" is subjective in nature.
- For objective assessment of equipment, the technical issues may be more useful.

- Good imaging performance is a complex term. Its definition has to cover parameters, which some of them are technical and/or physical and others are human observer dependent. The first part has advantage of being objective measurable, the second one is represented by subjects like contrast resolution, the sonograph presettings, display adjustment and being more close to observer's mind might be more important for subjective impression about image quality. This fact derives the difficulties of the imaging performance measurement.

- The main goal of imaging quality is to determine the detectability of definable structures or objects. Cyst and vessels are ideal but patient depending difficult to scan. The cysts are echo free and may be produced as spherical or cylindrical voids The detectability of cysts maybe expressed by signal to noise ratio measurement

- Five doctors - ten opinions: No objective definition possible. Good performance is given when device allows for easy adaptation of parameters to the actual diagnostic needs

- This is the problem: good or sufficient imaging performance was when the diagnosis has been to 100% correct. This means that specific requirements are necessary for every specific application case. This is not possible. So the problem is to find quantitative measures to define a performance quality which cover most of all application cases.

- resolution, geometric accuracy, and contrast

- technical issues, resolution and contrast

- when the contrast is good

- It might depend upon the purpose of use. By the way, I think that the contrast ratio is common for all use.

- The ultimate measure of imaging performance is the "lesion-signal-to-noise ratio", because it an objective measure of contrast sensitivity (Computational Observer) that contains the speckle size (related to spatial resolution), the contrast resolution and the noise of the images (SW Smith et al. 1983)

- I am in the group of "that which gives the best diagnosis". We don't sell boxes of circuitry and buttons, we sell diagnostic images. How those images are created is somewhat secondary. If we could make diagnostic images with a shoebox and a piece of wood on a string, we would be selling shoeboxes, string and wood. The technical parameters listed in the question are valuable, but their importance and contribution to what constitutes a "good" image is HIGHLY dependent on both application and operator. For example: contrast range is more important in OB than in Cardiac applications. In fact cardiac users LIKE contrasty images. Resolution is a GREAT parameter - but resolution where? Over what range? Again, isn't that dependent on application? Depth of resolution is less important when imaging a 1-month foetus vs. a fatty liver in a 350 pound patient.

- Excellent low contrast resolution for small focal lesions.

- Diagnostic accuracy should be the ultimate measure, which involves both the system performance and the reader performance. For the system performance, we do have a set of comprehensive tests that can measure various characteristics of the system. However, the correlation between the system characteristics and the clinical diagnostic accuracy is not very clear. But I believe the system characteristics measurement provides a bottom line.

- "Image Quality" is a personal choice like beauty or delicious. The key is to be able to differentiate between normal and disease. Ultrasound imaging is very angle dependent, in both Doppler and B-mode. Compound Imaging and Vector Doppler (and maybe colour power) can reduce angle dependence. Speckle suppression by multi-frequency imaging can reveal otherwise indistinguishable boundaries. Hmmm, maybe image an anechoic rod perpendicular to the image plane with impedance not equal to the anechoic surrounding material. Then, at the specular angles within the aperture you'd see the echo.

- The only repeatable and objective criteria are engineering metrics--e.g., penetration depth, resolution metrics, etc.

A.1.2 Concerning: Who should define imaging performance?

Responses from UK:

- Users

- Yes. Each group is clearing assessing something slightly different and there is insufficient evidence of how the various results are related.

- sonographers will define good imaging performance by looking at contrast throughout the image. Physicists by nature quantify measurements, and it is hard to quantify contrast.

- yes, because they use different references to assess the image quality, e.g. Physicists and Engineers use phantoms which are different to patients!

- The viewpoint depends on the reason for defining imaging performance. The physicist may define in terms of variables measured in a phantom, which may be valuable for serial perfomance measurement (QC) but not for clinical comparisons. The user may not be able to define performance in terms of measurable variables, but may be able to do so in a more qualitative, clinical way which is not easily translated into measurements. The 2 may be inappropriately linked by the user making value judgements in 'physics' terms, e.g. the resolution is good - they just mean they like the image better.

- All that matters is the clinical outcome, whether imaging is successfully applied to diagnosis. A familiar style of image may lead to more successful diagnosis than a technically superior image if insufficient or zero training is involved. A complicated operating system will have the same effect.

- The direct link between the parameters currently measured by technical staff and the apparent clinical performance of systems is tenuous. The main reason for this is both in the variety of settings available on the machines and in clinical preference for a brands characteristic image presentation (mostly down to the factory starting points for presets). Testing the machines at the clinical extremes (limits of performance) is unlikely to happen in side by side testing. The configuration of controls and ease of use (may be linked to familiarity) can allow clinicians to drive one brand more effectively to get 'best image' than another.

- Much of the difference has been allowed to fester because Physics have taken too long to come up with solutions to assist the clinicians who have become more and more frustrated and taken matters into their own hands. I would like to point out that the SNR tests from Mr Satrapa have been around for the last 10 years.

- Traditionally, medical physicists have tended to think in terms of "lateral resolution", "axial resolution" and "slice thickness resolution". Usually, these parameters have been measured using nylon filaments inside TM phantoms. Unfortunately, measurements of this type are not very good at picking up problems with scanners and transducers. This is partly because the measurements are very difficult to make properly, and partly because the system performance depends a lot on sidelobes etc., which are often "lost" in the background speckle.

- I refer to my answer in (13) above. In my experience, the perception of “different views” often arises through the use of different vocabularies and exemplars, without any fundamental difference of opinion existing with regard to the underlying facts. It is natural for the views of each category of user to be framed in their own professional vocabulary, and weighted by their particular professional experience and priorities.

Responses from outside the UK:

- They do have different views. My research has shown the major reason for this is lack of understanding and common terminology among the various groups.

- Both sonographers and manufacturers should define imaging performance.

- To achieve comparability of imaging performance measurements the definitions has to be based on using objective measuring methods. Therefore the imaging performance evaluation methods and parameters should be defined by engineers and physicists. I am not able to answer the second part of the question.

- There are some methods known to determine image quality. Manufacturer and scientific institutions prefer e.g. PSF measurements and phantoms. But PSF is not possible at point of installation. Standard phantoms offering only axial and lateral resolution and objects to display depth penetration. These measurements are irreal and do have no relation to the clinical image. Ultrasound is three dimensional and it is necessary to measure all three axis together with snr.

- They have, because they have different backgrounds and expectations. The interpretation of images is not based on physics alone, but also on haptic experience during image pickup and the anatomic knowledge of the user

- We need a set of physical quantities to define quantitative criteria. This should be done by technical staff. In addition more physiological tests would be helpful, for example structures of organs that should be recognized from medical staff or a computer program. So, both should be incorporated.

- Essentially, the image that is most successful at diagnosing a particular condition is possible only when it has got the desired resolution, geometric accuracy, and contrast. So, the views may differ with category of user, but basically they all seem to say the same thing.

- Physicists

- Because they have different approach: - the sonographer doesn't know the physics of the matter (ultrasound propagation in tissue) - The physicist knows the physics of the matter - the manufacturer knows the limit of their machine. ect,ect, there are different points of view

- Actually, I have no idea. But I agree the need to define the imaging performance. How about to make a consensus among all categories of users?

- International professional organizations of end users (WFUMB, IOMP), in cooperation with the manufacturers

- Absolutely. Different users and applications demand very different settings and parameter optimization. Companies spend an enormous amount of resources researching and creating "default" settings by application. And they vary CONSIDERABLY. To be honest, the opinions of physicists, manufacturers, Royal Colleges and regulators are not so important. The equipment is not made for them, they do not use it for what it is intended, and they don't buy it. It is the sonographer and the scanning MD whose opinion is paramount. If you go to the hospital for an ultrasound exam, who do you want deciding what constitutes a good image? Your doctor/sonographer or a regulator, or physicist, or even a manufacturer? I know how I would answer that question.

- Differences occur because of the backgrounds of these different people. The only category that is important are the clinical users of ultrasound. Physicists have gotten too much involved in the numbers game (the above questions that I did not answer).

- Since diagnostic accuracy is the ultimate goal, defining the image performance should be a team work of all that's involved, eg, engineers, physicists, sonographers, radiologists, and etc. Each group has its different perspective. That's why developing a reproducible and more quantitative QC test approach is important to provide a bottom line. Sort of, OK, these are good, what's next?

- One group is attempting to develop objective measures that don't relate to clinical tissue and the other is attempting to see disease that is not well defined. If you showed a diagnostician an edge detected image, like a line drawing (cartoon of a face), and a gray scale image (photograph of a face) the diagnostician would recognize the friend, but not remember which type of image that they saw. A physicist would be busy measuring the dB difference of gray levels, a parameter irrelevant to the diagnostician.

- Physicist, physician, and sonographer

- We physicists reduce image performance to an assemblage of individual measurements, but we may not know how to properly weight and combine these metrics into an overall "image performance".

A.1.3 Concerning: Should the performance be defined as application specific

Responses from UK:

- A fair amount of overlap. May need to allow different tolerances for different applications. Some applications require specific tests eg Doppler for pv

- I think it is possible to asses image performance over all the applications using the Edinburgh resolution integral. The graphical results make this easy to compare similar systems

- the decision is purely economical! For it determines the cost of producing equipment for either examples above. The technology can do both at probably equal efficiency, but it is the cost factor.

- This depends on the purpose of the assessment of imaging performance. For QC purposes I would suggest that a single scheme could be devised. For assessment of clinical performance relative to other scanners, if it actually possible to devise a scheme, it may well be different for some (but not all) applications.

- It would be nice to have a grand unified theory! I think the safer approach would be to concentrate on two or three different scenarios.

- We use a general purpose assessment for a machine that may be used for a wide range of purposes and a specific assessment for breast imaging. For other purposes we use individually tailored assessment set up with the help of a sonographer to pinpoint clinical need.

- Requirements should be grouped to match so there may be considerable crossover but for specific tasks in each area a particular level of performance may be required. These areas should be identified separately.

- they should be considered separately?

- See above: they need to be assessed separately.

- I am confident that it is possible to incorporate the vast majority of soft tissue imaging applications within a single general scheme for assessing imaging performance. Our research group has demonstrated the feasibility of achieving this (Journal of Physics, Conference Series (1) 2004, 187-192 ; UK Patent GB2396213). For routine assessments, we employ an “Edinburgh Pipe Phantom” to characterise a scanner’s capability of imaging anechoic structures with dimensions 0.2 mm to > 100 mm

Responses from outside the UK:

- Two answers: There are common performance parameters across every clinical discipline related to ultrasound. There are also clinical application specific parameters that should be addressed as well.

- It is possible to incorporate main applications within a single scheme. For different frequency may be considered different requirements.

- I see two main schemes of examinations. One is related more to spatial and contrast resolution and second to temporal resolution. But the set of measurement methods would be identical.
- Detectability of anechoic cysts with the 3D phantom. Size to be defined in regard to the structure.
- Final goal should be ONE simple test method
- The technical performance check should be device specific and not application specific. Physiological tests (see question 15) need special application targets.
- considered separately
- obstetric: this application present the most risk, the foetal must be protect. The exam must be very fast and at the minimum power necessary to see him. The 4D imagine is not necessary but the 3D imaging can be used. Convex and internal probe.
- The scheme should be performed at the settings and with the transducer commonly used for these applications. The tests (at least our tests) can be applied for any system/transducer combination and for each application
- Absolutely not. There are hundreds of parameters that manufacturers optimize for each and every application and they vary SIGNIFICANTLY from application to application. Anyone who say otherwise simply does not understand how modern ultrasound systems work. Certainly there are some over-arching performance issues which are common to all applications (ie: is the machine "broken") but they aren't sufficient by themselves to fully define performance for all applications and all users. Not by any stretch of the imagination.
- They should be considered separately using different phantoms.
- There're differences in QC criteria because different clinical applications may emphasize different aspects of the system characteristics.
- Requirements for each application are different. For instance, Doppler performance and B-mode performance are not necessarily related.
- No (see above).

A.1.4 Concerning: Computational or human observer

Responses from UK:

- CAD is becoming accepted in other areas and should have a limited role in ultrasound too. Human input also required.
- I would like to see a computerised evaluation system put into place as this removes subjectivity and would also be able to carry out more tests given the same space of time.
- human judgment will always be needed!
- Computational analysis has significant advantages for serial QC phantom measurements. In the assessment of clinical performance for equipment selection I believe that we do not yet have the means to replace the human observer (which is far from ideal). If we can define clinical performance in terms that enable measurement, we may then be able to develop computational methods.
- I prefer a computerised approach. Anyway one can always go back to looking at the image 'manually' if there are doubts.
- Computerised analysis leads to very reproducible measurements but for linking to clinical outcomes human input is essential, and this is unavoidably subjective. Also comparative measurements to assess, for example, if an upgrade has improved imaging need careful human analysis based on some understanding of how measured parameters link to clinical outcome
- I think a mix of observers is best with the most effective being used for each parameter.
- In practice it may be that even an ideal 'computational observer' will require human input. An example is where the SNR is so poor in the far field that the computer throws up false voids. An interpretation is then required.
- When possible, digital images should be captured from the scanner (via a PACS system, a magneto-optical disk or a DVD). These images should then be analysed using a computer (preferably on site, using a laptop). When it is not possible to capture images in this way (e.g. on smaller scanners), the analysis should be performed on site by a human observer. In my opinion, we should not use scanners' analogue video outputs to capture image data - the quality of the captured images just isn't good enough.
- I don’t have any strong preference for human or computational evaluation. I think either may be appropriate, depending on the nature of the evaluation. The choice will depend on the physical quantity being measured, and the limits of uncertainty that are acceptable.

Responses from outside the UK:

- Once objective measures of performance have been standardized, then Computer analysis would remove the major component of the human subjectivity.
- For more objective, I prefer computerised evaluation.
- Computational observer and evaluator using human control of measuring conditions and process.
- SNR measurement offer a objective evaluation together with human observing of defined organs
- Computational observer is ok if validated by multi-observer trials, but it should not be the last instance for decisions
- We need objective measures which means a human observer. Nevertheless, I can imagine that a physician could be easier convinced that his 30 years old machine should be replaced when an example shows him how bad it is!
- The computational observation requires human input. How can we separate them?
- human evaluation
- I think that the human evaluation of ultrasound imagine is necessary because the computerised evaluation can be affect of bugs the system interpretation
- For diagnosis, doctors or sonographers analyze the image by using their knowledge on inside-body. But this type of performance can not be shared well with other users. Hence, my preference is combination of human as well as computerized evaluation.
- The human observer is introducing bias and noise into the performance measurement and should NOT be part of it
- Show 10 sonographers/doctors an image, and you will likely get 10 different opinions as to its performance or image quality. A computational observer would simply be another observer no more right or wrong than one of the human observers. The bottom line is that different people like different things. Some people like compounded images, others don't. Some people like one kind of contrast curve, another likes something different. These differences are based on experience, expertise, perception, cognitive understand of the imaging process, knowledge of anatomy, and so on. Who decides that the "opinion" of a computational observer is better than that of any other human observer. Not me.
- Human evaluation of course. There are other things of importance to the clinical users than the quality of the image.
- Computerized QC is more reproducible and quantitative. However, the current existing automated QC softwares are very limited. There need more studies on correlating the computerized QC findings to problems in system performance characteristics.
- I don't think that a satisfactory computational observer will be developed for years. Currently it is nearly impossible to do reliable automatic edge detection. But human edge detection is pretty good when the edge is perpendicular to the ultrasound scan lines.

A.1.5 Concerning: Is imaging performance linked to safety?

Responses from UK:

- There are at least two points here. Firstly, as noticed above, deteriorating image quality introduces risk of misdiagnosis. Secondly, it's been noticed that improved performance of the scanners is quite often achieved by increasing output power, which again is related to safety.
- Yes because less than optimum image quality may result inadequate diagnosis eg missed pathology
- I think performance is a safety issue; the term fit for purpose should be used here.
- the IEC 60601 is the electrical safety standard written for designers and manufacturers of medical equipment. I don't think it relates to image quality as essential performance parameter.
- I believe that the image quality at acceptance by the user is "essential performance", as the user has determined in some way that the equipment is fit for purpose. A degradation in performance may increase the risk of an incorrect diagnosis or measurement. Adequate imaging performance is a safety issue for ultrasound.
- I think in theory it should. However most clinical decisions and patient pathways are not decided by the results of one imaging modality alone. It really depends on the importance of ultrasound in the patient decision making chain. This particular issue may be more important in Duplex studies where the velocity estimates and plaque identification are used together to arrive at a clinical decision.
- adequate imaging performance is key to the safe use of ultrasound, as it directly relates to the safety of the patient

- Adequate image performance is probably the largest safety issue for ultrasound. Especially with the massive expansion in the casual user where an awareness of acceptable performance for the diagnostic task could be limited. Even for full time ultrasound professionals the monitoring of system quality is difficult in a clinical environment. This nearly always comes to light when new systems are purchased and suddenly a system is unacceptable. How much may be attributed to degradation and how much to the improvements in technology is often unclear.

- Without clear guidance that is enforceable some imagers will be used that are sub standard in terms of IQ due to commercial pressures on the organisation concerned. So yes, inadequate imaging performance should be considered a safety issue for ultrasound.

- Yes, absolutely. However, the importance of image quality depends a lot on the clinical application. A scanner used to detect focal liver lesions in an X-ray Department should perform very well; a scanner used to check foetal presentation in a Delivery Suite does not need to perform very well at all.

- This is a grey area (if you'll excuse the pun!) My own opinion is that a case can indeed be made for adequate imaging performance being considered as a safety issue. For example: in liver & breast imaging, low-contrast lesions below a certain size will not be detected because of the partial volume effect. This will result in a proportion of false-negative results. A machine fault, such as element drop-out, may degrade the beam quality of a scanner and result in a higher percentage of false-negative outcomes than would otherwise be the case. A similar outcome would arise if the clinician used a scanner that was functioning correctly, but whose imaging performance was not "fit-for-purpose" (i.e. did not meet the clinician's expectations, because it was unable to image low-contrast lesions down to a particular size). This is certainly an issue of patient safety and clinical liability.

Responses from outside the UK:

- Only if people want to take ultrasound seriously. If an equivocal ultrasound study makes the physician refer the patient on to a more risky study, then yes - you have now exposed the patient to a risk that he would otherwise have not been exposed to had the ultrasound system provided the appropriate diagnostic image.

- No, the imaging performance not linked to safety.

- Definitely it is! A correct diagnosis is related to the two main factors - skills, experiences and professionalism of the physician and quality of supporting information. The problem is to specify relationship between quality parameters and kind of diagnosis. This is challenge for future.

- Yes, it is necessary to control the performance of scanner and related probes by periodical automated test devices like 3D phantom with data acquisition and process. The manufacturer or distributor should offer SNR measurement of probes by installation at customers office. This is the base for control of performance over time of usage.

- Yes, if imaging equipment is coupled to therapy equipment, e.g. in lithotripsy, HIFU ...

- YES! Good imaging performance means diagnosis safety and this is safety.

- I shall like to say that there can be no compromise on safety in the name of imaging performance. Safety is first. But, then poor quality of image is equally dangerous. Wrong diagnosis (because of poor image) can not be ruled out.

- Perhaps I don't understand the question. I don't think that there are some unacceptable risk when you do an echo graph exam. The quality of image in the exam will be able to give to the patient the complete or partial sentence about the diagnoses

- Most simple way to increase the signal to noise ratio of image is to increase the acoustic output of imaging device. Hence, I think it would be linked to safety.

- If the performance (overall sensitivity, contrast sensitivity) of the equipment/transducer combination is degraded, the user will tend to set the transmit power control always to the maximum, i.e., the ALARA principle will become useless.

- Certainly a device that produces artefacts that cannot be distinguished from clinical pathology could constitute a safety problem and this is covered in the -2-37 "essential performance" list. But I question whether any one person or even committee could define what constitutes a "good" diagnostic image for all applications and all users. At some point we MUST allow the users the freedom to exercise their professional capability. We as a society bestow a certain authority upon medical professionals (due to their training and - in some regions - their licensing)to make risk/benefit decisions that can directly affect the health and safety of their patients. If some regulatory body restricts the use of a particular device by a medical professional who can gain diagnostic value from it simply based on some predetermined set of parameters, is not that restricted use a type of safety concern? Isn't withholding treatment (or making it so expensive as to be unavailable to a percentage of the population) also not a safety problem? No matter how "safe" a device is, if it is unavailable to the population that needs it, we have created a serious unsafe situation.

- Of course.

- Uncertainties in ultrasound studies "due to poor image quality" often lead to more invasive tests which may cause "unacceptable risk".

- Sure, but the examiner performance is the important feature and the instrument performance is not so important. To test this, take patients with a disease (breast cancer or arterial stenosis) and controls. Have them evaluated 4 times: 1) Expert Examiner with Superior Instrument, 2) Expert Examiner with Inferior Instrument, 3) Novice Examiner with Superior Instrument, 4) Novice Examiner with Inferior Instrument. I think that sensitivity and specificity will be in order 1, 2, 3, 4 with a BIG gap between 2 and 3. I also think that the results will be much more dramatic for breast cancer than for arterial disease.

- No

- In general, I don't think imaging performance should be linked to safety. Ultrasound imaging is inherently fuzzy, and there are too many application and patient specific issues to make an objective rating of the risk that the image might be misleading--the user needs to be well trained and make these decision on his or her own. The only exceptions I can think of are: 1) issues related to ultrasound needle-guidance (e.g., amniocentesis), 2) caliper measurements, which can be objectively certified for accuracy.

A.1.6 Concerning: What are the benefits?

Responses from UK:

- It keeps the users on their toes! We have found problems with crystal dropout or decoupling of lens and damage to the cables where no action has been taken until we report. Conversely, two new probes appeared on a scanner using for breast imaging recently because radiologist dissatisfied with images although the old images had passed all the technical tests.

- having recently carried out a baseline test on a new system and found ghosting in one of the probes already signed off by the supplier. It was signed off as the image looked ok clinically. The test phantom showed up the problem the clinical image missed. I think QA testing is worthwhile

- it is in my opinion worthwhile. Certainly there are parameters which need checking routinely such as resolution and caliper accuracy. I have come across brand new scanners with faulty callipers!

- Our current testing adds value. We have very cheap, quick and simple tests for noise and sensitivity that detect real faults and I believe are worth doing monthly. We also have automated image analysis of phantom images (still under evaluation), performed biannually, that demonstrate faults, usually through low contrast penetration or contrast changes. We have published evidence that traditional manual/visual testing with phantoms has no benefit.

- I think it is true that users do often identify faults, however this may be because the tests we use (e.g. test objects) are not more accurate or sensitive than everyday imaging on patients. If we could identify small de-gradations in performance before they are obvious it would be beneficial in terms of financial and replacement strategies. This is partly why I am in favour of developing tests that can be analysed using software techniques.

- QA testing of ultrasound systems is essential. It sets clear reference points as windows between which the machine is deemed to have been operating normally. It does identify faults with the machines and potential faults which may be developing (eg probe front face wear or cable sheathing damage) modified behaviour can significantly extend the life of the system saving money. Regular recorded QA reduces the risk of litigation as it shows a duty of care to the equipment performance. There are large numbers of scanners in use that are used by many people and so the awareness of the normal working state for the machine does not apply.

- Systems are now software based and rapidly developed with updates frequently applied in the field. QA testing often identifies shortfalls in user awareness of equipment functions and capacity (training issues). This results in the systems being used closer to their maximum effect. To focus on QA testing just for the identification of faults is wrong but it does that as well.

- If the right tests are carried out, testing significantly benefits the organisation

- I feel strongly that our QA work is very valuable to the ultrasound users in the North East and Cumbria. We *do* find lots of problems with the scanners in our region, even though many of them are on manufacturers' service contracts. If people doing ultrasound QA aren't finding problems with the scanners that they look after, they're not doing the measurements properly! In my opinion, it is very important that users have access to NHS-based Medical Physics Departments with ultrasound expertise.

- My opinion is that the performance testing of ultrasound scanners can have substantial benefits. Our group delivers a testing programme, involving the measurement of Resolution Integral and associated parameters, to a UK health region with ~100 ultrasound scanners. A database of Resolution Integral measurements is used to define the specification of new equipment, and to inform pre-purchase evaluation, equipment acceptance, post-repair assessments, and twilight testing. Imaging performance assessment is one of the few objective criteria underpinning these activities, and is thus a valuable tool in enabling good clinical and corporate governance. The priority these deserve can be judged by the capital cost of the imaging equipment involved (£6M in our health region) and the volume and nature of the clinical activity (150,000 scans p.a. in the same region, including obs& gynae, gen radiology, cancer, MS, cardiac, intra-op, brachy).

Responses from outside the UK:

- Evidence-based QA testing is very worthwhile and will become ever more important as ultrasound is used in a more quantitative fashion.
- The QA testing is worthwhile. But the QA period can be once a year or half a year.
- Testing saves patients and sonographers from troubles because it decreases probability of wrong diagnosis. It also saves money by eliminating some un-useful examinations which could be done by insufficient sonographer.
- Only tests with standard phantoms are irrelevant, offering optimal axial and lateral resolution and display of defined cysts or other structures. With none of these phantoms it is possible to detect side lobes which affects the image quality.
- Too many insufficient Ultrasound devices are used to make insufficient diagnoses --> Economical problem, but also problem for the patients
- I'm not an opinion about it
- No idea
- Testing not only is important for finding problems with the equipment during its life cycle, but also for checking the results of up-grading and repair, as well as for deciding on the acquisition of new equipment (comparing various brands)
- I believe that regular preventative maintenance is required to ensure that the device is working according to the manufacturer's specifications. And if the manufacturer does not provide a PM protocol, that should be a factor that the buyer should consider when purchasing equipment. Service is also a fundamental part of any system. We are not talking about a television set here, we are talking about a complex piece of medical diagnostic equipment. Just as you consider service when you buy a new car, so too should you consider it for your medical devices. A regular service plan by people trained to detect anomalies in that particular system will ensure that the device continues to perform to the manufacturer's new product specifications. You do regular maintenance on your $50,000 car. Doesn't a $200,000 ultrasound machine deserve the same care?
- At present, I don not think that routine QA testing of ultrasound equipment is worthwhile. The physicists in the USA involved in this area have blindly been fixated on producing a "complete solution" or bible of QA testing. Other technical organizations produce short, specific standards covering individual aspects of equipment performance. And these standards have sunset clauses to limit their existence to a specific time frame. It should be embarrassing that the only ultrasound standard in the USA is the AIUM 100 mm test object that was flawed from the beginning due to mechanical resonances in the 3/4 mm diameter stainless steel rods.
- If a routine QA/QC testing includes: acceptance testing and establishing the baseline performance values, periodic quality control testing to detect changes in equipment performance and verification after corrective actions are taken to fix certain defects, it is very different from a preventative maintenance (PM) program. Some clinical ultrasound users argue that ultrasound QA/QC testing is not necessary because the defects can be identified during clinical evaluation. While it is true for some severe defects, gradual degradations may go un-noticed until the image quality is significantly deteriorated. QA/QC testing is able to reveal "small" problems before they become "big" problems.
- QA testing has never revealed an instrument fault undetected by a sonographer in my 30 years of shallow experience. Objective testing has found many defects in phantoms and test objects. However, I do strive for objective tests.

A.2 Questions on the shortcomings of current practice.

A.2.1 Concerning: Most important shortcoming

Responses from UK:

- results from IPEM report 71 are hard to interpret.
- image contrast is difficult to quantify and is very subjective. should also mention the need for a high frequency phantom(s) to cope with higher frequency probes above 10 MHz
- standard / objective tests seem to be fine (results considered normal/ acceptable) even though operator perceives a problem! standard tests not sensitive enough to detect subtle changes in performance
- Variability between staff undertaking testing, and with the increasing complexity of scanners, it is difficult to ensure identical set-up from year to year. A simple software upgrade can significantly change the appearance of the image, but it is not something that we are always aware of.
- Lack of evidence that phantom tests actually detect changes in performance. The mistaken belief amongst many that measurements in a phantom correlate with clinical performance.
- Subjective

- No direct tests on the transmit and receive characteristics of individual crystal elements. Related to this the lack of knowledge on the beam shape. There are also shortcomings with test objects. The stability of acoustic properties with time and temperature is not well known and this leads to uncertainty in the cause of changes in scanner performance. The acoustic properties of tissue mimics are not well known at the higher diagnostic frequencies used e.g. above 8 MHz. Current tests are time consuming and there are no well defined (nationally agreed?) scanner settings that should be used for common tests. Often a deterioration in clinical performance (as judged by a clinician) is not manifest in TO measurements of resolution, cysts or contrast targets.
- Contrast or Temporal resolution are not provided for in the current range of test objects
- We have now in place a method to measure SNR and can give a measure of artificial cyst detectability. We need a method to look at high echoic structures and to improve what we can offer cardiology.
- Of the above, we can only do (B) properly at the moment. Our spatial resolution measurements use high-contrast objects, i.e. nylon filaments in commercial TM phantoms. In addition, the contrast objects in our phantoms are "too contrasty".
- 1) Limited traceability of the measurements made using tissue mimicking test objects / phantoms. 2) The reproducibility of test object/ phantom characteristics. 3) The time required to carry out testing, in an environment where scanners are often in use 10 sessions per week.

Responses from outside the UK:

- none, we have developed a full range of products that objectively test the performance of both the probe and the system
- Depend on the technician. Two person may provide different results for the same equipment. Even one person may give out different results.
- Mostly I do a PSF test. The most reflected shortcoming is its time consumption – the measurement needs from half till few hours to complete – depends on measuring points density and area measured.
- Spatial resolution doesn't change over time and aging of probes. Only SNR changes essentially caused by side lobes. The side lobes generated from eroding of acoustical lens by usage of aggressive gel and depolarisation of piezo ceramics
- Time consuming Lots of equipment Cannot be done by medical personnel Expensive
- I'm not an hospital doctor, I'm an ultrasound measurement expert
- Our test is only concern the array transducer's performance. The most important shortcoming is that the relationship between the inter-channel uniformity of ERS and Image performance is unclear and much subjective.
- need to adjust TGC settings by user for estimating the penetration depth/overall sensitivity , because not all equipment has the facility of AGC (automatic gain compensation)
- I don't do image performance tests as part of my function with Philips. Nevertheless, the greatest shortcoming is the use of Phantoms as an absolute measure of image quality. Phantoms age and change properties. They do not accurately represent complex tissue structures. What makes a good image in a phantom does not necessarily make a good clinical image for all users and applications.
- Ultrasound phantoms do not adequately test the performance of modern ultrasound equipment.
- Phantom-based tests are not "well-computerized." Phantom manufacturers over design their phantoms, making this difficult. Computer programmers are not well directed to produce software that tests for likely flaws. Phantoms and test objects are expensive, not well designed (too many targets that have no use, for example; what person really, seriously thinks 'dead zone' is a useful test, for example?). Current phantoms provide inadequate surface geometries for curvilinear arrays (ever try looking for non-uniformities/subtle signs of element dropout in a curvelinear array while scanning through the rigid, smooth, flat windows of a phantom?). Phantom manufacturers do not team up with software engineers to provide meaningful objective tests of equipment. Electronic probe testers do not provide capabilities to test all transducers from all manufacturers. Adapters and "probe definition files" are very inadequate. Specular targets provided by probe test manufacturers seem to be adequate for linear arrays but are totally inadequate for curvilinear arrays.
- Due to the lack of standard acceptance criteria, existing guidelines typically suggest the measurements must be compared to baseline or previous measurements in order to validate consistency of the ultrasound system performance. Although this is helpful in monitoring system performance trends, questions remain with regard to how to accurately measure the performance of a system and compare it to others. Where should the action criteria be drawn if indeed variations are observed in QA/QC measurements? Survey studies have been published to represent typical ranges of the QA/QC parameters. However, criteria are usually dependent upon the model and vendor, not to mention the intra- and inter-observer variability.
- Being able to compare the performance between models, scan heads and manufacturers. I'd like to be able to gather B-mode gray scale values from tissue and convert them into echogenicity that could be reproduced from instrument to instrument.

- Probe tests are not built in to the system 3rd party probe testers inadequate for testing a large fraction of probes because of lack of transducer definition files, adapters, etc. Phantom/image analysis software not available.

A.2.2 Concerning: How to address the shortcomings in future

Responses from UK:

- a single figure of merit/graph (resolution vs depth) as produced by the Edinburgh resolution integral would be simpler to explain. as this gives a easily interpreted result and also allows for better comparisons
- image contrast forms the basis of any clinical judgment. As said above it is difficult to quantify, so I don't know how to address this.
- include semi-objective tests that factor in subjective observer assessment - would be helped by electronic image storage facilities that allow standard images taken under controlled / protocolled conditions (with standard phantoms) to be stored and compared under the same viewing conditions, with those taken at later dates.
- I think that the manufacturers should take some responsibility for continued image quality. They could provide a QA program in the software that remains constant, despite software upgrades
- Publish the evidence from the respondents to part 1 that backs up their responses about correlation with clinical performance, results being technically useful and being value for money! Further testing of the various systems in use in the field to assess their value in demonstrating degradation in performance. Try to isolate the most useful variables for measurement - we may be measuring too many interdependent variables. We need to be clear about the purpose of measurements; evaluation of clinical performance and testing for QA purposes are quite different tasks, requiring different approaches.
- A check sheet for each user to fill out from time to time.
- There are some new QA tools for looking at crystal behaviour, however at this stage I am not aware of their success. Measurements of crystal behaviour using in-air measurements with digital image capture and software analysis of the reverberation pattern seem promising. Again this is not an established method. A test phantom with stable acoustic properties would be good. The phantom would have several contrast targets with a variation of 2/3 dB at several depths. It would have a number of cyst targets (1mm diameter) at several depths. It would also be good to have a region of the phantom with an inhomogeneous (back scattering and velocity but well characterised material.
- Computerised analysis
- Design of a suitable contrast resolution test object. Temporal resolution may be evaluated using cineloop capture in conduction with a contrast target movement could be made manually as the sequence will allow timings to be analysed.
- I do not know. One would like to think that leading scientists in this field could cooperate but I am becoming a cynic in my later years and do not believe this to be the case.
- We are starting to use the Sonora FirstCall transducer testing system. We are thinking about moving over to the Edinburgh pipe phantom to measure something akin to (A). We are unlikely to use any of the commercial "spherical cyst" phantoms, or the Satrapa phantom. We are aiming to use a photometer to routinely measure the luminance of scanner monitors. In addition, when possible, we are aiming to do computer analysis of digitally captured images to improve reproducibility.
- Traceability - addressed by establishing the traceability of standard tissue mimicking materials (attenuation, backscatter, B/A , scatterer size and small-scale distribution within the TMM). 2) Reproducibility - partly addressed by resolving (1), and also by quality procedures at manufacture. (3) Difficult to address this problem, given that the quality of images are inherently operator dependent, and top-range scanners have a multitude of controls & applications. May need to continue to plan to make optimum use of the machine-time available.

Responses from outside the UK:

- Develop computer image analyse software may be a appropriate method.
- The measuring time is given by the sonographer frame rate and number of measured points.
- With standard phantoms it is not possible to detect side or grating lobes which effect the image quality. We have brought via TC 87 WG 9 a 3D anechoic phantom as IEC Technical Specification during a meeting march 20th in Stuttgart
- Built-in quality test functions in the devices International standardization --> High volume production of tools --> low prizes
- I'm not an hospital doctor, I'm an ultrasound measurement expert

- Installing AGC facility

- If I knew the answer to that question, we wouldn't be having this discussion I think. Probably the simplest approach is to follow the manufacturer's preventative maintenance steps to make sure the machine is simply not "broken". Are all the channels working? Are the power supplies at their correct voltages? Is the amplifier signal-to-noise ratio within spec? Is the gain performance of each channel within spec? And so on.... When tests get into the subjective nature of what people think is a "good image", performance testing collapses since it becomes a matter of opinion and what a particular person likes or is used to.
- Don't think so. For example, the present phantoms only test the performance of scanners in the "liver" tissue specific mode. Other scanning modes could be compromised and liver equivalent phantoms would not pick up their difficulties. If we had a complete compliment of phantoms for all tissue specific modes, then it would take forever to completely QA a scanner.

- Simpler and better phantoms, which have tightly coupled computer routines for analysis of images. Ultrasound equipment manufacturers provide on-board tests of array channels. Doppler phantoms enable performance capabilities in Doppler and flow imaging modes to be tested in a meaningful fashion.

- More reproducible and quantitative QC test results can be achieved with the aid of computerized analysis of phantom images. The momentum and pace of digital imaging has supported the development of computerized ultrasound QC.

- Frequency, scatterer size, focus and depth independent reflectors, distributed at depths, to compare echogenicity/gray scale curves (gamma curves) between ultrasound systems. I've tried glass phantoms for this with some success, but I'd like to do better.

- Scanner manufacturers provide probe self test diagnostics Phantom manufacturers provide decent software to test signal-to-noise

A.3 Questions on improvements to current practice

A.3.1 Concerning: Main cause of image degradation

Responses from UK:

monitors	probes		
piezoelectric element	CRT monitor lifetime		
coupling in the transducer head			
equipment age	lack of routine maintenance		
software faults / deterioration	changes in operator subjective perception		
Increased Noise	Decreased Penetration		
Probe deterioration	Lack of user awareness of controls	Electronic faults	
Transducer faults			
Age of machine	certain makes		
crystal drop-out	lens de-lamination	crystal degradation leading to beam aberrations	
Wear and tear on machine			
Probe Damage/Faults	Software Errors/Bugs	Setting Errors	
probe defects	monitor failure	system failures	user familiarity or knowledge
Transducer element drop-out	Other transducer faults & "wear and tear"	Scanner port & channel faults	Monitor degradation
The perception of the user (who has just been using a better machine elsewhere...)	Transducer faults (element drop-out, lens/membrane problems, failed multiplexer, break in electrical screen, damaged pins in probe plug, air in mechanical probe head, etc)	Scanner software applications not set-up appropriately (ie not optimised)	Internal hardware faults (Tx or Rx board, etc)

Responses from outside the UK:

Probes			
probe degradation	equipment degradation	monitor degradation	
CRT monitor luminiscence degradation	Transducer elements failure due to cable or connector break due to mishandling or wrong construction.	Ageing of elements	Acoustic lens damage mostly by wrong handling of transducer
difficult patient			

Narrow probe bandwidth(coarse specle)poor resolution mainly azimuthal and elevation, low snr caused from side lobes	degradation of monitors		
Transducer element degradation/failure	Lens (contact) degradation	Insufficient application of coupling media	Inadequate equipment and monitor settings
ageing of screen	defect of single elements in the transducers		
Transducer's Parameters Drift	Pc Monitor		
electrical noise			
Array probe (transducer) performance	Especially, the non(?)-uniformity of echo relative sensitivity among the channels (elements) of the array transducer	Unequalness of output powers from every elements of an array transducer	Unequalness of matched electric impedance values
transducer degradation	settings of pre-processing	settings of monitor	
dead elements, bad cables	ageing monitors	PZT ceramic ageing -> sensitivity loss	operator training
element dropout in transducer			
Transducer degradation	Analog pulser-receiver degradation	Monitor performance	
degradation through usage such as probe defects by transducer drop, cable damage.	presets change unintentionally	display monitor drifts	
Refraction in superficial tissues	Speckle	Attenuation	Image Thickness
Deterioration of transducer performance	Power supply degradation	Inappropriate preset setup	Deterioration of analog preamps, etc
lens damage	cable damage	connector damage	

A.3.2 Concerning: Clinically most relevant image performance measures

Responses from UK:

resolution	penetration	contrast sensitivity	calipers accuracy
low contrast resolution at depth	the high contrast resolution down through the image		
image contrast	resolution	caliper accuracy	
spatial resolution	contrast resolution		
Penetration	Calliper Accuracy	Cystic Lesion Detection	Resolution
Noise	Sensitivity	Contrast performance	Resolution
Contrast resolution & SNR	Lateral Resolution		
every day use			
uniformity of crystal sensitivity	depth of penetration	cyst detection	contrast detail
point spread function			
Contrast Resolution	Temporal Resolution	Penetration	Noise
systems ability to give good image quality	users ability to interpret what a system can do to produce a good image	test procedures from manufactures and QA to ensure the possible IQ is maintained.	
(Ideally) visualisation of spherical cysts	Depth of penetration	Monitor contrast response curve	(Ideally) visualisation of cylindrical (or spherical) low contrast objects
Resolution Integral and derived parameters (Depth of Field / Characteristic Resolution)	Other measures of lateral/elevation resolution	Contrast-detail analysis	Noise

Responses from outside the UK:

Don't forget Doppler!	Objective measures of Probe performance	any parameter that effects the coherent summation of the signals to and from the array elements	-20dB pulse width and the fractional bandwidth of the array. Also the effective dynamic range of the ultrasound system itself
resolution	dead zone	penetration	

Periodical testing using Tissue mimicking phantoms combined with objective method of evaluation. E.g. Thijssen's QA4US method, Satrapa's 3D signal/noise integral method, UltraIQ program from RAMSOFT and transducer parameters evaluation with ForstCalll 2000 from Sonora.	Simple daily routine check out system – like Kollmann's AUStrian test kit or more sophisticated and expensive Sonora – Nickel system, which is limited to check transducer and receiving module only.	Uniformity of channels gain parameters, dynamic focussing stability and affectivity, lines density and lateral and transversal focusses position and resolution evaluation using Point Spread Function analysis	Watching transmitted energy limits and TI and MI accuracy.
Lesion and cyst detectability			
Signal/Noise in anechoic cysts	Point spread function in different depths	Low contrast phantom	Monitor and ambient light (adjustment) quality
Resolution as seen on phantom	Sensitivity of far off wire on phantom		
axial and lateral resolution			
clearance of contrast ratio of the image	accuracy of tumour size detection		
overall system sensitivity varies by application acoustic output	contrast resolution varies by user	contrast sensitivity	displayed dynamic range
Image uniformity in a phantom	Electronic probe tests (such as the Sonora First Call)	Tests of penetration, signal-to-noise	
image uniformity test that reveals array transducer defects, such as element dropouts	spherical lesion detectability	soft copy/hard copy image fidelity	physical and visual inspection of the whole system
Image Thickness	Side Lobes	Transmit Spectrum	Receive Spectrum
Probe tests, done either electronically or with a phantom	For QA: penetration/signal-to-noise		
penetration	contrast	axial resolution	lateral resolution

A.3.3 Concerning: Other types of imaging equipment; adequately tested

Responses from UK:

- Same problems exist, with different groups having different ideas of what makes a good image. There is more active development of test objects for X-ray based imaging to keep abreast of new applications and developments. Ultrasound seems to be more static.
- Other modalities are easier, e.g. an x-ray is just an attenuation map, whereas a US image depends on reflection, scattering, absorption, etc. In other modalities there doesn't seem to be this perception that "physics" testing correlates well to clinical performance; we do constancy checks. If there is a lesson maybe it's that we should focus on testing for QC purposes, move away from trying to mimic the patient and develop systems to measure "physics" variables like resolution in the absence of any specific tissue mimic. Obviously we need to match SoS and maybe attenuation, but not necessarily backscatter.
- Changing from film to digital currently underway, we could learn lessons from image analysis now emergent in ultrasound
- Yes. The legislation requirements for X-ray has meant that greater commercial resources have become available to provide test equipment to meet the challenge. Enforcement is the key to get the commercial interest in providing solutions.
- My understanding of the tests carried out on MRI, diagnostic X-ray equipment and gamma cameras is that they normally provide objective results that (a) can be used to compare different imaging systems of the same modality, and (b) are clinically relevant in terms of setting minimum demonstrable standards of performance.

Responses from outside the UK:

- yes - standards of performance
- Should consider the monitor and hard copy which can influence imaging quality.
- Ultrasound could get an ispiration from nuclear medicine imaging systems.
- Ultrasound is complety different
- Better look at image quality of monitors, More look at ambient light conditions

- Performance tests for different modalities are optimized for that particular modality. Perhaps at a very high level like EMC, or temperature there is some commonality (see IEC 60601-1). But is a contrast test used in X-ray directly applicable to US? Probably not. The people designing these devices have been doing it for a very long time and there is a lot of institutional memory to draw from. Plus there IS a certain amount of cross-technology sharing of best practices. That said, I think that the simple answer to your question is probably "no".

- Yes they are adequate. A lesson would be to just test the machine characteristics that would not be evident to the operators. Proper QA should be ongoing and performed by the operators of the imaging equipment. When I come in to test x-ray machines I concentrate on those machine parameters that can drift without the operators noticing this drift (such as KVp).

- On other imaging modalities, the correlation between QC test results and their effect on images may be better studies and understood. E.g., an enlarged focal spot size will cause more geometrical unsharpness. Most tests are more quantitative and reproducible, e.g., kVp, mAs linearity, output. Ultrasound computerized QC will provide some of those merits.
- Nope, the things that I do are never adequate, I'm always learning.

A.3.4 Concerning: Any comment on the aspect of Performance evaluation

Responses from UK:

- I think the Edinburgh resolution integral is the future of QA in ultrasound. Though it takes a long time to acquire images, computerisation of this test would speed up the process

- It is difficult to assess the image quality of a scanner when using tissue harmonic mode on a phantom. This needs to be addressed.

- We should do acceptance testing of safety critical features, largely measurement accuracy, then choose some performance parameters we can reproducibly measure that can are most likely to detect performance change and use these for constancy checks. We shouldn't try to compare scanners at present - this is a field for research. The lack of evidence base is a real problem; those who strongly believe in what they are doing should disseminate their evidence. Even professional bodies make statements about the efficacy of QA/QC without referencing evidence - very disappointing.

- Test phantoms commercially available need redesigning

- There needs to be a marked improvement in both the scope and application of performance and QA testing. There is a huge amount of money invested in ultrasound each year with a wide range of users some with extensive training some with only a little. Without some form of over arching input advising and monitoring use large sums of money will be wasted. Equipment may be both over and underspecified for use the former being a waste of money and the later potentially representing a clinical risk. A classic example of a monitoring issue is with software options these are purchased with the system often for several thousand pounds. During the life of many software based machines they can require reinstalls of the software. It is not unusual for options to not be reinstalled during this process and for users to be unaware that the option is lost. Particularly where there are a number of system users and a number of differently configured machines.

- Why has it taken so long to get Mr Satrapas work acknowledged?

- I feel very positive about recent developments in ultrasound QA.

- I am a medical physicist specialising in medical ultrasound. However, when it comes to giving a professional opinion about the image quality and performance of scanners, I don't have any widely accepted way of doing this at my disposal. I can stand beside the scanner, watch a few patients getting scanned, and make encouraging noises. I can speak to the manufacturers and get to grips with new technologies and how they work. I can get hold of the probe myself and image various commercial test objects containing filaments and voids. What I haven't got is a standard or widely accepted way of objectively assessing imaging performance. My colleagues who deal with MRI, CT, diagnostic X-ray and nuclear medicine images can all do better than I when it comes to assessing imaging performance. That's a problem for me, and a problem for the health sector generally given that ultrasound scanning is the second most common imaging procedure after plain X-ray. Clinicians often rely on scanning just a few patients before deciding how to spend tens or hundreds of thousands of pounds on ultrasound equipment. That's not a good idea. As a scientist, I am looking for better evidence than this when it comes to choosing imaging equipment. My employer owns ultrasound equipment worth around £6M. Are they getting good value for money? Have they paid over the odds for bells-and-whistles that contribute little to better imaging performance and better patient outcomes? Could they purchase less expensive equipment and still achieve the same level of performance? These are pertinent questions for a health provider currently in a very challenging financial position. So, would better measurement techniques benefit providers and patients? My own view is that they would, because they would underpin the processes of evaluation, procurement and in-life assessment. New techniques need to be able to objectively summarise the imaging performance of a scanner in ways that allow it to be compared with other scanners, as well as with its own performance over time. The ability to do this would also benefit manufacturers, who, like users, have no objective way of assessing the imaging performance of their products. Our group has successfully employed the concept of a "Resolution Integral" measurement for the last 5 years. It has become apparent that the main source of uncertainty in making this measurement is the characterisation of individual test objects. The ability to obtain traceable measurements of attenuation, backscatter, B/A, and small-scale particle distribution within the TMM would add greatly to the robustness and reproducibility of the technique. We see no benefit at all to the user community if this, or any other performance assessment technique, is solely "artefact" based. Good science is underpinned by measurements that are objective and traceable. In ultrasound imaging, objective and traceable performance measurements would ensure reproducibility, and provide a depth of scientific understanding that is sadly lacking from the purely artefact-based ultrasound QA techniques which have been commonplace for many years.

Responses from outside the UK:

- Should consider the monitor and hard copy which can influence imaging quality.
- The QA has to be considered as a regular method of medical instrumentation checking in period of 2 years. Daily routine methods of simple checking of the ultrasonograph status performed by scanner operator (medical doctor or nurse) should be implemented in hospitals as a part of Hospital Quality Certificate.
- We are starting in Germany a large project of testing a high amount of installed scanner with a 3D phantom and automated data processing, processing and documentation. The reason is, that according to MDD (MPG in Germany) there is no request for periodic tests of ultrasound scanner. But we learned, that mainly alteration effects of probes are the main problem of image quality. Also Monitors
- It should be easy to use for every medical person and it should be easy to include in common QA systems
- The most important problem of ultrasound quality assurance is the political one: Nobody is willing to accept that it is really necessary. And the question for us is whether we should further try to convince the others or whether we should use our limited budget for other things.
- Periodicity for performance evaluation must be decided
- The QA of ultrasound equipment should be integrated in a general quality assurance program of the hospital for medical equipment(like in an ISO certification protocol)
- I am very nervous about creating a standard for such a nebulous concept as ultrasound performance. Such standards have a habit of appearing in some well-meaning country's regulatory process with the outcome that either certain diagnostically useful equipment is no longer available in that country or the cost of equipment is increased such that an additional percentage of the population can no longer afford the procedure. Availability of medical services MUST be considered in the safety equation. And it has not to date.
- I try to teach the sonographers to be knowledgeable in ultrasound QA tests and to make their own decisions concerning machine performance.
- Despite the shortcomings many argued, ultrasound QC should be done and its efficacy in revealing "small" problems before they become "big" should be studied.
- We do performance tests of new ultrasound methods, which are often quite helpful, and usually surprising. The results help us to measure such things as displacement sensitivity (currently at about 100 nanometres for signals 60 dB above thermal noise).
- Periodic safety checks should be done to make sure that the system is within safe emissions levels (e.g., acoustic power, Mechanical Index, etc.)