How does the technical status of medical ultrasound equipment affect image quality?

Studies based on human observer experiments and a novel method for automatic detection of defective transducers

Robert Lorentsson

Department of Medical Radiation Sciences
Institute of Clinical Sciences at
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2023

Cover: Two electrical sensitivity measurements of a defect ultrasound transducer (left) and SDR curves of the same transducer (right).

robert.lorentsson@gu.se
robert.lorentsson@vgregion.se

ISBN 978-91-8069-307-3 (PRINT)
ISBN 978-91-8069-308-0 (PDF)
E-publication: http://hdl.handle.net/2077/76803

Printed in Borås, Sweden 2023
Printed by Stema Specialtryck AB

SVANENMÄRKET
Trycksak
3041 0234

To my family, colleagues, and friends.

ABSTRACT

Ultrasound as an imaging modality has a fundamental role in modern healthcare. The equipment, especially transducers, is fragile and handled daily, and transducers are known to sometimes become defective. The overall aim of this thesis was to evaluate how the technical status of the equipment affects clinical image quality.

In study 1 a novel automatic method to detect defects in linear transducers by analysing the images that are saved for documentation was developed which makes it possible to monitor many transducers without interference with clinical activity.

The method was evaluated in study 2 by comparing the result obtained with the automatic method to the well-established in-air method, which is used for quality control, for 81 linear and curvilinear transducers. The methods showed good agreement, the area under curve for the receiver operating characteristics evaluation being 0.88 (SD 0.06).

There is a variety of different scanners, from regular to high-end. In study 3 a comparison of a regular and a high-end scanner was made, the images from the clinically relevant objects in a greyscale phantom being assessed by six observers in a detection study. Visibility was substantially better for the advanced scanner in four out of 16 combinations of depth, contrast and size.

The method from study 1 was also used in study 4 to find images from defective transducers that had been used clinically. 160 images from four defective transducers, with defects that had different grades of severity, were collected, and were compared to 160 images from non-defective transducers in an observer study with four experienced radiologists. There were four mandatory questions concerning whether the defects were detectable, whether the possible defects might affect the diagnosis, the visibility of structural details and total image quality. For three of the transducers the defects were detectable, the visibility of structural details being assessed as worse for all the four defective transducers. The total image quality was assessed as worse for three of the defective transducers. Out of the assessments of the images produced by the defective transducers, 19 % were “confident that the artifacts could affect the diagnosis”.

This thesis shows that differences in image quality between different kind of ultrasound scanners can be detected using human observers. The thesis also shows that clinical image quality can be affected by defective transducers and that there is a risk of misdiagnosis if transducers with severe defects are used. The developed method for automatic detection of non-uniformities in clinical images could be used as a complement to normal quality control for earlier detection of defective transducers.

Keywords: ultrasound, defective transducer, image quality

ISBN 978-91-8069-307-3 (PRINT)

ISBN 978-91-8069-308-0 (PDF)

SAMMANFATTNING PÅ SVENSKA

Användningen av ultraljud som bildgivande modalitet är av fundamental betydelse för modern sjukvård och används inom många områden. Ultraljudsproberna hanteras dagligen och är mycket känsliga för exempelvis stötar. I denna avhandling beskrivs i arbete ett en ny automatisk metod för att upptäcka defekter på linjära och kurverade prober genom att analysera de bilder som lagras för dokumentation på radiologiska avdelningar. Detta gör att man kontinuerligt kan övervaka många prober samtidigt utan att störa det kliniska arbetet. I arbete ett visades att metoden hade visuellt god överensstämmelse med elektriska mätningar som gjorts för tre prober. I arbete två gjordes en utvärdering av metoden på 81 prober i klinisk drift, som referensmetod användes in-air metoden. Resultatet visade på god överensstämmelse mellan de båda metoderna.

Det finns allt från enkla till mer avancerade ultraljudsmaskiner. I arbete tre gjordes en jämförelse mellan en reguljär och en avancerad maskin där bilderna av kliniskt relevanta lågkontrastobjekt i ett gråskalefantom bedömdes av sex observatörer i en detektionsstudie. Synbarheten var bättre för den avancerade maskinen i fyra av 16 kombinationer av djup, kontrast och storlek.

I arbete fyra användes den egenutvecklade metoden för att hitta bilder från defekta prober som använts i klinisk drift. 160 bilder från fyra prober med olika kraftiga defekter valdes ut och jämfördes med 160 bilder från icke-defekta prober i en observationsstudie med fyra erfarna radiologer. De fick svara på fyra frågor som gällde om de såg artefakterna, om de trodde att artefakterna kunde påverka diagnosen, synbarheten för strukturella detaljer samt total bildkvalité. För tre av proberna var artefakterna detekterbara, för alla proberna bedömdes synbarheten för strukturella detaljer vara sämre för de defekta proberna. Den totala bildkvalitén bedömdes vara sämre för tre av proberna. Av bedömningarna som gjordes av bilderna från de defekta proberna som gällde påverkan av diagnos fick 19% bedömningen: *säker på att artefakten skulle kunna påverka diagnosen*. I denna avhandling visades att det gick att detektera skillnader mellan maskiner med hjälp av mänskliga observatörer. Det har också visats att defekta prober kan påverka bildkvaliteten och att det finns risk för feldiagnos om prober med grövre fel används. Den egenutvecklade metoden kan användas som komplement till normal kvalitetskontroll för tidigare upptäckt av defekta prober.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Lorentsson R, Hosseini N, Johansson J-O, Rosenberg W, Stenborg B, Månsson L G, Båth M. Method for automatic detection of defective ultrasound linear array transducers based on uniformity assessment of clinical images – A case study.
J Appl Clin Med Phys 2018; 19:265–274
https://doi.org/10.1002/acm2.12248

II. Lorentsson R, Hosseini N. Månsson L G, Båth M. Evaluation of an automatic method for detection of defects in linear and curvilinear ultrasound transducers.
Phys Med 2021; 84:33–40
https://doi.org/10.1016/j.ejmp.2021.03.025

III. Lorentsson R, Hosseini N, Johansson J-O, Rosenberg W, Stenborg B, Månsson L G, Båth M. Comparison of the low-contrast detectability of two ultrasound systems using a grayscale phantom.
J Appl Clin Med Phys 2016; 17:366–378
https://doi.org/10.1120/jacmp.v17i6.6246

IV. Lorentsson R, Hosseini N, Aurell Y, Collin D, Frösing E, Szaro P, Månsson L G, Båth M. Investigation of the impact of defective ultrasound transducers on clinical image quality in grayscale 2-D still images.
Ultrasound Med Biol 2023; 49:2126-2133
https://doi.org/10.1016/j.ultrasmedbio.2023.06.004

CONTENTS

ABBREVIATIONS

2AFC	Two alternative forced choice
2D	Two-dimensional
3D	Three-dimensional
4AFC	Four alternative forced choice
A-Mode	Amplitude mode
ACR-AAPM	American College of Radiology – American Association of Physicists in Medicine
AFROC	Alternative free-response receiver operating characteristics
AIUM	American Institute of Ultrasound in Medicine
AUC	Area under curve
B-Mode	Brightness mode
BMUS	British Medical Ultrasound Society
CI	Confidence interval
DICOM	Digital imaging and communications in medicine
EFSUMB	European Federation of Societies for Ultrasound in Medicine and Biology
FPF	False positive fraction
FROC	Free-response receiver operating characteristics
FOM	Figure of merit
IEC	International Electrotechnical Commission
IPEM	Institute of Physics and Engineering in Medicine

JAFROC	Jackknife alternative FROC
LL	Lesion localisation
LLF	Lesion localisation fraction
MAFC	Multiple alternative forced choice
MRI	Magnetic resonance imaging
NL	Non-lesion localisation
NLF	Non-lesion localisation fraction
PC	Percent correct
QA	Quality assurance
QC	Quality control
ROC	Receiver operating characteristics
SDR	Systematic dark region
TGC	Time gain compensation
TMM	Tissue mimicking material
TPF	True positive fraction
US	Ultrasound
VGA	Visual grading analysis
VGC	Visual grading characteristics
VNA	Vendor neutral archive

1 INTRODUCTION

Ultrasound (US) has been used in the animal kingdom, for example by dolphins and bats, for millions of years. By sending an ultrasound pulse they can locate prey by calculating the distance using the time taken for the echo to return. This pulse-echo technique is the same one used in modern medical ultrasound systems.

In 1880, Pierre and Jacques Curie observed the piezoelectric effect in crystals of quartz or Rochelle salt. When a pressure was applied to a crystal an electric charge was created, and if a rapidly changing voltage was applied to the crystal the mechanical movements made it vibrate[1]. This discovery could be used to generate and receive pulses just as the dolphins and bats do. This provided the foundation to enable for humans to develop different ultrasound applications. Newman and Rozycki[1] describe developments in ultrasound in the 20th century. Ultrasound instruments for the detection of flaws in metal for industrial purposes were developed in 1941. From the late 1930s onward, various medical applications were tested. In 1937, Karl Theodore Dussik and his brother Friedrich scanned the human brain in an attempt to detect brain tumours. Later, in 1952, it was established that the "hyperphonograms" produced by the Dussik brothers were merely variations in bone thickness and not, as imagined, the lateral ventricles. In the 1950s, many developments took place that were then used in applications in the 1960s and 1970s. For example, the first echocardiogram was produced in Sweden by Inge Edler and Helmut Hertz in the early 1950s.

In modern healthcare, ultrasound is used in several areas. The most familiar use of ultrasound is in obstetrics and gynaecology. Most of us may have seen an ultrasound image of a foetus. One area where modern ultrasound is used is in general imaging, where everything from muscles and ligaments in the foot, leg, arm and shoulder to organs in the abdomen such as the kidneys and liver are examined. Other examinations that are performed in radiological departments (where general imaging scanners are mostly used) include the scrotum and pancreas. Doppler ultrasound is also used to measure blood flow in the heart and in the vascular tree. In clinical physiology, heart function and vascular function in vessels and organs are investigated. Today ultrasound is used for US-guided interventional procedures[2, 3], such as US-guided percutaneous biopsy and US-guided needle and catheter drainage[4]. Contrast-enhanced ultrasound, where micro-bubbles are injected as contrast agents to obtain enhanced contrast in the image, is also commonly used[5].

Ultrasound systems contain sensitive parts and can be damaged by incorrect use and handling. For example, the most fragile part, the transducer, may be dropped or knocked. There is also a risk of the transducer cable being run over by the heavy scanner during transport. If the transducer casing is cracked or the cable cover is broken, there is an immediate risk of electric shock both for patients and for the operator. Cracks in the transducer casing can be small and hardly visible, and it is therefore important that the transducer is immediately checked by qualified personnel if dropped or damaged in any way.

In Sweden there are no mandatory demands from the authorities regarding how often and in what way to perform quality control other than according to the instructions of the manufacturer, often once a year. However, there are recommendations and standards from various organisations of ultrasound experts such as EFSUMB[6] (European Federation of Societies for Ultrasound in Medicine and Biology), BMUS[7] (British Medical Ultrasound Society), IEC[8] (International Electrotechnical Commission), IPEM[9] (Institute of Physics and Engineering in Medicine), ACR-AAPM[10] (American College of Radiology – American Association of Physicists in Medicine) and AIUM[11] (American Institute of Ultrasound in Medicine). The recommendations are similar and only differ on a few points. The tests are divided into different groups, where a more comprehensive check should be performed once a year, and it is recommended that lighter tests are performed more often. It has been shown that US equipment, especially transducers,[12-15] is often found to be defective when checked.

This thesis investigates the relationship between the technical status of the ultrasound equipment and the resulting image quality. The focus is on the topic of a novel method for automatic detection of transducer defects and how image quality can be affected by these defects or by using a simpler scanner instead of a high-end one. The first two papers describe how the horizontal uniformity of the clinical brightness mode (B-mode) images can be monitored. The last two papers describe how image quality can be affected either by using a simpler scanner or by using defective transducers.

2 BACKGROUND

In the following sections an overview of the ultrasound technique is given, but also of other relevant subjects that have been used in Papers I-IV. Paper I concerns detecting defective transducers, and the transducer parts and tests are therefore described in detail. In Papers II, III and IV, different observer study techniques are used, and these are described in the section on human observer studies.

2.1 THE ULTRASOUND EQUIPMENT

Modern ultrasound equipment for diagnostic imaging consists of several essential parts: a transducer for sending and receiving the ultrasound pulses, a beam forming unit, a signal processor unit, an image processor unit, and a display for presentation. The transducers are developed to fit a certain purpose, for example one transducer may be for the abdomen and another for the heart. The scanner that contains the beam forming unit, the signal and image processor units and sometimes an integrated display has software specially developed for the purpose. For example, there are special scanner and transducer combinations for cardiovascular use or for primary care. In Paper IV the scanners and transducers used were developed for general imaging.

2.1.1 PULSE-ECHO TECHNIQUE

In the pulse-echo technique, a pulse is sent from the transducer and the time it takes for an echo to return is used to calculate distance. When an echo occurs, it occurs because the pulse reaches a medium with different acoustic impedance than the previous one, a small amount of the pulse is returned, and the rest continues deeper into the tissue. The bigger the difference in acoustic impedance, the more of the energy is reflected to the transducer. When the difference in acoustic impedance is too large, almost all the energy is reflected, and objects beyond this can be hard to see. Bone and air reflect much or all of the energy in the pulse. All information in the pulse-echo technique is based on borders between tissues with different acoustic impedances. If there is a border between two tissues with the same acoustic impedance, no echo is generated.

There are different ways to present an echo from a sent pulse. In amplitude mode (A-mode), a pulse with an amplitude that corresponds to the strength of the echo is presented on the screen. A-mode is used for a single scan-line.

In B-mode ultrasound, bright points of different intensities are presented on the screen where the ultrasound pulses are reflected, the more powerful the echo the more intense the point. When the time for the possible deepest echo has passed, a new pulse is sent beside the previous one and pulses all the way along the transducer are eventually sent and received (beam by beam technique). Each pulse is sent by a small group of piezoelectric elements. There have recently been some manufacturers that use all elements to send a pulse at the same time to increase frame rate (plane-wave imaging). In the greyscale image, areas where no echoes have returned are black. Small objects, scatterers, much smaller than the wavelength of the pulse, are too small to be reproduced individually. Echoes from those scatterers are very weak. Interference from many echoes from many small objects may, however, be detectable. These are what are known as speckles that appear on the B-mode image.

2.1.2 THE ULTRASOUND TRANSDUCER

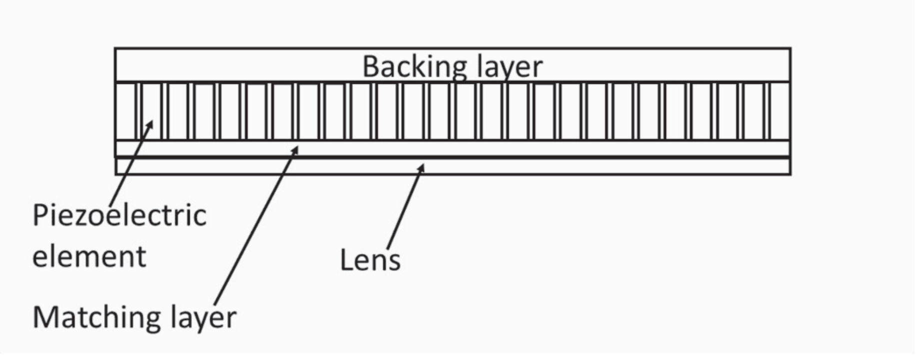

Figure 1. The internal parts in an ultrasound transducer.

The transducer is the part of an ultrasound system that converts electrical energy to mechanical ultrasound waves, normally pulses in the range of 1 to 20 MHz[16] for conventional transducers, depending on the application. Different transducers and frequencies are used for different applications. The higher the frequency the better the resolution but at the expense of penetration depth. There are even specialised ultra-high frequency transducers in the range of 30-100 MHz[17]. Transducers for diagnostic medical applications consist of the same basic components: a backing layer, piezoelectric elements and a matching layer. Normally there is a lens in front of the matching layer, see Figure 1.

2.1.2.1 BACKING LAYER

When an electric pulse is applied to a piezoelectric element, undesirable ringing occurs when the element vibrates for a longer time than wanted. To reduce the pulse duration and the spatial pulse length, a damping material is applied on the side of the piezoelectric elements that is not facing towards the patient. This is called a backing layer and is a mixture of metal powder and a plastic or epoxy resin[18]. The disadvantage of the backing material is that it reduces the amplitude of the pulse and hence the efficiency of the transducer.

2.1.2.2 PIEZOELECTRIC ELEMENT

In the transmitting phase, a voltage is applied to the piezoelectric element and a mechanical wave is transmitted into the patient. In the receiving phase, the returning echo is converted from mechanical energy back to electrical energy in the same piezoelectric element. In modern ultrasound transducers, materials such as lead metaniobate, lead titanate, zirconate or barium titanate are used for the elements[16, 19]. These ceramics are not naturally piezoelectric but are heated in a strong electric field to achieve piezoelectric characteristics. The type of transducer and its application determines the number of elements and their arrangements. There are typically hundreds of elements in linear and curvilinear transducers. In modern three-dimensional (3D) matrix transducers there can even be thousands of elements[20].

2.1.2.3 MATCHING LAYER

The difference in acoustic impedance between the piezoelectric element and the tissue of a human body is large. This means that most of the transmitted energy would be reflected if the transducer had no matching layer between the piezoelectric element and the human body. The matching layer often has several layers with a gradual transition in acoustic impedance. The matching layer often has a thickness of ¼ of the wavelength to reinforce the amplitude of the resultant pulse. Outside the matching layer a lens is usually incorporated for focusing purposes.

2.1.2.4 CABLES AND CONNECTORS

The cable that connects the elements to the connector consists of many small conductors. When there is a wide transducer with a large number of elements a multiplexer is sometimes used to control the elements. This means that two (or more) elements can be connected to the same conductor in the main cable but in different time slots, since they are not activated at the same time (for example for linear array transducers that use traditional the beam-by-beam technique). Otherwise, there are (at least) as many conductors in the main cable as there are elements in the transducer. The connectors are mostly either pintype or flat-type and are often locked with a handle to keep a stable pressure. There are normally several ports on the ultrasound machine, so it is easy to switch between the connected transducers.

2.1.2.5 TRANSDUCER TYPES

Most of the transducers are two-dimensional (2D) transducers that produce a 2D image and have a thickness in the elevational plane as thin as possible. Linear array transducers are straight transducers using many elements in a row, a small group of elements being activated for each beam along the transducer (except for plane-wave imaging). The image for a linear array transducer is a rectangular image as wide as the width of the transducer. Curvilinear transducers also have a row of many elements, but the array is curved. These transducers are used for deeper examinations, such as of the abdomen. Since the beams are sent along a curve, the resulting image is wider at the bottom than at the top. The top of the image is as wide as the width of the transducer. The internal of a curved linear array is shown in Figure 2.

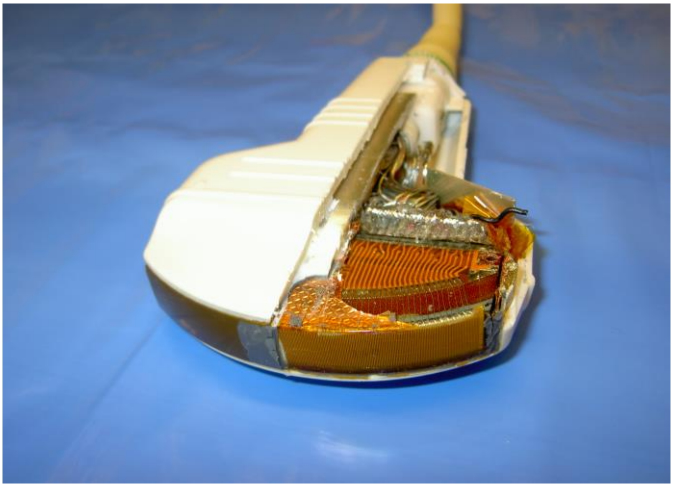

Figure 2. A curved linear array transducer with the housing sawn apart. Photograph by Benny Stenborg (reproduced with permission).

Phased array transducers have a small linear array but instead of activating a small group of elements at a time like the linear array, all or almost all the elements are activated with a small time difference between the elements. In that way, the beam can be steered in different directions. Since the array is very small, the beam is almost steered from one point, and the beams can pass between the ribs. This transducer type is used in particular for heart examinations. The image produced by a phased array transducer has the same shape as a fan.

There are also several specialised transducers such as endo-cavity transducers, which can be radial and produce circular images. 3D transducers have either a mechanical sweep or a matrix array to obtain a 3D image.

2.1.3 TRANSDUCER DEFECTS

In 2008 Mårtensson et al.[12] investigated 676 transducers that were in clinical use from 7 manufacturers in 32 hospitals using an electric transducer tester, FirstCall, (Sonora Medical Systems, Inc., Longmont, CO, USA). The maintenance that was used by the departments on a regular basis was either regular service by the manufacturer or checks of the scanners by the test protocols of the scanners. Despite this, 269 transducers (39.8 %) were found to be defective when the Sonora FirstCall test system was used. Delamination was the most common defect (n=179, 26.5%), followed by break in the cable (n=57, 8.4%). Twenty-three (3.4%) of the transducers had a short circuit, while weak (10-75% of sensitivity) and dead (below 10% of sensitivity) elements only were found in 6 (0.9%) and 4 (0.6%) transducers respectively.

Two hundred and ninety-nine transducers that were classified as operational one year before and had been in clinical use were tested with FirstCall equipment in a follow-up study. Eighty-one (27.1%) of the examined transducers were found to be defective. The distribution of the error types was similar to the previous study: delamination and cable break were the most common defects. These two studies focus only on the transducers and not on other parts of the scanner that can partly cause the same disturbances in the image if broken, for example channel or port connection problems. Other studies that also include the scanner have been performed[14, 21].

Hangiandreou et al.[15] describe a clinical ultrasound quality control program. Forty-five scanners and more than 256 transducers were included over a period of four years. The check-up points in their quality control program were maximum depth of penetration, mechanical integrity, distance measurement accuracy and image uniformity. In the mechanical integrity evaluation, careful inspection of all components such as transducers, scanners and scan table took place. To ensure that there was no risk of electrical hazard, the transducer housing and cable were examined. In image uniformity evaluation, the image of a tissue mimicking phantom was made[22]. The in-air reverberation method[23-25] was also used. The origin of the problems with image uniformity may be cable break, weak or dead elements and delamination. These problems are all situated in the transducer. There may also be oxide or other contact issues in the contact between the transducer and a scanner port. Internally in the scanner there are, for instance, channel faults that can be present in an image uniformity test or be shown in an in-air image. Hangiandreou et al. found that the distance measurement accuracy did not result in any deviations in the four years, and for the maximum depth of penetration there were only three (1.6%) detected failures. The most effective evaluations for finding defective equipment were image uniformity, 124 (66.3%), and mechanical integrity, 47 (25.1%). The number of defects in the equipment detected by clinical personnel was 13 (7.0%). Among all 187 detected failures, only 22 (11.8%) were related to the scanners, the remaining 165 (88.2%) being related to the transducers. Their conclusion was that uniformity evaluations and mechanical integrity were most efficient in finding defects, and their recommendations were to perform these tests quarterly.

There are many ways a transducer can become defective, the following defects being the most common ones according to Mårtensson et al.[12] and Hangiandreou et al.[15]

- *Mechanical integrity*. Cracks can cause leakage current into the patient or the sonographer. Since the excitation voltage of the piezoelectric elements is rather high, for example >100 V_{pp}[26], the current can be dangerous. If a transducer has cracks or holes in the housing it is important to immediately take it out of use and perform an electrical measurement of the isolation. Physical damage of the cables can occur if they are run over by the heavy scanner.
- *Delamination.* If the glue that binds either the lens and matching layer or the matching layer and the piezoelectric elements comes loose, a small gap of air occurs. Since the acoustic differences between air and the human body are large, a large proportion of the transmitted signal is reflected even before it enters the skin layer of the patient.
- *Break in the cable*. If a multiplexed transducer is used, two or more elements can be affected by one conductor break. A break in a conductor makes excitation of the connected element(s) impossible. Cable breaks are often intermittent and can depend on the position of the cable. When it is checked whether a transducer is functional, a wiggle of the cable is appropriate. To verify whether it is a cable break, a measurement of the capacitance between two conductors can be made; this is done in an electric transducer tester.

- *Short circuit*. When the conductors to an element are unintentionally connected before the element.
- *Weak element*. When a piezoelectric element is functional both in transmitting and receiving phase but the response is weaker than normal when measuring the sensitivity in the receiving phase.
- *Dead element.* No response from the element at all.

2.1.4 QUALITY CONTROL AND QUALITY ASSURANCE

ISO 9000, consisting of several standards, defines quality assurance (QA) as “part of quality management focused on providing confidence that quality requirements will be fulfilled” and quality control (QC) as “part of quality management focused on fulfilling quality requirements”[27]. Quality assurance is a method to avoid mistakes and failures in products and comprises the actions to provide trust that the service satisfies the specifications. Quality control refers to a process to ensure certain measurable items that can be compared to a defined set of criteria to maintain quality over time[28]. This may, for example, be different aspects of image quality for ultrasound. QC checks if the quality of the product reaches expected standards.

There are some examples of guidelines regarding quality assurance for ultrasound, e.g. IPEM[9], BMUS[7], EFSUMB[6], and AIUM[11], while ACR-AAPM[10] has focused on a guideline regarding quality control. Test techniques should be objective measures of performance, be relevant to the clinical application, be reproducible over a long time and be so sensitive that changes are detected before they become clinically relevant[29]. Absolute and relative controls are available. In the absolute controls the result of the measurement can be compared to another machine or compared to a purchase specification. The relative controls are only suitable for deciding whether the performance of one specific machine has changed from one time to another.

2.1.4.1 ABSOLUTE MEASURES OF TRANSDUCER PERFORMANCE

There is a wide range of ways to measure performance of an ultrasound machine, from using phantoms and subjective measurement[15, 22, 30, 31] to computerised measurements of a digitised image[32-37]. The performance parameters that are usually measured are[29]: axial, lateral and slice thickness resolution; noise; penetration depth; geometric and measurement accuracy; and contrast resolution.

To measure these parameters using test objects, tissue mimicking material (TMM) phantoms are used. These phantoms must have a speed of sound in the material that is the same as that used by the ultrasound machine when measuring depth in a patient, 1540 m/s[38]. The attenuation must match the attenuation for soft tissue at the frequencies that are used. The attenuation is normally 0.5 and 0.7 dB/cm/MHz[39]. The properties should be stable for a long period of time and not change much when the ambient temperature changes. The values of the parameters that are measured should be reproducible and not depend on the operator. Echo patterns from the objects inside the TMM should be similar to those from soft tissues in human patients. Graphite powder is often used by the phantom manufacturer to create a speckle pattern. The resolution is tested by scanning nylon filaments that are moulded into the TMM. Spherical or cylindrical objects in the phantom serve the purpose of simulating regions with less echogenic characteristics such as cysts. Spherical objects are preferable since cylindrical objects are normally longer than the slice thickness of the beam and the visibility of a cylindrical object can be better than for a spherical (cyst) object [40-45]. The penetration depth depends on frequency, and the phantoms are often designed to cover the lowest frequencies (1-5 MHz), which means a depth of 15-18 cm. When the speckle becomes noise, it indicates that the penetration depth has been reached. Contrast resolution measurements are done on objects in the TMM phantom that have small differences in echogenicity. The objects are often divided into 3 dB steps and can be either more or less echogenic than the surrounding material.

There are at least two companies that have developed similar equipment for a comprehensive evaluation of the transducer, FirstCall and Probehunter (BBS Medical AB, Stockholm, Sweden). The transducer head to be examined is placed in a water tank with the scanning surface facing a metal target. The target is flat for linear array transducers and curved for curved array transducers. The type and model of transducer decide the depth of the target. The connector is connected to a box on which different transducer adapters fit, and the box is connected to a computer. Adjustments are made to keep the beams perpendicular to the target. The elements are activated one at a time, the echoes are recorded and sensitivity can be calculated for each element. With Firstcall this is done once, but Probehunter keeps sending pulses continuously so that the cable can be wiggled to discover cable defects that only appear in certain cable positions. Capacitance tests are also performed to test the conductors in the cable. Both Probehunter and FirstCall produce diagrams that are presented for every element that shows pulse width, centre frequency and fractional bandwidth. Some of the elements are selected to show their pulse waveform and frequency spectrum. A sensitivity bar graph and capacitance bar graph for a transducer with 7 defective elements from a FirstCall measurement are shown in Figure 3.

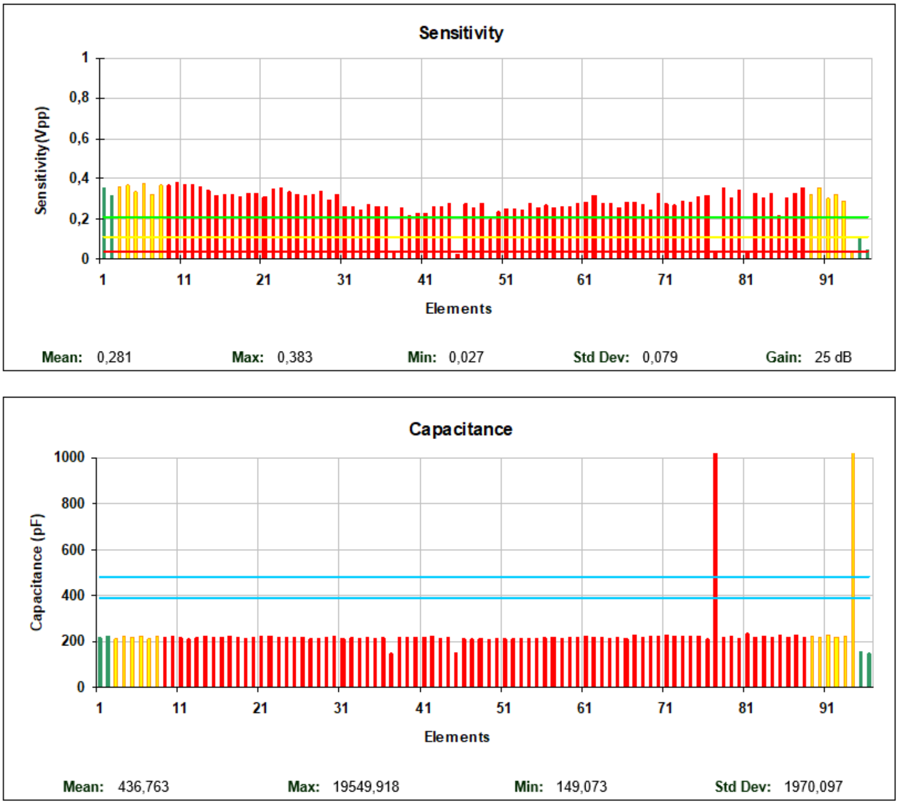

Figure 3. A transducer report from an electrical measurement with several weak elements, some of which are due to cable problems.

2.1.4.2 RELATIVE MEASURES OF TRANSDUCER PERFORMANCE

There are several ways to perform relative measures of transducer performance. A simple but effective check to detect defective elements (or broken conductors) in a transducer is the "paperclip" method[46]. The transducer scan surface is covered in a thin layer of gel, and a paperclip or other thin metal object is moved in the gel while the machine is operating in air, and the B-mode image is observed. The signal will bounce back and forth between the paperclip and the piezoelectric elements and produce a strong pattern just under the paperclip if everything is normal. If instead one or several elements are weak or dead the pattern becomes weaker or disappears just when the paperclip is passing that place.

The in-air test is simple and easy to perform and needs no extra equipment besides the ultrasound machine and the transducers. The transducer to be examined should be free from gel, clean, dry and be placed in its holder so that the transducer face operates in air. A pre-set normal clinical setting is chosen, and the gain is set to a value to optimise the image of any defects in the reverberation lines[25]. Time gain compensation (TGC) is set to the default position. Harmonic imaging, spatial compounding and other advance image processing are disabled. The air-image that appears on the screen shows the reverberations between the different layers (matching layer and lens) in the transducer. If something is defective, such as dead elements or broken cables, this appears as dark streaks in the image originating from the transducer head. For delamination errors, a brighter area appears before the dark streak. The air test is suitable for beam-stepping arrays (where the beams are activated by groups of elements sequentially along the array). A different method can be performed for phased arrays[47]. As well as crystal dropout and uniformity, the in-air test can be used for sensitivity and noise. How deep the reverberations reach shows the sensitivity when the gain is set at max. Adjustment of the gain and TGC are done to see where the noise disappears in the image just below the deepest reverberation. An example of an in-air image can be seen in Figure 4.

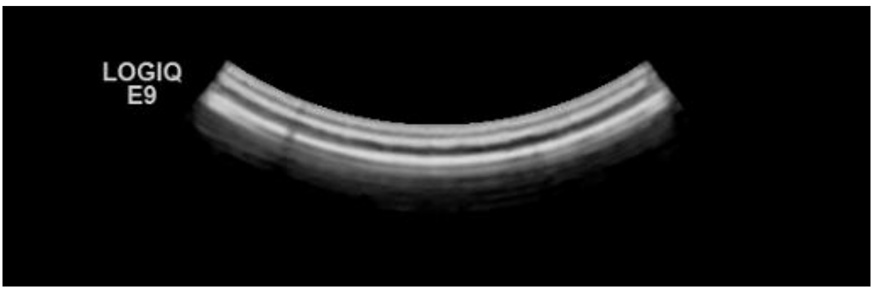

Figure 4. An in-air image of a curved array transducer, showing a defect on the left side.

Another common way to verify transducer function is to use a TMM phantom and scan a uniform part of the phantom and identify deviations from the tissue mimicking texture. The gain and TGC is adjusted to obtain a brightness that can be compared to a clinical image and spatial compounding, harmonics, etc. is turned off. Vertical darker bands are signs of transducer defects. Horizontal darker bands can indicate focus problems or problems with the circuits[22].

2.1.5 SAFETY

Ultrasound is generally considered to be a safe modality, however there are some safety aspects to consider. There are two risks in particular that come with the use of ultrasound: heating and cavitation. Both depend on the settings on the ultrasound machine. There are therefore two indexes that are presented at the ultrasound screen: thermal index (TI) and mechanical index (MI)[48, 49]. For TI there are three different indexes depending on which tissue is examined, the soft-tissue thermal index (TIS), the bone-at-focus thermal index (TIB) and the cranial thermal index (TIC). The TI is a guide to how much the tissue is supposed to rise in temperature when thermal equilibrium is achieved. If TI is 2.0, then the tissue can be expected to rise 2 °C at the warmest point along the ultrasound beam if the tissue is exposed for a longer time.

MI is a measure indicative of the risk of cavitation, for example, when the gas bubbles violently collapse in the tissue[50]. There are recommendations regarding whether or for how long it is suitable to scan with certain values for TI and MI. This is especially important for obstetric examinations and scanning of the eye.

Other recommendations to keep ultrasound examinations safe are the guidelines from various organisations, The British Medical Ultrasound Society recommends[51] that the only purpose to use medical ultrasound should be medical diagnosis. The people who use the ultrasound equipment should be fully trained in safe and proper operation and they should be able to appreciate the potential mechanical and thermal bio-effects of ultrasound. They should also have a full awareness of equipment settings and understand the effects on power levels that are caused by scanner settings. The examination times should be kept as short as necessary to produce a useful diagnostic result and the operator should aim to keep the scan times within those recommended by the BMUS. Output levels should be minimised while still achieving a meaningful diagnostic outcome. Finally, pregnancy scans for the sole purpose of souvenir images or videos should not be carried out.

2.2 HUMAN OBSERVER STUDIES

In medical imaging, image quality evaluation is an important matter. For image modalities that use ionising radiation, where optimisation is an important issue, image quality evaluation is essential. For ultrasound the absolute measurements that are carried out during quality assurance and quality control, such as axial and lateral resolution, are examples of physical measurements for image quality evaluation. These physical evaluations are important, although, in the end, there is always a human observer interpreting the images that are crucial for the diagnostic outcome. Therefore, human observer studies are also important to perform when comparing image quality in different modalities.

There are many types of human observer studies, and some of them will be discussed here. Human observer studies can be divided into two major branches, observer performance studies and visual grading studies[52]. Observer performance studies measure the ability of the observer, to find pathology while visual grading studies concern the opinion of the observer of different aspects of image quality. Observer performance studies are generally more accepted since they are objective and there is an established ground truth. Visual grading studies are subjective and dependent on the opinion of the observer. However, visual grading studies are frequently used, and it has been suggested that they may be as adequate as observer performance studies when it comes to evaluating image quality[53].

2.2.1 OBSERVER PERFORMANCE STUDIES

2.2.1.1 MAFC

One observer performance experiment that is commonly used is the multiple alternative forced choice (MAFC) experiment[54-56]. The simplest of MAFC experiments is the two alternative forced choice (2AFC) detection task. Two images are shown side by side to the observer, one of which contains a signal, position and shape which are known, plus noise, and the other one containing just noise. The location of the signal image is randomly assigned to one of the image locations. The task for the observer is to choose which of the images contains the signal. The observer must choose one, hence the F in 2AFC for forced. This is repeated for many image pairs and P, the proportion of correct responses can be decided and after that d`, the detectability index can be determined[57].

In a MAFC setup there are M images. M-1 images contain only noise and only one of the M images contains the signal. The location of the signal is randomised and the prior probabilities are usually equal[57]. Normally a reference image, that contains an object of the same contrast and size as the signal, is presented as an aid to the observer in knowing what to look for[57-59].

2.2.1.2 ROC

Receiver operating characteristics (ROC) originates from World War II, when the ability of radar operators to discover enemy aircraft was tested[60]. When the gain of the received radar signal was turned down to zero no signals (aircraft) were detected. If the gain was turned up, the possibility of detecting signals increased, but also the amount of noise that could be mistaken for real signals. A radiologist has a similar task to the radar operator, but the task here is to find pathology in a lumpy background and to distinguish it from normal variations.

ROC was introduced for medical imaging more than 50 years ago[61]. The observer is asked to answer whether an image contains pathology; the answer must be rated on a scale from, for example "I am confident that there is no pathology" to "I am confident that there is pathology", typically with several steps in-between. Each scale step corresponds to a specific decision level, describing the confidence of the observer, for which the observed signal is between two specific decision thresholds (Figure 5). Based on the distributions of ratings for a set of images, pairs of sensitivity and specificity are calculated for the different decision thresholds. The pairs are plotted as an ROC curve in an ROC diagram, where the X-axis represents the false positive fraction (FPF) (1-specificity), while the Y-axis represents the true positive fraction (TPF) (sensitivity). (The fractions are sometimes referred to as the false positive rate and the true positive rate.) An ROC curve describes how well an observer can distinguish a normal case from a non-normal case, for example. An example of different decision levels, the true healthy and diseased distributions and the corresponding ROC curve is presented in Figure 5.

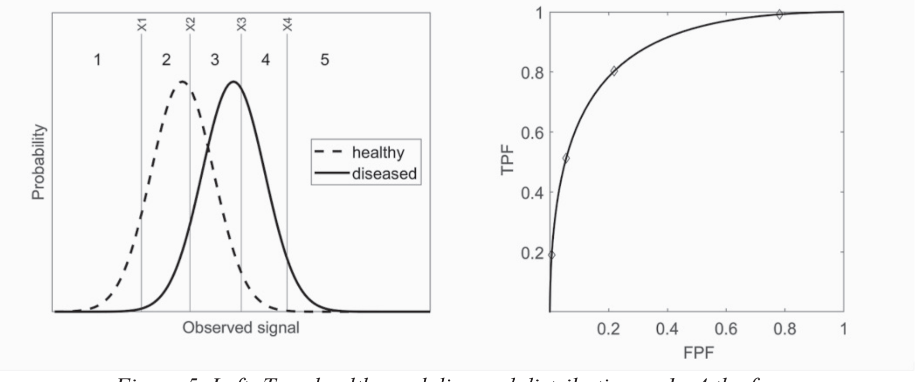

Figure 5. Left: True healthy and diseased distributions, x1-x4 the four decision thresholds for the five answer options. Right: The corresponding ROC curve and the four points that are used to draw the curve based on the four decision thresholds.

If the observer is no better than chance, the ROC curve becomes a straight line between origin and (1,1). The area under curve (AUC) for ROC shows the skill of an observer where everything above 0.5 is better than guessing and 1 is perfect. By using ROC analysis, different modalities can be compared to each other, for example. A delicate problem is when the AUC becomes numerically equal, but the shape of the curve is different, which can mean that one of the modalities is better for a lower decision level and the other for a higher decision level. It has been shown that P in a 2AFC experiment is numerically equal to AUC in the corresponding ROC experiment[62].

A widely known issue with ROC is the fact that the observer is restricted to answering how convinced the observer is whether there is pathology in the image or not, irrespective of the location of the pathology. There is an obvious risk that the observer looks at the image and interpretates an area as pathological, but it happens to be in an image where the pathology is located in a different place than the observer believes. The outcome of this is a correct answer but the observer had the wrong position in the image in mind when answering positive. The false negative and the false positive "cancel" each other and the result is a true positive. As described in the next section, various strategies have been developed to address this problem[63].

2.2.1.3 FROC AND AFROC

In a free-response receiver operating characteristics (FROC) study the number of lesions can be zero or more and the problem mentioned for ROC is taken care of. The observer marks suspected lesion(s) in the image material. The marking must be near an actual lesion to be counted as a lesion localisation (LL) – a true positive – alternatively it becomes a non-lesion localisation (NL) – a false positive. The statistical power is higher in FROC than in ROC [64]. One of the reasons for this is that the possibility of getting a right answer while looking in the wrong part of the image is removed, another reason is that the localisation results in more data.

Bunch et al.[65] defined the FROC curve as the graphical representation of LL fraction (LLF) at the Y-axis versus NL fraction (NLF) at the X-axis. The NLF can be large and extend the X-axis far. Observers, modalities and tests may differ between different experiments, and based on the FROC curve, it is hard to find a figure of merit (FOM) since observers may find different number of NLs (there is no upper limit). The alternative FROC curve (AFROC), is another measure. This is the plot of LFF versus FPF. Both LFF and FPF are probabilities, and the curve therefore fits in the unit square (only the NL with highest rate of the observer is used to calculate the AFROC curve). For FROC experiments, the AUC for the AFROC curve can serve as FOM, just like the AUC is used for ROC experiments[66, 67].

One method for statistical analysis of FROC experiments is Jackknife alternative FROC (JAFROC). JAFROC is used mostly for analysis of multi-case, multi reader FROC data[68]. As FOM, AUC_{AFROC} is used. In JAFROC the FOM is repeatedly recalculated using jackknifing; for each calculation one case is excluded[69]. In this way an estimation of the uncertainty of the FOM is obtained

Normally ROC and other observer performance techniques are used to measure human observer performance. They can also be used for measuring the performance of numerical observers[70]. In Paper II, ROC and AFROC were used to measure the performance of the automatic method from Paper I in terms of finding defects in ultrasound transducers.

2.2.2 VISUAL GRADING STUDIES

ROC and MAFC studies can be troublesome and time-consuming since a ground truth must be established. For radiological images there are European guidelines on how important anatomical structures of the body should be reproduced to be good enough for a correct diagnosis[71]. Visual grading is an established method for evaluating image quality in radiological images – especially those mentioned in the guidelines since the structures are clinically relevant and have a relevant anatomical background[72]. There are extensive descriptions on how to perform visual grading studies[52, 73-75]. The relation between visual grading and ROC has been investigated, and the correlations for the FOM for the two methods were sometimes high[53, 76, 77]. In visual grading analysis the observer can be asked about how clinical landmarks are visualised in an image. The observers are asked to rate their assessments of the visibility of the landmark on an ordinal scale. The European guidelines are produced in a way that makes it is possible to optimise the radiation to a level where the radiation is as low as possible, and the image is still usable for the diagnostic purpose. Visual grading analysis (VGA) is suitable for this purpose[78, 79]. Where radiation is not an issue, for ultrasound for example, visual grading can be used as well but there are no strict guidelines to follow regarding how an image should be evaluated. Examples from magnetic resonance imaging (MRI) and ultrasound where VGA is used are Zarb el al.[80], Ulm et al.[81] and Cüneyitoğlu el al.[82].

Båth and Månsson have developed a method called visual grading characteristics (VGC)[72], which handles the ordinal data from a visual grading study in a correct manner. Two modalities a and b can be compared to evaluate which one is assessed as having the best image quality. The images from a and b are scored by the observer(s) in several steps on an ordinal scale and two image quality distributions are formed. The VGC curve is plotted based on these two image quality distributions in the same way as an ROC curve is based on the signal (disease) and noise (no disease) distributions. The area under the VGC curve (AUC_{VGC}) is used to decide which of a and b is best. If the confidence interval (CI) for AUC_{VGC} does not cover 0.5, a and b can be separated, otherwise they cannot be separated.

3 AIMS

The overall aim of the thesis was to evaluate the relationship between the technical status of an ultrasound machine and the resulting clinical image quality produced by it. The use of ultrasound as an imaging modality is of fundamental importance to modern healthcare. In clinical use, a deterioration in image quality is often difficult to assess because gradual deterioration and wear of an ultrasound transducer is difficult to detect with the naked eye. However, it is known that deterioration can lead to reduced diagnostic options[83]. The thesis also aimed to improve the options for identifying ultrasound equipment that needs repair or replacement. This has the potential to lead to improved diagnostics, for example through reduced risk of incorrect clinical assessments due to imaging artifacts and better utilisation of existing equipment.

The four specific aims were:

Paper I: To describe the concept of a novel method that uses clinical B-mode images from linear transducers for automatic detection of defective transducers.

Paper II: To evaluate the performance of the automatic method developed in Paper I against an established method of detecting defective transducers.

Paper III: To compare the ability of two ultrasound systems to reproduce clinically relevant low-contrast objects using a commercially available phantom and considering human observer variability and other methodological issues related to observer performance studies.

Paper IV: To investigate whether defective transducers affect image quality and the risk of misdiagnosis in clinical images.

4 SUMMARY OF PAPERS

4.1 PAPER I

4.1.1 CONCEPT

In Paper I a novel method for detecting transducer defects by analysing clinical images was developed. The idea for the method occurred when another method of detecting transducer defects was being used. This method involves slowly moving a linear transducer (in gel) along the arm and watching the dynamic B-mode image at the same time. With a moving background it is easier to see possible dark vertical streaks (dark regions near the transducer) than in a static image. The darker streaks remain at the same places while the tissue in the arm is moving. When using this detection method, the question arose as to whether it was possible to use several different static clinical B-mode images as a "dynamic" background instead of a moving arm. Some simple tests were done using a linear transducer. The linear transducer had some plastic tubes on some of the pins in the connector to simulate broken elements or conductors. Several images of a general-purpose phantom were collected. A median image of the collected images was constructed and the darker streaks from the disconnected elements appeared clearer than in the single images. When the same was done with images on a human object, the result was the same or even slightly better.

4.1.2 IMAGE COLLECTION

In Paper I the development of the method and its implementation on images in the region of Västra Götaland was described. The goal was to collect images produced by a certain transducer and place them in an image stack. From this stack a median image should be constructed and the dark regions near the transducer detected. To make the method functional for images collected from a large image archive there were certain issues that needed to be addressed:

- Ensure that a certain image is taken by a certain transducer or transducer/scanner combination
- Ensure that only images that use the full width of the transducer is used
- Extract the B-mode images from the surrounding background
- Address the issue that images from linear transducers have different proportions depending on depth
- Ensure that no spectral flow images is used

The vendor-neutral archive (VNA) in the region of Västra Götaland is used for storage of all kinds of digital imaging and communications in medicine (DICOM) images, and all ultrasound images must be in the DICOM format to be stored in the VNA. To ensure that a certain image was taken by a certain scanner the DICOM tag Station Name was used. For some scanners this tag was empty and there were two machines that had the same Station Name, otherwise the names were unique. To ensure that an image was produced by a certain transducer/scanner combination, the Station Name had to be correct, but the transducer model also had to be the same as the one to be checked. Several of the manufacturers used the DICOM tag Transducer Data for description of the transducer model. GE (GE Healthcare, Milwaukee, WI) did not use the Transducer Data DICOM tag for the scanners used at the time, so for the GE transducers the information in the images concerned was used instead. The small part of the image where it says which transducer model is used was compared with part of images with different models and the best match decided the transducer model for the image concerned. Different scanners stored the images in different size formats, but it turned out to be a functional solution. Using the Station Name in combination with the transducer model did not address the fact that a specific transducer could be moved between scanners. This solution was as close as possible to identify an individual transducer as long as the manufacturer did not have a unique identifier such as serial number for the transducer in a DICOM tag.

To ensure that the full width of the transducer was visible in the image, and that no horizontally zoomed images were used, all transducer models had their physical width recorded, to which the width in the image was compared. Other data stored for each transducer model were the number of elements. The number of elements was used for the number of columns in the median image.

4.1.3 EXTRACTION OF B-MODE IMAGE

The images stored in the VNA did not just have the B-mode image in the frame, but also the surrounding information such as patient data above the B-mode image and settings on the sides. To extract just the B-mode image, the connected non-black areas were chosen and the largest one was the B-mode image. This was extracted and put in an image stack.

The fact that linear array images of different depths had different proportions was addressed by spatially scaling all extracted images to the same size. With an image stack filled with B-mode images produced with a certain scanner and a certain transducer, it was possible to create a median image from a given number of images in the stack.

The median image can be used to visually assess the dark streaks near the transducer, just like in a TMM phantom which is one way to check the horizontal uniformity[22]. These dark streaks visible in the median image were called systematic dark regions (SDR).

4.1.4 THE AUTOMATIC ALGORITHM

To automatically identify the systematic dark regions an algorithm was constructed with a goal that the output should be a curve (SDR curve) with peaks at the positions where the dark regions were positioned. The automatic algorithm was designed based on three paths illustrated in the flow chart in Figure 6.

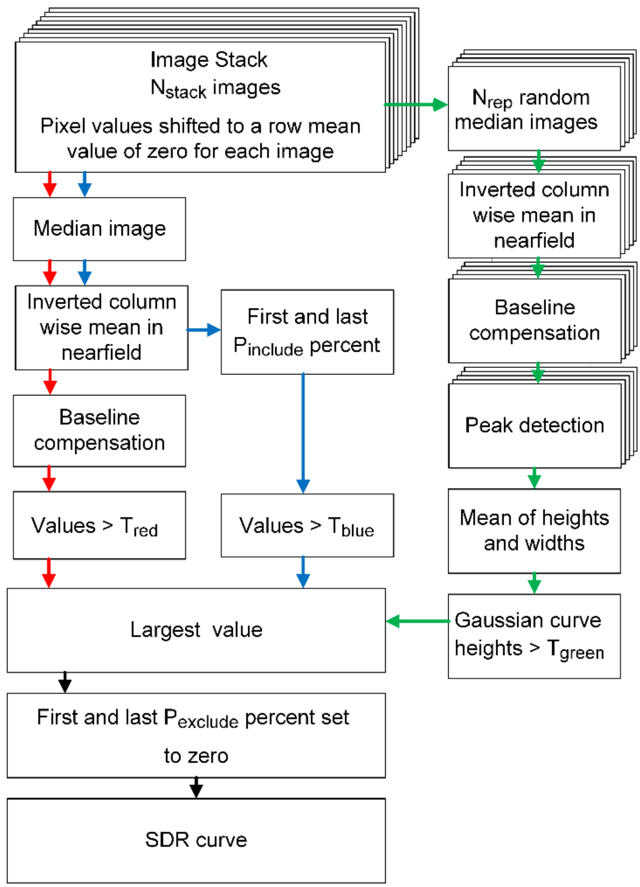

Figure 6. The different paths in the construction of an SDR curve from a stack of N clinical B-mode images. (Adapted from Paper I.)

The most obvious path was to take the top of the median image and check whether any pixels were darker than the rest (red path in Figure 6). To do this, the row mean was subtracted for every row (or at least the top ones) in every image in the image stack. These shifted images were used to calculate a median image. The top of the median image (19 rows out of 500 used in Paper I) was selected and inverted to get darker values at the higher amplitude. For the selected rows, column-wise means were calculated. The curve was composed of a vector with the same number of elements as the number of transducer elements. The baseline of the curve sometimes drifted; this drift was the result of anatomical variations in the images. Defects are often confined to a limited part of the transducer, and the dark part is surrounded by brighter parts. Therefore, a baseline compensation was made for this path of the algorithm. The baseline compensation was a subtraction of a polynomial fit of the curve. In Paper I the degree of the polynomial fit was 6. This first path was suitable for finding medium-sized defects. For defects that had a high amplitude, a second path was constructed where the central of the curve was used without baseline compensation (blue path in Figure 6). For narrower defects, such as single defective elements, a third approach was taken (green path in Figure 6). The large stack was divided into smaller groups of images, randomly chosen without replacement. The idea was to check for every smaller group whether the defect was there and then to check the other groups. As for the first path, the pixel values were shifted to a row mean value of zero for all images. A column-wise mean was calculated for a given number of pixels in the top of the images, then the signal was inverted, and baseline compensation was performed. These steps were similar for the smaller groups as for those for the larger group described in the first path. For the smaller groups, a peak detection that returned the local maxima was performed for each of the curves from all groups. The amplitudes and the widths of all peaks were stored, and all peaks that had an amplitude over a certain value were chosen. From these chosen peaks, new peaks were constructed and added to a clean signal that was zero if no peaks were found. The highest value of the three paths was chosen for every position of the curve. Finally, the beginning and end of the SDR curve were set to zero since the images were darker there with non-defective transducers.

The method was applied to old images from three transducers that had been found defective when electrical measurements (FirstCall) were done. New SDR curves were calculated for every newer image that replaced the oldest in the image stack that consisted of 150 images (in Paper I). By working this way and showing the SDR curves in a 3D plot, it was easy to follow when the defects appeared and compare the defects to the FirstCall measurements. The day of the measurement was known, and the SDR curve when the oldest image in the stack had the same date as the FirstCall measurement was used for comparison.

To exemplify how an automatic alarm could be set, the area under the SDR curve was calculated for the three cases. When the area under the SDR curve reaches a specific level, an alarm could be activated that the transducer should be further checked. A user interface in the form of a console was developed. A screenshot from the console can be seen in Figure 7.

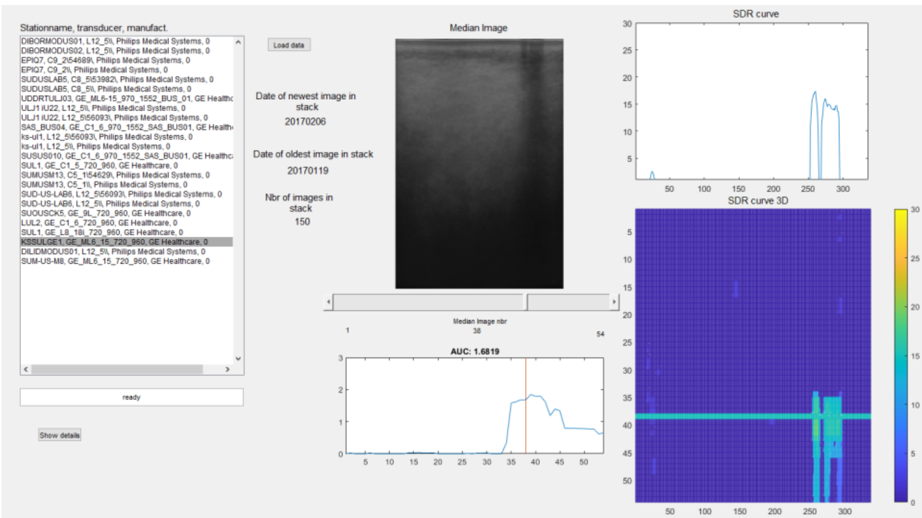

Figure 7. A screenshot from the console used to manage the information from the SDR curves for several transducers, left: scanner/transducer, middle: median image(top) and AUC_{SDR}, (bottom), right: Single SDR curve (top) SDR curves 3D over time (bottom). (Adapted from Paper II)

4.1.5 RESULTS

Images that did not fulfil certain criteria were rejected by the algorithm that extracted the B-mode image. The non-used images were either Doppler-curve images or rejected by the extracting algorithm due to discrepancy of the image width and transducer width. The number of used images compared to the total number of images for the three cases was 94%, 96% and 66% for Cases 1-3 respectively. The low percentage for Case 3 was partly due to a disturbing logo. The SDR curves are shown in Figure 8 for the three cases.

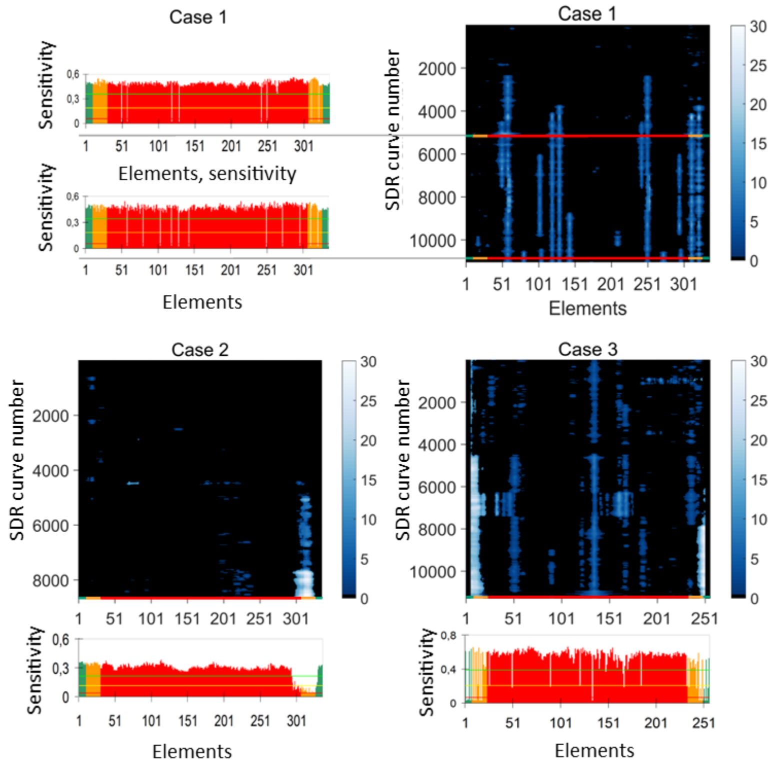

Figure 8. 3D plots of the SDR curves and the element sensitivity reports from FirstCall measurements for Cases 1-3. The corresponding time points for the four FirstCall measurements are marked in the SDR curves with multicoloured lines. The SDR curves are presented with the lowest curve number for the oldest curves. (Adapted from Paper I)

The time periods that were analysed were approximately 30 months, 25 months and 55 months for Cases 1-3, respectively, with one new SDR curve for every new approved image. (The oldest SDR curves are the ones with low SDR curve numbers in Figure 8.) This was done retrospectively, and as ground truth for comparison there were four electrical measurements of the transducers, two for Case 1 and one each for Case 2 and Case 3. These electrical measurements were compared to the SDR curves corresponding to the time points the electrical measurement was conducted. In Region Västra Götaland, quality control in most hospitals is performed once a year, either by performing an electrical measurement or by checking the transducer (and scanner) using other methods. In Figure 8 the advantages of having continuous monitoring can be seen. For Case 2, for example, the slowly graduating defect was detected by the method approx. 9 months before the defect was actually detected by an electrical measurement.

The four comparisons to the electrical measurements visually corresponded well to the SDR curves in the three cases, as can be seen in Figure 8. Both narrow defects such as single elements as in Case 1 and greater spread that covers several elements as in Case 2 were detected by the algorithm. Over time, the same defects were detected in the SDR curves although the image material was completely replaced every 150 images.

Paper I was a case study exemplifying the automatic method, for three transducers, although no evaluation other than visual comparison between FirstCall measurements and SDR curves was performed. From now on, the automatic method will be referred to as the SDR method in this thesis.

4.2 PAPER II

In Paper II a more thorough evaluation of the SDR method of detection of defective transducers developed in Paper I was made. The arrangement was to compare the result from the method to the result from an established method used to detect defects in transducers. In this paper curvilinear transducers were also included.

4.2.1 MATERIAL AND METHODS

An inventory of the scanners and linear and curvilinear transducers that regularly stored their images in the VNA of Västra Götaland was conducted. A total of 152 linear and curvilinear transducers used by 37 scanners were identified. To compare the SDR method to an established method, the in-air method was chosen to be the ground truth. This is an established method that has been recommended for QA for many years[9]. The in-air method is a subjective method but had the advantage that it could be used for all linear and curvilinear transducers, which was not the case for the electrical measurements. Of the 152 transducers, 71 were excluded and 81 remained, see Figure 9.

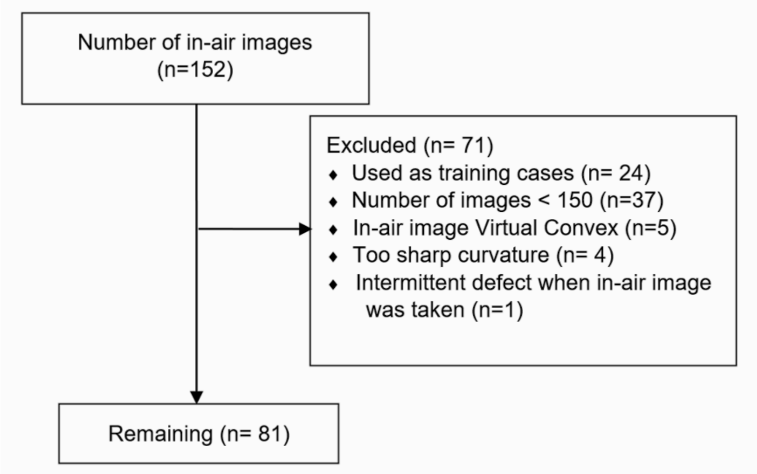

Figure 9. Exclusion reasons for the transducers in the study described in Paper II.

Two observers trained their skills in assessing in-air images by comparing 24 in-air images to electrical measurements that were performed on the transducers. These in-air images were excluded, like the ones for the transduces that had not produced 150 images or more in the last 9-12 months (the SDR method required 150 images to produce one SDR curve according to the values used in Paper I). The observers marked at the top of the in-air images on the remaining transducers where transducer defects were found. This was used as the ground truth when the performance of the SDR method was evaluated.

The SDR curves were updated once a day and not once for every new image as in Paper I. The reason this was done was to simulate a real situation where it is reasonable to update the SDR curves once a day. If many images were produced that day the impact on the curve was greater than if few images were produced. All the parameters chosen in Paper I for calculating the SDR curve were the same in this study. To approve an artefact from the SDR method to be an artefact, there had to be a signal for 20 consecutive days around the date for the in-air image. To evaluate the result, an AFROC analysis was conducted, this was a good method since the defects had locations on the transducers and there could be several in the same transducer. The FOM from an AFROC study may be difficult to interpret, and a classic ROC study was therefore also carried out. When a signal was categorized as a defect, the amplitude was recorded to be utilized as input for the ROC and AFROC analyses. The software Rjafroc (Pittsburgh, PA) v1.2.0.9000 was used to calculate both the FOM for ROC and

AFROC and the associated curves. Rjafroc is a statistical software; available from https://dpc10ster.github.io/RJafroc/index.html. Both for ROC and AFROC the AUC was used as the FOM. The closer the AUC is to 1.0 the better the agreement between the SDR method and the reference method.

There was also a test of different parameter settings in the SDR method. For the evaluation of the method the same settings as in Paper I were used, but additionally some settings were changed to evaluate how this affected the result. The depth that was used originally, 19 of 500 pixels, was changed to 30 of 500. The image stack was tested to have fewer images, down to 50 from the original 150. Baseline compensation was done with a polynomial degree for two of the three paths in the algorithm for the SDR curve. In Paper I a degree of six was used, in Paper II a degree of three was also tested.

4.2.2 RESULTS

The FOM for the ROC evaluation, when the ground truth was the assessment of the collected in-air images, was 0.88 (SD 0.06). Since several defects could be present at one transducer (the observers of the in-air images marked as many defects as they detected and marked them at the right location), the AFROC curve was also evaluated, since AFROC takes location and whether there are more than one defect into account. The FOM for the AFROC evaluation was 0.71 (SE 0.07). For the 69 transducers that were assessed as healthy by the two observers using the in-air method, the SDR signals were negative in 60 cases and false positive in 9 cases, resulting in a specificity of 87%. Twelve in-air images were assessed as being produced with defective transducers; the SDR method detected 8 of them, giving a sensitivity of 67%.

Different parameter values were tested and changed the result in different ways. The change of amount of depth in the images that was used from 19/500 to 30/500 caused the AFROC FOM to be significantly smaller (0.64, $p = 0.014$). Using 50 images in the stack instead of 150 resulted in some false positives that lasted for one day or a few days. Since the limit of 20 consecutive days for the SDR curves was set, the short false positive signals did not affect the AFROC FOM more than when there were 150 images in the stack. The change of the polynomial degree from six to three for the baseline compensation did not affect the AFROC FOM either.

4.3 PAPER III

In Paper III a comparison between a general and a high-end scanner was made to evaluate whether there are differences in their ability to reproduce clinically relevant low-contrast objects (e.g. lesions) embedded in a homogenous organ in the abdomen. The clinical background was simulated by a TMM greyscale phantom, which also contained low contrast objects. The images of the low-contrast objects and from the background were acquired from the phantom. An observer study was conducted to compare the ability to reproduce low-contrast objects for the two scanners.

4.3.1 SETTINGS

The two machines that were compared were one general GE Logiq LP5 (LP5) and one high-end GE Logiq 9 (L9). The same transducer was used for both scanners, an Abdominal transducer C4, multifrequency (2-5 MHz). The settings for the scanners were chosen to be the default abdominal settings for each scanner. These are presented in Table 1.

Table 1. The settings on the L9 and LP5 when the images were acquired. (From Paper III)

Setting	L9	LP5
Acoustic power output	100%	100%
Compounding	Low	None
Fundamental or harmonics	Fundamental	Fundamental
Smoothing	None	Low
Frequency	4 MHz	5 MHz
Speckle reduction imaging	3	0
Greyscale map	D/0/0	C/0/0
Dynamic range	72	72

4.3.2 IMAGE ACQUISITION AND PREPARATION

The phantom used in paper III was a greyscale phantom CIRS 047 (Computerized Imaging Reference Systems, Norfolk, VA). Three sizes of cylinders, 2.4, 4 and 6.4 mm, were situated at gradually increasing depths. The finest cylinders had a possible scanning depth from 1 to 6 cm and the thickest from 3 to 12 cm. Seven contrast levels were available for every size: anechoic, -9, -6, -3, +3, +6 and +9 dB. The speed of sound in the TMM phantom was 1540 m/s, and the attenuation was 0.5 dB/cm/MHz. A router table was used to move the phantom in fine steps under the transducer which was fixed in a holder above the phantom. Water was used as coupling medium. Images were collected from two depths for each of the two larger objects at the contrasts -6 dB, -3 dB, +3 dB and +6 dB. 30 images of each object size and depth were acquired for both scanners, 240 images in total. For the 4 mm objects, the depths used ranged from 35 to 42 mm and 90 to 98 mm. As for the 6.4 mm objects, the depths utilized ranged from 53 to 64 mm and 110 to 121 mm. The phantom was moved in both directions (15 mm×3 lateral positions and 2 mm×10 along the cylinders) to obtain fairly independent images for both the speckle and the objects. The images were stored in DICOM format and were moved to a computer. The objects were cut out of the images as a square and together with three squares of just speckle to create 4-AFC image sets. The location of the objects was randomised. At the top of each 4-AFC image set there was a reference square with a computer-generated object of the same contrast and size as the actual one as an aid for the observer[57-59].

4.3.3 OBSERVER STUDY

Six observers who had experience of ultrasound imaging performed the four alternative forced choice (4AFC)-study consisting of 960 image sets, and the instruction was to decide which one of the four squares that contained the object from the phantom for every image set. A DICOM-calibrated screen, EIZO Radiforce RX320 (EIZO Corporation, Ishikawa, Japan), was used in a room where the lighting was kept at a constant low level. ViewDEX[84-86] was used for presenting the image sets in a randomized sequence for each observer and for collecting the answers, see Figure 10.

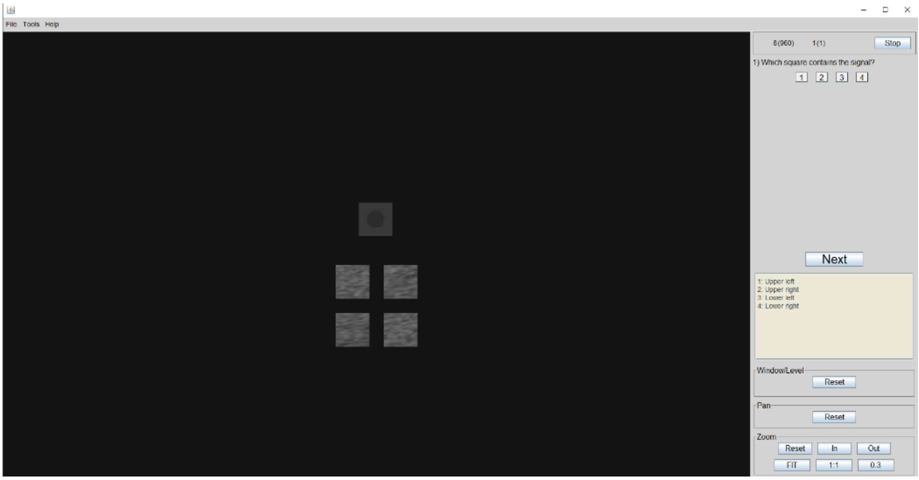

Figure 10. Screenshot of the presentation of the 4AFC images in the software ViewDEX. The software was also used for collecting the answers from the observers.

4.3.4 STATISTICAL ANALYSIS

The FOM that was used was percent correct (PC) averaged over the six observers for each combination of 30 images for every combination of machine, size, contrast and depth. To determine the uncertainty of FOM bootstrapping was used. The resampling technique created several resamples from the original sample (with replacement), and a bootstrap distribution was constructed from 10000 resamples. The 2.5 and 97.5 percentiles of the bootstrapped distribution were used for 95% CI.

4.3.5 RESULTS

There were 32 combinations of depth, contrast size and scanner that were compared. For one of the objects, the visibility was so high at the depth of 53-64 mm that all observers answered correctly for both scanners; it was the 6.4 mm object that had a contrast of -6 dB. The proportion of correct answers is shown in Table 2. For four of the 16 combinations, the difference between the two scanners was significant, all in favour of the high-end machine, and all objects were located in the superficial part of the phantom (in the superficial range of the actual object). To investigate whether expected differences showed significance, comparisons were made between the contrast levels ± 6 dB and ± 3 dB while the other parameters were fixed. All these 32 contrast differences showed significant differences. These were comparisons that were made to confirm that the used method worked and showed differences where the signal had different strength. Comparisons of signals with the same strength of positive and negative contrast levels were also made. The result of these comparisons showed that three out of 16 had significant differences.

*Table 2. The percent correct answers for the two machines at different contrast, sizes and depths. P-value <0.01 is marked *. (Adapted from Paper III)*

Con-trast dB	Object size mm	Depth mm	L9 P (95% CI)	LP 5 P (95% CI)	Significant difference
6	4	35-42	0.98 (0.92 - 1.0)	0.90 (0.78 - 0.98)	*
		90-98	0.76 (0.57 - 0.91)	0.82 (0.63 - 0.96)	
	6.4	53-64	0.99 (0.95 - 1.0)	0.97 (0.9 - 1.0)	
		110-121	0.93 (0.83 - 0.99)	0.92 (0.83 - 0.98)	
3	4	35-42	0.83 (0.68 - 0.94)	0.66 (0.51 - 0.78)	*
		90-98	0.50 (0.36 - 0.66)	0.55 (0.37 - 0.72)	
	6.4	53-64	0.89 (0.69 - 1.0)	0.76 (0.53 - 0.94)	*
		110-121	0.42 (0.31 - 0.54)	0.51 (0.39 - 0.63)	
-3	4	35-42	0.59 (0.47 - 0.71)	0.55 (0.43 - 0.68)	
		90-98	0.58 (0.42 - 0.73)	0.57 (0.46 - 0.69)	
	6.4	53-64	0.84 (0.74 - 0.93)	0.81 (0.67 - 0.92)	
		110-121	0.54 (0.39 - 0.69)	0.50 (0.34 - 0.64)	
-6	4	35-42	0.97 (0.92 - 1.00)	0.89 (0.80 - 0.96)	*
		90-98	0.76 (0.63 - 0.87)	0.88 (0.79 - 0.95)	
	6.4	53-64	1.0 (1.0 - 1.0)	1.0 (1.0 - 1.0)	
		110-121	0.86 (0.73 - 0.96)	0.91 (0.82 - 0.97)	

4.4 PAPER IV

Paper IV concerns how transducer defects affect the clinical B-mode images. Images that were produced with defective transducers were selected and compared to images produced with non-defective transducers. Four experienced radiologists conducted an observer study including questions about image quality and whether the artifacts from the transducers could affect the diagnosis. The SDR method was used on the VNR by analysing over one hundred transducers during a period of several years. In that way, clinical images from transducers that had been in use although they were defective could be found. Of the four transducers that were selected to be part of the study, three had severe defects and one had four defective non-adjacent elements. The transducer with four defective elements had been approved in that condition by the manufacturer at the time of the previous annual check. The transducers were Philips (Philips Healthcare, Amsterdam, Netherlands) 12-5 linear transducer for Case 1 and Case 3, Case 2 was a GE ML6-15D multirow transducer and Case 4 was a GE curvilinear transducer C1-6-D. The scanners were Philips EPIQ 7 for Case 1, and Philips iU22 for Case 3. Case 2 and Case 4 were both GE LOGIQ E9. The number of piezoelectric elements that were defective was estimated from the median images based on 150 images. Cases 1-4 had 4 of 256, 17 of 336 (in central row), 26 of 256 and 22 of 192 defective elements respectively. For Case 4 the elements were adjacent in two clusters, and for Cases 1-3 the defective elements were scattered. The four transducers had been in clinical use for several months even if they were defective.

4.4.1 IMAGE COLLECTION

Forty images from each transducer were selected from when they were defective, and 40 images were selected from when they were not defective or from another equivalent scanner/transducer combination that was non-defective. To determine whether a transducer was non-defective the SDR method from Paper I (with the same settings) was used, and if there were no signals above zero in the SDR curve, the transducer was classified as non-defective for the examined period. All the transducers were used in radiological departments. The images were chosen to be from the different Study Descriptions (DICOM tag) that were mostly used for the different transducers. Ten images from each of four study descriptions for both defective and non-defective transducers were taken to obtain similar comparisons. The study descriptions are presented in Table 3.

Table 3. The study descriptions (DICOM tag) for Cases 1-4 (Adapted from Paper IV)

	Study description 1	**Study description 2**	**Study description 3**	**Study description 4**
Case 1	Thyroid	Shoulder	Scrotum	Lower extremity
Case 2	Bone, joints and discs	Soft parts of the neck	Scrotum	Arm
Case 3	Thyroid	Shoulder	Scrotum	Lower extremity
Case 4	Bone, joints and discs	Hip	Abdomen	Kidneys

One image per examination was chosen in the majority of the 320 instances. In the rest, two or more images were collected from the same examination. Since defective transducers can be intermittent, a visual check that there were dark streaks right under the transducer in the chosen images was made. When there were several images to choose from in the same examination, an image was chosen where the defect was clearly apparent.

Examples of clinical images from the four cases are shown in Figure 11A-D.

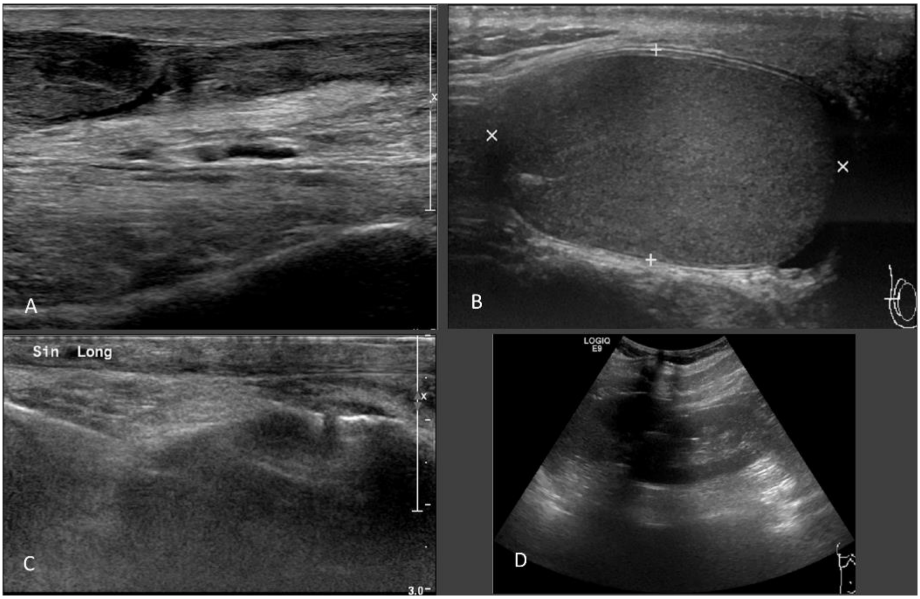

Figure 11. Examples of images produced by the defective transducers. A was acquired with Case 1 and show a low extremity image. B:is a testis image from Case 2. C is a low extremity image acquired with Case 3, D shows a kidney image from Case 4. (From Paper IV.)

4.4.2 OBSERVER STUDY

Four experienced radiologists participated in an observer study where there were four mandatory questions. The questions and the options for the answers in Paper IV were:

1. Do you see regions with altered greyscale level that emanate from the transducer (top of the image) and that you take for artificial artifacts?

1. Confident that there is at least one such region
2. Somewhat confident that there is at least one such region
3. Indecisive whether there is such a region
4. Somewhat confident that there is no such region
5. Confident that there is no such region

2. Could these artificial artifacts, if any, affect the diagnosis if the artifacts would coincide with the region for the diagnosis?

1. Confident that the artifacts could affect the diagnosis
2. Somewhat confident that the artifacts could affect the diagnosis
3. Indecisive whether the artifacts could affect the diagnosis
4. Somewhat confident that the artifacts could not affect the diagnosis
5. Confident that the artifacts could not affect the diagnosis or that there are no artifacts

3. How do you assess the visibility of structural details?

1. Poor
2. Restricted
3. Sufficient
4. Good
5. Excellent

4. How do you assess the overall image quality?

1. Poor
2. Restricted
3. Sufficient
4. Good
5. Excellent

ViewDEX[84-86] was used to display the pseudonymised images in a randomised sequence and to collect the answers. The observers were unaware of whether the images were from healthy or defective transducers. A DICOM part 14[87] calibrated medical display EIZO Radiforce RX 350 was used. Question one was formulated to be evaluated as an ROC study where the artificial artifacts were positive signals. Question 2 was presented as number of scores for the different answer options for the images produced with the defective transducers. Questions 3 and 4 were evaluated as visual grading studies. The evaluation of Questions 1, 3 and 4 was done using the software VGC Analyzer release 3[88-90]; the results were valid for fixed-reader analysis.

4.4.3 RESULTS

Question 1 was evaluated as an ROC study and concerned whether the artificial artifacts in the images caused by the transducer defects were detectable by the observers. The ROC curves are shown in Figure 12.

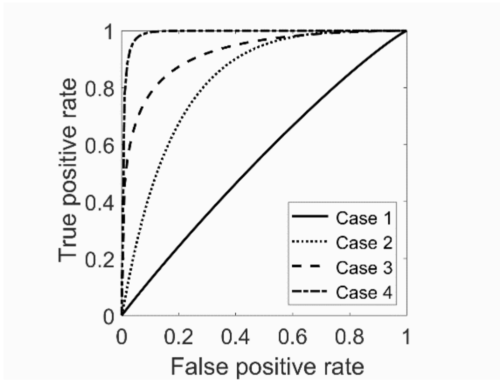

Figure 12. The ROC curves regarding Question 1, detection of artifacts. (Adapted from Paper IV.)

Table 4. The AUC_{ROC} for Question 1 and the 95% CI in brackets.

	AUC_{ROC}
Case 1	0.56 (0.50, 0.61)
Case 2	0.80 (0.74, 0.86)
Case 3	0.91 (0.88, 0.94)
Case 4	0.96 (0.94, 0.98)

The artifacts were detectable in the images by the observers in Cases 2-4 but not in all images. The AUC_{ROC} are presented in Table 4.

Question 2 was about whether the diagnoses could be affected by the artificial artifacts. The options ranged from 1 “Confident that the artifacts could affect the diagnosis”, to 5 “Confident that the artifacts could not affect the diagnosis or that there are no artifacts”. The question was formulated in such a way that neither ROC nor VGC was an appropriate evaluation method. Figure 13 shows the answers for Question 2 from all observers for the images produced with the defective transducers. In total there were 121 assessments out of 640 for the four cases that received answer no. 1– that the observer was confident that the diagnosis could be affected if the artifact and the area for the diagnosis coincided.

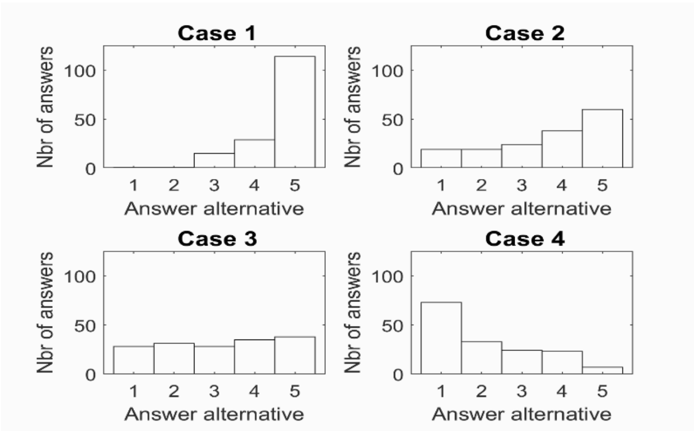

Figure 13. The distribution of answers for the option for Question 2. The lower the number, the more confident the observer was that the artificial artifact could affect the diagnosis. Only the answers from the images produced with the defective transducers are presented. (Adapted from Paper IV.)

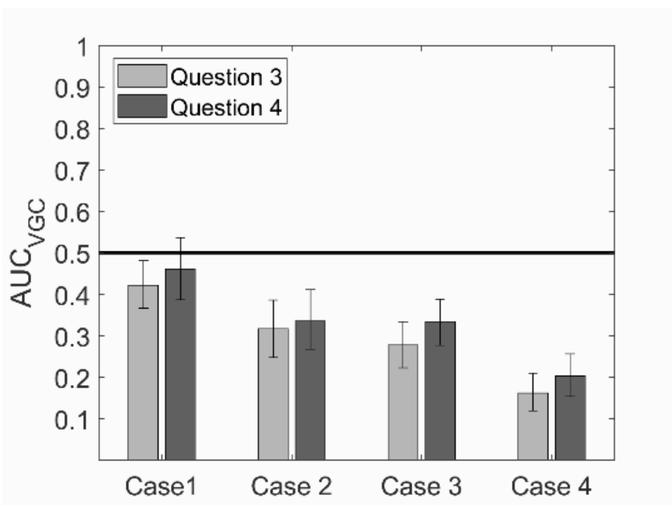

Figure 14. The values of the AUC$_{VGC}$ for Question 3 and Question 4. Values significantly under 0.5 show that the images from the defective transducers were assessed as being worse than the non-defective transducers. The 95% asymmetric CI is shown by the error bars. (Adapted from Paper IV.)

The result from the VGC analysis is shown in Figure 14. The image quality was measured in Questions 3 and 4. Questions 3 was about how well the structural objects were seen in the image. In all four cases the visibility of structural details was inferior for the defective transducers compared to the non-defective ones. Question 4 was about how the overall image quality was assessed. For Case 1 there was no measurable difference between defective and non-defective transducer, while the total image quality was assessed as being worse for the defective transducers for Cases 2-4.

5 DISCUSSION

Ultrasound is used in many different examinations where the equipment is specialised for the purpose. Detecting lesions is one of the areas of use. In Paper III two scanners in different price ranges were compared in a 4AFC study. The more expensive one performed slightly better at shallower depths. Different equipment may be one factor that can distinguish images, another may be defective equipment. A common vital but fragile part of all ultrasound equipment is the transducer, and it is well known that transducers are prone to being defective. One type of transducer defect is one in which sending and receiving the ultrasound beam is affected. This can result in a dark region shallow in the image like a vertical dark band; a TMM phantom can be used to detect this band. Since the dark region is visible in a phantom image it is also visible in the clinical image, even if the anatomic structures make it harder to detect. To avoid impact on the clinical use of the ultrasound scanners, the SDR method that uses many clinical images to detect the dark regions was developed in Paper I. This method is automatic and makes it possible to check many linear and curvilinear transducers simultaneously. A new evaluation is made every day new images are available. In Paper I the result was presented as three cases and visually compared to electrical measurements of the transducers, and the results subjectively conformed well. In Paper II a more thorough evaluation of the method was made. Eighty-one linear and curvilinear transducers were evaluated by the SDR method and the in-air method was used as reference. The result for the ROC analysis was 0.88 (SD 0.06), which can be categorised as “good”[91, 92].

The SDR method from Paper I and a VNA that included ultrasound images from many years made it possible to find clinical images that were actually produced with defective transducers. In Paper IV four defective transducers were evaluated, three of which had severe defects. Despite this, the transducers had been in clinical use for several months each. The defective transducers were compared to non-defective transducers by four experienced radiologists in an observer study. The questions whether the defects affect clinical image quality and whether the defects could affect the diagnosis were investigated. In all cases, even for the transducer that only had four defective elements, the visibility of structural details was affected. For the three transducers that had more severe defects, a considerable number received the answers “Confident that the artifacts could affect the diagnosis” and “Somewhat confident that the artifacts could affect the diagnosis” in response to the question “Could these artificial artifacts, if any, affect the diagnosis if the artifacts would coincide with the region for the diagnosis?” This shows that defective transducers do affect the image quality and sometimes also the diagnosis and that it is important to keep track of the ultrasound equipment and especially monitor that the transducers are fully functional.

5.1 PAPER I

In Paper I the SDR method for detecting a defective transducer was presented. The obvious advantage of the method is that it can be done automatically for many transducers at the same time without any interference with clinical activity. The method has a lag of 75-150 clinical images (approved by the selection algorithm) until the defect appears in full in the SDR curves. For general imaging scanners that have several different transducers that are used for different examinations, the lag in time may correspond to 11-23 days for the cases used in Paper IV. If the SDR method is used on a scanner that is used for mammography, the number of transducers is lower. As an example, images from a scanner used for mammography that used two transducer models were collected for a period of 35 days. During that period 816 images were collected (23.3 images per day) for the most frequently used transducer model. During the most productive day during this period, 77 images were stored in the VNA. In this example the SDR method would have had a delay of less than a week for the most frequently used transducer, which is the one that is most at risk due to its frequent use.

In the European Union, all medical devices are regulated by Regulation (EU) 2017/745. "*The characteristics and performance of a device shall not be adversely affected to such a degree that the health or safety of the patient or the user and, where applicable, of other persons are compromised during the lifetime of the device, as indicated by the manufacturer, when the device is subjected to the stresses which can occur during normal conditions of use and has been properly maintained in accordance with the manufacturer's instructions.*" The instructions of the manufacturer are often that checks must be performed once a year or sometimes due to the QA programme of the facility. Hospitals and other healthcare institutions may, in addition to the instructions of the manufacturer, follow recommendations from, among others, special-interest associations such as BMUS[7], IPEM[9], ACR-AAPM[10], AIUM[11] and EFSUMB[6]. Some of these associations recommend that the transducers be checked for defects every month, and two of them recommend that this be done every day. To have such frequent checks, it is proposed that sonographers conduct the check instead of an engineer or physicist. In practice, there are often longer intervals between transducer checks. In the region of Västra Götaland the periodic checking of the scanners and the transducers is performed once a year at most hospitals according to the instructions of the manufacturer, which according to Mårtensson[13] and Paper IV is too seldom. Quarterly assessment is recommended for image uniformity by Hangiandreou et al.[15], and the same interval is used by Vitikainen et al.[21] for the in-air image checking. Even quarterly checks might be too infrequent when it is considered that a transducer can produce many images a day. In the mammography example above more than 2100 images can be produced using a defective transducer if the transducer defect occurs right after a quarterly check and four times as many if the check is performed annually. If the transducer is checked monthly in the same mammography example, up to 700 images might be produced with a defective transducer. Defects that are prominent can sometimes be discovered directly in the clinical image, as shown in Paper IV, Question 1. However, these transducers are not always replaced immediately, as seen in Paper IV. Checks of the uniformity for a transducer that are performed more often than once a year and even quarterly seem to be necessary considering the quantity of images produced. The SDR method that monitors many linear and curvilinear transducers at the same time could be one option. Another is to let the sonographers perform the in-air method as recommended by some of the associations mentioned above.

Use of the SDR method in research required approval from the Regional Ethical Review Board for Papers I, II and IV. The reason is that the method uses clinical images. To launch such a method in ordinary healthcare requires national data protection legislation to be followed. The Swedish Patient Data Act, 2008:355, covers use of patient data by healthcare providers. One of the approved purposes of personal data processing according to that Act could be systematically and continuously developing and assuring quality in the clinical setting. This purpose suits the use of the SDR method well, so the use of patient data to evaluate defects in ultrasound transducers should not pose a problem in Sweden.

5.2 PAPER II

In the evaluation of the SDR method, the SDR curves from 81 transducers were compared to a known method as reference. An objective and precise method would have been preferable as reference method, like the electrical measurements from Probehunter, which were available. However, it was not possible to test all transducers at the time of data collection, for example ML6-15 from GE and L12-5 from Philips could not be handled by Probehunter. Another aspect of using electrical measurement as reference method is that defects such as channel defects in the scanner will be missed, which is not the case for the SDR method. The method that was chosen was the in-air method, which is subjective but widely used and it is suitable for all transducer models and includes defects in the scanner.

To classify a defect using the SDR method, a time limit was set that the defect should be present for at least 20 SDR curves. The SDR curves were updated once a day (if there were new images) and the new images replaced the oldest ones, so that the most recent images were included for the latest SDR curve. The limit of 20 days was set to ensure that the defects were persistent. The image data being updated daily was a consequence of the design of our system where the new images were imported once per night. This design made it likely that if few images were produced per day for a transducer, it was more likely that the SDR curves were more correlated than for a transducer that produced many images per day. A limit where all image data had been replaced X times would probably have been a more equal limit.

Some defects (classified as such by the observers using the in-air method) were missed by the SDR method. The possible reason for this will be discussed here.

– The in-air method had certain settings when the images were produced where the harmonic imaging, spatial compounding and time averaging settings were disabled. This makes defects clearly apparent. In the clinical situation these functions are probably used and may conceal the dark region caused by the defect in the image. How much this may affect the result from the SDR method has not been investigated.

– There are three different thresholds that are used in the algorithm. If the difference between the dark region and the mean of the rest of the image (the shallowest part of the image) is not large enough to cross one of these thresholds, no dark region will be registered. The same thresholds were used in Papers I, II and IV and were set according to the cases in Paper I – these levels correctly identified the dead elements in the cases but only sometimes registered single weak elements, see Case 3 in Figure 8. If the thresholds are lowered, a higher sensitivity will be reached at the cost of lower specificity.

– Even if the observers had trained on 24 in-air images where there were electrical measurements to compare to, the reference method was associated with uncertainties. Furthermore, using a single method for reference has limitations since many methods complement each other and no method covers all defects[14]. It is important to note that if the reference method makes an incorrect decision (compared to the unknown truth) and the evaluated method makes a correct decision (compared to the same unknown truth), this is manifested as an incorrect decision by the evaluated method in a study such as this.

– In the studies by Mårtensson et al.[12, 13] and Sipilä et al.[14] delamination was the most common defect. Delamination defects may be confirmed when pressure is applied on the transducer face[25]. The defect can disappear when pressure is applied to some delamination defects. Among the different methods that exist to evaluate transducer errors, there are some where the transducer has pressure applied to the transducer head, as in normal use. The in-air method and electric transducer testing do not use any pressure on the transducer during the test. When the paperclip method[46] is used, some pressure is applied right under the paperclip (or similar metal object). For flat TMM phantoms and linear array transducers, the pressure is like in the patient situation. If a curved array transducer is used on a flat phantom with water as coupling medium, only a part of the transducer has pressure applied to it. There are, however, phantoms where the whole curved array surface is in contact with the phantom[43, 93]. In the blinded comparison between in-air test and electric transducer test, Dudley and Woolley[25] confirmed suspicious defects in the in-air pattern by using the paperclip test. This was not performed in Paper II and may be part of the reason why the SDR method missed defects that were present in the in-air images. In the SDR method the pressure is always there, and delamination defects may be compressed and be rendered harmless when the same defect in the in-air method is unpressurised and hence visible. If the pressure used to press the transducer to the patient results in the image being unaffected, the SDR method will not detect the defect, nor will a linear array transducer that is used on a flat TMM phantom.

Two methods were chosen as evaluation metrics, the well-known ROC method and the more rarely used AFROC. ROC has sometimes been used with a set scale from fail (0.5-0.6) to excellent (0.9-1.0)[91, 92] The AUC_{ROC} was 0.88 (SD 0.06) for the SDR method when the in-air method was used as ground truth. The SDR method as well as the in-air method use localisation of the defects and the possibility of handling several defects, which makes AFROC suitable. The AUC_{AFROC} was 0.71(SE 0.07). In the AFROC method, it is more difficult to interpret the result than in the ROC method, and both methods were therefore used in the analysis.

5.3 PAPER III

In Paper III, comparisons between the ability to reproduce low contrast objects of two ultrasound scanners were made. There are many techniques to distinguish one ultrasound machine from another or a new ultrasound technology from another. One way of doing this is to examine one or more organs in many patients and let experienced observers evaluate the images and see whether one technology outperforms another[94, 95]. Another way is to focus on the ability to reproduce special objects in a single image, for example a synthetic detail in a TMM phantom, and evaluate for example different beamforming techniques and how much they affect lateral and axial resolution[96]. A third way, the one that was chosen in Paper III, was to collect many statistically independent images of low-contrast objects and let several observers perform a detection study. One of the low-contrast objects that is interesting for ultrasound examinations in the human body is lesions. The size of found liver lesions begins, according to Hall et al.,[97] at 4-5 mm. Hill and Sanders[98] investigated the characteristics of intra-abdominal cystic masses and found variations in the amount of echoes in the cystic masses. Taking the size of and variation in internal echoes into account, a TMM phantom was chosen that had 4.5 mm and 6 mm objects and both positive and negative contrast levels for the objects. The cylindrical objects were acquired at different depths. Regarding the shapes of the object in TMM phantoms, there are phantoms that have spherical objects instead of cylinders, and that is preferable since the slice thickness of the ultrasound beam is not considered when using cylinders[40, 42]. When the study was performed there were no commercially available phantoms of the six investigated phantom manufacturers that could fulfil all criteria regarding spherical objects in different sizes and contrasts and depths, so the one chosen, CIRS 047, fulfilled all criteria except spherical objects.

The background in the images collected for the 4AFC study was taken from two regions above and one region below the object concerned. This was done to obtain image squares just containing independent speckle to the 4AFC images. It would have been preferable to collect the speckle squares from the same depths as the objects as the size of the speckles depends on the depth and also transmitted focus[99]. However, the selected regions were situated directly above and under the objects, as close to the objects as possible, and this approach was deemed to have minimal impact on speckle size.

How the settings would be chosen for the scanners was an interesting issue. Since both machines had features such as harmonics, smoothing, compounding and speckle reduction imaging, it was difficult to know how much the different features would affect the visibility of the objects. Turning off all features was one option that could be used, but the features that the manufacturer had developed would then be unused. The option that was chosen was to use the default Abdomen setting that each machine had for the transducer concerned, a C4 curvilinear array transducer.

Statistical uncertainty is an important issue when performing observer studies. In this study the non-parametric bootstrap plug-in principle[100] was used. Bootstrapping is a way of simulating a distribution for the particular situation. By using the percentiles for the bootstrapped distribution, confidence intervals can be determined. If the results are to be valid for all observers or just the ones who participated in the study, or just valid for the particular cases, there are different approaches to performing bootstrapping. For a fixed-case situation where the observer population is simulated, representing how generalised observers are assessing the cases concerned, the observers are sampled with replacement from an observer pool containing all observers, the sample size is equal to the number of observers for every draw. For a result only valid for the observers who participated but for a general case–a fixed reader situation, the cases are put in a case pool and sampled with replacement, the number of used cases is used as sample size. By combining these two approaches and drawing samples from both cases and observers, the simulated distribution contains both observer and case variability and can be valid both for general observers and for general cases. This was the approach in Paper III. The result was that in only 25% of the 16 comparisons significant differences were found in favour of the high-end L9 scanner, which may be surprising. The number of observers and the image material were after all relatively large. When the variability of all sources is taken into account, it is more difficult to find differences between the scanners. Nevertheless, when it comes to comparing different imaging modalities it is important to include human observers in order to include the whole chain even if the variability becomes high. Thilander-Klang et al.[101] and Tapiovaara and Sandborg[102] also conclude that there are wide variations in performing constancy control using human observers for low-contrast detectability.

5.4 PAPER IV

Two of the tasks for an ultrasound transducer used on a modern scanner are to produce a good B-mode image and to measure blood flow via Doppler. The fact that defective elements can cause erroneous Doppler measurements has been shown[83, 103] and will not be further discussed. How defects affect the B-mode image has been investigated by Weigang et al.[83] and Rosenfeld et al.[104]. Weigang et al. showed that there was a decrease in brightness and a slight shadow in the B-mode image right below six consecutive dead elements. Rosenfeld et al. tested disconnecting an increasing number of elements and found that deterioration of the images began at five consecutive inoperable elements. The impact on clinical image quality and the possible effect on the diagnosis caused by defective transducers were investigated in Paper IV. The four cases had different grades of transducer defects. The least affected transducer had 1.6 % defective elements and the others 5.1%, 10.1% and 11.5%. In all four cases the visibility of structural details was assessed as being worse than for the non-defective transducers, even for Case 1, which only had four defective elements, and the impact on clinical images was demonstrated at least in the fixed-reader case. Regarding the question about possible effect on diagnosis, almost none of the answers were 1, “confident that the artifact could affect the diagnosis” or 2 “somewhat confident that the artifact could affect the diagnosis” for Case 1. On the other hand, for the three cases that had more severe defects, there were many answers where the observer was confident that the artifact could affect the diagnosis.

An important question is when a transducer is disqualified for clinical use. One example of when to disqualify a transducer from clinical use is given in IPEM report No 102[9], where no axial banding is acceptable: if the banding is visibly affecting the image the transducer should be taken out of use and be technically assessed. Two weak elements in consecution or four weak elements in total are recommended by Sonora FirstCall to be considered as unacceptable. FirstCall also has different colours for elements that are in the centre or more peripheral, (see Figure 3) although it is not stated how these should be interpreted. For a linear transducer, for example, the whole image can be filled with tissues, that are of interest for the diagnosis, where the peripheral elements are of equal importance to the central elements (see e.g. Figure 11A or Figure 11C). Furthermore, Figure 3 shows a 96-element transducer, and the transducer in Case 3 in Paper IV has 256 elements. Having the same limits for these two transducers can be questioned since the contribution to the image for one element is larger for the transducers that have fewer elements. Based on the results for Case 1 in Paper IV, it seems that four of 256 affected elements can cause misdiagnosis as described in Figure 4A in Paper IV. However, when all the images used were considered, the assessment for effect on diagnosis was small. Both the limit of no axial banding that is visibly affecting the clinical image and four weak elements in aggregate or two consecutive weak elements seem reasonable, but if the affected elements cause banding visible in the clinical image, taking the transducer out of use should be considered.

There are several ways in which transducer defects can affect the clinical image. The visibility of structural details can be affected, as well as the overall image quality, as shown in Paper IV. In addition, narrow darker vertical streaks can be mistaken for ruptures, for example in a tendon, (see Figure 11 A and 11 C). These shallow artifacts can also make it difficult to evaluate whether the texture of the tissue is uniform or not (see Figure 11 B). When the defects become wider and deeper, the dark streak can shadow entire parts of organs, (see Figure 11D), and in the worst case cause a tumour to be missed.

Ultrasound, unlike conventional radiography or MRI, is a modality that is used live instead of first producing an image and then performing diagnosis on that image. The sonographer can move the transducer and view the tissues and organs from different angles. Some sonographers argue that if a part of the transducer is defective it can be compensated for by moving the transducer to cover for the affected part of the image. This is a way to compensate for missing image parts if the result of a defect is obvious in that a dark band appears in the image. If the dark streak(s) appear in the outer edge of the image the result just becomes a narrower image and is not as disturbing as if the transducer is damaged more in the middle, as in Case 4 in Paper IV(see Figure 11D). However, if a narrower image is as good as a wider, it can be argued that the transducer would have been constructed that way from the beginning. So, this workaround by moving the transducer is feasible as long as the defect is severe and obvious when using the transducer, but the examination time becomes longer and it takes focus from the examiner. When it comes to defects that are more spread across the transducer as in Case 2 in Paper IV, it is not as easy to move the transducer to avoid the defect. Possibly it can be decided if there are ruptures in the tissue since the defects do not move when the transducer is moved laterally, while real ruptures do. However, to do this workaround it has to be suspected that there are small defects in the transducer. In Paper IV it is shown that the ability to reproduce structural details is affected even for small defects (as in Case 1 in Paper IV, which had 4 defective elements out of 256), and even if it sometimes is possible to work around by moving the transducer to see if the defects are moving, this is not always the case nor possible in the stored images. Still images are stored for documentation and are representative for each examination. The images contain relevant anatomy, and any findings, if applicable.

5.5 SUMMARY

How does the technical status of medical ultrasound equipment affect image quality? That defective equipment and especially the transducers affect the image quality is known but has previously not been thoroughly investigated. The SDR method described in Paper I was developed to have a continuous monitoring over linear and later also curvilinear transducers for early detection of defects. It can be handled by a computer that imports the images from a VNA once a day. In Paper II the method was evaluated against an established method for defect detection in ultrasound transducers and the conformity was "good".

Different ultrasound scanners can also be considered to have different technical status depending on how advanced (and expensive) they are. A comparison of how well a high-end and a regular ultrasound scanner reproduced clinically relevant objects was made in Paper III. For some of the objects there was a significant difference in favour of the more advanced scanner. This indicates that it is important to use a scanner that is advanced enough for the purpose. In Paper IV clinical images that had been produced with defective transducers were evaluated and compared with images produced with non-defective transducers in an observer study. The defective transducers were found by using the SDR method on old clinical image material. Even if three of the four transducers had severe defects they had been in clinical use for several months. The total image quality was assessed to be worse for three of the four defective transducers compared to the non-defective and the visibility of structural details was worse for all four transducers. In 19 percent of the assessments of the images produced with the defective transducers the observers (radiologists) were convinced that the artifacts could affect the diagnosis if the artifacts from the defective transducers would coincide with the region for the diagnosis. This shows that it is important to check the equipment and especially the transducers often to detect defects.

6 CONCLUSIONS

This thesis concerns how different aspects of ultrasound equipment status affect image quality. The thesis also concerns a novel method for automatic detection of defective transducers and an evaluation of this method. The specific conclusions drawn from the studies in this thesis are presented here.

Paper I

A method for assessment of horizontal uniformity in clinical ultrasound images produced with linear array transducers was introduced. An algorithm for automatic detection (SDR method) of horizontal non-uniformities was presented and tested on three ultrasound systems in a case study. Subjectively, the detected defects were persistent over time (when different images were used as input) and had visual agreement with electrical measurements of the transducers. This indicates that stored clinical images can be used for early detection of defective transducers, as a complement to quality control.

Paper II

The developed SDR method for automatic detection of defects in ultrasound systems using clinical images has a good agreement with the well-established in-air method.

Paper III

It is possible to use a greyscale phantom and perform an observer study to discriminate the ability of ultrasound systems to reproduce lesion-like objects. In a comparison between a high-end scanner and a regular scanner, statistically significant differences were obtained in favour of the high-end scanner for four of sixteen combinations of object size, contrast and depth. However, the image material and number of observers were larger than for normal quality control.

Paper IV

The image quality in greyscale 2D still ultrasound images and the risk of misdiagnosis can be affected by the use of defective transducers. This highlights the importance of frequent quality control of the transducers to avoid decreased image quality and even misdiagnosis.

7 FUTURE PERSPECTIVES

The SDR method has been used with the same values of the parameters in Papers I, II and IV. These were concluded by testing different values. However, the values were only tried out on a few transducers and manufacturers. In Paper II other parameter values were tested: smaller image stack, different depth, and a different degree of polynomial fit for one of the paths for the SDR curve. These were just isolated tests of single values and no real optimisation. Future work could be to optimise the different parameters, especially the limits for the three paths, which decides whether the SDR curve becomes positive or not. In the future it would also be interesting to compare how different scanner settings such as spatial compounding, harmonics and so on affect the SDR curve.

It would be interesting to use the SDR method on many transducers over a long period of time and compare the result with different methods when the SDR method flags a defect. It would also be interesting to measure how many are detected by normal QC, detection by sonographers and the SDR method. For this, a large VNA like the one in the Region of Västra Götaland would be suitable for the implementation of a system that monitors the linear and curvilinear transducers in radiological departments. When a defect occurs, the responsible technician or physicist should be notified.

The software for the SDR method was developed in MATLAB (MathWorks, Inc., Natick, MA, USA) and requires some knowledge to run. There are special scripts for importing images from the VNA, there are specific settings for each transducer model, and so on. However, if it was possible to build usable software for free distribution with minimum settings this would be a future goal.

Paper IV included three transducers with severe defects and one that had been approved by the manufacturer in the annual checks. All images were taken from general imaging scanners in clinical use. A future investigation could be a visual grading study with defective transducers on images from mammography where certain structures can be classified as a tumour. This is relevant since the ability to reproduce structural object has been shown to be affected by defective transducers (in Paper IV) and this type of diagnostics is very important.

ACKNOWLEDGEMENTS

Denna avhandling hade inte varit möjlig utan stöd och hjälp från många andra.

Min ständigt positive huvudhandledare **Magnus Båth,** utan ditt engagemang och kunnande hade detta aldrig varit möjligt. Du har guidat mig på ett utmärkt sätt genom mina forskarstudier. Du har trott på mig i alla lägen även när jag själv har tvivlat. Tack från hela mitt hjärta!

Min kollega och bihandledare **Nasser Hosseini** vars kunnande i det vetenskapliga språket har betytt mycket. Dina kunskaper i MATLAB och att du tagit fram ett specialanpassat program för hämtning av ultraljudsbilder från BFR har varit avgörande för doktorandprojektet. Tusen tack!

Stort tack även till min bihandledare **Lars Gunnar Månsson** för dina gedigna kunskaper inom forskningsmetodik och hur man uttrycker sig i vetenskapliga publikationer.

Vid starten på denna resa var **Morgan Andréasson** min enhetschef, tack för att du trodde på mig och att jag fick chansen att göra detta.

Min nuvarande enhetschef **Peter Embervik** som givit mig utrymme att utföra och slutföra forskarstudierna-ditt stöd har betytt mycket, stort tack.

MFT:s verksamhetschef **Annette Lövefors Daun** - tack för att du prioriterar forskning i MFT:s verksamhet vilket givit mig denna möjlighet.

Tack till alla kollegor på **MTMIS** och sektionsledare för Bild & Funktionsmedicin **Thomas Anderson**.

MT-kollegorna **Wiebke Rosenberg**, **Benny Stenborg** och **Jan-Olof Johansson** som lärt mig mycket om ultraljud och bidrog till arbete 1 och arbete 3, varmaste tack. Tack även till nuvarande **ultraljudsgrupp**.

Ylva Aurell, tack för ditt värdefulla kliniska kunnande och guidning i arbete 4. Tack även till övriga granskare och medförfattare till arbete 4: **David Collin**, **Eva Frösing** och **Pawel Szaro**.

Johan Fors som ordnat så att jag fått tillgång till ultraljudsapparater, utan dem hade inte arbete 3 kunnat genomföras, tack.

Tack även till **Angelica Svalkvist** och **Jonathan Arvidsson** som hjälpt till med värdefulla synpunkter på ramen.

Tack alla i **RALF-gruppen** för att ni agerat värdefullt bollplank genom forskarstudierna.

Tack till min läkare **Dimitra Bachlava** som så skickligt ställt in min medicin så att jag kunnat genomföra doktorandstudierna trots min diagnos.

Jag vill även tacka min sjuksköterskekontakt **Tomas Karelöf** i Älvängen.

Till sist vill jag tacka min älskade familj **Åsa**, **Amelie** och **Anton** som stöttat mig i detta, ni betyder allt.

REFERENCES

1. Newman PG, Rozycki GS. The history of ultrasound. Surg Clin North Am. 1998;78(2):179-95.

2. Holm HH, Skjoldbye B. Interventional ultrasound. Ultrasound Med Biol. 1996;22(7):773-89.

3. Holm HH, Skjoldbye B. Interventional procedures. Ultrasound Med Biol. 2000;26 Suppl 1:S131-4.

4. Nielsen MB, Søgaard SB, Bech Andersen S, Skjoldbye B, Hansen KL, Rafaelsen S, et al. Highlights of the development in ultrasound during the last 70 years: A historical review. Acta Radiol. 2021;62(11):1499-514.

5. Eckersley RJ, Chin CT, Burns PN. Optimising phase and amplitude modulation schemes for imaging microbubble contrast agents at low acoustic power. Ultrasound Med Biol. 2005;31(2):213-9.

6. Kollmann C, deKorte C, Dudley NJ, Gritzmann N, Martin K, Evans DH. Guideline for technical quality assurance (TQA) of ultrasound devices (B-Mode)--version 1.0 (July 2012): EFSUMB Technical Quality Assurance Group--US-TQA/B. Ultraschall Med. 2012;33(6):544-9.

7. Dudley N, Russell S, Ward B, Hoskins P. BMUS guidelines for the regular quality assurance testing of ultrasound scanners by sonographers. Ultrasound. 2014;22(1):8-14.

8. IEC. TS 62736 Ed 1.0:2016 Ultrasonics - pulse-echo-scanners- simple methods for periodic testing to verify stability of an imaging system´s elementary performance. 2016.

9. Russel S, Dudley NJ, Evans T, Hoskins P, Watson A, Starrit H. Report No 102, Quality assurance of ultrasound imaging systems: IPEM Report Series; 2010.

10. Ann KM, A. ES, Christopher F, Geiser W, M JL, Hangiandreou N, et al. ACR–AAPM technical standard for diagnostic medical physics performance monitoring of real time ultrasound equipment. rev. 2021.

11. Boote EJ, Forsberg F, Garra BS, Ophir J, Ophir KJ, Zagzebski JA. Routine quality assurance for diagnostic ultrasound equipment. American institute of ultrasound in medicine 2008.

12. Mårtensson M, Olsson M, Segall B, Fraser AG, Winter R, Brodin LÅ. High incidence of defective ultrasound transducers in use in routine clinical practice. Eur J Echocardiogr. 2009;10(3):389-94.

13. Mårtensson M, Olsson M, Brodin LÅ. Ultrasound transducer function: annual testing is not sufficient. Eur J Echocardiogr. 2010;11(9):801-5.

14. Sipilä O, Mannila V, Vartiainen E. Quality assurance in diagnostic ultrasound. Eur J Radiol. 2011;80(2):519-25.

15. Hangiandreou NJ, Stekel SF, Tradup DJ, Gorny KR, King DM. Four-year experience with a clinical ultrasound quality control program. Ultrasound Med Biol. 2011;37(8):1350-7.

16. Noce JP. Fundamentals of diagnostic ultrasonography. Biomed Instrum Technol. 1990;24(6):456-9.

17. Izzetti R, Vitali S, Aringhieri G, Nisi M, Oranges T, Dini V, et al. Ultra-high frequency ultrasound, a promising diagnostic technique: review of the literature and single-center experience. Can Assoc Radiol J. 2021;72(3):418-31.

18. Kremkau FW. Sonography principles and instruments. Eight ed. Missouri: Elsevier Saunders; 2011.

19. Bakaric M, Fromme P, Hurrell A, Rajagopal S, Miloro P, Zeqiri B, et al. Measurement of the temperature-dependent output of lead zirconate titanate transducers. Ultrasonics. 2021;114:106378.

20. Yang HS, Bansal RC, Mookadam F, Khandheria BK, Tajik AJ, Chandrasekaran K. Practical guide for three-dimensional transthoracic echocardiography using a fully sampled matrix array transducer. J Am Soc Echocardiogr. 2008;21(9):979-89; quiz 1081-2.

21. Vitikainen AM, Peltonen JI, Vartiainen E. Routine ultrasound quality assurance in a multi-unit radiology department: a retrospective evaluation of transducer failures. Ultrasound Med Biol. 2017;43(9):1930-7.

22. Goodsitt MM, Carson PL, Witt S, Hykes DL, Kofler JM, Jr. Real-time B-mode ultrasound quality control test procedures. Report of AAPM Ultrasound Task Group No. 1. Med Phys. 1998;25(8):1385-406.

23. Dudley NJ, Griffith K, Houldsworth G, Holloway M, Dunn MA. A review of two alternative ultrasound quality assurance programmes. Eur J Ultrasound. 2001;12(3):233-45.

24. Dudley NJ. Ultrasound probe acceptance testing using the in-air reverberation pattern. Ultrasound Med Biol. 2019;45(6):1475-82.

25. Dudley NJ, Woolley DJ. Blinded comparison between an in-air reverberation method and an electronic probe tester in the detection of ultrasound probe faults. Ultrasound Med Biol. 2017;43(12):2954-8.

26. Hsia C, Hsiao YC, Huang YC. A single-chip high-voltage integrated actuator for biomedical ultrasound scanners. Sensors (Basel, Switzerland). 2019;19(23).

27. International Organization for Standardization ISO 9000, Quality management systems - Fundamentals and vocabulary. 2015.

28. Sreedher G, Ho ML, Smith M, Udayasankar UK, Risacher S, Rapalino O, et al. Magnetic resonance imaging quality control, quality assurance and quality improvement. Pediatr Radiol. 2021;51(5):698-708.

29. Hoskins P, Martin K, Thrush A. Diagnostic ultrasound physics and equipment. second ed. New York: Cambridge University Press; 2010.

30. Grazhdani H, David E, Ventura Spagnolo O, Buemi F, Perri A, Orsogna N, et al. Quality assurance of ultrasound systems: current status and review of literature. J of Ultrasound. 2018;21(3):173-82.

31. Mannila V, Sipilä O. Phantom-based quality assurance measurements in B-mode ultrasound. Acta Radiol Short Rep. 2013;2(8):2047981613511967.

32. Thijssen JM, Weijers G, de Korte CL. Objective performance testing and quality assurance of medical ultrasound equipment. Ultrasound Med Biol. 2007;33(3):460-71.

33. Gibson NM, Dudley NJ, Griffith K. A computerised quality control testing system for B-mode ultrasound. Ultrasound Med Biol. 2001;27(12):1697-711.

34. Long Z, Tradup DJ, Stekel SF, Gorny KR, Hangiandreou NJ. Evaluations of UltraiQ software for objective ultrasound image quality assessment using images from a commercial scanner. J Appl Clin Med Phys. 2018;19(2):298-304.

35. MacGillivray TJ, Ellis W, Pye SD. The resolution integral: visual and computational approaches to characterizing ultrasound images. Phys Med Biol. 2010;55(17):5067-88.

36. Rownd JJ, Madsen EL, Zagzebski JA, Frank GR, Dong F. Phantoms and automated system for testing the resolution of ultrasound scanners. Ultrasound Med Biol. 1997;23(2):245-60.

37. Doyle AJ, Cody D, Knight S, King DM, Sullivan FJ, Browne JE. Contrast-specific spherical lesion phantoms and ancillary analysis software for the objective evaluation of transrectal ultrasound system contrast detectability. Ultrasound Med Biol. 2022;48(8):1615-27.

38. Zagzebski J. WE-G-211-03: Development of ultrasound imaging equipment. Med Phys. 2012;39(6Part28):3964.

39. Browne JE, Ramnarine KV, Watson AJ, Hoskins PR. Assessment of the acoustic properties of common tissue-mimicking test phantoms. Ultrasound Med Biol. 2003;29(7):1053-60.

40. Madsen EL, Zagzebski JA, Macdonald MC, Frank GR. Ultrasound focal lesion detectability phantoms. Med Phys. 1991;18(6):1171-80.

41. Kofler JM, Jr., Madsen EL. Improved method for determining resolution zones in ultrasound phantoms with spherical simulated lesions. Ultrasound Med Biol. 2001;27(12):1667-76.

42. Kofler JM, Jr., Lindstrom MJ, Kelcz F, Madsen EL. Association of automated and human observer lesion detecting ability using phantoms. Ultrasound Med Biol. 2005;31(3):351-9.

43. Madsen EL, Song C, Frank GR. Low-echo sphere phantoms and methods for assessing imaging performance of medical ultrasound scanners. Ultrasound Med Biol. 2014;40(7):1697-717.

44. Browne JE, Cannon LM, Fagan AJ. Comparison of in-house development cylindrical and spherical anechoic target phantoms for performance testing of breast ultrasound scanners. Phys Med. 2014;30(6):718-24.

45. Browne JE, Cannon LM, Fagan AJ, Cournane S. Development of a focal lesion phantom with clinically relevant lesion characteristics for image quality evaluation of breast ultrasound scanners. Phys Med. 2022;94:65-74.

46. Goldstein A, Ranney D, McLeary RD. Linear array test tool. J Ultrasound Med. 1989;8(7):385-97.

47. Dudley NJ, Woolley DJ. A simple uniformity test for ultrasound phased arrays. Phys Med. 2016;32(9):1162-6.

48. Fatahi Asl J, Farzanegan Z, Tahmasbi M, Birgani SM, Malekzade M, Yazdaninejad H. Evaluation of the scan duration and mechanical and thermal indices applied for the diagnostic ultrasound examinations. J Ultrasound Med. 2021;40(9):1839-50.

49. Harris GR, Church CC, Dalecki D, Ziskin MC, Bagley JE. Comparison of thermal safety practice guidelines for diagnostic ultrasound exposures. Ultrasound Med Biol. 2016;42(2):345-57.

50. Nowicki A. Safety of ultrasonic examinations; thermal and mechanical indices. Med Ultrason. 2020;22(2):203-10.

51. BMUS. Guidelines for the safe use of diagnostic ultrasound equipment 2009 [cited 2022 1215].

52. Månsson LG. Methods for the evaluation of image quality: a review. Radiat Prot Dosim. 2000;90(1-2):89-99.

53. Tingberg A, Herrmann C, Almén A, Besjakov J, Mattsson S, Sund P, et al. Comparison of two methods for evaluation of the image quality of lumbar spine radiographs. Radiat Prot Dosim. 2000;90:165-8.

54. Burgess AE, Wagner RF, Jennings RJ, Barlow HB. Efficiency of Human Visual Signal Discrimination. Science. 1981;214(4516):93-4.

55. Burgess AE, Ghandeharian H. Visual signal detection. II. Signal-location identification. J Opt Soc Am A. 1984;1(8):906-10.

56. Burgess A. Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation. J Opt Soc Am A. 1985;2(9):1498-507.

57. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. Med Phys. 1995;22(5):643-55.

58. Timberg P, Båth M, Andersson I, Mattsson S, Tingberg A, Ruschin M. In-plane visibility of lesions using breast tomosynthesis and digital mammography. Med Phys. 2010;37(11):5618-26.

59. Timberg P, Båth M, Andersson I, Mattsson S, Tingberg A, Ruschin M. Visibility of microcalcification clusters and masses in breast tomosynthesis image volumes and digital mammography: a 4AFC human observer study. Med Phys. 2012;39(5):2431-7.

60. Streiner DL, Cairney J. What's under the ROC? An introduction to receiver operating characteristics curves. Can J Psychiatry. 2007;52(2):121-8.

61. Goodenough DJ, Rossmann K, Lusted LB. Radiographic applications of signal detection theory. Radiology. 1972;105(1):199-200.

62. Green DM. General prediction relating yes-no and forced-choice results. J Acoust Soc Am. 1964;36:1042-.

63. Chakraborty DP. New developments in observer performance methodology in medical imaging. Semin Nucl Med. 2011;41(6):401-18.

64. Chakraborty D. Statistical power in observer-performance studies: comparison of the receiver operating characteristic and free-response methods in tasks involving localization. Acad Radiol. 2002;9(2):147-56.

65. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH, editors. A free response approach to the measurement and characterization of radiographic observer performance. Application of Optical Instrumentation in Medicine VI; 1977: International Society for Optics and Photonics.

66. Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. Med Phys. 1989;16(4):561-8.

67. Chakraborty DP, Winter LH. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology. 1990;174(3 Pt 1):873-81.

68. Chakraborty DP. A brief history of free-response receiver operating characteristic paradigm data analysis. Acad Radiol. 2013;20(7):915-9.

69. Chakraborty DP. A status report on free-response analysis. Radiat Prot Dosim. 2010;139(1-3):20-5.

70. Zhang L, Cavaro-Ménard C, Le Callet P, Tanguy JY. A perceptually relevant channelized joint observer (PCJO) for the detection-localization of parametric signals. IEEE Trans Med Imaging. 2012;31(10):1875-88.

71. European Commision, Directorate-General for Research and Innovation, Carmichael J, Moores B, Maccia C. European guidelines on quality criteria for diagnostic radiographic images. Publications Office; 2000.

72. Båth M, Månsson LG. Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. Br J Radiol. 2007;80(951):169-76.

73. Månsson LG. Evaluation of radiographic procedures. Investigations related to chest imaging. (Thesis) Göteborg: Göteborg University. 1994.

74. Tingberg A. Quantifying the quality of medical X-ray images. An evaluation based on normal anatomy for lumbar spine and chest images. (Thesis) Lund: Lund Univerity. 2000.

75. Båth M. Imaging properties of digital radiographic systems. Development, application and assessment of evaluation methods based on linear-systems theory. (Thesis) Göteborg: Göteborg University. 2003.

76. Sund P, Kheddache S, Månsson LG, Båth M, Tylen U, editors. Comparison of two methods for evaluating image quality of chest radiographs. Medical Imaging 2000: Physics of Medical Imaging; 2000: International Society for Optics and Photonics.

77. Redlich U, Hoeschen C, Doehring W. Assessment and optimisation of the image quality of chest-radiography systems. Radiat Prot Dosim. 2005;114(1-3):264-8.

78. Jansson M, Geijer H, Persliden J, Andersson T. Reducing dose in urography while maintaining image quality-a comparison of storage

phosphor plates and a flat-panel detector. Eur Radiol. 2006;16(1):221-6.

79. Cristofaro M, Busi Rizzi E, Piselli P, Pianura E, Petrone A, Fusco N, et al. Image quality and radiation dose reduction in chest CT in pulmonary infection. Radiol Med. 2020;125(5):451-60.

80. Zarb F, McNulty J, Gatt A, Formosa C, Chockalingam N, Evanoff MG, et al. Comparison of in vivo vs. frozen vs. thiel cadaver specimens in visualisation of anatomical structures of the ankle on proton density magnetic resonance imaging (MRI) through a visual grading analysis (VGA) study. Radiography. 2017;23(2):117-24.

81. Ulm B, Dovjak GO, Scharrer A, Muin DA, Zimpfer D, Prayer D, et al. Diagnostic quality of 3Tesla postmortem magnetic resonance imaging in fetuses with and without congenital heart disease. Am J Obstet Gynecol. 2021;225(2):189.e1-.e30.

82. Cüneyitoğlu Özkul M, Mumcuoğlu Ü E, Sancak İ T. Single-image bayesian restoration and multi-image super-resolution restoration for B-mode ultrasound using an accurate system model involving correlated nature of the speckle noise. Ultrason Imaging. 2019;41(6):368-86.

83. Weigang B, Moore GW, Gessert J, Phillips WH, Schafer M. The methods and effects of transducer degradation on image quality and the clinical efficacy of diagnostic sonography. J Diag Med Son. 2003;19(1):3-13.

84. Börjesson S, Håkansson M, Båth M, Kheddache S, Svensson S, Tingberg A, et al. A software tool for increased efficiency in observer performance studies in radiology. Radiat Prot Dosim. 2005;114(1-3):45-52.

85. Håkansson M, Svensson S, Zachrisson S, Svalkvist A, Båth M, Månsson LG. ViewDEX: an efficient and easy-to-use software for observer performance studies. Radiat Prot Dosim. 2010;139(1-3):42-51.

86. Svalkvist A, Svensson S, Håkansson M, Båth M, Månsson LG. ViewDEX: a status report. Radiat Prot Dosim. 2016;169(1-4):38-45.

87. NEMA PS3.14/ISO 12052 Digital Imaging and Communications in Medicine (DICOM) Standard-Grayscale Standard Display Function. Rosslyn, VA, USA (Available free at http://medicalnemaorg/).

88. Båth M, Hansson J. VGC analyzer: a software for statistical analysis of fully crossed multiple-reader multiple-case visual grading characteristics studies. Radiat Prot Dosim. 2016;169(1-4):46-53.

89. Hansson J, Månsson LG, Båth M. The validity of using ROC software for analysing visual grading characteristics data: an investigation based on the novel software VGC analyzer. Radiat Prot Dosim. 2016;169(1-4):54-9.

90. Hansson J, Månsson LG, Båth M. Evaluation of VGC analyzer by comparison with gold standard ROC software and analysis of simulated visual grading data. Radiat Prot Dosim. 2021;195(3-4):378-90.

91. El Khouli RH, Macura KJ, Barker PB, Habba MR, Jacobs MA, Bluemke DA. Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. J Magn Reson Imaging. 2009;30(5):999-1004.

92. Lüdemann L, Grieger W, Wurm R, Wust P, Zimmer C. Glioma assessment using quantitative blood volume maps generated by T1-weighted dynamic contrast-enhanced magnetic resonance imaging: a receiver operating characteristic study. Acta Radiol. 2006;47(3):303-10.

93. King DM, Hangiandreou NJ, Tradup DJ, Stekel SF. Evaluation of a low-cost liquid ultrasound test object for detection of transducer artefacts. Phys Med Biol. 2010;55(23):N557-70.

94. Shapiro RS, Wagreich J, Parsons RB, Stancato-Pasik A, Yeh HC, Lao R. Tissue harmonic imaging sonography: evaluation of image quality compared with conventional sonography. AJR Am J Roentgenol. 1998;171(5):1203-6.

95. Rezaei-Dalouei H, Seilanian-Toosi F, Nekooei S, Davachi B, Azarian A, Abbasi B. Comparing the image quality of tissue harmonic and conventional B-mode ultrasound of kidney in over-obese individuals. Electronic physician. 2018;10(7):7095-100.

96. Xie HW, Guo H, Zhou GQ, Nguyen NQ, Prager RW. Improved ultrasound image quality with pixel-based beamforming using a Wiener-filter and a SNR-dependent coherence factor. Ultrasonics. 2022;119:106594.

97. Hall TJ, Insana MF, Harrison LA, Soller NM, Schlehr KJ. Ultrasound contrast-detail analysis: A comparison of low-contrast detectability among scanhead designs. Med Phys. 1995;22(7):1117-25.

98. Hill M, Sanders RC. Gray scale B scan characteristics of intra-abdominal cystic masses. J Clin Ultrasound. 1978;6(4):217-22.

99. Oosterveld BJ, Thijssen JM, Verhoef WA. Texture of B-mode echograms: 3-D simulations and experiments of the effects of diffraction and scatterer density. Ultrason Imaging. 1985;7(2):142-60.

100. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Boca Raton (FL): CRC Press LLC; 1998.

101. Thilander-Klang A, Ledenius K, Hansson J, Sund P, Båth M. Evaluation of subjective assessment of the low-contrast visibility in constancy control of computed tomography. Radiat Prot Dosim. 2010;139(1-3):449-54.

102. Tapiovaara MJ, Sandborg M. How should low-contrast detail detectability be measured in fluoroscopy? Med Phys. 2004;31(9):2564-76.

103. Vachutka J, Dolezal L, Kollmann C, Klein J. The effect of dead elements on the accuracy of Doppler ultrasound measurements. Ultrason Imaging. 2014;36(1):18-34.

104. Rosenfeld E, Jenderka KV, Kopp A, Keim V. How perfect are you with defective probes? Information on the results of the mini-trial on technical quality assurance during the "Ultraschall 2012" conference in Davos. Ultraschall Med. 2013;34(2):185-8.