Original paper

Evaluation of an automatic method for detection of defects in linear and curvilinear ultrasound transducers

Robert Lorentsson^{a,b,*}, Nasser Hosseini^b, Lars Gunnar Månsson^{a,b}, Magnus Båth^{a,b}^a Department of Radiation Physics, Institute of Clinical Sciences at Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden^b Department of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, Gothenburg, Sweden

ARTICLE INFO

Keywords:

Transducer

Probe

Ultrasound

ABSTRACT

Purpose: The high incidence of defective ultrasound transducers in clinical practice has been shown in several studies. Recently, a novel method using only stored images for automatic detection of defective transducers was presented. The method makes it possible to remotely monitor many transducers at the same time and send a notification when a defective transducer is found. The purpose of the present study was to evaluate the novel method and assess how well it performs when compared to an established method as reference.

Methods: To evaluate the novel method, in-air images were collected from 81 transducers in radiologic departments in nine hospitals. Two observers assessed the in-air images and marked the defects. Receiver operating characteristic (ROC)- and alternative free response receiver operating characteristic (AFROC)-curves and their figures of merit (FOM) were calculated for the novel method, using marked defects in the in-air images as reference truth.

Results: The area under the ROC curve was 0.88 (SD 0.06), and the AFROC FOM was 0.71 (SE 0.07).

Conclusion: The result shows that the novel method has a good agreement with the in-air method for detecting defects in ultrasound systems. This indicates that the novel method could be a complement to the normal quality control for early, and automatic detection of defects.

1. Introduction

Ultrasound transducers are by nature exposed to harm, and the high incidence of defective transducers in clinical practice has been shown in several studies [1–6]. Mårtensson et al. [1] tested 676 transducers from seven manufacturers using an electronic tester (FirstCall (Sonora Medical Systems, Inc., Longmont, CO, USA)) and found that 39.8% exhibited some kind of transducer error. In a follow-up study [2] 299 transducers that were classified as fully functional the previous year were tested again. 27.1% of the transducers were found defective and the conclusion was that annual testing is not sufficient. Sipilä et al. [3] tested 151 transducers using FirstCall, of which 135 also were tested using a tissue mimicking phantom. Transducers and scanners were also visually checked. For the FirstCall and the phantom test the proportion of defective transducers was 17% and 16% respectively. The tested methods produced partly complementary results, and all methods seemed to be necessary. One reason why the methods complement each other is that the electronic test of the transducer cannot find faults that are located in the scanner.

In a 2011 study including 265 transducers over a 4-year period, mechanical integrity and uniformity evaluations were most effective in detecting equipment defects [4]. The annual scanner component and transducer failure rates were 10.5% and 13.9%, respectively. The mechanical integrity and uniformity evaluations together with defects detected by clinical sonographers accounted for 98.4% of all detected failures. Dudley and Woolley [5] performed a multicenter survey of the condition of ultrasound transducers. The only method used was the in-air reverberation method [7]. When a dropout was seen or delamination was suspected, it was checked with the paperclip method [8]. When these simple methods were used, 37% of the investigated 219 transducers were found faulty, and for 13% immediate replacement was recommended. The same authors did a blinded comparison between an in-air reverberation method and an electronic probe tester (FirstCall) in the detection of transducer faults [9]. A total of 62 transducers were investigated, of which 28 were detected as faulty with the two methods. The in-air reverberation and the electrical measurement detected 93% and 89% of the faults, respectively. The studies show that there is a high rate of defects and it is desirable to detect these defects as early as

* Corresponding author at: Sahlgrenska sjukhuset, Blå stråket 7, plan 2, 413 45, Göteborg, Sweden.

E-mail address: robert.lorentsson@gu.se (R. Lorentsson).

<https://doi.org/10.1016/j.ejmp.2021.03.025>

Received 26 November 2020; Received in revised form 18 March 2021; Accepted 21 March 2021

Available online 6 April 2021

1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

possible.

The existing methods of transducer quality control all require access to the ultrasound equipment, or at least the transducers, and this access for testing takes valuable time from the clinical use. These quality control tests are recommended to be performed every three months for mobile and emergency room systems and every six months for others in the report of AAPM Ultrasound Task Group No.1 [10]. Recently, a novel method for detecting defective ultrasound linear transducers by analyzing clinical images was introduced in a case study [11]. The method uses the information in the clinical images to find defects that can be seen by assessing the horizontal uniformity. A number of images are averaged and darker streaks in the superficial part of the images are identified as defects. By using clinical images, no access to either the transducer or the ultrasound scanner is required, and the method can be used to automatically monitor many transducers remotely at the same time and to get a notification, when a defective transducer is found. Intermittent defects, in the meaning of defects appearing and disappearing during long time periods, are easy to follow by looking at the history, although defects that appear in single images now and then are not detectable by the method.

The defects that were identified and visualized by the novel method [11] using clinical images showed good visual agreement with FirstCall measurements for a small selection of transducers in the proof of concept study [11], but a thorough evaluation of the method has not up till now been performed. The main purpose of the present study was therefore to perform an extensive evaluation of the novel method and to evaluate it against an established method. Another purpose was to test the different parameters that are used by the method and to investigate how much their settings affect the results. In the previous study only linear array transducers were reported, while in the present study also curvilinear transducers were included.

2. Methods

This retrospective study using clinical images was approved by the Regional Ethical Review Board. The requirement for informed consent was waived since the study was based on previously collected clinical images and since the analysis was performed on non-identifiable images, created from data from a large number of clinical images. In the Region Västra Götaland, Sweden, there is a Vendor Neutral Archive (VNA) for radiological and ultrasound images. In the present study, clinical ultrasound images from the radiological departments in nine hospitals were used as input for the evaluation of the novel method. For comparison, in-air images were collected from the ultrasound scanners and the associated transducers. Two observers established the reference truth by assessing the in-air images. Receiver Operating Characteristics (ROC)- and Alternative Free Response Receiver Operating Characteristics (AFROC) curves were calculated as measures of the level of agreement between the novel method and the in-air method. The Area Under the Curve (AUC) is the Figure Of Merit (FOM) both for ROC and AFROC [12]. A complete agreement between the tested method and the reference method would result in an AUC of 1.0.

2.1. The novel method

If a part of a transducer is defective, the ability to send and receive signals is affected. The origin of the defect can be e.g. short circuit, oxide at the connector, cable break, dead- or weak elements, or delamination. For linear and curvilinear transducers, this results in a vertical dark streak in the image just under the defective part of the transducer. The idea of the novel method is to use the fact that every clinical image produced with a defective transducer has these diffuse vertical darker streaks. In a given clinical image, it may be difficult to perceive this defect, since it may be hidden in the inhomogeneous anatomical background. However, by averaging a number of clinical B-mode images, the defect will emerge, since it is present in all images, whereas the

Fig. 1. An example of an in-air reverberation image of a curvilinear transducer with a small defect.

anatomical variations tend to cancel each other. In the method proposed to implement this idea, the clinical images are piled in an image stack, which is used to create a Systematic Dark Region (SDR) curve [11]. The SDR curve has a positive value where dark regions are detected in the superficial part of the images and is zero where no dark regions are detected. The position of the detected dark streak in the SDR curve is the same as the position of the defect on the transducer. All steps required can be automated and performed by a computer for many transducers at the same time. This method for automatic detection of defective linear ultrasound transducers was, as mentioned before, presented in a previous paper [11], where a detailed description of how the SDR curve is calculated is given.

2.2. Other methods for detecting defects

The in-air reverberation method has been recommended for use in quality assurance of ultrasound equipment for several years [13]. By using the appropriate settings on the scanner, it is possible to detect defects normally located in the transducer but also in the scanner. The transducer is held in open air and a dark streak appears, where there is probably an element or data channel defect (Fig. 1). The method can also be used for sensitivity tests [14]. To get a more objective evaluation of a transducer, an electronic transducer tester such as FirstCall or Probehunter (BBS Medical AB, Stockholm, Sweden) can be used. Electrical measurements of a transducer are performed by connecting the connector of the transducer to the equipment. The head of the transducer is mounted at the surface of a water bath and is directed towards a reflecting metal target. There are different targets depending on whether the transducer is flat (linear or phased array) or curved. Pulses are sent elementwise to the target and the echoes are evaluated. A report is created containing, among other parameters, the sensitivity of individual elements and a capacitance plot. As an alternative, a manufacturer can include a self-check of the transducer and the scanner data channel in their equipment. One manufacturer included in the present study has an internal sensitivity check in some of their scanners. The check is performed while the transducer is in its holder and contains element sensitivity for all elements very similar to the bar plot from FirstCall or Probehunter. In the present study, both Probehunter and Philips (Philips Healthcare, Amsterdam, the Netherlands) internal checks were used to train and calibrate in-air assessments by two observers, as described later.

Table 1

The 81 transducers and the number of each model used in the study.

Brand	Transducer	Type	Number
General Electric	ML6-15	Linear	23
General Electric	9L	Linear	11
General Electric	C2-9	Curvilinear	4
General Electric	C1-5	Curvilinear	2
General Electric	C1-6	Curvilinear	14
General Electric	L8-18i	Linear	4
Philips	C5-1	Curvilinear	4
Philips	C7-2	Curvilinear	1
Philips	C9-2	Curvilinear	1
Philips	L12-5	Linear	3
Philips	L17-5	Linear	2
Toshiba	PVI-475BX	Curvilinear	3
Toshiba	PLT-1005BT	Linear	5
Toshiba	PLT-705BT	Linear	1
Toshiba	PLI-1205BX	Linear	1
Toshiba	PVT-712BT	Curvilinear	1
Toshiba	PLT-1204BT	Linear	1

2.3. Data collection

As an initiating point for the study, a survey among the scanners in the region using the VNA was made. A total of 37 scanners and 152 linear and curvilinear transducers were found. The settings for the scanners when collecting the in-air images were decided as follows:

- Choose a setting that the transducer normally is used with.
- To reflect the clinical use, choose a frequency as low as possible for curvilinear transducers and as high as possible for linear transducers [6].
- Choose a suitable depth that makes the whole transducer head visible in the image.
- Set one focus as shallow as possible.
- Put the TGC (Time Gain Compensation) levers in a neutral position.
- Make the image acquisition as basic as possible, by disabling (or minimizing) harmonic imaging, spatial compounding and time averaging controls.
- Turn up gain, until the brightest part in the image does not get saturated.
- Save the image in DICOM (Digital Imaging and Communications in Medicine) format.

A total number of 152 single-frame in-air images from both linear and curvilinear transducers were collected from radiological departments in nine hospitals. For 24 of the transducers, electrical measurements were performed as well. These in-air images together with the electrical measurements were used for training of the two observers, who would assess the in-air images for defects. These 24 images were then excluded from the material.

In [11], 150 images were used to produce one SDR curve. Therefore, 37 transducers, that had not been used for 150 clinical images (that were approved by the extraction algorithm) during the 9–12 months prior to the study, were also excluded. Of the remaining 91 transducers, four had a very sharp curvature (two Philips C8-5 and two GE C3-10) and were deemed not suitable for the novel method, because many clinical images were missing 100% skin contact for the full curvature. This made the median images dark at the edges, so these four transducers were also excluded. Five of the in-air images were collected in Virtual Convex mode. As this affects the beam steering at the ends of the arrays, these transducers were excluded. Finally, one transducer showed a very strange pattern in the in-air image for more than half of the transducer. The day after the in-air image was normal. This transducer was also excluded. When the excluded transducers were removed, the number of the remaining transducers was 81. The models and numbers of the remaining transducers are shown in Table 1.

Fig. 2. Illustration of the image extraction lines of curvilinear transducers when the B-mode image was extracted from the surrounding information.

2.4. Image extraction for curvilinear transducers

For the novel method to be able to use the clinical images, the B-mode images must be extracted from the surrounding information (such as patient name, logos etc.). In the previous study [11], this was described for linear transducers. In the present study, curvilinear array transducers were included as well. An in-house developed MATLAB (MathWorks, Inc., Natick, MA, USA) application was used for this purpose. To extract curvilinear images the largest area of non-black pixels was identified, the B-mode area (this technique using non-black pixels would probably not work if the images are irreversibly compressed). The borders for the top arc were automatically detected. A circle was constructed using coordinates from the top arc. The MATLAB function `improfile` was used to collect the image material along the lines crossing the origin of the circle, starting at the top arc and ending at the lower arc (Fig. 2). If the angle was wider than the widest angle for the actual transducer or narrower than three degrees below the widest transducer angle, the extracted image was discarded. The collected pixels were then used in a rectangle the same way as for the linear transducers.

2.5. Assessment of the SDR curves

Clinical images were collected retrospectively for at least 10 months back from when the in-air images were gathered, and they were sorted on a day-to-day basis. The images were extracted and placed in an image stack, that was updated with new images every day in a first in- first out queue system. This was made for every transducer. The same parameter settings (like depth, number of images etcetera) for calculating the SDR curves were used as in the previous study. To detect possible defects, for each transducer the SDR curve for which the date of the last image in the stack was nearest the date for the in-air image was first selected. This SDR curve was then assessed for signals indicating defects. If a signal was present for 20 consecutive days in adjacent SDR curves, it was assessed as lasting and classified as a defect. If a signal was classified as a defect, the amplitude of the signal in the originally selected SDR curve was recorded and used as input (signal level) in the ROC and AFROC analyses.

The SDR curves are created using three different built-in thresholds [11], meaning that the signal level of possible defect must exceed a certain value to contribute to the SDR curve and be reported. Although these thresholds can be altered, in the present study the same settings as in the previous study were used. The SDR curves were inspected manually, and the median image was not used. Fig. 3 shows an example of the in-air image, one SDR curve and the median image from the clinical images.

Fig. 3. An example of in-air image (a), the red arrow is what two observers assessed as a defect. The SDR curve (b) and the median image (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.6. Test of parameter settings

Different parameter settings used for calculating the SDR curves were tested to investigate to what extent the result was affected. Firstly, the depth of the portion of the images from which the information to the SDR curve was collected. This was tested using pixels 1–30 (of 500) instead of 1–19 as used in the previous study. A decreased number of images in the stack was also tested down to 50 instead of 150. The polynomial degree of the two polynomials that are used for baseline compensation (Opolyred and Opolygreen [11]) was tested with three instead of six.

2.7. Training of the observers and establishment of reference truth

The observers had 24 in-air images and 24 electrical measurements from the same transducers to use for training purposes. Even if subjective assessment is a well-established method for quality control, the result is depending on the threshold of the observer. This threshold was calibrated against the objective method by the observers by comparing the 24 in-air images with the electrical measurements. To establish the reference truth, the two trained observers then separately evaluated the 81 in-air images and marked the assessed defects. Reference truth in this case was just the identification of a transducer (or channel) defect. It did not matter if the location of the defect was in the center of the image or how severe the defect was. In this study, the goal was just to identify defects and no consideration was taken to if the defect was judged to be clinically significant. The observers were blinded to the results of the novel method. Where there were differences in the assessments, the observers met to reach a consensus. In no case the observers had difficulties in reaching consensus, indicating that observer variability and not systematic effects was the reason for the originally different assessments. The observers finally established 15 discrepancies to be defects in the images and marked the positions; several defects could appear in the same in-air image. Two of the in-air images were assessed to have two and three defects, the rest of the defects were singular. Six of the defects were located in six linear transducers and nine of the defects

in six curved transducers. The result from the observers was used as reference truth when calculating the ROC and AFROC curves for the novel method.

2.8. Evaluation

ROC is often used in task-based evaluations, where detection of lesions or other focal abnormalities is the main task [15]. The task for the observer (a human or, as in the present study, an algorithm) is to answer the question, if there is abnormality for each image. The ROC method is based on a case-level assessment and it makes no difference if the observer e.g. has marked all lesions or if their location is right [16]. The area under the ROC curve is a measure of how well the observer performs the task, where 1 is perfect and 0.5 is no better than chance. One criticism against the ROC method is, that the observer can get a positive case right, even if the assessment, that there is a lesion, is done in a non-lesion region of the image. Defects in the transducers can be several and their locations can vary. In AFROC, localization and number of lesions are included in the analysis of the observer's performance. Therefore, the results for both ROC and AFROC are presented as measures of the ability of the novel method to find the defects in the ultrasound transducers (or systems) in the present study.

The result from the SDR curves (the SDR signal levels for all classified defects) and the reference truth from the observers were used as input to the software Rjafroc (Pittsburgh, PA) v1.2.0.9000 to calculate ROC and AFROC curves, as well as the FOM for ROC and AFROC. Rjafroc is a statistical software; available from <https://dpc10ster.github.io/RJafroc/index.html>, last accessed 20210203. The ROC curve is a plot of the true positive fraction (case-level sensitivity) vs. the false positive fraction (1-specificity) as the decision threshold is altered, here corresponding to the proportion of actually defective transducers (according to the reference truth) accurately reported as defective by the novel method vs the proportion of actually healthy defective transducers inaccurately reported as defective as the SDR signal level threshold is altered. The AFROC curve is a plot of the lesion-localization fraction (lesion-level sensitivity) vs. the false positive fraction as the decision

Fig. 4. The binormal ROC curve for the novel method. The area under the curve was 0.88 (SD 0.06).

Fig. 5. Results from the AFROC analysis of the novel method. The FOM was 0.71 (SE 0.07).

threshold is altered, here corresponding to the proportion of actual defects (according to the reference truth) accurately reported as defects by the novel method vs the proportion of actually healthy defective transducers inaccurately reported as defective as the SDR signal level threshold is altered. Additionally, the case-level sensitivity and specificity were determined based on all defects classified by the novel method, irrespective of their SDR signal levels (SDR signal level > 0).

3. Results

Using the same settings for the SDR curves as reported in the previous study, the FOM for ROC (Fig. 4) was 0.88 (SD 0.06) and the FOM for the AFROC (Fig. 5) was 0.71, (SE 0.07). Fig. 6 shows the distribution of the case-level SDR signal level (the highest reported SDR-signal level for each case) for the cases (transducers) established as defective by the reference truth, whereas Fig. 7 shows the corresponding distribution for the healthy cases (transducers). Table 2 presents the SDR result compared to the in-air result on a case level, showing that the novel method achieved a case-level sensitivity of 67% at a specificity of 87%

Fig. 6. The distribution of the case-level SDR signal level (the highest reported SDR signal for each case) for the 12 cases (transducers) established as defective by the reference truth. An SDR signal level of 0 corresponds to the novel method not reporting any defects for the case.

Fig. 7. The distribution of the case-level SDR signal level (the highest reported SDR signal for each case) for the 69 cases (transducers) established as healthy by the reference truth. An SDR signal level of 0 corresponds to the novel method not reporting any defects for the case.

Table 2

A 2×2 contingency table for the case-level outcome of the novel method (SDR), based on all reported defects irrespective of SDR signal level (>0), and the reference truth (in-air). The resulting sensitivity and specificity of the novel method is 67% and 87%, respectively.

	Positive in-air	Negative in-air	
Positive result SDR	8	9	17
Negative result SDR	4	60	64
	12	69	81

when all reported defects were included.

The change of the settings only marginally affected the result. The increased depth resulted in darker regions in the edges of the extracted

part of the images and made the AFROC FOM value significantly smaller (0.64, $p = 0.014$) than the case, when the original depth was used. The decreased number of images resulted in some temporary false SDR curves, but since the limit was 20 days, this did not affect the AFROC FOM. The decreased degree of the polynomial used in the curve fitting for baseline compensation, did also not affect the AFROC FOM.

4. Discussion

Recently a novel method [11] for detecting defects automatically in ultrasound transducers by analyzing the statistics in the clinical B-mode images was developed. By analyzing images from the clinical workflow, it is possible to monitor the equipment without interference with the clinical work. The main purpose of the present study was to evaluate the novel method against another known method, where assessment of the in-air image was chosen. Visual subjective assessment of in-air images for detection of defects has been used in several studies [3–5,7,9,17]. The in-air method has also been suggested for computerized evaluation for detection of transducer defects in in-air images [18,19]. The method is applicable to all linear and curvilinear transducers; therefore the in-air method was chosen as reference method for the present study. 81 in-air images from 81 transducers and 33 scanners were assessed by two observers, who marked the locations for suspected defects. The result of these assessments was compared with the result from the novel method by using ROC and AFROC curves and their figure of merits. A good agreement (ROC AUC = 0.88) between the novel method and the in-air method was found.

There are several established methods to use as reference to choose from, all with their own drawbacks and advantages. Electrical measurements are very precise and objective but do not include defects that are located in the scanner. Transducer-reference records and adapters must be available for all transducers to be tested, which was not the case for the transducers in the present study (Probehunter could not handle the multiplexed GE ML6-15 or Philips L12-5 for example at the time of the data collection). Goodsitt et al. [10] recommend to use a tissue mimicking phantom for visual inspection of the screen to detect both vertical and horizontal nonuniformities. Phantom measurements are similar to the in-air method, both includes scanner defects, but the assessments are subjective. Phantom measurements and the novel method is not functional for phased arrays transducers, for these another method can be used [20].

The choice to use two different metrics for the evaluation was made to use one classic (ROC) and one more suitable for the fact that both the in-air image method and the novel method can use localization of the defects (AFROC). The difference in the results was expected, since some of the lesion-level false positives were interpreted as true positives by the ROC method, analyzing the data only on a case level.

The output from the novel method was the amplitude of the SDR signal at the location of the detected defects. Fig. 6 and Table 2 show that not all actual defects were detected by the method, even at the lowest SDR signal level. One reason for this could be that there are three built-in thresholds in the algorithm that calculates the SDR curve. [11] The smallest value of deviation from the mean of the layer nearest the transducer was 2 (Tgreen in the previous study) of the 8-bit images. This way, a threshold level of 2 for the SDR curve is effectively used. If a higher sensitivity is desired, the parameters of the algorithm must be changed, whereas if a higher specificity is wanted, an additional threshold can be applied to the calculated SDR signals. Although the ROC curve in Fig. 4 shows the compromise between sensitivity and specificity for the novel method, the number of defective transducers in the present study was too small for an analysis of optimal sensitivity/specificity settings.

The 20-day requirement on the 150 images data was added to decrease false positive SDR signals that appeared for short periods, and to imitate a real situation, where the positive SDR signal are followed for some days to see if the defect persists. The fact that the SDR curves were

updated once a day was a result of our design, where the images are fetched once a day. If few images were replaced each day, it is possible that the SDR curves would be highly correlated, and the artifacts would be more likely to be persistent. To have a condition where the image stack would have been replaced by X images Y times and a SDR signal is present all the time would probably have been a better condition for a fair comparison than the 20-day role. The result on a case-level without the 20-day requirement was a 67% sensitivity and 80% specificity.

The test of 50 images instead of 150 showed no difference in the result when the requirement of 20 consecutive days was applied. Without the 20 days requirement there was several shorter positive SDR signals. When fewer images were used there were naturally more positive SDR signals, both false positives and true positives. In a real monitoring situation, a check-up of a suspected defective transducer probably would be done before 20 days have passed, maybe after a few days. In such a situation too many false positives should be avoided. 150 images in the previous study were chosen to get good visual agreement with the firstCall measurements and could probably be decreased to 100 images in a monitoring situation for a faster response.

To use the setting virtual convex when collecting in-air images is not optimal, since the detection at the flank of the arrays might be affected. There were five images that were collected using virtual convex, these were excluded for this reason. The exclusion had minor impact on the result.

It may be difficult to interpret the achieved FOM values, especially the AFROC FOM of 0.71. To the best of the authors' knowledge, no categorization of obtained AFROC FOM values has been proposed and the AFROC FOM is mostly used for its efficiency in finding statistical differences between compared modalities or settings. The ROC is more common and the FOM is sometimes used with the following scale; 0.5–0.6 fail, 0.6–0.7 poor, 0.7–0.8 fair, 0.8–0.9 good and 0.9–1.0 excellent [21,22]. According to this scale the obtained result 0.88 is good, and closer to excellent than to fair. The ROC FOM also has a well-known interpretation in that it corresponds to the percentage correct decisions in a two-alternative forced choice experiment. For the present study, this means that if the novel method would be applied to a randomly chosen defective transducer and to a randomly chosen healthy transducer, with the task of determining which one that is defective, in 88% of the cases the method would report the correct one.

The easiest way to detect defective transducers in a timely manner in clinical use is to have a user check at the beginning of each session, for example by performing an in-air image check. However, as mentioned before there are several studies that show that a large number of defective transducers are found at periodic quality assurance, indicating that routine user checks are not as common as they should be, and it is as a complement to these periodic tests the novel method has been developed.

4.1. Limitations of the study

The present study has several limitations. When an in-air image is acquired, the settings are important. For some transducers, it was difficult to get one single focus at a shallow depth. In these cases, multiple focuses were chosen to get at least one shallow focus. Some of the transducers in the study are multi-row arrays. When collecting in-air image using one shallow focus all rows may not be used. It is also less likely, that the evidence of element failure will be seen in clinical use unless all elements across the slice are faulty. Thus, for both methods single elements may have been hard to detect for multi-rows arrays. The fact that there were only 15 classified defects (in 12 transducers) by the reference method is of course a limiting factor. 15% of the transducers were defective, which is in the lower range compared to previous studies [1–5]. However, the fact that 81 transducers from nine radiological departments were included makes the material quite extensive for an investigation, where stored images are required. Another limitation is the fact that the outcome of the reference method is dependent on two

Fig. 8. A screenshot of the software tool used for the novel method. Available transducers (a), the median image for the user to evaluate (b). Historical SDR curves (d) and (c) the chosen SDR curve. The historical area under SDR curve where it is possible to follow a defect from the start (e). The transducer in the example has several defects and it is possible to follow when they arose.

observers' subjective assessments. Even if the observers were "calibrated" against an objective method, the subjectivity still could be a limiting factor.

4.2. Clinical experience of the novel method

The novel method has been tested over a period of time of two years on clinical images, and our practical experience of the method is that it is no problem to monitor a large number of transducers (in our case 152) at a time by using one single computer. To import the images takes about 0.5 h, to extract the images and update the image stacks takes one hour, and to make new SDR curves for new images takes another hour. These activities can be automated and carried out at night, and in the morning the SDR curves are updated. An interface has been developed, that presents the latest SDR curve, the latest area under the SDR curve, previous SDR curves in a 3-D plot, previous areas under the SDR curves, and a median image. For every transducer, it is possible to scroll back in time to follow defects, intermittent defects are easy to follow this way. Fig. 8 shows a screenshot of the software tool, that is used for the novel method. The novel method has not been implemented in the routine workflow yet, it is mainly used as a complement to the normal quality assurance. To assess the current SDR curve, the historical SDR curves and the current median image, it only takes a few seconds per transducer when assessing them one by one manually. The median image can be used as manual verification when the SDR signal is positive. It is also possible to set an alarm for when the area under the SDR curve reaches a certain limit for automatic detection, although this has not been implemented yet.

The method is mainly applicable for transducers that are frequently used and where there are many images saved. For mammography, for example, a transducer usually generates 150 or more usable images in a typical week. For these transducers, completely independent 150-image medians are produced approximately weekly.

For transducers that produce few usable images, the method is not

suitable. In the present study, there were 37 transducers that had not produced 150 usable images for 9–12 months. It was not known, if these transducers were used without saving the images or if they were just not used frequently. For transducers that are seldom used, traditional quality control testing covers the need. Whether the novel method can replace the uniformity check in the normal quality assurance is an interesting question. A defect in form of dark vertical streak can have its origin in the transducer or internal in the scanner. Therefore, a second manual check is needed, for example two in-air images tested on two different ports or an electric test of the suspected transducer. If the novel method produces no positive findings, the manual uniformity check could probably be omitted.

5. Conclusion

The present study shows, that the novel method for automatic detection of defects in ultrasound systems using clinical images has a good agreement with a well-established method for quality assurance. This indicates, that the novel method could be used as a complement for early and automatic detection of defective transducers between the normal quality controls. The method could also be used to supervise minor defects to see, if they grow or keep steady. The advantages of the method are that it can be fully automated, that it is objective and can be used on many transducers at the same time, that the interference of the clinical examinations is non-existent, and that the method has potential to decrease the time from when a defect occurs until it is detected.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by The Healthcare Board, Region Västra Götaland, [grant number VGFOUREG-842791]. The authors would like to thank Jesper Stenström, Anders Dahlgren, Andreas Magnusson and Reinhold Sandelin for valuable help with collection of in-air images.

References

- [1] Martensson M, Olsson M, Segall B, Fraser AG, Winter R, Brodin L-A. High incidence of defective ultrasound transducers in use in routine clinical practice. *Eur J Echocardiogr* 2009;10(3):389–94.
- [2] Martensson M, Olsson M, Brodin L-A. Ultrasound transducer function: annual testing is not sufficient. *Eur J Echocardiogr* 2010;11(9):801–5.
- [3] Sipilä O, Mannila V, Vartiainen E. Quality assurance in diagnostic ultrasound. *Eur J Radiol* 2011;80(2):519–25.
- [4] Hangiandreou NJ, Stekel SF, Tradup DJ, Gomy KR, King DM. Four-year experience with a clinical ultrasound quality control program. *Ultrasound Med Biol*. 2011;37: 1350–7.
- [5] Dudley NJ, Woolley DJ. A multicentre survey of the condition of ultrasound probes. *Ultrasound* 2016;24(4):190–7.
- [6] Vitikainen A-M, 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.
- [7] 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.
- [8] Goldstein A, Ranney D, McLeary RD. Linear array test tool. *J Ultrasound Med*. 1989;8:385–97.
- [9] 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.
- [10] 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;25:1385–406.
- [11] Lorentsson R, Hosseini N, Johansson J-O, Rosenberg W, Stenborg B, Månsson LG, et al. 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(2):265–74.
- [12] Chakraborty DP. A brief history of free-response receiver operating characteristic paradigm data analysis. *Acad Radiol* 2013;20(7):915–9.
- [13] 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.
- [14] Dudley NJ. A proposed enhancement to the in air sensitivity test for ultrasound quality assurance. *Phys Med* 2018;53:1–3.
- [15] Chakraborty DP, Haygood TM, Ryan J, Marom EM, Evanoff M, McEntee MF, et al. Quantifying the clinical relevance of a laboratory observer performance paradigm. *Br J Radiol* 2012;85(1017):1287–302.
- [16] Chakraborty DP. Clinical relevance of the ROC and free-response paradigms for comparing imaging system efficacies. *Radiat Prot Dosim* 2010;139(1-3):37–41.
- [17] Quinn T, Verma PK. The analysis of in-air reverberation patterns from medical ultrasound transducers. *Ultrasound* 2014;22(1):26–36.
- [18] Rosenfeld E, Kopp A, Liebscher E, Jenderka K-V. Quick test of ultrasonic transducer arrays radiating in air using B-mode-images. *Biomedizinische Technik/Biomedical Engineering* 2014:47.
- [19] van Horsen P, Schilham A, Dickerscheid D, van der Werf N, Keijzers H, van Almere R, et al. Automated quality control of ultrasound based on in-air reverberation patterns. *Ultrasound* 2017;25(4):229–38.
- [20] Dudley NJ, Woolley DJ. A simple uniformity test for ultrasound phased arrays. *Phys Med* 2016;32(9):1162–6.
- [21] 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.
- [22] 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.