

Shear wave elastography for breast masses is highly reproducible

David O. Cosgrove · Wendie A. Berg · Caroline J Doré ·
Danny M. Skyba · Jean-Pierre Henry · Joel Gay ·
Claude Cohen-Bacrie · the BE1 Study Group

Received: 22 June 2011 / Revised: 13 September 2011 / Accepted: 1 October 2011 / Published online: 31 December 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract

Objectives To evaluate intra- and interobserver reproducibility of shear wave elastography (SWE) for breast masses. **Methods** For intraobserver reproducibility, each observer obtained three consecutive SWE images of 758 masses that were visible on ultrasound. 144 (19%) were malignant. Weighted kappa was used to assess the agreement of qualitative elastographic features; the reliability of quantitative measurements was assessed by intraclass correlation coefficients (ICC). For the interobserver reproducibility, a blinded observer reviewed images and agreement on features was determined. **Results** Mean age was 50 years; mean mass size was 13 mm. Qualitatively, SWE images were at least reasonably similar for 666/758 (87.9%). Intraclass correlation for SWE

diameter, area and perimeter was almost perfect (ICC \geq 0.94). Intraobserver reliability for maximum and mean elasticity was almost perfect (ICC=0.84 and 0.87) and was substantial for the ratio of mass-to-fat elasticity (ICC=0.77). Interobserver agreement was moderate for SWE homogeneity (κ =0.57), substantial for qualitative colour assessment of maximum elasticity (κ =0.66), fair for SWE shape (κ =0.40), fair for B-mode mass margins (κ =0.38), and moderate for B-mode mass shape (κ =0.58), orientation (κ =0.53) and BI-RADS assessment (κ =0.59).

Conclusions SWE is highly reproducible for assessing elastographic features of breast masses within and across observers. SWE interpretation is at least as consistent as that of BI-RADS ultrasound B-mode features.

Key Points

- Shear wave ultrasound elastography can measure the stiffness of breast tissue
- It provides a qualitatively and quantitatively interpretable colour-coded map of tissue stiffness
- Intraobserver reproducibility of SWE is almost perfect while intraobserver reproducibility of SWE proved to be moderate to substantial
- The most reproducible SWE features between observers were SWE image homogeneity and maximum elasticity

D. O. Cosgrove
Imperial College,
Hammersmith Campus,
London, UK

W. A. Berg
Magee-Womens Hospital,
University of Pittsburgh School of Medicine,
Pittsburgh, PA, USA

C. J. Doré
MRC Clinical Trials Unit,
London, UK

D. M. Skyba · J.-P. Henry · J. Gay · C. Cohen-Bacrie
SuperSonic Imagine,
Aix en Provence, France

D. O. Cosgrove (✉)
Imaging Sciences Department, Imperial College,
Hammersmith Hospital,
Du Cane Road,
London W12 0HS, UK
e-mail: d.cosgrove@imperial.ac.uk

Keywords Breast diseases · Ultrasound · Elasticity imaging · Shear wave elastography · BI-RADS · Biopsy

Introduction

While breast ultrasound is a valuable adjunct to mammography and other breast imaging methods, it suffers from low specificity [1–3]. Elastography is a method of imaging tissue stiffness. Insofar as many benign masses are soft and malignancies tend to be stiff, elastography may be able to

improve patient management for masses with a low index of suspicion on B-mode ultrasound [4, 5].

Various methods for performing elastography are available on many current ultrasound systems from a variety of manufacturers. Most of them rely on some form of deformation of the tissue, i.e. strain, either by hand pressure and release, or by exploiting normal cardiac or respiratory motion. While individual images acquired using these approaches can be interpreted consistently [6], it can be difficult to capture the same information across acquisitions or users, and this may hamper clinical utility [7].

In ShearWave™ Elastography (SWE™, Supersonic Imagine, Aix-en-Provence, France) an acoustic pressure wave induces slow-moving lateral waves within the tissue, and the speed of propagation of the shear wave is proportional to the square root of the tissue's elastic modulus [8]. Shear waves travel more slowly in softer tissue and faster in stiffer tissue. Ultrafast™ imaging of the propagation of shear waves allows measurement of the small changes in velocity that occur when the waves pass through tissues of different stiffness. The velocity information can be mapped to create an image of the stiffness, with the option of measuring SWE features such as the minimum, mean, and maximum elasticity in a region of interest. Deformation of tissue leading to shear waves is created by an acoustic impulse that is generated electronically. As such, we expected that more similar results would be obtained with SWE if the same region of tissue was evaluated repeatedly than with other elastographic techniques.

For any elastographic approach to become widely accepted, consistent image generation and interpretation must be demonstrated. Therefore, the purpose of this study was to evaluate the intra- and interobserver reproducibility of shear wave elastography for breast masses.

Materials and methods

Ethics

This HIPAA-compliant study was approved by the Ethics Committees/Institutional Review Boards of each of the 16 participating sites in the United States and Europe (France, Germany, Italy and the UK). The study was funded by SuperSonic Imagine (Aix en Provence, France), the manufacturer of the prototype ultrasound system equipped with SWE. All SWE images were provided free of charge to participants. The integrity of the database management was audited by an independent guarantor (Alison Roberts, Qualitus, Strasbourg, France).

Participants

Women 21 years of age or older with a breast mass detected by palpation, mammography, ultrasound or MRI, who were referred for breast ultrasound, were invited to participate. We excluded pregnant and lactating women, those with breast implants, women receiving chemotherapy or radiotherapy for any cancer, skin masses and any that had been biopsied, and patients with a history of ipsilateral breast surgery. All participants provided written informed consent. Reference standard of biopsy, fine needle aspiration or at least 1 year of follow-up was available for all masses assessed as BI-RADS® 3 or higher [9]. BI-RADS 2 masses were presumed to be benign. Patient management was based on the standard B-mode ultrasound available at each site (using their “house system”).

Study protocol

All participants underwent diagnostic breast ultrasound on the house breast ultrasound system used in routine practice (with a minimum of 12-MHz peak frequency linear array transducer). All investigators were specialists in breast diagnostic imaging with experience in acquiring and interpreting a minimum of 300 breast ultrasound examinations in the prior 2 years. The participants then underwent an additional B-mode ultrasound examination using the experimental ultrasound system (RUBI, a prototype of the Aixplorer® ultrasound system [Supersonic Imagine, Aix-en-Provence, France]), and B-mode features and BI-RADS assessments were recorded. After identifying the mass of interest using the prototype system, SWE was performed. Three separate images of the mass were acquired in SWE mode. Upon completion, the SWE images were recalled and the user was asked to assess the qualitative similarity of all three images and the quantitative SWE features of the mass were measured. In addition, an expert observer (WAB) performed a blinded review of one set of static images of all masses in B-mode and SWE mode.

Image acquisition: B-mode

The experimental RUBI system was equipped with a 15–4 MHz linear array transducer with default software pre-sets for breast imaging. Initial house and B-mode RUBI images were reviewed. Mass size (longest diameter, perimeter and area), BI-RADS features of shape (oval, round, irregular), orientation (parallel or not) and margin (circumscribed or not) were recorded for each mass, together with a BI-RADS assessment (2, benign; 3, probably benign; 4a, low index of suspicion; 4b, intermediate index of suspicion; 4c, moderate index of suspicion; or 5, highly suggestive of malignancy) [9].

Image acquisition: SWE mode

After recording the B-mode features, SWE imaging was performed with the same transducer on the RUBI system [10]. The recommended technique for SWE was to image the mass with no pressure induced by the transducer before switching to SWE mode. After a few seconds of immobilisation to allow the SWE image to stabilise, the SWE image was frozen and saved. At each site, investigators were first trained on phantoms and then on at least 10 non-study patients before enrolling their first participant.

Three consecutive and distinct SWE images of the mass were acquired by the same operator in the same imaging plane, the probe having been removed and re-applied between each acquisition. In each image, the tissue stiffness of each pixel in the SWE image was displayed as a semi-transparent colour overlay with a range from dark blue, indicating the lowest stiffness (at just over 0 m/s), up to red, indicating the highest stiffness (set at 7.7 m/s for this study) (Fig. 1). Areas of black on the SWE images represent tissue in which no shear wave was detected. Using the experimental system's review capability, the perimeter of the mass (including surrounding tissue stiffness) was traced and the operator determined the largest diameter on each SWE image. Where no zone of altered stiffness was apparent and the mass was homogeneously soft (blue) on SWE, the dimensions recorded were those of the mass seen on the underlying B-mode image.

Each operator visually assessed two features on the colour overlay SWE images: the shape of the mass including the surrounding areas of stiffness (oval, round,

irregular), and stiffness homogeneity (not homogeneous, reasonably homogeneous, very homogeneous). The qualitative similarity among the three SWE acquisitions was recorded as: all three images very similar, all three images reasonably similar, some images similar (2 out of 3) or all three images very dissimilar.

In addition to the size measurements, quantitative elasticity was measured on each of the three SWE images using the system's quantification tool, known as the "Q-Box™", which defined a 2 mm×2 mm region of interest (ROI) that was positioned over the stiffest part of the mass or surrounding tissue on the SWE image (Fig. 2). For each of the three acquisitions, an additional Q-Box was placed in the adjacent fatty tissue; the intention was that this be located in breast fatty tissue, though we recognised that this was not always feasible and was not always performed. The system calculated the minimum (E_{min}), maximum (E_{max}) and mean (E_{mean}) elasticity values in kPa or m/s within the mass and in the fatty tissue, as well as the ratio between the mean elasticity in the mass and the fatty tissue (E_{ratio}).

Interobserver agreement

After an initial training review of 11 proven cases, a second observer with 20 years' experience in breast ultrasound (WAB) reviewed the static images of all of the masses. This observer was blinded to the diagnosis, other imaging findings, and the clinical history. One or two representative B-mode images, or a B-mode and Doppler image, of each case were reviewed and the BI-RADS features of the mass

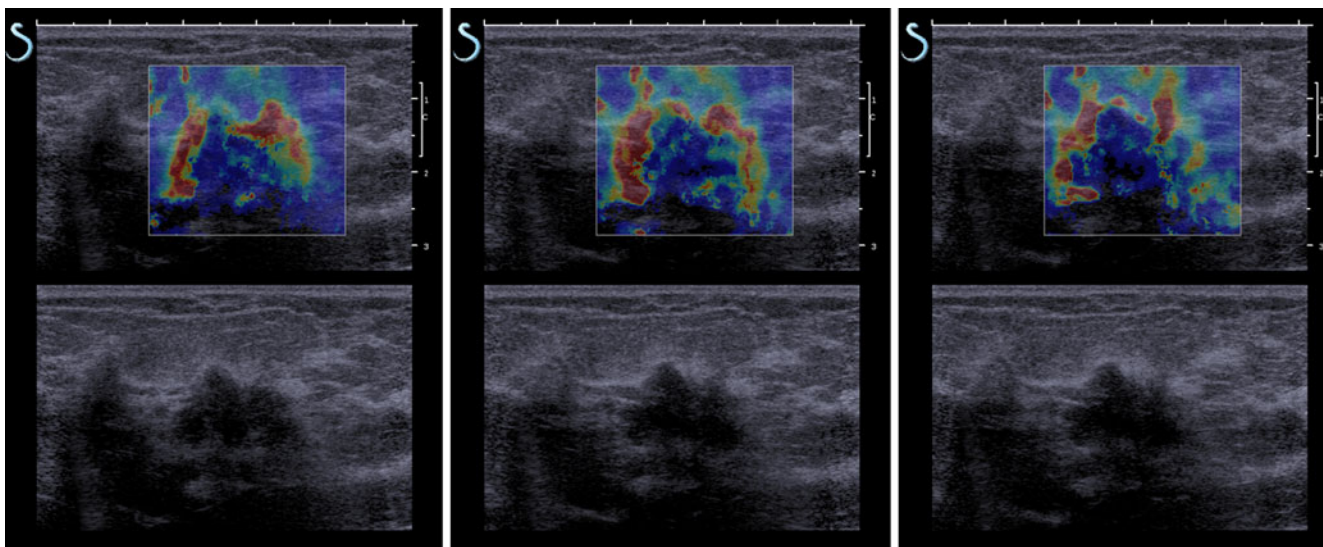


Fig. 1 Three sequentially acquired Shearwave Elastography (SWE™) images of a breast cancer showing the small changes registered over time. The repeatability in this case was scored as "all images very similar". The upper portion of each image shows the tissue elasticity

as a colour overlay where red represents the highest stiffness and blue represents the lowest stiffness. The lower image shows the registered B-mode image

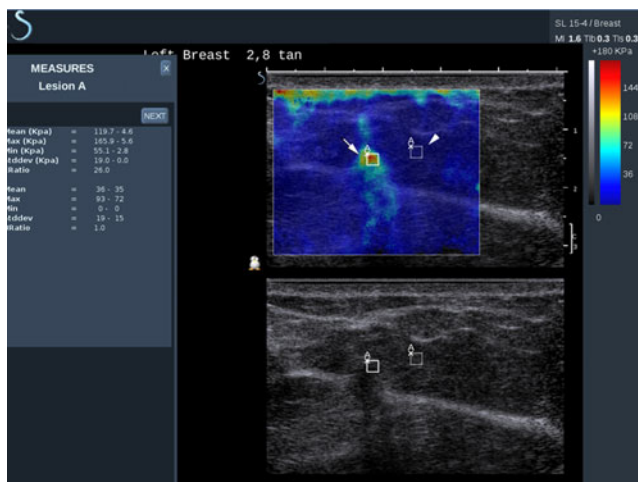


Fig. 2 A 52-year-old woman with invasive ductal carcinoma. Shear wave elastography shows high stiffness values around the lesion (red tints in the upper frame). One ROI is in the stiffest part of the lesion (arrow) and another is in the adjacent subcutaneous fat (arrowhead). The quantitative values for the two ROIs are shown in kPa in the panel on the left: the lesion value is given first (here, mean 6.3 m/s), followed by the value in fat (here 1.2 m/s). This was available on the prototype RUBI system. The commercially available system, Aixplorer, can display the elasticity scale in kPa or m/s

shape (oval, round, irregular), orientation (parallel or not), and margins (circumscribed or not), were assessed, followed by an overall BI-RADS assessment (2, 3, 4a, 4b, 4c, 5). A representative SWE image and its corresponding B-mode image were then reviewed; when all three SWE images had been scored as “very similar”, the first was shown to the observer, and when they had been scored as “some images similar” then the first of the two similar images was reviewed. For cases in which images were assessed as being dissimilar, one image was arbitrarily chosen (by J-PH) for observer review. The shape of the mass and surrounding stiffness on the SWE image (oval, round, irregular), together with the homogeneity of the elasticity (very homogeneous, reasonably homogeneous, not homogeneous), were recorded. In addition, the second observer assessed the qualitative maximum elasticity in the masses using a six-level visual colour scale (Fig. 3). Interpretations from the second observer were compared to the initial mass assessment performed by investigators on the house and prototype systems.

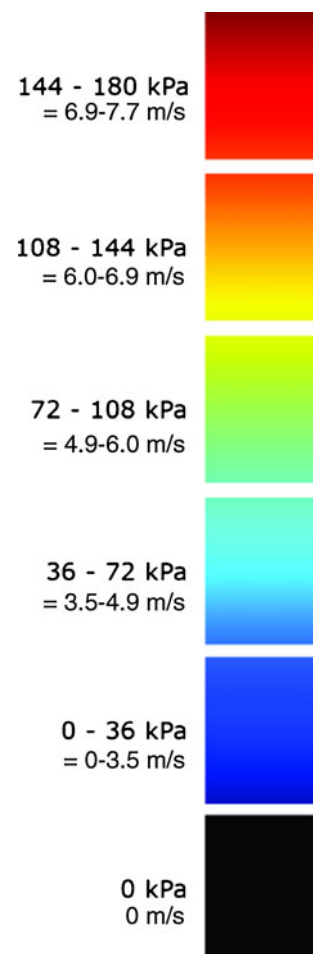
Masses included and excluded

From September 2008 until September 2010, 958 women with 1000 breast masses were initially enrolled. When a participant had multiple masses, only the malignant mass was retained. There were 42 cases with multiple malignant or multiple benign masses; in these, one was selected at random. Seventeen cases with a missing reference standard

(2 papillomas not excised and 15 BI-RADS 3 masses without follow-up), 1 skin lesion, and 1 case with no mass were excluded. Clinical outcomes with SWE for the remaining 939 masses have been reported [10].

The prototype system did not support measurements above 7.7 m/s; this occurred in 113 masses, which were excluded from the reproducibility analysis because there was no possibility of variance as their Emax exceeded 7.7 m/s. An additional 61 cases were excluded because identical rounded values of Emax were observed for one of two reasons: 1) the same SWE acquisition was used more than once for different measurements or 2) the pre-set maximum SWE scale value was reached and was lower than that specified by the protocol (e.g. 7.3 m/s) because of investigator error. Also excluded were 3 cases in which all values of Emax were identical; 2 with missing values; and 2 cases in which the same image acquisition had been used for more than one set of measurements (i.e. data were not independent). In total, 758 masses were retained. Among the 181 excluded masses were 145 malignancies: we repeated the evaluation of agreement with these masses included (i.e. among 939 total masses, including 289 malignancies).

Fig. 3 Colour scale used by the blinded observer. The SWE scale corresponds to that in the colour bar of the RUBI image shown in Fig. 2 but it has been divided into bands as indicated. The blinded reader chose the range corresponding to the stiffest portion of the mass and/or surrounding tissue on the colour overlay. 0 kPa = 0 m/s, 0–36 kPa = 0–3.5 m/s, 36–72 kPa = 3.5–4.9 m/s, 72–108 kPa = 4.9–6.0 m/s, 108–144 kPa = 6.0–6.9 m/s, 144–180 kPa = 6.9–7.7 m/s



Statistical methods

The data were analysed using the statistical software package Stata/IC, release 11.0 for Windows (StataCorp LP, College Station, TX, USA). For continuous variables, intraclass correlation coefficients (ICC) were calculated as a measure of reliability using a one-way analysis of variance (ANOVA). The ICC is the proportion of the total variability (between masses plus measurement error) explained by variation between masses. A $\log_e(X+1)$ transformation was used for the analysis of all continuous variables to make the distribution more symmetrical; and a 1 (one) was added to each value before taking the logarithm to handle observed values of zero, as the logarithm of zero is undefined.

For ordered categorical variables such as qualitative SWE features, weighted kappa (κ) [11] was used as a measure of agreement using quadratic weights [12]. Weighted kappa is preferable to unweighted kappa for an ordered categorical variable, where disagreements can be weighted according to the extent of the discrepancy. Using the quadratic weighting scheme, weighted kappa is identical to the intraclass correlation coefficient [12]. Weighted kappa is interpreted in the same way as kappa and has a maximum value of 1 (perfect agreement). No weighting was used to assess agreement for individual BI-RADS assessment categories. Kappa of 0–0.2 indicates slight agreement; 0.21–0.4, fair agreement; 0.41–0.6, moderate agreement; 0.61–0.8, substantial agreement; and 0.81–0.99 almost perfect agreement [11]. Kappa for BI-RADS assessments were calculated using six categories: 2, 3, 4a, 4b, 4c and 5. Maximum quantitative elasticity across three acquisitions was used to assess agreement with the second observer's rating of the visual maximum within six pre-specified colour ranges from one acquisition.

Kruskal–Wallis one-way ANOVA was used to compare independent groups for ordered categorical variables.

Spearman rank correlation (ρ) was used to assess correlation of ordered categorical variables.

Results

Participants and masses

A total of 758 women, each contributing a single mass, were available for analysis. Their mean age was 50.0 years (median 48.9, SD 13.9, range 21.2–89.5). Of the 758 masses, 144 (19.0%) were malignant, including 8 pure ductal carcinomas in situ. Mass size on B-mode with the house system (mean 12.9 mm, SD 7.5, range 1.5–53) was slightly larger than on the RUBI system (mean 12.4 mm, SD 7.2, range 2.6–50.2 [mean difference 0.6 mm, CI 0.3 to 0.8, $P < 0.001$]). Of the 758 masses, 102 were classified as BI-RADS 2 by site investigators (all presumed benign), 285 as BI-RADS 3 (6 [2.1%] malignant), 180 as BI-RADS 4a (13 [7.2%] malignant), 79 as BI-RADS 4b (27 [34%] malignant), 38 as BI-RADS 4c (27 [71%] malignant), and 74 as BI-RADS 5 (71 [96%] malignant).

Intraobserver reproducibility of BI-RADS and B-mode features

The intraobserver agreement of BI-RADS assessments between the house and RUBI systems was almost perfect ($\kappa = 0.93$, CI 0.86 to 1.0). Near perfect agreement for the BI-RADS 2, 3 and 5 assessments was observed and substantial agreement was found for the BI-RADS 4a, 4b and 4c assessments (Table 1). Intraobserver agreement for breast mass features between the house and the RUBI systems was near perfect for shape ($\kappa = 0.82$, CI 0.75 to 0.89), substantial for margins, orientation, echogenicity and posterior features ($\kappa \geq 0.67$) and was moderate for calcifications ($\kappa = 0.59$ CI 0.52 to 0.65) (Table 2)

Table 1 Intraobserver agreement of BI-RADS assessments performed on the house and RUBI ultrasound systems for 614 benign and 144 malignant breast masses

House BI-RADS assessment	N	RUBI BI-RADS assessment						Kappa ^a	Kappa 95% CI
		2	3	4a	4b	4c	5		
2 – Benign	102	92	9	1	0	0	0	0.90	0.82 to 0.97
3 – Probably benign	285	6	267	11	1	0	0	0.86	0.79 to 0.93
4a – Low suspicion	180	2	16	137	22	2	1	0.73	0.66 to 0.81
4b – Intermediate suspicion	79	0	6	15	50	3	5	0.58	0.50 to 0.65
4c – Moderate suspicion	38	0	1	1	5	25	6	0.67	0.60 to 0.74
5 – Highly suggestive of malignancy	74	0	1	0	4	5	64	0.84	0.76 to 0.91
Total	758	100	300	165	82	35	76	0.93 ^a	0.86 to 1.00

^a Quadratic weighting was used for overall kappa values; unweighted kappa values are presented for individual assessment categories

Table 2 Intraobserver agreement of House and RUBI B mode ultrasound features for 614 benign and 144 malignant breast masses

Mass feature or assessment	N for house B mode	Kappa ^a (95% CI)
Shape		0.82 (0.75 to 0.89)
Oval	464	
Round	87	
Irregular	207	
Margins		0.78 (0.71 to 0.85)
Circumscribed	449	
Not circumscribed	309	
Orientation		0.75 (0.68 to 0.82)
Parallel	599	
Not parallel	159	
Echogenicity		0.79 (0.72 to 0.86)
Anechoic	68	
Hyperechoic	13	
Complex	102	
Hypoechoic	552	
Isoechoic	23	
Posterior features		0.67 (0.59 to 0.74)
None	468	
Enhancement	148	
Shadowing	108	
Combined	34	
Calcifications		0.59 (0.52 to 0.65)
None	702	
Macrocalcifications	5	
Microcalcifications in a mass	43	
Microcalcifications out of a mass	8	

^a Quadratic weighting was used for kappa values

Qualitative intraobserver reproducibility of SWE features

Site investigators assessed all three consecutive SWE acquisitions as being very similar in 433 cases (57.1%) and reasonably similar in 233 cases (30.7%); some images were similar (two of the three) in 84 cases (11.1%) and all images were very dissimilar in 8 cases (1.1%) (Table 3). All images had very or reasonably similar shapes in 666 cases (88%). Of 758 masses, 391 (51.6%) were scored as very homogeneous on SWE, 212 (28.0%) as reasonably homogeneous and 155 (20.4%) as not homogeneous. The malignancy rate for the very homogeneous group was 17/391 (4.4%), for the reasonably homogeneous group 35/212 (16.5%) and for the not homogeneous group 92/155 (59.4%). There was a strong correlation between qualitative reproducibility and homogeneity ($\rho=0.63$, $P<0.001$). All images had very or reasonably similar shapes in 666 cases (88%). Irregular masses were significantly ($P<0.001$) less similar than oval or round masses. Benign masses were significantly more similar across the three acquisitions than malignant masses ($P<0.001$).

Quantitative intraobserver reproducibility of SWE features

Intraobserver reliability of size measurements on SWE using the RUBI system was almost perfect, with $ICC\geq 0.94$ for diameter, perimeter and area across the three acquisitions (Table 4). Intraobserver reliability of quantitative elasticity measurements on the SWE images performed with the Q-Box quantification tool was also almost perfect for the mean and maximum elasticity values. Emean was the most reliable measurement of elasticity, with $ICC=0.87$, and Emin was the least reliable ($ICC=0.71$).

The reliability of SWE measurements of fat was lower than that for SWE measurements of masses.

Table 3 Qualitative intraobserver reproducibility of shear wave elasticity (SWE) related to homogeneity, shape and malignancy for 614 benign and 144 malignant breast masses

Similarity score	Homogeneity ^c			Shape ^c			Malignancy ^c	
	Very homogeneous N (%) ^a	Reasonably homogeneous N (%) ^a	Not homogeneous N (%) ^a	Oval N (%)	Round N (%)	Irregular N (%)	Benign N (%) ^a	Malignant N (%) ^a
All images very similar	341 (87.2)	66 (31.1)	26 (16.8)	279 (66.1)	12 (60)	142 (44.9)	395 (64.3)	38 (26.4)
All images reasonably similar	44 (11.3)	110 (51.9)	79 (51.0)	108 (25.6)	5 (25)	120 (38.0)	164 (26.7)	69 (47.9)
Some images similar ^b	5 (1.3)	34 (16.0)	45 (29.0)	32 (7.6)	3 (15)	49 (15.5)	52 (8.5)	32 (22.2)
Images very dissimilar	1 (0.3)	2 (0.9)	5 (3.2)	3 (0.7)	0 (0)	5 (1.6)	3 (0.5)	5 (3.5)
Total	391	212	155	422	20	316	614	144

^a Percentages are column percentages for the similarity of three SWE acquisitions per mass

^b “Some images similar” was used when two of the three images were similar

^c $P<0.001$ from Kruskal–Wallis one-way ANOVA

Table 4 Intraobserver reliability of quantitative shear wave elastography measurements for 614 benign and 144 malignant breast masses

Variable measured	Overall		Benign		Malignant	
	ICC	95% CI	ICC	95% CI	ICC	95% CI
Diameter ^a	0.94	0.94 to 0.95	0.95	0.94 to 0.96	0.92	0.90 to 0.94
Area	0.95	0.94 to 0.95	0.95	0.94 to 0.96	0.92	0.90 to 0.94
Perimeter ^b	0.95	0.94 to 0.95	0.95	0.95 to 0.96	0.92	0.90 to 0.94
E minimum mass ^c	0.71	0.68 to 0.74	0.70	0.67 to 0.73	0.54	0.44 to 0.63
E maximum mass ^d	0.84	0.82 to 0.86	0.82	0.80 to 0.84	0.66	0.59 to 0.74
E mean mass ^e	0.87	0.85 to 0.88	0.85	0.84 to 0.87	0.71	0.64 to 0.77
E ratio (mass:fat) ^f	0.77	0.74 to 0.79	0.73	0.70 to 0.76	0.56	0.48 to 0.65
E minimum fat ^g	0.57	0.53 to 0.61	0.58	0.53 to 0.62	0.52	0.43 to 0.61
E maximum fat ^h	0.62	0.58 to 0.65	0.63	0.60 to 0.67	0.51	0.42 to 0.61
E mean fat ⁱ	0.71	0.68 to 0.74	0.73	0.70 to 0.76	0.58	0.50 to 0.67

^a Diameter was the largest measurement of the mass

^b Perimeter was the length of the border of the mass

^c E minimum mass was the minimum value in the Q-box of the mass as calculated by the system

^d E maximum mass was the maximum value in the Q-box of the mass as calculated by the system

^e E mean mass was the mean value in the Q-box of the mass as calculated by the system

^f E ratio (mass:fat) was the ratio between the mean elasticity value in the mass divided by the mean elasticity value in the fat

^g E minimum fat was the minimum value of the Q-box in the fat

^h E maximum fat was the maximum value of the Q-box in the fat

ⁱ E mean fat was the mean value of the Q-box in the fat

Intraobserver reliability was consistently higher for all measured SWE features of benign masses than of malignancies (Table 4).

Interobserver reproducibility of B-mode and SWE features

The interobserver agreement for the B-mode features of the shape and orientation of the mass was moderate with κ of 0.58 (95% CI 0.51 to 0.65) and 0.53 (95% CI 0.46 to 0.60) respectively (Table 5). Agreement that the margin was circumscribed or not was only fair ($\kappa=0.38$, 95% CI 0.32 to 0.45).

Interobserver agreement on BI-RADS assessments based on B-mode images between site investigators and the second observer was moderate ($\kappa=0.59$, 95% CI 0.53 to 0.66), although agreement for each of the individual assessment categories 3, 4a, 4b and 4c was poor ($\kappa=0.17$, 0.04, 0.004 and 0.07 respectively) (Table 6).

Fair interobserver agreement was found for SWE shape, moderate agreement for SWE homogeneity, and substantial agreement for maximum elasticity quantification using six colour-coded categories with $\kappa=0.66$ (95% CI 0.59 to 0.73). The second observer's visual assessment of Emax based on SWE colours correlated very strongly with quantitative Emax ($\rho=0.63$, $P<0.001$).

Analysis of reproducibility across all 939 masses

There were no substantial changes in the conclusions regarding reproducibility when all 939 masses (including all 289) malignancies were retained in the analysis set. Qualitatively, all three images were at least reasonably similar for 815/939 lesions (86.8%). Intraclass correlation coefficients for lesion diameter and perimeter were nearly perfect ($\text{ICC}\geq 0.95$ for each). ICCs for maximum lesion elasticity, mean elasticity and ratio of lesion-to-fat elasticity were also nearly perfect ($\text{ICC}=0.88$; 0.91; and 0.83 respectively). Interobserver agreement for SWE homogeneity ($\kappa=0.66$) and for the qualitative assessment of maximum elasticity ($\kappa=0.80$) was substantial.

Discussion

In this prospective multicentre, multinational study, intraobserver reproducibility of shear wave elastography was shown to be highly reliable for both quantitative and qualitative assessments. Importantly, the level of interobserver agreement seen for SWE features was at least as high, if not higher, than agreement seen for BI-RADS: Ultrasound B-mode features and assessments in this and previous studies [13–15]. As in prior studies [14, 15],

Table 5 Interobserver agreement of B mode ultrasound and SWE features for 614 benign and 144 malignant breast masses

Mass feature or assessment	Overall		Benign		Malignant	
	Kappa	95% CI	Kappa	95% CI	Kappa	95% CI
B-mode shape ^a	0.58	0.51 to 0.65	0.46	0.38 to 0.53	0.29	0.16 to 0.43
B-mode orientation ^b	0.53	0.46 to 0.60	0.47	0.39 to 0.55	0.40	0.23 to 0.56
B-mode margin ^c	0.38	0.32 to 0.45	0.26	0.19 to 0.33	0.29	0.14 to 0.43
B-mode BI-RADS assessment ^d	0.59	0.53 to 0.66	0.34	0.27 to 0.41	0.42	0.26 to 0.58
SWE shape ^a	0.40	0.33 to 0.47	0.27	0.20 to 0.34	0.25	0.14 to 0.36
SWE homogeneity ^e	0.57	0.51 to 0.64	0.47	0.40 to 0.54	0.27	0.17 to 0.37
SWE Emax ^f	0.66	0.59 to 0.73	0.47	0.39 to 0.54	0.60	0.45 to 0.74

SWE=shear wave elastography

^a Shape was described as oval, round, or irregular

^b Orientation was described as parallel to the skin surface or not

^c Margins were described as circumscribed or not

^d BI-RADS was assessed as benign, probably benign, low suspicion, intermediate suspicion, moderate suspicion or highly suggestive of malignancy and the kappa describes the agreement between the house and the second read BI-RADS assessments

^e Homogeneity was described as very homogeneous, reasonably homogeneous, or not homogeneous

^f Comparison of maximum elasticity across three SWE acquisitions within a 2- x 2-mm region of interest versus maximum derived by a six-level visual colour overlay (Fig. 3) (usually) of the entire mass on a representative image

interobserver agreement for BI-RADS assessments 4a, 4b, and 4c was low, and we also found poor agreement on BI-RADS 3 assessments, probably due to lack of predefined criteria for such classification in this protocol. Importantly, despite this variability, rates of malignancy in each category fell within expected ranges [10].

Qualitatively, benign lesions tend to be homogeneously soft (blue or blue-black) on SWE, whereas malignancies have a heterogeneous appearance within and across SWE images. Indeed, the heterogeneity of the SWE stiffness was moderately predictive of malignancy in the full set of 939 masses, with 70.6% of heterogeneous masses malignant compared with 4.3% of very homogeneous masses ($P < 0.001$) in a population with 30.8% prevalence of malig-

nancy [10]. There may have been some confusion on the part of investigators in assessing the “similarity” of SWE images in that three acquisitions each showing heterogeneous stiff and soft areas in a given lesion and its surrounding tissue are indeed similar to each other but were sometimes rated as “dissimilar” by investigators.

Despite the qualitative variability of stiffness within and adjacent to a given malignancy on a single acquisition, quantitative assessment of maximum and median elasticity were highly reliable across multiple SWE acquisitions. Further, visual colour assessment of maximum stiffness in the lesion or surrounding tissue on a single acquisition was highly correlated with measured maximum Emax across three acquisitions. For clinical use, we suggest that orthogonal SWE acquisitions be

Table 6 Interobserver agreement of BI-RADS assessments performed on the house and RUBI ultrasound systems for 614 benign and 144 malignant breast masses

House BI-RADS	Second observer RUBI BI-RADS							Kappa ^a	Kappa 95% CI
	N	2	3	4a	4b	4c	5		
2 – Benign	102	65	14	13	8	0	2	0.43	0.36 to 0.50
3 – Probably benign	285	57	82	62	46	27	11	0.17	0.11 to 0.24
4a – Low suspicion	180	18	40	34	39	31	18	0.04	–0.04 to 0.11
4b – Intermediate suspicion	79	7	9	9	12	22	20	0.004	–0.07 to 0.08
4c – Moderate suspicion	38	0	0	6	1	10	21	0.07	0.01 to 0.14
5 – Highly suggestive of malignancy	74	0	0	1	6	13	54	0.48	0.41 to 0.54
Total	758	147	145	125	112	103	126	0.59 ^a	0.53 to 0.66

^a Quadratic weighting was used for overall kappa values; unweighted kappa values are presented for individual assessment categories

performed, with clinical management of the lesion based on its most suspicious SWE features, together with its B-mode features. A lesion that is highly suspicious on B-mode should still be biopsied, regardless of its SWE appearance.

Because we used shear wave elastography, which measures the shear wave velocity in m/s and estimates Young's modulus in kPa (which is not possible with strain-based systems), quantitative values of tissue stiffness were available. Both mean and maximum elasticity showed almost perfect intraobserver reliability (ICC=0.87 and 0.84, respectively). Use of a maximum elasticity value, E_{max} , of less than 5.2 m/s to downgrade BI-RADS 4a lesions to surveillance, and E_{max} of 7.3 m/s or greater to upgrade BI-RADS 3 lesions to biopsy was shown to have great potential clinical impact, improving the specificity of breast ultrasound while retaining or possibly improving sensitivity [10].

It is important that the second observer's assessment of E_{max} based on SWE colours correlated very strongly with quantitative E_{max} , because this was based on a visual assessment of the stiffest colour seen in the colour overlay of the entire lesion and surrounding tissue (on a single image), which is similar to what would be done in a clinical setting. E_{max} from site investigators was based on the placement of a 2×2 mm region of interest (Q-Box) on each of three images. We noted that at times the Q-Box was placed within the soft (sometimes necrotic) centre of cancers or not in the stiffest portion of the lesion or surrounding tissue. Further, compression of the tissue can artificially increase the measured stiffness [16], and the Q-Box for lesion and/or fat was occasionally placed over areas of compression artefact. These errors point to a learning curve even after training on a minimum of 10 cases at each site and suggest that visual colour assessment of E_{max} may be easier to adopt in practice than precise quantification, although provision for colour-blind individuals would be necessary.

Minimum elasticity assessed using the Q-Box positioned on the stiffest part of the mass or surrounding tissue was the least reproducible of the quantitative measurements. On images where any pixel within the Q-Box lacked SWE information, a zero value was returned as the minimum; this probably accounts for the lower agreement for E_{min} .

The exact method used for comparison of different elastographic images could markedly affect the intraobserver reproducibility reported across the literature. For example, the approach where one person collects the images for subsequent repeated scoring by one or more observers would always give more reproducible results than separate acquisitions as performed in this series. In a paper by Thomas et al. [17], still images from the original real-time sequences were selected by the person performing the imaging for later interpretation by two blinded reviewers

who were presented with the B-mode and strain elastography images in a dual display and substantial agreement was found (weighted kappa=0.73). Selection of the best images from a cine loop is not required with SWE, a fact that should improve reliability.

Using a different system, Burnside et al. [5] presented strain imaging video clips of masses to three radiologists and found significant interobserver differences. Strict selection of cases according to the recruitment and quality of the elastograms meant that only 98 out of 445 cases (22.0%) were able to be included and kappa values were not given.

A single-centre study of 52 cases using the Aixplorer system, which was developed from the RUBI SWE prototype used in our study, compared two observers who independently measured the mean SWE from two previously acquired pairs of images and found an interobserver ICC of 0.99 [18]. In a subset of 15 cases, they also assessed interobserver reliability with a second observer re-imaging the mass, and found an ICC of 0.80 for mean elasticity.

In general, reproducibility is important for the widespread adoption of any imaging technique. Shear wave elastography can improve the management of BI-RADS 3 and 4a masses [10]. Standardisation is a central tenet of the BI-RADS lexicon for consistent interpretation and communication of results, and elastography will be added to the next edition of BI-RADS: Ultrasound [19]. Consistent image generation is critical for serial studies of masses, for example, follow-up of probably benign masses, or to monitor the response of breast carcinomas to neoadjuvant chemotherapy, and the reliability of SWE could be helpful in these clinical situations. Standardising the colour display across elastography systems so that, for example, soft tissues are consistently blue and stiff tissues consistently red, would avoid confusion in centres with multiple different systems or on receipt of images from outside facilities, and should urgently be addressed.

There are limitations to our study. One hundred and seventy-four masses, including 141 malignancies (81%), were excluded from reproducibility analysis because the SWE values exceeded the (default) 7.7 m/s upper limit that was pre-set for the scale. The prototype RUBI software did not allow SWE measurements of values higher than 7.7 m/s, although analysis including these masses showed similarly high reliability. Difficulties and errors in the placement of the Q-box could have had a negative impact on the reproducibility of the measurements. The immediate reapplication technique that we used for the intraobserver reliability study likely improves intraobserver reproducibility but any other method is hard to implement and the immediate reapplication approach is more clinically pertinent than two or more readers re-reading previously acquired images. The blinded second read was retrospective, with only two static B-mode images and one SWE image available for each mass; ideally, a

prospective interobserver comparison would use real-time image acquisition. A factor that is likely to affect reproducibility is the two-dimensional nature of conventional ultrasound, both B-mode and elastography, with resulting differences in slice selection across acquisitions, and we did not compare the reliability of SWE imaging in all planes. Three-dimensional SWE acquisition is possible but was not available during this study. Despite this, we showed high intraobserver reproducibility across multiple acquisitions.

In conclusion, in both intra- and interobserver comparisons, breast SWE was highly reproducible, as might be expected of a technique that minimises operator input. This underlies its value in improving the BI-RADS classification of breast masses and suggests that SWE should be useful for serial comparative studies. It should also aid in standardising reporting of the elasticity of breast masses.

Acknowledgement We thank the following investigators for their participation in the data collection for the protocol: Helmut Madjar, MD, Deutsche Klinik für Diagnostik, Wiesbaden, Germany; Lora Barke, MD, Sally Jobe Breast Center, Denver, CO; Alda Cossi, MD, Boston Medical Center, Boston, MA; Annina Wilkes, MD, Thomas Jefferson University School of Medicine, Philadelphia, PA; Linda Hovanessian Larsen, MD, Keck School of Medicine, University of Southern California, Los Angeles, CA; Alexander Mundinger, MD, Marienhospital and Breast Centre, Osnabrück, Germany; Ingrid Schreer, MD, University Hospital, Schleswig-Holstein, Kiel, Germany.

In addition we thank the following employees at SuperSonic Imagine, Aix-en-Provence, France, for contributing to the education and training of investigators, to the collection of participants' data, and to the development and maintenance of the database: Arnaud Perrin, Product Clinical Specialist; Carmen Perri, RDMS, Product Clinical Specialist; Marion Meckelburg, RDMS, Product Clinical Specialist; Sharon Bruce, Research Coordinator; and Julien Salenson, Database Manager.

D. Cosgrove has received consulting fees and travel expenses from SuperSonic Imagine; W. Berg received consulting fees and travel expenses from SuperSonic Imagine; Consultant to Naviscan, Inc.; Compensated to prepare educational materials for Gamma Medica; Medical advisory board of Philips Ultrasound; C Dore received consulting fees from SuperSonic Imagine. Employees of SuperSonic Imagine.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Berg WA (2004) Supplemental screening sonography in dense breasts. *Radiol Clin North Am* 42:845–851
- Corsetti V, Ferrari A, Ghirardi M et al (2006) Role of ultrasonography in detecting mammographically occult breast carcinoma in women with dense breasts. *Radiol Med (Torino)* 111:440–448
- Houssami N, Irwig L, Simpson JM, McKessar M, Blome S, Noakes J (2003) Sydney Breast Imaging Accuracy Study: comparative sensitivity and specificity of mammography and sonography in young women with symptoms. *AJR Am J Roentgenol* 180:935–940
- Krouskop TA, Wheeler TM, Kallel F, Garra BS, Hall T (1998) Elastic moduli of breast and prostate tissues under compression. *Ultrason Imaging* 20:260–274
- Burnside ES, Hall TJ, Sommer AM et al (2007) Differentiating benign from malignant solid breast masses with US strain imaging. *Radiology* 245:401–410
- Itoh A, Ueno E, Tohno E et al (2006) Breast disease: clinical application of US elastography for diagnosis. *Radiology* 239:341–350
- Chang JM, Moon WK, Cho N, Kim SJ (2011) Breast mass evaluation: factors influencing the quality of US elastography. *Radiology* 259:59–64
- Athanasioiu A, Tardivon A, Tanter M et al (2010) Breast lesions: quantitative elastography with supersonic shear imaging—preliminary results. *Radiology* 256:297–303
- Mendelson EB, Baum JK, Berg WA, Merritt CRB, Rubin E (2003) *Breast Imaging Reporting and Data System, BI-RADS: ultrasound*, 1st edn. American College of Radiology, Reston
- Berg W, Cosgrove D, Doré C, et al (2011 accepted, pending revision) Shear wave elastography improves the specificity of breast ultrasound: the BEI Multinational Study of 939 Masses. *Radiology*
- Landis JR, Koch GG (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33:363–374
- Fleiss J, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *J Educational and Psychological Measurement* 33:613–619
- Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS (2006) BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. *Radiology* 239:385–391
- Lee HJ, Kim EK, Kim MJ et al (2008) Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. *Eur J Radiol* 65:293–298
- Abdullah N, Mesurrolle B, El-Khoury M, Kao E (2009) Breast Imaging Reporting and Data System lexicon for US: interobserver agreement for assessment of breast masses. *Radiology* 252:665–672
- Garra BS, Cespedes EI, Ophir J et al (1997) Elastography of breast lesions: initial clinical results. *Radiology* 202:79–86
- Thomas A, Fischer T, Frey H et al (2006) Real-time elastography—an advanced method of ultrasound: first results in 108 patients with breast lesions. *Ultrasound Obstet Gynecol* 28:335–340
- Evans A, Whelehan P, Thomson K et al (2010) Quantitative shear wave ultrasound elastography: initial experience in solid breast masses. *Breast Cancer Res Treat* 12:R104
- Mendelson E, Böhm-Vélez M, Berg W, et al (2012 in press) *Breast Imaging Reporting and Data System, BI-RADS: Ultrasound*, 2nd edition. American College of Radiology, Reston