Evaluation of Inter-System Variability in Liver Stiffness Measurements

Bewertung der Intersystem-Variabilität bei Lebersteifigkeitsmessungen

Authors

Giovanna Ferraioli[1], Annalisa De Silvestri[2], Raffaella Lissandrin[1], Laura Maiocchi[1], Carmine Tinelli[2], Carlo Filice[1], Richard G. Barr[3]

Affiliations

1 Clinical Sciences and Infectious Diseases Department, Fondazione IRCCS Policlinico San Matteo, Medical School University of Pavia, Italy

2 Clinical Epidemiology and Biometric Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

3 Radiology, Northeastern Ohio Medical University, Rootstown, United States

Key words

liver stiffness, shear wave elastography, variability studies, cirrhosis, liver fibrosis

received 02.07.2017

accepted 16.11.2017

Bibliography

DOI https://doi.org/10.1055/s-0043-124184

Published online: March 22, 2018

Ultraschall in Med 2019; 40: 64–75

© Georg Thieme Verlag KG, Stuttgart · New York

ISSN 0172-4614

Correspondence

Dr. Giovanna Ferraioli

Clinical Sciences and Infectious Diseases Department, Fondazione IRCCS Policlinico S. Matteo, Medical School University of Pavia, Viale Camillo Golgi 19, 27100 Pavia, Italy

Tel.: ++ 39/03 82/50 27 99

giovanna.ferraioli@unipv.it

ABSTRACT

Aim The primary aim of this study was to determine the inter-system variability of liver stiffness measurements (LSMs) in patients with varying degrees of liver stiffness. The secondary aim was to determine the inter-observer variability of measurements.

Materials and Methods 21 individuals affected by chronic hepatitis C and 5 healthy individuals were prospectively enrolled. The assessment of LSMs was performed using six ultrasound (US) systems, four of which with point shear wave elastography (p-SWE) and two with 2 D shear wave elastography (2D-SWE) systems. The Fibroscan (Echosens, France) was used as the reference standard. Four observers performed the measurements in pairs (A-B, C-D). The agreement between different observers or methods was calculated using Lin's concordance correlation coefficient. The Bland-Altman limits of agreement (LOA) were calculated as well.

Results There was agreement above 0.80 for all pairs of systems. The mean difference between the values of the systems with 2D-SWE technique was 1.54 kPa, whereas the maximum mean difference between the values of three out of four systems with the pSWE technique was 0.79 kPa. The intra-patient concordance for all systems was 0.89 (95 % CI: 0.83 – 0.94). Inter-observer agreement was 0.96 (95 % CI: 0.94 – 0.98) for the pair of observers A-B and 0.93 (95 % CI: 0.89 – 0.96) for the pair of observers C-D.

Conclusion The results of this study show that the agreement between LSMs performed with different US systems is good to excellent and the overall inter-observer agreement in "ideal conditions" is above 0.90 in expert hands.

ZUSAMMENFASSUNG

Ziel Das Primärziel dieser Studie war es, die Intersystem-Variabilität von Lebersteifigkeitsmessungen (LSM) bei Patienten mit unterschiedlichen Lebersteifigkeiten zu bestimmen. Das Sekundärziel bestand in der Bestimmung der Inter-Beobachter-Variabilität der Messungen.

Material und Methoden Einundzwanzig Personen mit chronischer Hepatitis C und fünf gesunde Personen wurden prospektiv aufgenommen. Die Bewertung der LSM erfolgte mit sechs Ultraschallsystemen (US), vier davon verwendeten Punkt-Scherwellen-Elastografie (p-SWE) und zwei 2D-Scherwellen-Elastografie (2D-SWE). Der Fibroscan (Echosens, Frankreich) wurde als Referenzstandard verwendet. Vier Beobachter führten jeweils in Paaren die Messungen durch (A-B, C-D). Die Übereinstimmung zwischen verschiedenen Beobachtern oder Methoden wurde unter Verwendung des Konkordanz-Korrelationskoeffizienten nach Lin berechnet. Die Bland-Altman Übereinstimmungsgrenzen (LOA) wurden ebenfalls bestimmt.

Ergebnisse Die Übereinstimmung betrug mehr als 0,80 für alle Systempaare. Der mittlere Unterschied zwischen den Werten der Systeme mit 2D-SWE-Technik betrug 1,54 kPa, während der maximale mittlere Unterschied zwischen den

Werten von drei der vier Systeme mit pSWE-Technik 0,79 kPa betrug. Die Intra-Patienten-Konkordanz für alle Systeme war 0,89 (95 % CI: 0,83 – 0,94). Die Interobserver-Übereinstimmung betrug 0,96 (95 % CI: 0,94 – 0,98) für das Paar der Beobachter A-B und 0,93 (95 % CI: 0,89 – 0,96) für das Paar der Beobachter C-D.

Schlussfolgerung Die Ergebnisse dieser Studie zeigen, dass die Übereinstimmung zwischen LSMs, die mit unterschiedlichen US-Systemen durchgeführt wurden, gut bis hervorragend ist und die globale Übereinstimmung zwischen den Beobachtern, sofern diese Expertise besitzen, unter „idealen Bedingungen“ über 0,90 liegt.

Introduction

Diffuse liver disease is one of the major health problems in the world. Chronic liver damage results in hepatic fibrosis characterized by an increase in extracellular matrix material produced by fibroblast-like cells [1]. Liver fibrosis can progress to cirrhosis with distortion of the normal liver architecture and portal hypertension.

With increasing fibrosis, the liver becomes stiffer which can be monitored using shear wave elastography (SWE) [2, 3]. This technology is FDA-approved for several vendors. The technology is used widely worldwide and has resulted in a decrease in the number of liver biopsies performed in Europe and Asia. Recently, the United Kingdom’s National Institute for Health and Care Excellence (NICE) updated their clinical guidelines for managing patients with chronic viral hepatitis B infection and included liver SWE in the workup of patients [4]. However, there is significant inter-system variability in liver stiffness measurements that can preclude meaningful comparison of measurements performed with different systems [3, 5, 6].

The RSNA Quantitative Imaging Biomarker Alliance (QIBA) ultrasound shear wave speed (SWS) committee has developed elastic and viscoelastic phantoms to evaluate system dependencies of SWS estimates used for the noninvasive staging of liver fibrosis. Previous elastic phantom studies demonstrated inter-system variability ranging from 6 – 12 % in elastic phantoms with a nominal SWS of 1.0 and 2.0 m/s [7]. In visco-elastic phantoms that more accurately simulate the liver, the median SWS estimates for the greatest outliers in each phantom/focal depth combination ranged from 12.7 – 17.6 % [8].

There are many factors that affect liver stiffness measurement, including but not limited to the amount of subcutaneous fat, liver depth, breathing, motion from heartbeat, reverberation from the liver capsule, and fasting [5]. There are also bile ducts and blood vessels present in the liver that need to be avoided when obtaining measurements. All these factors are not assessed with phantoms. Our hypothesis was that inter-system variability assessed in-vivo is similar to or greater than in viscoelastic phantoms.

The main aim of this study was to determine the inter-system variability of liver stiffness measurements (LSMs) in volunteer patients with varying degrees of liver stiffness using the FibroScan system as the reference standard. The secondary aim was to determine the inter-observer variability of measurements in each patient for each system and in “ideal conditions”.

Materials and Methods

Systems

The assessment of SWS was performed using multiple systems with acoustic radiation force impulse (ARFI) shear wave technology, either point shear wave elastography (p-SWE) or 2 D shear wave elastography (2D-SWE). We contacted all the manufacturers that have commercially released the SWE technology on their system, and the design of the study was fully explained to them. The systems of the manufacturers that agreed to participate were used in this study. The systems with a pSWE technique were, in alphabetical order, Acuson S2000 (Siemens Medical Systems, Erlangen, Germany), EPIQ7 (Philips Medical Systems, Bothell, WA, USA), Hi-Vision Ascendus (Hitachi Ltd., Japan), MyLab Twice (Esaote SpA, Genoa, Italy). The systems with a 2D-SWE technique were Aixplorer (SuperSonic Imagine, Aix-en-Provence, France) and Aplio 500 (Canon/Toshiba, Japan). The FibroScan 502 Touch system (Echosens, Paris, France) was used as the reference standard. The results were anonymized before performing the statistical analysis. The two systems with a 2D-SWE technique received identification number 1 or 2 by randomization using a randomization table. The four systems with a pSWE technique received an identification number from 3 to 6 by randomization as well. The FibroScan was identified as system 7.

Study population and design of the study

Volunteer subjects with varying degrees of liver stiffness (normal subjects and patients with known liver fibrosis from prior studies) were enrolled in this prospective study performed in October 2016. Inclusion criteria were age greater than 18 years and ability to give informed consent. 26 individuals (14 males; mean age: 57 yrs. (15.8); 12 females; mean age: 56.3 yrs. (19.2)) were studied. 21 individuals were affected by chronic hepatitis C (13 males (mean age: 59.5 yrs. (13.1), 8 females (mean age: 66.7yrs (13.7)) and 5 individuals were healthy volunteers (one male; age, 24 yrs., four females (mean age: 35.5yrs (7.3)). The stage of liver fibrosis was based on transient elastography (TE), which was the reference standard, using the cutoffs of a published meta-analysis [9]. ▶ **Table 1** shows the characteristics of the patients enrolled in the study. No patients were on antiviral treatment or had previously been treated. Patients with decompensated liver cirrhosis were not included. Healthy volunteers were hospital staff members who were regularly followed with laboratory investigations, including testing for infection by hepatotropic viruses. All of them had normal values of transaminases and none of them had a history of liver disease or were using medication. Their alcohol

▶ **Table 1** Characteristics of the patients with chronic hepatitis C enrolled in the study.

characteristics	n = 21
sex, men (%)	61.9 %
age, yrs., (SD)	56.7 (16.9)
BMI, kg/m^2 (SD)	23.8 (3.9)
AST, IU/L (IQR)	24 (21 – 36)
ALT, IU/L (IQR)	23 (15.5 – 30)
INR	1.1 (0.11)
albumin, g/dL	4.3 (0.4)
GGT, IU/L (IQR)	31.5 (22 – 67)
ALP, IU/L (SD)	72 (65 – 85)
platelet count, 10^3/mm^3, (SD)	145 (62.5)
fibrosis stage (as assessed with TE)	
mild/no fibrosis (F0-F1)	5 (23.8 %)
significant fibrosis (F2)	4 (19.0 %)
advanced fibrosis (F3)	5 (23.8 %)
liver cirrhosis (F4)	7 (33.4 %)

SD: standard deviation; IQR: interquartile range; BMI: body mass index; AST: aspartate aminotransferase; ALT: alanine aminotransferase; INR: international normalized ratio; GGT: gamma-glutamyl transferase; ALP: alkaline phosphatase; TE: transient elastography.

intake was less than 20 g/day for all of them. All healthy subjects had normal liver morphology on conventional ultrasound.

Four observers (R.G.B., G.F., R.L., L.M.) with at least 3 years of experience in liver stiffness measurements on multiple systems performed the measurements. We estimated that each set of measurements per system per observer was 5 minutes. Therefore, to complete scanning, each patient would require approximately one hour – one hour and a half of participation. Each observer was anonymized with an alphabetical letter assigned in a random order. Each pair of observers (A-B or C-D) studied 13 subjects with the six systems in random order. A randomization list was used to match pair of observers and subjects. The TE measurements were the last to be performed and were taken by three of the four observers (G.F., R.L., L.M.) with at least three years of experience using the FibroScan system.

The Hospital Ethics Committee approved the study, which was performed according to the Declaration of Helsinki (revision of the year 2000, Edinburgh, appendix 12.4) and the current norms of Good Clinical Practice. Patient's written informed consent was required for inclusion in the study.

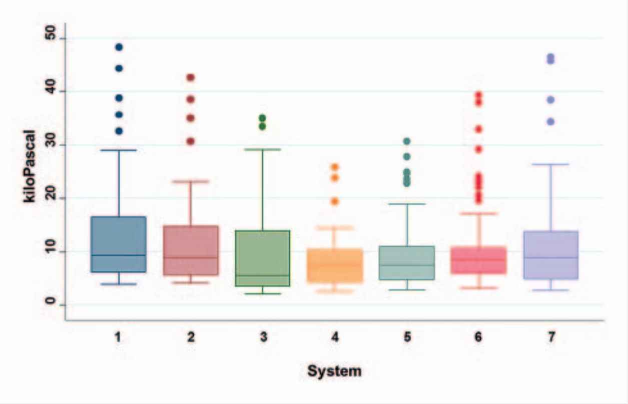
Scanning protocol

All patients were studied in fasting condition. The scanning technique recommended by the consensus document of the Society of Radiologists in Ultrasound (SRU) was followed [3]. Measurements were taken in the right intercostal spaces with the patient in the supine position with the right arm raised above the head. The location was determined by identifying the "best" sonographic window to visualize the right lobe of the liver. We tried to reduce the variability between observers to a minimum by following a strict protocol in order to focus on the variability between systems. For this reason, the first observer marked the location of the probe on the skin with a magic marker and the same location was utilized by the following observer. The following items were controlled: (a) measurement was taken with the patient in neutral breathing for choosing the acoustic window and in suspended respiration during the measurement; (b) the transducer was placed so that the ARFI pulse and measurement were perpendicular to the liver capsule; (c) the measurements were taken between 1.5 and 2.0 cm from the liver capsule (top of ROI box); (d) the sample box was maintained at a distance of around 4 cm from the skin as long as it was always at least 1.5 cm below the liver capsule; (e) ten valid measurements in the same location by each scanner for each patient; (f) the measurement of skin-to-liver capsule distance was obtained from the frozen image; (g) a measure-

▶ **Table 2** Failures and unreliable results observed with the six ultrasound systems and the FibroScan.

system	failures (%)	unreliable results (%)
1	0 (0)	4 (7.7)
2	2 (3.8)	2 (3.8)
3	0 (0)	0 (0)
4	4 (7.7)	5 (9.6)
5	0 (0)	0 (0)
6	0 (0)	1 (1.9)
7[1]	0 (0)	0 (0)

There were four observers, and each observer studied 13 subjects.
[1] FibroScan.



▶ **Fig. 1** Values obtained with the seven systems. The central box represents values from the lower to upper quartile (25 to 75 percentile). The horizontal line inside the box represents the median. The circles represent the outside values.

▶ **Table 3** Mean and median values, as well minimum and maximum values, and coefficient of variation of the observations obtained with the six ultrasound systems and the FibroScan.

system	observation #	mean, kPa (SD)	median, kPa (IQR)	minimum value, kPa	maximum value, kPa	CV
1	48	13.53 (11.23)	9.3 (6.1–16.5)	3.95	48.3	0.80
2	48	12.16 (9.21)	8.8 (5.7–14.7)	4.15	42.6	0.80
3	52	10.3 (9.5)	5.6 (3.6–14.0)	2.1	35.0	0.90
4	43	7.98 (5.28)	7.4 (4.3–10.2)	2.5	25.78	0.70
5	52	9.90 (7.05)	7.5 (4.7–11.0)	2.88	30.64	0.70
6	51	10.77 (8.47)	8.0 (5.4–10.5)	3	37.4	0.80
7[1]	39	12.68 (11.57)	8.9 (4.8–13.8)	2.8	46.5	0.90

kPa: kilopascal; SD: standard deviation; IQR: interquartile range; CV: coefficient of variation.

[1] FibroScan.

▶ **Table 4** Concordance two by two between measurements performed with the six ultrasound systems and the FibroScan. The concordance is estimated using the concordance correlation coefficient (CCC), the r Pearson's correlation coefficient, the bias-correction factor (Cb, a measure of accuracy) and the Bland-Altman method.

system	system	Obs. #	CCC (95 %CI)	Pearson's r	Cb	mean difference, kPa (95 % limits of agreement)
7[1]	1	37	0.97 (0.94–0.99)	0.97	0.99	−0.78 (−6.35–4.80)
7	2	33	0.92 (0.87–0.96)	0.94	0.98	0.44 (−8.06–8.94)
7	3	38	0.86 (0.79–0.94)	0.90	0.96	1.88 (−8.26–12.01)
7	4	33	0.80 (0.70–0.89)	0.90	0.88	2.07 (−6.40–10.55)
7	5	38	0.81 (0.75–0.88)	0.95	0.86	2.40 (−8.21–13.00)
7	6	37	0.89 (0.83–0.95)	0.91	0.97	0.68 (−7.77–9.13)
1	2	46	0.95 (0.92–0.97)	0.97	0.97	−1.54 (−7.59–4.50)
1	3	47	0.84 (0.76–0.91)	0.90	0.93	3.19 (−6.77–13.15)
1	4	41	0.74 (0.65–0.84)	0.92	0.81	−3.49 (−11.28–4.30)
1	5	47	0.78 (0.71–0.85)	0.95	0.82	−3.64 (−13.85–6.58)
1	6	46	0.88 (0.83–0.94)	0.93	0.95	−2.20 (−10.61–6.21)
2	3	48	0.87 (0.80–0.94)	0.89	0.98	1.91 (−6.51–10.34)
2	4	43	0.79 (0.69–0.88)	0.89	0.88	−2.54 (−9.20–4.13)
2	5	48	0.87 (0.81–0.92)	0.95	0.91	2.26 (−4.80–9.31)
2	6	47	0.94 (0.91–0.97)	0.95	0.99	1.00 (−4.72–6.72)
3	4	43	0.85 (0.78–0.92)	0.90	0.95	−0.69 (−7.44–6.07)
3	5	52	0.91 (0.88–0.95)	0.96	0.96	−0.40 (−7.13–6.33)
3	6	51	0.89 (0.84–0.95)	0.90	0.99	0.83 (−7.08–8.74)
4	5	43	0.95 (0.93–0.98)	0.96	0.99	−0.79 (−3.62–2.03)
4	6	51	0.85 (0.77–0.92)	0.90	0.94	−1.64 (−7.36–4.08)
5	6	51	0.88 (0.83–0.94)	0.91	0.97	−1.13 (−8.18–5.92)

Obs.: observations; CI: confidence interval; kPa: kilopascal.

[1] FibroScan.

▶ **Table 5** Concordance two by two between measurements performed with the six ultrasound systems and the FibroScan and with IQR/M ≤ 0.30. The concordance is estimated using the concordance correlation coefficient (CCC), the r Pearson's correlation coefficient, the bias-correction factor (Cb, a measure of accuracy) and the Bland-Altman method.

system	system	Obs. #	CCC (95 %CI)	Pearson's r	Cb	mean difference, kPa (95 % limits of agreement)
7[1]	1	35	0.97 (0.94 – 0.99)	0.97	0.99	–0.70 (–6.39 – 4.98)
7	2	29	0.98 (0.96 – 0.99)	0.98	1.00	–0.57 (–3.48 – 2.35)
7	3	17	0.90 (0.82 – 0.98)	0.94	0.96	2.29 (–6.11 – 10.69)
7	4	17	0.78 (0.64 – 0.91)	0.91	0.85	2.80 (–8.04 – 13.63)
7	5	26	0.86 (0.80 – 0.92)	0.97	0.89	2.02 (–7.36 – 11.40)
7	6	4	0.72 (0.22 – 1.22)	0.82	0.88	0.43 (–3.85 – 4.70)
1	2	38	0.95 (0.92 – 0.98)	0.97	0.98	–1.02 (–6.04 – 4.00)
1	3	26	0.90 (0.84 – 0.97)	0.95	0.95	2.83 (–4.51 – 10.17)
1	4	23	0.79 (0.69 – 0.90)	0.94	0.84	–3.17 (–11.53 – 5.18)
1	5	35	0.85 (0.80 – 0.91)	0.98	0.87	–2.80 (–10.65 – 5.04)
1	6	5	0.86 (0.62 – 1.09)	0.94	0.91	–1.05 (–3.70 – 1.60)
2	3	25	0.90 (0.84 – 0.97)	0.95	0.96	2.30 (–3.21 – 7.80)
2	4	21	0.78 (0.65 – 0.92)	0.93	0.85	–1.98 (–5.43 – 1.46)
2	5	30	0.83 (0.75 – 0.91)	0.94	0.89	1.57 (–4.48 – 7.62)
2	6	5	0.89 (0.69 – 1.10)	0.94	0.95	0.83 (–1.13 – 2.79)
3	4	14	0.97 (0.94 – 1.00)	0.98	0.99	0.44 (–2.41 – 3.30)
3	5	22	0.90 (0.85 – 0.95)	0.96	0.94	0.53 (–5.55 – 6.62)
3	6	3	0.97 (0.84 – 1.09)	0.97	1.00	0.14 (–1.65 – 1.94)
4	5	17	0.96 (0.91 – 1.00)	0.97	0.99	–0.83 (–4.61 – 2.94)
4	6	5	0.91 (0.75 – 1.07)	0.97	0.94	–0.82 (–2.21 – 0.57)
5	6	3	0.91 (0.61 – 1.22)	0.92	0.99	0.04 (–1.93 – 2.00)

Obs.: observations; CI: confidence interval; kPa: kilopascal.

[1] FibroScan.

ment known to be inaccurate (patient breaths, observer moves while obtaining measurement) was deleted and another measurement was taken; (h) the measurement images were recorded. Examinations with 10 validated measurements were defined as reliable. On systems with a manufacturer's quality measure, even if a numerical value was obtained, the measurement was taken as a technical failure if the quality measure was poor. Technical failure was defined as no successful measurement after 10 attempts. Unreliable results were those obtained with less than 10 measurements. The ratio between interquartile range (IQR; 25th – 75th percentile) and the median value (M) of ten measurements (IQR/M) ≤ 0.30 was defined as a quality factor [3]. An attempt was made to collect 10 valid measurements with an IQR/M ≤ 0.30 [3].

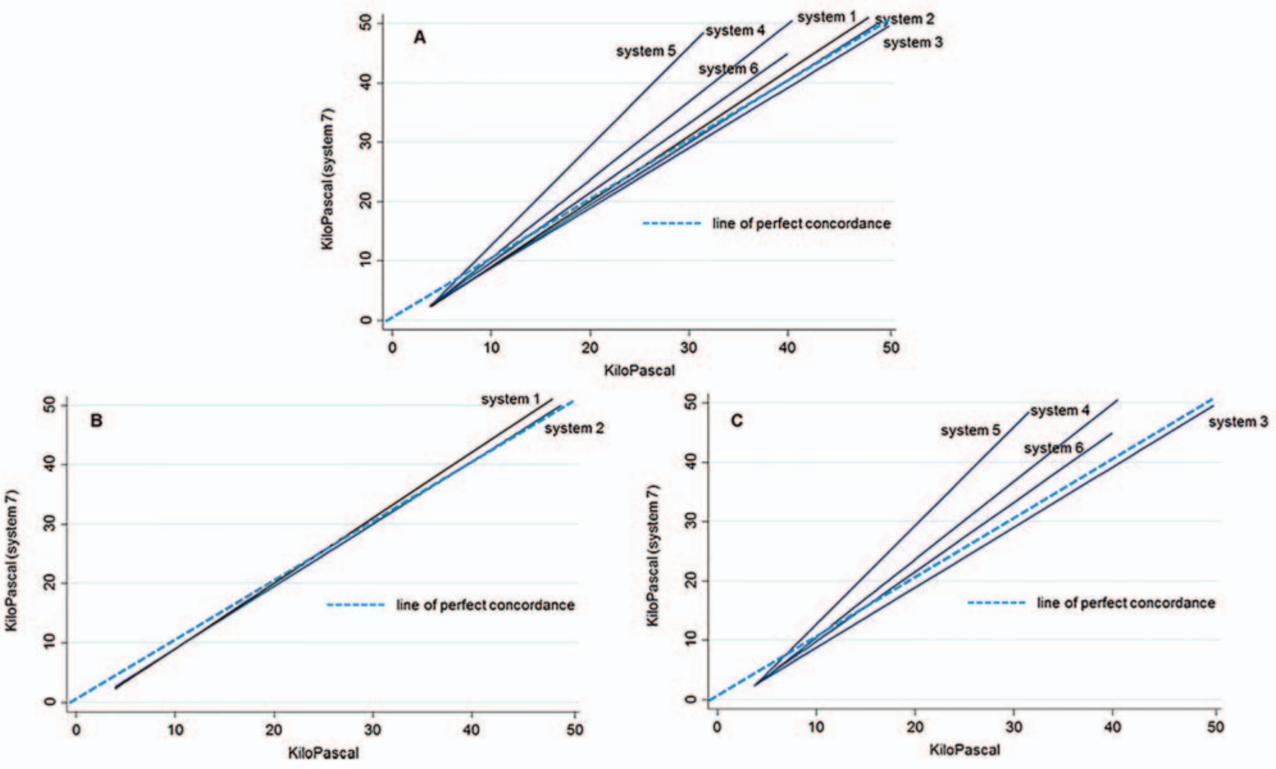
The measurements were reported in kilopascal (kPa) which is the measurement of the Young's modulus $E = 3\ \rho v^2$ where ρ is the density of the tissue and v the shear wave speed. The choice of reporting the Young's modulus unit instead of shear wave speed in m/s was due to the use of the FibroScan, which only reports results in kPa, as the reference standard.

Each observer was blinded to the other observers' results while measurements were taken.

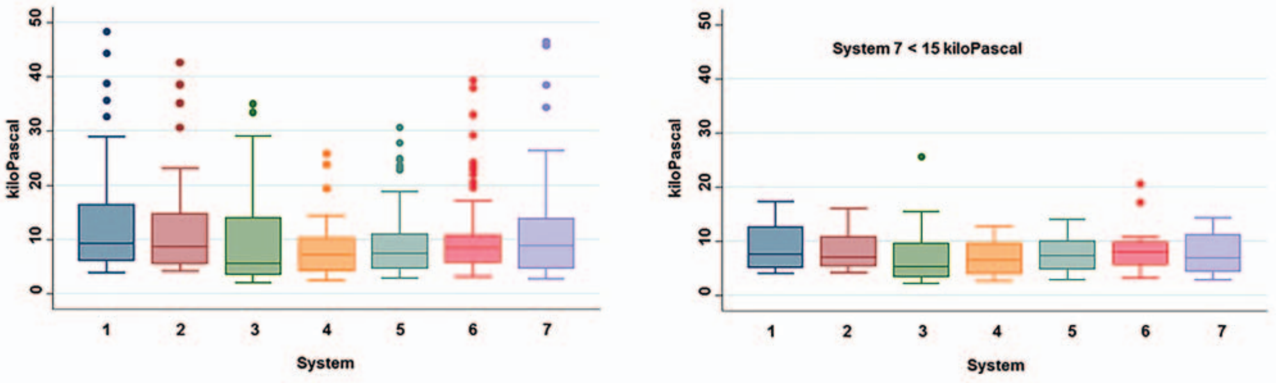
Statistics

A sample size of 26 subjects with 120 measurements (10 measurements for each of the 6 systems taken by two pairs of different operators) per subject achieves 90 % power to detect an intra-class correlation of 0.95 under the alternative hypothesis when the intra-class correlation under the null hypothesis is 0.88 using an F-test with a significance level of 0.05.

Descriptive statistics were produced for demographic characteristics for this study sample of patients. The Shapiro-Wilk test was used to test the normal distribution of quantitative variables. If the quantitative variables were normally distributed, the results were expressed as the mean value and standard deviation (SD). Otherwise the median and the IQR were reported. The coefficient of variation was reported as well. Qualitative variables were summarized as counts and percentages.



▶ **Fig. 2** **A** Agreement between measurements obtained with the six US systems and those obtained with system 7 considered the reference standard for this purpose. **B** Agreement between measurements obtained with the two US systems with a 2D-SWE technique and those obtained with system 7. **C** Agreement between measurements obtained with the four US systems with a pSWE technique and those obtained with system 7.



▶ **Fig. 3** On the left side: values obtained with the seven systems. The central box represents values from the lower to upper quartile (25 to 75 percentile). The horizontal line inside the box represents the median. The circles represent the outside values. On the right side: values obtained with the seven systems for values of the system 7 ≤ 15 kiloPascal.

To describe the agreement between continuous measurements obtained by different operators or methods, we calculated Lin's concordance correlation coefficient (CCC). It can be expressed as the product of Pearson's r (the measurement of precision) and the bias-correction factor (Cb, the measure of accuracy) [10]. CCC ranges in values from 0 to + 1. Agreement was classified as poor (0.00 – 0.20), fair (0.21 – 0.40), moderate (0.41 – 0.60), good (0.61 – 0.80), or excellent (0.81 – 1.00) [11]. The Bland and Altman limits of agreement (LOA), with their 95 % confidence interval (CI), within and between observers were reported as well: these represent the interval within which the absolute difference between two repeated test results, even with a high agreement or concordance, may be expected to lie with a probability of 95 %. If

▶ **Table 6** Concordance two by two between measurements performed with the six ultrasound systems for values of system 7 ≤ 15 kPa. The concordance is estimated using the concordance correlation coefficient (CCC), the r Pearson's correlation coefficient, the bias-correction factor (Cb, a measure of accuracy) and the Bland-Altman method.

system	system	Obs. #	CCC (95 %CI)	Pearson's r	Cb	mean difference, kPa (95 % limits of agreement)
7[1]	1	29	0.88 (0.81 – 0.96)	0.93	0.95	–1.16 (–4.19 – 1.88)
7	2	29	0.91 (0.85 – 0.97)	0.93	0.98	–0.61 (–3.31 – 2.10)
7	3	30	0.72 (0.56 – 0.88)	0.76	0.95	0.68 (–5.66 – 7.03)
7	4	28	0.83 (0.73 – 0.94)	0.88	0.95	1.02 (–2.38 – 4.41)
7	5	30	0.87 (0.80 – 0.95)	0.90	0.97	0.31 (–2.95 – 3.57)
7	6	30	0.74 (0.58 – 0.91)	0.75	0.99	–0.56 (–5.72 – 4.59)
1	2	29	0.93 (0.88 – 0.97)	0.96	0.97	–0.55 (–3.22 – 2.12)
1	3	29	0.73 (0.57 – 0.89)	0.80	0.91	1.79 (–4.10 – 7.68)
1	4	28	0.74 (0.62 – 0.87)	0.91	0.82	–2.18 (–5.75 – 1.40)
1	5	29	0.80 (0.70 – 0.90)	0.92	0.87	–1.45 (–5.12 – 2.23)
1	6	29	0.74 (0.57 – 0.91)	0.75	0.98	–0.63 (–6.19 – 4.94)
2	3	29	0.69 (0.52 – 0.85)	0.77	0.89	1.24 (–5.07 – 7.56)
2	4	28	0.78 (0.65 – 0.91)	0.88	0.89	–1.60 (–4.74 – 1.53)
2	5	29	0.84 (0.74 – 0.94)	0.89	0.95	0.90 (–2.17 – 3.97)
2	6	29	0.78 (0.63 – 0.92)	0.78	0.99	0.08 (–4.62 – 4.78)
3	4	28	0.73 (0.59 – 0.87)	0.81	0.90	–0.37 (–6.41 – 5.67)
3	5	30	0.76 (0.66 – 0.87)	0.87	0.88	0.38 (–5.15 – 5.90)
3	6	30	0.77 (0.64 – 0.90)	0.84	0.92	1.25 (–4.17 – 6.67)
4	5	28	0.92 (0.87 – 0.98)	0.95	0.97	–0.68 (–2.63 – 1.26)
4	6	28	0.70 (0.53 – 0.88)	0.79	0.89	–1.54 (–6.11 – 3.03)
5	6	30	0.79 (0.67 – 0.92)	0.84	0.94	–0.87 (–4.85 – 3.11)

Obs.: observations; CI: confidence interval; kPa: kilopascal.

[1] FibroScan.

the differences within mean ± 1.96 SD (LOA) are not clinically important, the two methods may be used interchangeably.

P < 0.05 was considered statistically significant. All tests were two-sided. The data analysis was performed with the STATA statistical package (release 14.0, 2015, Stata Corporation, College Station, Texas, USA).

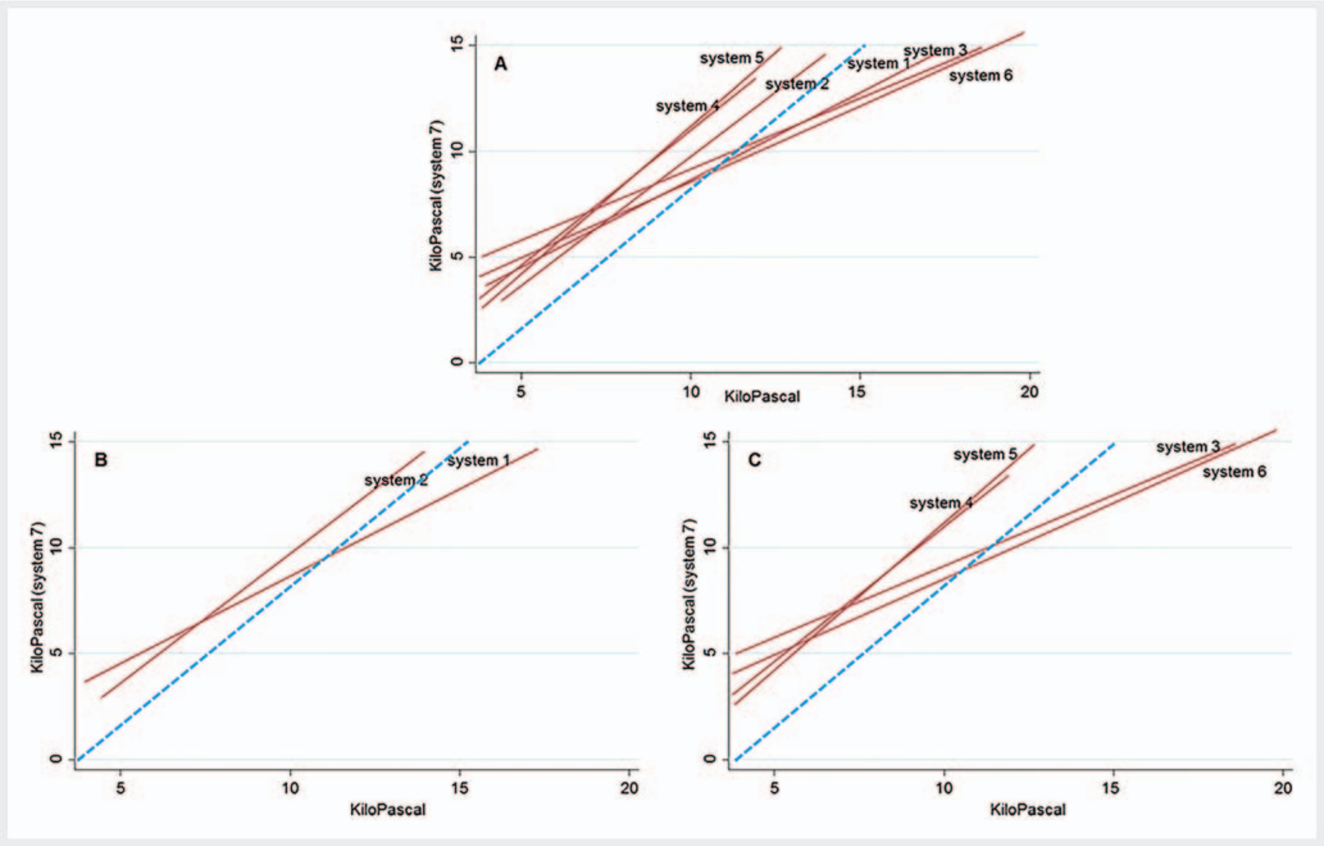
Results

The M probe of the FibroScan device was used for all subjects because all of them had a skin-to-liver capsule distance ≤ 25 mm. The failures and unreliable results observed with the seven systems are reported in ▶ **Table 2**. There wasn't any difference in the rate of failures or unreliable results between the four observers. Overall, 294 observations in 26 patients were made by the 4 observers with the 6 systems. The maximum median time for carrying out the examinations was less than 5 minutes.

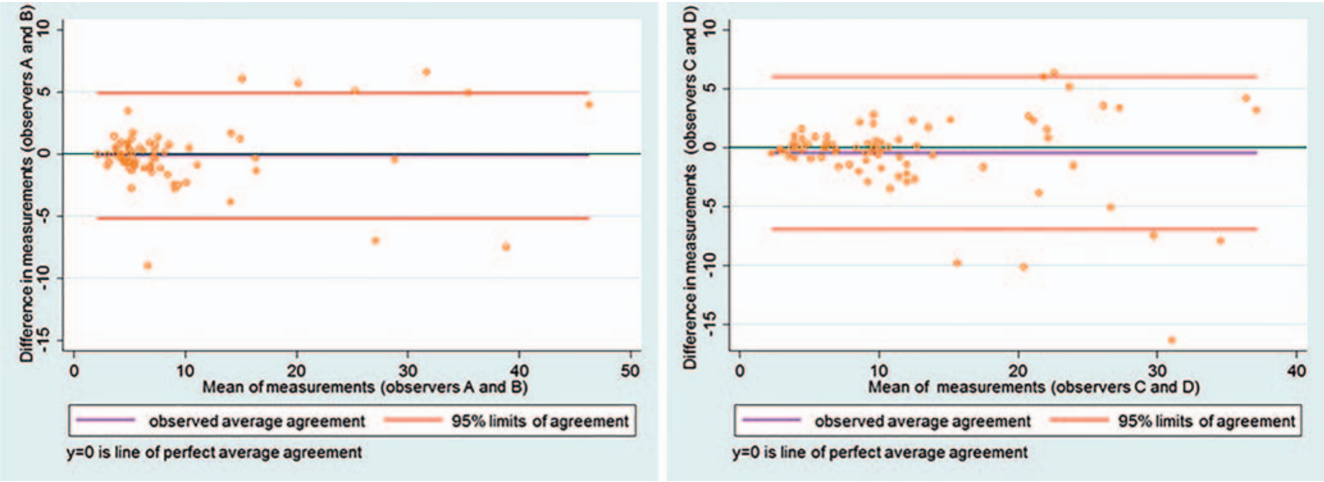
▶ **Table 3,** ▶ **Fig. 1** report the values obtained with the six ultrasound systems and with the FibroScan. The median values obtained with the pSWE systems were lower than the median values obtained with the 2D-SWE systems. The latter showed values closer to those of the FibroScan. The median value obtained with system 3 was lower than those obtained with the other pSWE systems because the low values were over-represented.

The intra-patient concordance for all systems was 0.89 (95 % CI: 0.83 – 0.94). The comparison between all systems and intra-patient and intra-observer showed a CCC of 0.88 (95 % CI: 0.83 – 0.93).

▶ **Table 4** shows the concordance two by two between measurements performed with the six ultrasound systems and the FibroScan and ▶ **Table 5** reports the concordance two by two assessed with measurements with IQR/M ≤ 0.30. The mean difference in kPa was between 0.40 and 3.64, and it slightly decreased for measurements with an IQR/M ≤ 0.30. There was an excellent agreement (CCC above 0.80) for all the pairs of systems except for system 1 and system 4 (CCC: 0.74), system 1 and system 5 (CCC: 0.78), and system 2 and system 4 (CCC: 0.79) that showed a good agreement. These lower CCCs were obtained for the agreement between systems with the 2D-SWE and pSWE tech-



▶ **Fig. 4** **A** Agreement between measurements obtained with the six US systems and those obtained with the FibroScan (system 7, reference standard) for values of system 7 up to 15 kiloPascal. **B** Agreement between measurements obtained with the two US systems with a 2D-SWE technique and those obtained with system 7. **C** Agreement between measurements obtained with the four US systems with a pSWE technique and those obtained with system 7.



▶ **Fig. 5** Bland and Altman plot of differences in measurements between observers **A, B** (left) and observers **C, D** (right). The green line (y = 0) is a line of perfect average agreement. The purple line represents the mean of the difference of measurements. The orange lines define the limits of agreement (mean of the difference (2 SD)).

nique. The agreement between systems with the pSWE technique was excellent (CCC: 0.85 – 0.95 for all measurements, and 0.90 – 0.97 for those with an IQR/M ≤ 0.30) for all pairs of systems. There was an increase in the agreement for measurements with IQR/M ≤ 0.30. However, with system 6 only very few measurements had an IQR/M ≤ 0.30, thus the CCC with this system is likely to be biased. The mean difference (in kPa) between the values of the two systems with the 2D-SWE technique was 1.54 kPa, whereas

▶ **Table 7** The inter-observer agreement between the measurements performed by the four observers with each system. The concordance is estimated using the concordance correlation coefficient (CCC), the r Pearson's correlation coefficient, the bias-correction factor (Cb, a measure of accuracy) and the Bland-Altman method.

observers	system	subjects #	CCC (95 %CI)	Pearson's r	Cb	mean difference, kPa (95 % limits of agreement)
A-B	1	10	0.98 (0.96 – 1.00)	0.99	0.99	1.25 (–2.80 – 5.30)
C-D	1	12	0.98 (0.96 – 1.00)	0.99	0.99	0.58 (–3.04 – 4.19)
A-B	2	11	0.96 (0.93 – 0.99)	0.98	0.98	–0.64 (–6.02 – 4.75)
C-D	2	12	0.95 (0.90 – 1.00)	0.97	0.98	–0.90 (–6.08 – 4.27)
A-B	3	13	0.93 (0.87 – 1.00)	0.95	0.99	–0.01 (–7.04 – 7.02)
C-D	3	12	0.91 (0.81 – 1.01)	0.94	0.97	–1.73 (–9.02 – 5.55)
A-B	4	10	0.86 (0.74 – 0.98)	0.98	0.88	–0.90 (–2.96 – 1.17)
C-D	4	10	0.88 (0.77 – 1.00)	0.94	0.94	1.28 (–3.50 – 6.07)
A-B	5	13	0.94 (0.87 – 1.00)	0.94	0.99	–0.73 (–5.81 – 4.36)
C-D	5	12	0.95 (0.89 – 1.01)	0.95	0.99	–0.02 (–4.26 – 4.22)
A-B	6	13	0.97 (0.94 – 1.00)	0.98	0.99	0.37 (–3.70 – 4.43)
C-D	6	12	0.79 (0.65 – 0.94)	0.94	0.84	–3.00 (-13.26 – 7.25)

CI: confidence interval; kPa: kiloPascal.

the maximum mean difference between the values of three out of four systems with the pSWE technique was 0.79 kPa. The outlier was system 6 which showed a maximum mean difference with the values obtained with the other pSWE systems of 1.64 kPa. Only a few measurements with IQR/M ≤ 0.30 were obtained with system 6, thus the estimation of the level of concordance for this system was not reliable. ▶ **Fig. 2** shows the agreement in measurements between systems. System 7 was considered the reference standard for this purpose.

▶ **Fig. 3** compares the values obtained with the six ultrasound systems and TE overall and for values of TE≤ 15 kPa. If the agreement was assessed for values of TE≤ 15 kPa, i. e. the measurements in very stiff livers that are already in the range of liver cirrhosis were not included, there was a decrease in the agreement among the different US systems. However, the number of observations was too low to allow a robust statistical analysis of the data (▶ **Table 6**, ▶ **Fig. 4**).

The inter-observer agreement was 0.95 (95 % CI: 0.93 – 0.96) overall, 0.96 (95 % CI: 0.94 – 0.98), with a difference between measurements of 0.13 kPa (95 % LOA:–5.14 – 4.90) for the pair of observers A-B, and 0.93 (95 % CI: 0.89 – 0.96), with a difference between measurements of 0.44 kPa (95 % LOA: –6.94 – 6.06) for the pair of observers C–D (▶ **Fig. 5**). The inter-observer agreement between the measurements performed by the four observers with each system is reported in ▶ **Table 7**.

A separate analysis of the data obtained in patients with stages of liver fibrosis F2 or higher showed that the intra-patient concordance for all systems was 0.84 (95 % CI: 0.74 – 0.94). ▶ **Table 8** reports the concordance two by two assessed for the measurements obtained in these patients. The mean difference in kPa was between 0.42 and 5.58. The comparison between all systems and intra-patient and intra-observer showed a CCC of 0.83 (95 % CI: 076 – 0.91). The inter-observer concordance was 0.93 (0.90 – 0.95).

Discussion

The results of this study show that the agreement between measurements of liver stiffness performed with different US systems is excellent. Of course, an excellent agreement doesn't mean that the values are the same but that there is concordance between them because they follow the same direction. The differences between values obtained with different systems may be higher than two kPa. Thus, in staging liver fibrosis with SWE, the cutoff values could not be interchangeably applied across different US systems. In fact, a difference of two kPa assigns the patient to a different stage of liver fibrosis. A recent study has compared liver stiffness findings, acquired with several US systems, with the results of the FibroScan in a series of patients with chronic hepatitis C and has shown only a moderate concordance between the results obtained with the US systems and those obtained with the FibroScan [12]. In our study, we evaluated several aspects of the concordance, including the estimate of the 95 % limits of agreement, and focused on the assessment of variability between systems, reducing to a minimum the variability between operators by applying a strict protocol for the acquisition of stiffness values. The Bland-Altman plots showed that the variability between measurements obtained with different systems was higher in a stiffer liver. In fact, the agreement decreased slightly when the data were analyzed for patients in stage F2 or higher. Using liver stiffness measurements for the evaluation of patients with chronic hepatitis C in everyday clinical practice, the complete

▶ **Table 8** Concordance two by two between measurements performed with the six ultrasound systems and the FibroScan in patients with fibrosis stage F2 or higher. The concordance is estimated using the concordance correlation coefficient (CCC), the r Pearson's correlation coefficient, the bias-correction factor (Cb, a measure of accuracy) and the Bland-Altman method.

system	system	Obs. #	CCC (95 %CI)	Pearson's r	Cb	mean difference, kPa (95 % limits of agreement)
7[1]	1	25	0.95 (0.92 – 0.99)	0.96	0.99	−0.64 (−7.36 – 6.08)
7	2	25	0.89 (0.81 – 0.96)	0.91	0.97	1.21 (−8.81 – 11.23)
7	3	25	0.81 (0.68 – 0.93)	0.86	0.94	2.52 (−9.84 – 14.88)
7	4	22	0.71 (0.56 – 0.86)	0.87	0.82	2.86 (−7.13 – 12.86)
7	5	25	0.75 (0.64 – 0.85)	0.94	0.80	3.92 (−8.02 – 15.86)
7	6	24	0.84 (0.72 – 0.96)	0.85	0.98	0.58 (−10.35 – 11.51)
1	2	29	0.92 (0.87 – 0.97)	0.96	0.96	−2.43 (−4.58 – 9.44)
1	3	29	0.77 (0.63 – 0.90)	0.84	0.91	4.15 (−8.16 – 16.47)
1	4	25	0.64 (0.49 – 0.80)	0.88	0.73	4.61 (−4.71 – 13.93)
1	5	29	0.70 (0.57 – 0.82)	0.93	0.75	5.58 (−5.85 – 17.01)
1	6	28	0.84 (0.74 – 0.95)	0.88	0.96	2.44 (−8.34 – 13.22)
2	3	31	0.82 (0.71 – 0.94)	0.84	0.98	2.04 (−8.47 – 12.55)
2	4	27	0.71 (0.56 – 0.86)	0.84	0.84	3.03 (−5.20 – 11.25)
2	5	31	0.82 (0.72 – 0.91)	0.93	0.88	3.24 (−4.83 – 11.31)
2	6	30	0.91 (0.85 – 0.97)	0.91	1.00	0.42 (−7.45 – 8.30)
3	4	27	0.78 (0.66 – 0.90)	0.86	0.91	−1.09 (−9.50 – 7.32)
3	5	32	0.88 (0.83 – 0.94)	0.94	0.94	−1.39 (−9.29 – 6.52)
3	6	31	0.88 ()0.79 – 0.96)	0.89	0.99	1.38 (−7.63 10.39)
4	5	27	0.95 (0.91 – 0.99)	0.96	0.99	0.53 (−2.74 – 3.81)
4	6	27	0.74 (0.60 – 0.88)	0.85	0.87	−2.71 (−10.254.83)
5	6	31	0.81 (0.70 – 0.91)	0.88	0.92	−2.63 (−11.69 – 6.41)

Obs.: observations; CI: confidence interval; kPa: kilopascal.

[1] FibroScan.

spectrum of liver fibrosis is represented and patients with not significant fibrosis account for a sizeable percentage. In our study, we enrolled patients with different degrees of liver fibrosis, including not significant fibrosis and healthy volunteers. Thus, the fibrosis spectrum was well-balanced. The difference in absolute stiffness values taken with different systems tends to increase with the increase in fibrosis. In patients at higher risk of cirrhosis, the difference in values with different systems is higher.

The Ultrasound SWS Technical Committee of the QIBA has quantified the differences in measurements between commercially available shear wave elastography US systems using elastic phantoms, and has shown that there is very good agreement among SWS estimations with different systems [7]. Our study confirms "in vivo" the results obtained in elastic phantoms. Moreover, in our study we reduced one source of bias identified in the phantom study, i. e. the depth of the sample box. In fact, it was maintained at a distance of around 4 cm from the skin as long as it was always at least 1.5 cm below the liver capsule as recommended in the statement of the SRU consensus conference, in agreement with the study in phantoms that has shown that the ARFI pulse has a "sweet spot" at a depth of 4 – 5 cm with most systems [3, 7].

The findings of our study confirm that, as suggested by the SRU consensus, an IQR/M ≤ 30 % should be regarded as a quality factor [3]. With the addition of this quality factor to ARFI systems, more measurements may be rejected due to poor quality. However, this should lead to improved accuracy in the final result.

The range of values obtained with the two 2D-SWE systems paralleled that of the FibroScan in cases of a very stiff liver (> 15 kPa), whereas the four systems with a pSWE technology gave lower values in the higher range of liver stiffness. However, as observed in phantoms for one of the systems used in this study, both 2D-SWE systems showed values higher than those of the FibroScan in softer livers [7].

On the other hand, the mean difference in values obtained with three out of four pSWE systems was as low as 0.79 kPa, even though the 95 % LOAs were large, whereas the mean difference between pSWE and 2D-SWE measurements reached a value of 3.64 kPa, which was obtained in the comparison between system 1 and system 5. The different trends observed for pSWE and

2D-SWE measurements compared to the FibroScan could be due to the fact that, like what happens with the FibroScan, the value obtained with 2D-SWE technology is the average of several measurements performed simultaneously in a larger area of the liver, whereas the pSWE assessment is made at a single location. Moreover, the agreement between values obtained with pSWE and 2D-SWE systems was below 0.80 for system 4 and system 5.

Several published studies have reported excellent inter-observer reproducibility between measurements performed in healthy subjects or patients with chronic liver disease with some of the US SWE systems used in this study [13 – 19]. The overall inter-observer agreement observed in our series was above 0.90. The inter-observer agreement for each system was excellent except for system 6 for the pair of observers C-D. On the other hand, system 6 was an outlier since it showed the lowest rate of values with an IQR/M≤ 30 %. When considering the results obtained by both pairs of observers, the highest agreement was observed for the values obtained with the two systems with 2D-SWE technology. It has been reported that the inter-observer agreement between 2D-SWE measurements is affected by the experience of the operator [13, 20]. However, in our study all operators were experts. Unlike pSWE, the assessment of stiffness with 2D-SWE is performed in an ROI that includes several points of the liver tissue. Thus, the possible variance due to heterogeneities between nearby points could be decreased by the average of the values obtained inside the ROI.

This study has limitations. First, the intra-observer variability was not assessed. We made this choice to avoid discomfort to the patient by prolonging the estimated time to complete scanning, which was likely to already be longer than one hour. Besides, the main objective of this study was to assess the variability between systems. Second, the inter-observer agreement for the measurements performed with system 7, i. e. the FibroScan, was not evaluated for both pairs of operators. However, one of the operators had limited experience with the system. Thus, he didn't perform the measurements with the FibroScan in order to avoid any source of bias. On the other hand, in this study the FibroScan was used as the reference standard and the agreement that we obtained for a pair of operators was 0.97 (95 % CI: 0.95 – 0.99), which was similar to the data in the literature [21]. Third, we didn't assess the concordance between systems for different stages of liver fibrosis because the small sample size didn't allow a robust statistical analysis. Fourth, since the operators had already investigated the same patients with other systems, it could be questioned whether the variability between systems could be biased. However, we used the TE as the reference standard, and the measurements with the FibroScan were taken at the end. Thus, the operators did not know the stiffness value from the "reference" system.

In conclusion, the results of this study show that the agreement between LSMs performed with different US systems is good to excellent and the overall inter-observer agreement is above 0.90 in expert hands. However, when staging liver fibrosis with shear wave elastography, the cutoff values could not be applied interchangeably across different US systems.

Conflict of Interest

A conflict of interest has been declared by the author(s).

Acknowledgments

The authors would like to thank Ms. Nadia Locatelli, Secretary of the Ultrasound Unit, for her valuable help in complying with the study protocol.

References

[1] Perepelyuk M, Terajima M, Wang AY et al. Hepatic stellate cells and portal fibroblasts are the major cellular sources of collagens and lysyl oxidases in normal liver and early after injury. Am J Physiol Gastrointest Liver Physiol 2013; 304: G605 – G614

[2] Ferraioli G, Parekh P, Levitov AB et al. Shear wave elastography for evaluation of liver fibrosis. J Ultrasound Med 2014; 33: 197 – 203

[3] Barr RG, Ferraioli G, Palmeri ML et al. Elastography Assessment of Liver Fibrosis: Society of Radiologists in Ultrasound Consensus Conference Statement. Radiology 2015; 276: 845 – 861

[4] Hepatitis B (chronic): diagnosis and management | Guidance and guidelines | NICE [Internet]. [cited 2017 Feb 2]. Available from: https://www.nice.org.uk/guidance/cg165

[5] Dietrich CF, Bamber J, Berzigotti A et al. EFSUMB Guidelines and recommendations on the clinical use of liver ultrasound elastography, Update 2017 (short version). Ultraschall in Med 2017; 38: 377 – 394. doi:10.1055/s-0043-103955

[6] Ferraioli G, Filice C, Castera L et al. "WFUMB Guidelines and Recommendations for Clinical Use of Ultrasound Elastography:Part 3: Liver". Ultrasound Med Biol 2015; 41: 1161 – 1179

[7] Hall TJ, Milkowski A, Garra B et al. RSNA/QIBA: shear wave speed as a biomarker for liver fibrosis staging. In: Ultrasonics Symposium (IUS) I.E. International. 2013: 397 – 400

[8] Palmeri M, Nightingale K, Fielding S et al. RSNA QIBA ultrasound shear wave speed Phase II phantom study in viscoelastic media. Proceedings of the 2015 IEEE Ultrasonics Symposium. 2015: 397 – 400

[9] Tsochatzis EA, Gurusamy KS, Ntaoula S et al. Elastography for the diagnosis of severity of fibrosis in chronic liver disease: a meta-analysis of diagnostic accuracy. J Hepatol 2011; 54: 650 – 659

[10] Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255 – 268

[11] Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1997

[12] Piscaglia F, Salvatore V, Mulazzani L et al. Differences in liver stiffness values obtained with new ultrasound elastography machines and Fibroscan: A comparative study. Dig Liver Dis 2017; 49: 802 – 808. doi:10.1016/j.dld.2017.03.001

[13] Ferraioli G, Tinelli C, Zicchetti M et al. Reproducibility of real-time shear wave elastography in the evaluation of liver elasticity. Eur J Radiol 2012; 81: 3102 – 3106

[14] Boursier J, Isselin G, Fouchard-Hubert I et al. Acoustic radiation force impulse: a new ultrasonographic technology for the widespread noninvasive diagnosis of liver fibrosis. Eur J Gastroenterol Hepatol 2010; 22: 1074 – 1084

[15] D'Onofrio M, Gallotti A, Mucelli RP. Tissue quantification with acoustic radiation force impulse imaging: Measurement repeatability and normal values in the healthy liver. Am J Roentgenol 2010; 195: 132 – 136

[16] Friedrich-Rust M, Wunder K, Kriener S et al. Liver fibrosis in viral hepatitis: noninvasive assessment with acoustic radiation force impulse imaging versus transient elastography. Radiology 2009; 252: 595 – 604

[17] Ferraioli G, Tinelli C, Lissandrin R et al. Point shear wave elastography method for assessing liver stiffness. World J Gastroenterol 2014; 20: 4787 – 4796

[18] Bota S, Sporea I, Sirli R et al. Intra- and interoperator reproducibility of acoustic radiation force impulse (ARFI) elastography–preliminary results. Ultrasound Med Biol 2012; 38: 1103 – 1108

[19] Yoon JH, Lee JM, Han JK et al. Shear wave elastography for liver stiffness measurement in clinical sonographic examinations: evaluation of intraobserver reproducibility, technical failure, and unreliable stiffness measurements. J Ultrasound Med 2014; 33: 437 – 447

[20] Hudson JM, Milot L, Parry C et al. Inter- and intra-operator reliability and repeatability of shear wave elastography in the liver: a study in healthy volunteers. Ultrasound Med Biol 2013; 39: 950 – 955

[21] Fraquelli M, Rigamonti C, Casazza G et al. Reproducibility of transient elastography in the evaluation of liver fibrosis in patients with chronic liver disease. Gut 2007; 56: 968 – 973