Research Article

EASL EUROPEAN ASSOCIATION FOR THE STUDY OF THE LIVER | JOURNAL OF HEPATOLOGY

Liver fibrosis evaluation using real-time shear wave elastography: Applicability and diagnostic performance using methods without a gold standard

Thierry Poynard[1,*], Mona Munteanu[2], Elena Luckina[1,3], Hugo Perazzo[1], Yen Ngo[1], Luca Royer[1], Larysa Fedchuk[1], Florence Sattonnet[1], Raluca Pais[1], Pascal Lebray[1], Marika Rudler[1], Dominique Thabut[1], Vlad Ratziu[1]

[1]*Hepato-Gastroenterology, APHP UPMC Liver Center, Paris, France;* [2]*Research Unit, BioPredictive, Paris, France;* [3]*ANRS, Paris, France*

Background & Aims: Real-time shear wave elastography (SWE) is a new two-dimensional transient elastography which had no assessment of factors associated with reliability, and had limited comparisons with other validated fibrosis biomarkers. The aim was to assess the applicability and performances of SWE for the diagnosis of fibrosis as compared with FibroTest (FT) and liver stiffness measurement (LSM) by transient elastography using two probes (TE-M and TE-XL).

Methods: Without a gold standard, the strength of concordance, discordance analysis and latent class analysis (LCM) were applied.

Results: 422 patients were included. The applicability of SWE (90.0%) was significantly lower than that of FT (97.9%; p <0.0001) and did not differ from those of TE-M (90.5%) and TE-XL (90.3%); it was higher though for SWE (86%) in 22 patients with ascites *vs.* 55% using TE-M ($p = 0.04$). For the diagnosis of all fibrosis stages as presumed by FT, the performance of SWE was highly significant (Obuchowski measure 0.807 ± 0.013 [m ± se]), but lower than those of TE-M (0.852; $p = 0.0007$) and TE-XL (0.834; $p = 0.046$). SWE had a low performance for discrimination between F0 and F1. For the diagnosis of cirrhosis using LCM, SWE specificities were all equal to 99%, and SWE sensitivities ranged from 0.47 to 0.64. For the diagnosis of non-cirrhotic stages, the results were heterogeneous.

Conclusions: The performance of SWE for the diagnosis of cirrhosis was similar to those of FT and TE. SWE applicability was lower than that of FT, but greater than that of TE in patients with ascites.

Keywords: Fibrosis; Obuchowski measure; Concordance; FibroTest; FibroSure; FibroScan; Aixplorer; Non-invasive methods.

Received 30 October 2012; received in revised form 8 December 2012; accepted 20 December 2012; available online 12 January 2013

* Corresponding author. Address: Hepato-Gastroenterology, APHP UPMC Liver Center, 47-83 Boulevard de l'Hôpital, 75651 Paris Cedex 13, France. Tel.: +33 1 42 16 10 22.

E-mail address: tpoynard@teaser.fr (T. Poynard).

Introduction

Liver fibrosis evaluation using real-time shear wave elastography (SWE) by Aixplorer™ is a new two-dimensional transient elastography technique [1], which has been used in few studies of liver disease, and only in patients with chronic hepatitis C [2,3].

Like transient elastography (TE) and acoustic radiation force imaging (ARFI), SWE evaluates the speed of a shear wave to provide a quantitative estimate of tissue stiffness. SWE has the advantage over TE of being able to image liver stiffness in real time because the shear waves are generated by ultrasound pushes. The SWE image, not limited to a single location, is also guided by a higher frame-rate B-mode image than TE and ARFI. SWE is providing a real-time quantitative map of liver tissue stiffness [1–3].

The applicability (failure rate and non-reliable rate) from these limited published studies of SWE in liver disease is unknown. Furthermore, no study has compared the applicability and performance of SWE for the diagnosis of fibrosis with the two most validated non-invasive biomarkers, the *in vitro* multivariate assay FibroTest (FT) and liver stiffness measurement based on TE using Fibroscan™ [4–6].

The first aim of this study was to estimate and compare the applicability of SWE with that of FT and the two Fibroscan probes (TE-M and TE-XL) in consecutive patients with chronic liver disease not restricted to chronic hepatitis C. We previously demonstrated that the applicability of fibrosis estimates directly impacts the performance of tests in an intention-to-diagnose analysis [7].

The second aim of this study was to compare the diagnostic performance of SWE to those of FT and the two Fibroscan probes. Liver biopsy is usually used as the reference when performing these comparisons, but this methodology has major limitations. Even a biopsy specimen 25 mm in length has more than 20% false positive or false negative results for fibrosis staging *vs.* large surgical biopsies [8], and there is a significant gray zone for intermediate stages [9]. Therefore classical estimates of diagnostic test accuracy (sensitivity, specificity, area under the ROC curves [AUROC] and predictive values) are false or very limited [10]. The magnitude of the impact of this error of biopsy is so great that AUROC determinations >0.90 for the diagnosis of advanced

fibrosis could not be achieved, even for a marker that perfectly measured the disease [11]. Statistical analysis using methods without a gold standard (“Truth in the Absence of a Gold Standard” [TAGS]) has been recently validated for the evaluation of fibrosis biomarkers and was applied in the present study [12,13].

Patients and methods

Patients

Consecutive patients undergoing chronic liver disease assessment at the “Groupe Hospitalier Pitié Salpêtrière” Hospital in Paris, France were prospectively recruited. We included patients aged 18 years or older who had undergone simultaneous serum sampling for FT and attempted liver stiffness measurements with SWE and TE-M and TE-XL. All patients gave written informed consent.

Standard definitions of chronic liver disease etiologies were used for alcoholic liver disease (ALD), chronic viral hepatitis B (CHB) and C (CHC) and non-alcoholic fatty liver disease (NAFLD) with standard definition of metabolic syndrome [14]. Anthropometric tests included body weight, body height, and waist circumference measurements. Liver histology was not prospectively performed, and did not serve as the gold standard; a common fibrosis scoring system, similar to the METAVIR, was used [15] (Supplementary methods).

Real-time shearwave elastography

Real-time SWE was performed using the Aixplorer™ ultrasound system (SuperSonic Imagine S.A., Aix-en-Provence, France) with a convex broadband probe (SC6-1) as recommended [2] (Supplementary methods). Shear waves are created in liver tissue from the acoustic radiation force generated by focalized ultrasound pulses. By placing a circular region of interest (ROI) in an SWE image, the mean and standard deviation of the elasticity within the ROI can be displayed. In this study, we used an SWE box size of 3.5 × 2.5 cm. SWE measurements were performed on the right lobe of the liver through intercostal spaces with the patient in the supine position and the right arm maximally abducted.

The same intercostal space was used for both TE and SWE measurements, with SWE successively performed after TE-M and TE-XL. The upper edge of the SWE box was placed 1.5–2 cm from Glisson’s capsule in the liver and in an area of parenchyma free of large vessels. The entire real-time SWE examination lasted approximately 5 min per patient.

Only experienced operators (more than 50 SWE, TE-M, and TE-XL measurements) participated in the study (EL, HP, YN, MM, and LF). The reproducibility of SWE measurements was estimated for elasticity by Ferraioli *et al.* in 42 healthy volunteers [16]. Healthy subjects showed real-time SWE values ranging from 4.92 (SD 0.71) kPa to 5.39 (SD 0.91) kPa.

Transient elastography

All consecutive patients had liver stiffness measurements using M and XL probes [17]. These were done in the same session before SWE using FibroScan™ (Echosens, Paris, France) according to the instructions and training provided by the manufacturer. The following standard recommended cut-offs were used to estimate the presumed fibrosis stages: 7.1 and 14.5 kPa for F234 and F4 staging, respectively [13,17].

FibroTest

FT was performed according to the manufacturer’s and health authorities’ recommendations [5–6,18]. The following standard recommended cut-offs were used to estimate the presumed fibrosis stages: 0.48 and 0.74 for stages F234 and F4, respectively.

Definition of applicability rate

In order to maintain consistency with large studies on validated fibrosis biomarkers, the following definitions of applicability rate (non-failure rate + reliability rate) were used. For FT, a measurement was classified as a failure when serum sampling was impossible; it was classified as non-reliable if one component in the measurement had an extreme value, which induced a change of more than 0.20 in the FT value when calculated using the usual median instead [18].

For TE-M and TE-XL measurements, signal absence was considered a failure, and 2 reliability definitions were used. The standard one was the IQR/liver stiffness measurement (IQR/M) >0.30, less than 10 measurements and/or a success rate lower than 60% [17]. Recently, a second definition using only IQR and LSM was proposed, as the number of measurements and the success rate had no independent significant value when biopsy was the reference [19]. Three categories were generated: “very reliable” (IQR/M $\leqslant$0.10), “reliable” (0.10 <IQR/M $\leqslant$0.30, or IQR/M >0.30 with LSE median <7.1 kPa), and “poorly reliable” (IQR/M >0.30 with LSE median $\geqslant$7.1 kPa).

For SWE, a measurement was classified as a failure when no signal was obtained. No clear definition of a non-reliable measurement has been published by the SWE manufacturer or in the first studies of liver disease [1–3,16]. One study only stated that a failure was defined as “no/little signal (lower than 1 kPa) obtained in the SWE box for all the acquisitions” [3]. Therefore we used the “strength of concordance” method to identify non-reliable SWE measurements, as detailed in the next paragraph. SWE measurements with a minimal value less than 0.2 kPa were considered non-reliable.

Design and modeling

Concepts

We applied the methods without a gold standard (“Truth in the Absence of a Gold Standard” [TAGS]), which had been previously validated for the evaluation of fibrosis biomarkers: the strength of concordance between estimates of fibrosis stages [12], the discordance analysis and the latent class model (LCM) [10,13,20,21].

Strength of concordance

When there are no perfect gold standards but only imperfect ones for estimating the truth, measurement of the strength of the concordance between these imperfect gold standards could be used as a tool for identifying factors of variability.

Any variability factor of one test should impact the strength of the association between the two tests, assuming that this variability factor is not also associated with the other test (independent tests). We have previously illustrated this concept [14] (Supplementary methods). When subjects with TE-M high variability factors (IQR/M >0.30, number of measurements <10 and success rate <60%) were included, the strength of association between LSM and FT decreased significantly as compared with that from a population excluding these subjects.

We apply this concept in the present study, first to improve the definition of non-reliable SWE measurements, and second to identify potential factors of SWE variability.

To improve the definition of not reliable SWE measurements, and because poor signals (<1 kPa) were considered to be non-reliable signals [3], we tested the strength of concordance between SWE measurements and FT in the lower categories of minimal signals (<1 kPa) to identify a security cut-off.

The potential factors of SWE variability were identified after excluding patients with non-reliable tests. The strength of concordance between SWE and FT was compared according to operators, gender, presence of significant steatosis (presumed with SteatoTest >0.57), significant necroinflammatory activity (presumed with ActiTest >0.52), thoracic fold measured by SWE, cause of liver disease and ethnic origin.

Analysis of discordant cases

For patients with high discordance (cirrhosis *vs.* F0–F1) between SWE and FT, the criteria used for attributing the cause of error were independent of both FT and SWE: LSM-M, LSM-XL, other ROI: SWE-Q2, -Q3 or -Q4, prothrombin time was less than 50%, platelet count lower than 100,000/mm^3, large varices present on endoscopy (Supplementary methods).

Latent class analysis

LCM uses a mathematical method (the standard maximum likelihood) to obtain a (unique) solution for constructing a reference standard [10,20,21]. For each test, this reference standard estimates sensitivities and specificities compatible with the observed distributions [15].

This method acknowledges that there is no gold standard, and that the available tests are all related to the unknown true disease status: fibrosis present or absent. These unobservable outcomes are named “latent classes”.