Research Article

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

Relative performances of FibroTest, Fibroscan, and biopsy for the assessment of the stage of liver fibrosis in patients with chronic hepatitis C: A step toward the truth in the absence of a gold standard

Thierry Poynard[1,*], Victor de Ledinghen[2], Jean Pierre Zarski[3], Carol Stanciu[4], Mona Munteanu[5], Julien Vergniol[2], Julie France[6], Anca Trifan[4], Gilles Le Naour[1], Jean Christophe Vaillant[1], Vlad Ratziu[1], Frederic Charlotte[2], The Fibrosis-TAGS group [†]

[1]*APHP UPMC Liver Center, France;* [2]*Service d'Hépato-Gastroentérologie, CHU de Bordeaux, France;* [3]*Clinique Universitaire d'Hépato-Gastroentérologie, Grenoble, France;* [4]*Gastroenterology and Hepatology Institute, Iasi, Romania;* [5]*Biopredictive, Paris, Franc e;* [6]*Centre d'Investigation Clinique INSERM CIC003, CHU Grenoble, France*

Background & Aims: Liver fibrosis stage is traditionally assessed with biopsy, an imperfect gold standard. Two widely used techniques, FibroTest®, and liver stiffness measurement (LSM) using Fibroscan® have been validated using biopsy, and therefore the true performances of these estimates are still unknown in the absence of a perfect reference.

The aim was to assess the relative accuracy of FibroTest, LSM, and biopsy using methods without gold standard in patients with chronic hepatitis C (CHC) and controls.

Methods: A total of 1289 patients with CHC and 604 healthy volunteers, with assessment of fibrosis stage by the three techniques, and alanine aminotransferase (ALT) taken as a control test, were analyzed by latent class method with random effects. In the volunteers, the false positive risk of biopsy was obtained from a large surgical sample of four normal livers.

Results: The latent class model with random effects permitted to conciliate the observed data and estimates of test performances. For advanced fibrosis, the specificity/sensitivity was for FibroTest 0.93/0.70, LSM 0.96/0.45, ALT 0.79/0.78 and biopsy 0.67/0.63, and for cirrhosis FibroTest 0.87/0.41, LSM 0.93/0.39, ALT 0.78/0.08 and biopsy 0.95/0.51. The analysis of the discordances between pairs suggested that the variability of the model was mainly related to the discordances between biopsy and LSM (residuals >10; p <0.0001).

Conclusions: A method without the use of a gold standard confirmed the accuracy of FibroTest and Fibroscan for the diagnosis of advanced fibrosis and cirrhosis in patients with chronic hepatitis C. The variability of the model was mostly due to the discordances between Fibroscan and biopsy.

Keywords: Liver biopsy; FibroTest; FibroSure; FibroScan; Transaminases; Latent class; Accuracy; Methods without gold standard; Hepatitis C; Non-invasive biomarker.

Received 14 April 2011; received in revised form 30 July 2011; accepted 9 August 2011; available online 1 September 2011

* Corresponding author. Address: 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).

† List of members of the Fibrosis-TAGS group (Truth in the Absence of a Gold Standard) is available in Supplementary data.

Introduction

New tests generally are evaluated in comparison with a reference test, often termed a "gold standard", whose sensitivity and specificity are both assumed to be 100%. If the reference test is not perfect, classical estimates of accuracy (sensitivity, specificity and AUROC) of the new diagnostic test are false [1].

One example of major debate surrounds the efforts to find the best means of evaluating and managing the increasing numbers of patients with chronic liver disease [2,3]. Liver biopsy, due to its risks and limitations, is no longer considered mandatory as the first-line indicator of liver injury, and several markers have been developed as non-invasive alternatives [2,3]. Among patients with chronic viral hepatitis, the assessment of liver fibrosis by two validated non-invasive techniques, biomarkers [FibroTest® (FT)] Biopredictive Paris, France [4] and liver stiffness measurements (LSM) by Fibroscan® Echosens, Paris, France [5], is now widely done in countries where these techniques are available and approved [6].

The true liver disease status, the "true gold standard", is the histological analysis of large surgical biopsies [7]. Therefore, the definitive diagnosis is impossible to obtain in routine practice, and liver biopsy, an "imperfect gold standard", is used as a standard against which new tests are evaluated.

In this situation with several tests and no perfect gold standard, latent class analysis has been recommended to better estimate the rate of false positives and false negatives [1], and we previously performed a pilot study using this methodology [8].

The aim was then to apply this methodology to estimate the relative accuracy of FT, LSM and biopsy for the diagnosis of fibrosis in the absence of a gold standard in a large group of patients,

Research Article

with CHC, independent of our institution, and in healthy volunteers. The reference was the model which fitted the best the observed distribution of the estimates of fibrosis.

Materials and methods

Patients

The final database included 1893 subjects retrospectively extracted from four prospective cohorts (Fig. 1): three populations of patients with CHC (n = 1289 out of 2675), and one population of apparently healthy volunteers (Healthy cohort, n = 604 out of 766). HCV patients belonged to one tertiary center in Bordeaux, France (Bordeaux cohort, n = 768) [9], one multicenter French study (Fibrostar cohort, n = 378) [10] and one multicenter Romanian study (Romanian cohort, n = 143) [11].

The inclusion criteria were retrospectively determined: patients had to have chronic hepatitis C, be PCR positive, and have the results of liver biopsy, FibroTest, LSM and alanine aminotransferase [ALT] interpretable according to the usual recommendations and precaution of use [4,9]. In all these cohorts, each of the four tests was performed without knowledge of the three others.

Controls

This group was analyzed in order to define the specificity of each test, as the probability of true advanced fibrosis was very low. Among a prospective cohort of healthy volunteers, a group of 604 subjects without any risk of liver disease was retrospectively selected [12]. The inclusion criteria were: no liver disease history, no or low alcohol consumption ($\leqslant$10 g/day for females, $\leqslant$20 g/day for males), HBsAg negative, HCV antibodies negative, and FibroTest and LSM results interpretable.

As it was not possible to perform liver biopsy in these healthy volunteers, we used large surgical biopsies obtained from four subjects without liver disease. From the digitized image of the whole section, 626 virtual biopsy specimens of 20 mm length were produced [13] (Supplementary Table 1).

FibroTest and ALT

FibroTest was performed according to published recommendations [4]. The following usual recommended cut-offs were used to estimate the presumed fibrosis stages: 0.48, and 0.74 for the F2 and F4 staging, respectively. ALT was used as a control liver test as a nonspecific biomarker of liver injury. As there is no consensual definition for the upper limit of normal for ALT, the following simple cut-offs were predetermined: 50 IU/L and 100 IU/L for F2 stage and F4 stage METAVIR, respectively.

Liver stiffness measurements

Patients were studied using transient elastography. The LSM results are expressed in kilopascals (kPa). For LSM reliability, the recommended criteria were a success rate greater than 60%, at least 10 valid LSM and interquartile range/median LSM <30% [9]. The following usual recommended cut-offs were used to estimate the presumed fibrosis stages: 8.8, and 14.5 kPa for the F2, and F4 staging, respectively [9,14,15].

Biopsy among patients with chronic hepatitis C

Staging and grading were performed blinded to the non-invasive methods. In the three groups, liver biopsies were performed with a 1.6 mm needle (Hepafix, Brown, Melsungen, Germany), and were formalin-fixed and paraffin embedded. Sections (4 mm) were stained with hematoxylin-eosin-saffron and picrosirius red. The liver fibrosis stage was evaluated according to the METAVIR scoring system [16] by one senior pathologist in the Bordeaux cohort and in the Romanian study, by two senior liver pathologists in Fibrostar. In Fibrostar, slides were simultaneously reviewed to reach a consensus in case of disagreement; to be eligible for scoring, biopsies less than 20 mm had to measure at least 15 mm and/or contain at least 11 portal tracts, except for cirrhosis. The reliability of biopsy was decided by each pathologist in the Romanian study and Bordeaux cohorts.

Design and modeling

Concept

The first concept was to estimate the performances of four estimates (tests) of liver fibrosis using methods without a gold standard.

The second concept was to use a control population without any risk of chronic liver disease, therefore with a very low risk of advanced fibrosis. This concept will permit to assess the performance of the fibrosis tests in screening strategies. As a biopsy cannot be directly performed in a large group of non-selected healthy volunteers, the distribution of subjects according to the results of a virtual biopsy (fibrosis present or absent) was calculated using the prevalence of fibrosis observed using large surgical biopsies from normal livers. For each eight possible combinations of FibroTest, LSM and ALT results (fibrosis present or absent), the number of virtual biopsy results (fibrosis present or absent) was calculated by multiplying the number of subjects in each eight possible combinations by the

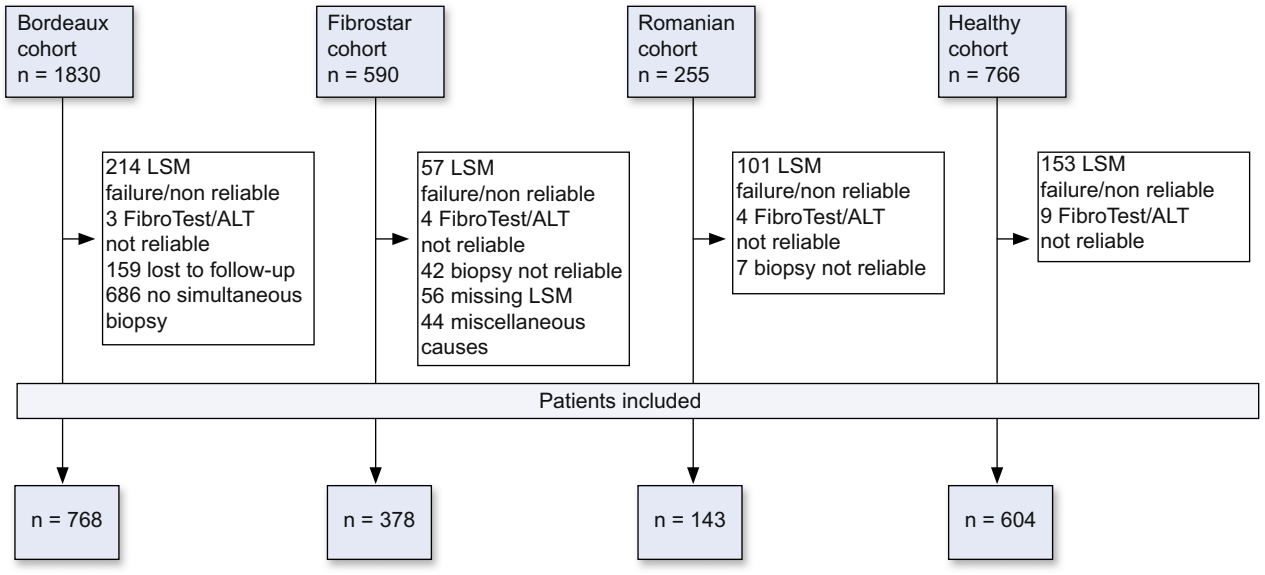

Fig. 1. Cohort and number of patients included and excluded.

mean prevalence of fibrosis observed using large surgical biopsies from normal livers. This method has permitted to generate the 16 distributions of subjects according to the four test results (Supplementary Table 3).

Basic model

Four different tests (FibroTest, LSM, ALT, biopsy) were applied in all patients, with each test producing a dichotomous test result (e.g. the test was either positive or negative). None of these tests was error-free. For a single test, the probability of obtaining a positive test result could be written as the sum of finding a positive test in a patient who has fibrosis and a positive test result in a patient without fibrosis. These probabilities can be written as a function of the following unknown measures: prevalence, sensitivity and specificity of the test. Therefore, nine parameters were unknown in this study: one prevalence parameter and the sensitivity and specificity for each of the four tests.

With four different dichotomous tests, there were 16 possible combinations. By using the probabilities for a positive or negative test result, the likelihood of observing each pattern of test results could be calculated. We observed the number of subjects for each of the 16 patterns of test results. Standard maximum likelihood methods could be used to obtain a (unique) solution [1,17,18].

Latent class analysis

Latent class uses the standard maximum likelihood method to combine the test results from each patient for constructing a reference standard [1,17–19]. This method acknowledges that there is no gold standard and that the available tests are all related to the unknown true status: fibrosis present or absent. These unobservable outcomes are named latent classes.

The fact that a two-class model might not fit the data is either seen as an artifact of the measurement instrument or as a result of within-class heterogeneity. To allow for local dependencies and within-class heterogeneity, we used a LCM model with a random-factor, the LCM-R model [1,17–19]. The LCM-R model incorporates random effects and thus relaxes the conditional independence assumption (see Supplementary statistical method details).

The specific assumptions for random effects were the following: the dependency between tests for FibroTest and LSM which were initially validated by biopsy; the intra-class heterogeneities for biopsy due to inter-observer variability and sampling error; for LSM, the inter-observer variability and the impact of inflammation and steatosis.

In LCM-R, it is assumed that the outcome of a diagnostic test is governed by two mechanisms or factors: the disease status of the subject, and the individual biological process or the diagnostic test technological characteristics.

Sources of fit impairment

We assessed which test dependency or heterogeneity significantly impaired the fit of the standard LCM without random effects by using bivariate residuals of the baseline latent class analysis. The pair of tests was excluded step by step up until a model fitting the observed results was obtained. The fit was reached when the likelihood-ratio goodness-of-fit value [likelihood squared (L^2)] L^2 significance was >0.05 [1,17–19].

Standard performance analysis using biopsy as a gold-standard

The standard performances of FibroTest, LSM and ALT were assessed using the fibrosis stage obtained by liver biopsy, the classical gold standard, expressed using the METAVIR scoring system. The thresholds for test positivity were the usual ones. The standard area under the Receiver Operating Characteristics Curves (AUROC) was estimated by the empirical (non-parametric) method, and compared using the paired method of Zhou *et al.* [20].

Sensitivity analyses

To assess possible variability due to the sampling population, we performed successive LCM-R models (excluding each populations): excluding false positives from each test, one without any false positive, one with lower cutoff for cirrhosis 10.1% of the area of fibrosis, and two with lower LSM cut-offs: 7.1 for advanced fibrosis and 12.5 kPa for cirrhosis. We performed also a meta-analysis using random effect model of weighted AUROCs (Obuchowski measure) to identify significant heterogeneity between the different populations of patients [21].

Statistical analysis and software

We used NCSS software (Kaysville, Utah, USA) [22] for standard statistics and LatentGold-4.5 software (Statistical Innovation, Belmont, MA, USA) for estimating the model parameters [19]. We used the following criteria to identify a good model: the *p*-value of the likelihood squared (L^2) had to be greater than 0.05, and the Bayesian information criterion (BIC), defined as $L^2 - \log(N) \times Df$ (degrees of freedom of the data), had to be the smallest among all competing models. Standard error of L^2 was calculated used bootstrap method [19].

This study was conducted according to the principles expressed in the declaration of Helsinki. Signed informed consent was obtained for all controls and for patients for whom tests were not routinely performed according to the standard of care.

Results

Failure and non-reliable results were observed in 15.3% (525/3441) of LSM and in 0.6% for FibroTest (20/3441).

Subjects included

The characteristics of included patients are described in Table 1. Healthy controls were more often female and older than HCV patients. Patients of the Romanian population were more often female, and had less cirrhosis at biopsy. The median length of biopsy was 17 mm in the Bordeaux group, 25 mm in the French multicenter group and 20 in the Romanian multicenter group.

Standard assessment of biomarker performance using biopsy as the reference (imperfect gold-standard)

Performances of FibroTest, LSM and ALT using the standard AUROCs (95% CI), observed among patients with biopsy, were similar to those of the extensive literature [8,10]; for the diagnosis of advanced fibrosis: 0.75 (95% CI 0.72–0.77), 0.76 (0.73–0.79) and 0.62 (0.59–0.65), and for cirrhosis 0.85 (0.82–0.88), 0.90 (0.87–0.92) and 0.61 (0.57–0.66) respectively. As expected, performances of ALT were significantly lower than those of FibroTest and LSM (p <0.0001).

Assessment of the specificity of liver biopsy using large surgical biopsies

The distribution of the area of fibrosis estimated by virtual biopsies of different lengths is shown in Fig. 2 and Supplementary Table 1. Cases with areas of fibrosis above 5.3% were considered to be false positives of biopsy for the diagnosis of advanced fibrosis, and those above 16.5% as false positives for the diagnosis of cirrhosis. The specificity of a 20 mm length biopsy for the diagnosis of advanced fibrosis was 83.71% (Supplementary Table 2).

Assessment of test performances in the absence of a gold standard

The distribution of the subjects according to the 16 possible combinations of the four test results are shown in Supplementary Table 3 for presuming advanced fibrosis, and in Supplementary Table 4 for cirrhosis. Perfect concordance between the tests for the diagnosis of advanced fibrosis was observed in 1059 (55.4%) subjects (728 all negatives and 321 all positives) and for the diagnosis of cirrhosis in 1340 (70.8%) (1292 all negatives and 48 all positives). Details of the assessment in healthy volunteers are given in Supplementary data 2 for the diagnosis of advanced fibrosis.

Models using LCM-R were interpretable as they fit (Table 2) the observed distribution of test results. For advanced fibrosis, the ranking for the specificities was LSM (0.96), FibroTest (0.93)

Research Article

Table 1. Characteristics of the 1893 included subjects.

Characteristics	HCV patients' group				Healthy volunteers
	Bordeaux	Multicenter France	Multicenter Romania	All patients	
	n = 768	n = 378	n = 143	n = 1289	n = 604
Age, yr[1]	48 (47-49)	50 (49-51)	49 (48-52)	49 (48-50)	58 (56-59)
Male, (%)	441 (57%)	239 (63%)	48 (34%)	728 (56%)	209 (44%)
Biopsy stage[2]	2 (2-2)	1 (1-2)	2 (2-2)	2 (2-2)	0 (0-0)
Presumed fibrosis	523 (68%)	176 (47%)	89 (62%)	788 (61%)	16%[3]
Presumed cirrhosis	136 (18%)	57 (15%)	6 (4%)	199 (15%)	3%[3]
FibroTest	0.47 (0.43-0.50)	0.58 (0.53-0.64)	0.48 (0.40-0.53)	0.50 (0.48-0.53)	0.16 (0.15-0.16)
Presumed fibrosis	370 (48%)	229 (61%)	69 (48%)	668 (52%)	19 (3%)
Presumed cirrhosis	171 (22%)	123 (33%)	18 (13%)	312 (24%)	2 (0.3%)
LSM, kPa (8.8/14.5)	7.0 (6.8-7.3)	7.0 (6.7-7.7)	7.7 (7.2-8.8)	7.1 (6.9-7.4)	5.4 (3.6-6.7)
Presumed fibrosis	251 (33%)	126 (33%)	58 (41%)	435 (34%)	19 (3%)
Presumed cirrhosis	124 (16%)	54 (14%)	28 (20%)	206 (16%)	2 (0.3%)
LSM, kPa (7.1/12.5)					
Presumed fibrosis	368 (48%)	185 (49%)	84 (59%)	637 (49%)	38 (6%)
Presumed cirrhosis	151 (20%)	70 (19%)	36 (25%)	257 (20%)	5 (1%)
ALT, IU/L	65 (61-68)	72 (65-78)	93 (83-105)	69 (66-74)	22 (21-23)
Presumed fibrosis	491 (64%)	271 (72%)	117 (82%)	879 (68%)	23 (4%)
Presumed cirrhosis	207 (27%)	105 (28%)	64 (48%)	376 (29%)	4 (0.7%)

[1]Median (95% confidence interval).
[2]METAVIR scoring system.
[3]False positive of a 20 mm length biopsy as assessed using large surgical specimens.

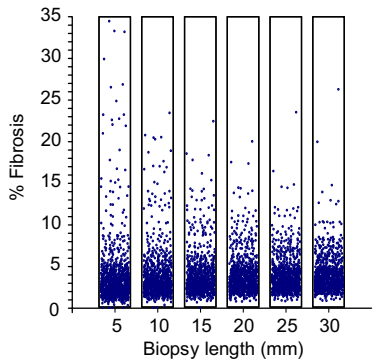

Fig. 2. Area of fibrosis estimated by biopsy according to its length (mm) in subjects scoring METAVIR F0 (no fibrosis) on the large surgical section. Cases with area of fibrosis above 5.3% were considered false positives of biopsy for the diagnosis of advanced fibrosis and those above 16.5% were false positives for the diagnosis of cirrhosis.

and biopsy (0.67); the ranking for the sensitivities was FibroTest (0.70), biopsy (0.63) and LSM (0.45). For cirrhosis, the ranking for the specificities was biopsy (0.95), LSM (0.93), and FibroTest (0.87); all sensitivities were low with the following ranking: biopsy (0.51), FibroTest (0.41), and LSM (0.39).

Compared to their performances assessed by biopsy, the performances of FibroTest and assessed by LCM-R were all increased for the diagnosis of advanced fibrosis and decreased for the diagnosis of cirrhosis. The performances of LSM were lower using LCM-R except for an increase in the specificity for advanced fibrosis (Table 3).

Models using LCM without random effects did not fit the observed distribution, suggesting a random effect due to dependency between tests (as expected due to previous validation of FibroTest and LSM by biopsy) and intra-class heterogeneity such as inter-observers variability for biopsy and LSM (Supplementary Table 5).

Assessment of significant sources of impairment in modeling

Biopsy-LSM and biopsy-ALT were identified as the two main sources of impairment in LCM models both for advanced fibrosis and cirrhosis. Bivariate residuals of LSM-ALT and biopsy-FibroTest were lower but also significantly impaired the model fit for advanced fibrosis (Table 4).

Sensitivity analyses

The population that impaired the goodness of fit the most was the healthy population results, since when excluded, the baseline BIC decreased from 34.4 to −17.7 for advanced fibrosis and from 21.6 to 9.6 for cirrhosis (Supplementary Table 6). The exclusion of healthy volunteers strongly modified the estimates, reducing specificities both for advanced fibrosis and cirrhosis and increasing sensitivities for advanced fibrosis (Supplementary Table 7). None of the other LCM-R analyses showed a major decrease of the fit assessed by BIC value (Supplementary Table 6). Results were not different when the diagnosis of cirrhosis used >10.1% area of fibrosis in healthy volunteers (Supplementary Table 8). When lower cut-offs (7.1 *vs.* 8.8 kPa) were used for LSM, this induced an expected dramatic increase in the sensitivity of LSM for advanced fibrosis from 0.45 to 0.88 but a decrease of specificity from 0.96 to 0.83 (Supplementary Table 9).

Table 2. Best latent class model with random effect of fibrosis estimate performances.

Best model for advanced fibrosis (n = 1893)		
L-Squared (standard error calculated using bootstrap)	3.2 (0.02)	
Goodness of fit likelihood ratio test statistics: *p* value[1]	0.20	
Bayesian information criterion	-11.9	
Performance of test	**Specificity[2]**	**Sensitivity[2]**
FibroTest	0.93	0.70
LSM	0.96	0.45
ALT	0.79	0.78
Biopsy	0.67	0.63
Best model for cirrhosis (n = 1893)		
L-Squared (standard error calculated using bootstrap)	0.61 (0.01)	
Goodness of fit likelihood ratio test statistics: *p* value[1]	0.74	
Bayesian information criterion	-14.5	
Performance of test	**Specificity[2]**	**Sensitivity[2]**
FibroTest	0.87	0.41
LSM	0.93	0.39
ALT	0.78	0.08
Biopsy	0.95	0.51

[1]Model fit when $p > 0.05$.
[2]No confidence interval for the LCM-derived sensitivity and specificity estimates because these estimates are calculated from combinations of conditional probabilities, which have individual maximum-likelihood estimated standard errors.

Table 3. Sensitivity and specificity of fibrosis biomarkers according to the choice of the reference: biopsy (an imperfect gold standard) or a model without gold standard (latent class model with random effect [LCM-R] as reference) in 1893 subjects.

Estimate[1]	Advanced fibrosis				Cirrhosis			
Reference	Biopsy		Latent class		Biopsy		Latent Class	
	Sp	Se	Sp	Se	Sp	Se	Sp	Se
FibroTest	0.85	0.66	0.93	0.70	0.89	0.68	0.87	0.41
LSM	0.93	0.48	0.96	0.45	0.95	0.65	0.93	0.39
ALT	0.70	0.73	0.79	0.78	0.83	0.42	0.78	0.08
Biopsy	1.00[2]	1.00[2]	0.67	0.63	1.00[2]	1.00[2]	0.95	0.51

The standard test cut-offs used for the diagnosis of advanced fibrosis and cirrhosis were 0.48 and 0.74 for FibroTest, 8.8 and 14.5 kPa for stiffness, 50 IU/L and 100 IU/L for ALT, and for biopsy in LCM-R model F2 stage and F4 stage METAVIR for real biopsy, and 5.3% and 16.5% area of fibrosis for virtual biopsies in healthy volunteers respectively.
[1]Standard errors or 95% confidence interval are not given as for the LCM-derived sensitivity and specificity estimates, because they are calculated from combinations of conditional probabilities.
[2]In this model, biopsy is considered as the reference ("gold standard") with 100% accuracy.

The meta-analysis using random effect model of weighted AUROCs showed no significant heterogeneity between the different populations of patients (Supplementary Table 10) contrarily to nonweighted AUROCs (Supplementary Table 11). The details of the 95% confidence intervals of standard sensitivity and specificities (using biopsy as reference) are given in Supplementary Table 12.

Discussion

This study is the first using appropriate methods for better reconciliation of the estimates of sensitivity and specificity of noninvasive fibrosis biomarkers, as well as those of biopsy, the former gold standard, which cannot be 100% accurate [23]. The main result is that a model without using reference is compatible with the distribution of biomarkers and biopsy results.

The high specificity (>0.85) of FibroTest and LSM was confirmed for the diagnosis of both advanced fibrosis and cirrhosis. As already observed in standard analysis and in a preliminary latent class study [8], the results confirmed that the sensitivity of FibroTest (0.70) was higher than that of LSM (0.48) for the diagnosis of advanced fibrosis. The performance for the diagnosis of cirrhosis was similar between FibroTest and LSM.

One original result of the present study is the relative lower level of biopsy performance, in comparison with FibroTest and LSM when evaluated similarly for the diagnosis of advanced fibrosis. For cirrhosis, biopsy had the best performance with the highest specificity, and the highest sensitivity but far from perfection, with 49% of presumed false negativity rate, as FibroTest and LSM.

Strengths of the study

Population included

The first strength was the wide spectrum of liver injury, from healthy volunteers to cirrhotic patients, with two multicenter studies in two different countries.

The second strength was the inclusion of a large healthy population with biomarkers, together with the presumed results of biopsies generated from normal livers. The inclusion of a healthy population in the model changed it very significantly. One major

Research Article

Table 4. Direct effects of pairs of variables that impaired the fit of the baseline latent class model. Effects are estimated by bivariate residuals of the baseline latent class analysis, without random effects. The effect of the most significant pair was excluded to achieve non-significance.

	Bivariate residuals			Model improvement after excluding residuals		
	FibroTest	LSM	ALT	Pair excluded (Cumulative)	Fit (L^2) (Cumulative)	Significance after pair exclusion[2]
Advanced fibrosis				None	79.7[1]	<0.0001
LSM	0.44			Biopsy-LSM	38.3	<0.0001
ALT	2.9	0.14		Biopsy-ALT	30.1	<0.0001
Biopsy	0.11	11.6	0.47	LSM-ALT	13.8	0.003
				Biopsy-FibroTest	0.32	0.85
Cirrhosis				None	66.0[1]	<0.0001
LSM	0.24			Biopsy-LSM	27.7	<0.0001
ALT	3.52	0.29		Biopsy-ALT	9.96	0.04
Biopsy	0.95	10.8	10.7			

[1]Baseline fit.
[2]Model fit when $p > 0.05$.

weakness of previous overviews of LSM performance was the absence of conciliation between the LSM accuracy estimated in patients [5] with the positive rate observed in healthy populations [12,15]. The 95th percentiles of LSM in a healthy non-obese population without metabolic syndrome, 7.8 kPa for females and 8.0 kPa for males, observed by Roulot *et al.* [15], were in accordance with the 3% positive rate of LSM (above 8.8 kPa) observed in our healthy volunteers (Table 1) and with the 4% of false positive for advanced fibrosis estimated by our LCM-R model (Table 2 and Supplementary Table 5). LSM should not be used at the 8.8 kPa cutoff for screening purposes, as the specificity was 96% but only applicable in 45% of patients.

Use of latent class with random effects

The third strength was the use of a latent class paradigm with random effect which introduces a random variability factor in the model. FibroTest and LSM were initially validated using biopsy, and therefore it was rational to use a method which takes into account this non-independence between tests.

All tests can then be compared without the systematic bias of the absence of error for biopsy. FibroTest performances were similar to that of a 20 mm biopsy for the diagnosis of advanced fibrosis.

As expected, performances of ALT were lower than those of FibroTest for the diagnosis of both advanced fibrosis and cirrhosis. The main interest of ALT used as a negative control test was to better understand the possible sources of variability among LSM and biopsy.

Sources of major variability among tests

The fourth strength was the identification of the major sources of test dependency and heterogeneity. Indeed, LCM failed to obtain a model that fits with the observed distribution, without including a "random factor" that is unknown sources of discordances not related to the diagnostic performance of tests (Table 3). As FibroTest and LSM were validated using biopsy, the first rational variability factor was this initial "dependency".

The variability was mainly related to the biopsy-LSM residual. It was rational to observe the greater variability for the biopsy-LSM pair, as these indicators have both significant intra- and inter-observer variability [7,8,9,15,24,25] in comparison with the smaller analytical variability of FibroTest [26]. Furthermore the biopsy-LSM pair variability is impacted by the (fibrosis stages) spectrum effect to a greater degree than the biopsy-FibroTest pair. LSM has no diagnostic value for the initial fibrosis stages (METAVIR F0 and F1), a limited accuracy between stages F1 *vs.* F2, and a higher accuracy between F2, F3 and F4. Contrary to LSM, FibroTest has a consistent accuracy between adjacent stages [3,4,8,10].

The biopsy-ALT pair was the second source of residuals for the diagnosis of advanced fibrosis, without obvious bias as the pathologists were not aware of the ALT value. However, a bias related to an overestimation of liver fibrosis stage cannot be ruled out during biopsy readings, when biopsies showed higher activity grades.

The LSM-ALT pair was the third most important residual with a documented rationale, as necrosis and inflammation increased LSM independent of fibrosis stage [8,27,28].

The various sensitivity analyses (LSM cut-offs, area of fibrosis cut-offs, population, false positive rate in healthy volunteers) did not induced any absence of fit (Supplementary Tables 6 and 8). In the LCM-R model, despite no change in the fit, there was indeed a "cutoff effect" of LSM on FibroTest performances but limited to the sensitivity for cirrhosis, which was lower to the impact observed on biopsy (Supplementary Table 10).

Limitations of the study

Biopsy estimates in healthy volunteers

The results of biopsy in volunteers were directly estimated in only 4 subjects with large, normal liver biopsies, the specificity being assessed using 626 generated virtual biopsies. This method is imperfect. However, the observed false-positive rates were compatible with other assessments using virtual biopsies, or surgical samples [7]. The distribution of area of fibrosis was similar to that of Bedossa *et al.* (Supplementary Table 1) [7]. Furthermore the change for another more sensitive cutoff for cirrhosis (Supplementary Tables 8 and 9) and the exclusion of all false-positive cases of biopsies (Supplementary Tables 6 and 7) did not impair the model. The model was constructed with a median of biopsy around 20 mm and if the length had been around 40 mm the expected performance of biopsy would have been better but less realistic [25].

Other test performances

The present study compared the accuracy of tests, which is considered only one part of the performance. The failure rates and

reliability were not assessed as well as the other features that could be provided by each test. For liver biopsy, pathologists recommend lengths of at least 20–25 mm [7], which could correspond to a reliability rate of 50% according to the length distribution in large cohorts [25]. For LSM using Fibroscan, the failure rate is 3.8% and the reliability rate 15.8% [9]. For FibroTest, the failure rate is 0% and reliability rate is 98% [28].

Biopsy has an obvious advantage by providing activity grade, steatosis grade and features of other liver diseases. FibroTest assessment includes ActiTest, validated for activity grade diagnosis [29]; SteatoTest, which assesses steatosis grade, can also be associated with FibroTest but has been less validated [30,31].

Variability factors not analyzed

We did not directly analyze the impact of factors from individual data, such as histological steatosis and activity, metabolic factors, age, gender, ethnicity or operator effects that could be related to diagnostic performance [9], and the pathologist variability [24]. As for LSM, the inter-observer variability is a pragmatic weakness of biopsy in comparison with serum biomarkers.

How can the comparisons between liver fibrosis indicators be improved?

First, clear guidelines must be provided defining the reliability criteria of each indicator. For FibroTest, pre-analytical and analytical recommendations must be applied [28]. Other studies have previously demonstrated for LSM that few changes in the precautions of use had a direct impact on its reliability rate or on its risk of false-positives or negatives [8,9]. Publications not applying the precautions of use concerning IQR/LSM and success rate made hazardous conclusions, such as the suggestion that five valid shots could be sufficient for cirrhosis diagnosis [14]. For liver biopsy, it would be wise to consider the results of specimens shorter than 20 mm reliable only after checking the concordance with the reliable results of a validated biomarker.

Second, the intra-indicator variability should be reduced. For FibroTest, the improvement of analytical calibration should reduce the inter-laboratory variability [4,26]. For LSM [8,9] and biopsy [24], the major concern is the operator variability, even if the results are reliable. New methodology such as the concordance rate between LSM and FibroTest can identify observers with too high variability [8]. This method could also be applied to pathologists.

Third, these results must be confirmed by independent groups. However, in the present study all the included cohorts of patients were independent of the FibroTest inventor.

Conclusions

In a model without gold-standard, the high specificity (>0.85) of FibroTest and LSM was confirmed for the diagnosis of both advanced fibrosis and cirrhosis. However, from the analysis of the tests that impaired the fit of the model, more studies should be performed to identify the causes of the high discordances rates between biopsy and LSM, including their intra- and inter-observers' variability.

If the accuracy paradigm cannot convince the users in this field, it is possible to replace it by a new one: the concept of the validation of medical tests [1]. The present results were consistent with the recent prognostic validation of fibrosis biomarkers. In patients with chronic hepatitis C [32,33] as well as in patients with chronic hepatitis B [34] and alcoholic liver disease [35], the prognostic value of FibroTest was at least similar to that of biopsy.

The present results confirm that balanced discussions are needed when discordances are observed between estimates of fibrosis. Biopsy, even of 20 mm, is no more the reference. This model confirms the first guidelines and reimbursement by French health authorities recommending either FibroTest or LSM as first line fibrosis estimates in adult patients with uncomplicated chronic hepatitis C [36]. Finally to move forward such models without gold standard should permit also to better estimate the forthcoming new test performances.

Authors' involvement

T.P. study concept and design, analysis and interpretation of data, drafting; statistical analysis; study supervision. V.dL., J.P.Z., C.S., J.V., J.F., A.T., G.L.N., J.C.V., V.R., F.C.: acquisition of data; V.R., M.M.: critical revision of the manuscript.

Conflict of interest

TP is the inventor of FibroTest and has a capital interest in Biopredictive the company marketing the test. The patents belong to Assistance Publique Hôpitaux de Paris, a public organization. M.M. is a full time employee of Biopredictive and participated in the critical analysis of the manuscript. Biopredictive has no role in the study design, in the collection, analysis, and interpretation of data.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhep.2011.08.007.

References

[1] Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11:1–51.
[2] Castera L, Pinzani M. Non-invasive assessment of liver fibrosis: are we ready? Lancet 2010;375:1419–1420.
[3] Poynard T. First-line assessment of patients with chronic liver disease with non-invasive techniques and without recourse to liver biopsy. J Hepatol 2011;54:586–587.
[4] Halfon P, Munteanu M, Poynard T. FibroTest–ActiTest as a non-invasive marker of liver fibrosis. Gastroenterol Clin Biol 2008;32:22–38.
[5] Friedrich-Rust M, Ong MF, Martens S, Sarrazin C, Bojunga J, Zeuzem S, et al. Performance of transient elastography for the staging of liver fibrosis: a meta-analysis. Gastroenterology 2008;134:960–974.
[6] Castera L, Denis J, Babany G, Roudot-Thoraval F. Evolving practices of non-invasive markers of liver fibrosis in patients with chronic hepatitis C in France: time for new guidelines? J Hepatol 2007;46:528–529.
[7] Bedossa P, Dargère D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. Hepatology 2003;38:1449–1457.
[8] Poynard T, Ingiliz P, Elkrief L, Munteanu M, Lebray P, Morra R, et al. Concordance in a world without a gold standard: a new non-invasive methodology for improving accuracy of fibrosis markers. PLoS One 2008;3:e3857.
[9] Castéra L, Foucher J, Bernard PH, Carvalho F, Allaix D, Merrouche W, et al. Pitfalls of liver stiffness measurement: a 5-year prospective study of 13,369 examinations. Hepatology 2010;51:828–835.
[10] Zarski, Sturm N, Guechot J, Paris A, Zafrani ES, Asselah T, et al. Comparison of 10 blood tests and transient elastography for liver fibrosis in chronic hepatitis C: the ANRS HCEP-23 study. J Hepatol 2012;56:55–62.

Research Article

[11] Sporea I, Sirli R, Deleanu A, Tudora A, Popescu A, Curescu M, et al. Liver stiffness measurements in patients with HBV vs. HCV chronic hepatitis: a comparative study. World J Gastroenterol 2010;16:4832–4837.

[12] Poynard T, Lebray P, Ingiliz P, Varaut A, Varsat B, Ngo Y, et al. Prevalence of liver fibrosis and risk factors in a general population using non-invasive biomarkers (FibroTest). BMC Gastroenterol 2010;10:40.

[13] Poynard T, Lenaour G, Charlotte F, Munteanu M, Vaillant JC, Ngo Y, et al. Comparison of Fibrotest–Fibrosure and liver stiffness measurement, using fibrosis area and large surgical samples. Hepatology 2010;52:282A, [abstract].

[14] Kettaneh A, Marcellin P, Douvin C, Poupon R, Ziol M, Beaugrand M, et al. Features associated with success rate and performance of FibroScan measurements for the diagnosis of cirrhosis in HCV patients: a prospective study of 935 patients. J Hepatol 2007;46:628–634.

[15] Roulot D, Czernichow S, Le Clésiau H, Costes JL, Vergnaud AC, Beaugrand M, et al. Liver stiffness values in apparently healthy subjects: influence of gender and metabolic syndrome. J Hepatol 2008;48:606–613.

[16] Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology 1996;24:289–293.

[17] Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 1996;52:797–810.

[18] Hadgu A, Qu Y. A biomedical application of latent class models with random effects. Appl Stat 1998;47:603–616.

[19] Vermunt JK, Magidson J. Technical guide to latent gold 4.5. Belmont, Massachusetts, USA: Statistical Innovations Inc.; 2007.

[20] Zhou X, Obuchowski N, McClish D. Statistical methods in diagnostic medicine. John Wiley and Sons; 2002.

[21] Lambert J, Halfon P, Penaranda G, Bedossa P, Cacoub P, Carrat F. How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. Clin Chem 2008;54:1372–1378.

[22] Hintze JL. NCSS 2007 user guide. Number cruncher statistical systems software. Kaysville, Utah: NCSS; 2007.

[23] Mehta SH, Lau B, Afdhal NH, Thomas DL. Exceeding the limits of liver histology markers. J Hepatol 2009;50:36–41.

[24] Rousselet MC, Michalak S, Dupré F, Croué A, Bedossa P, Saint-André JP, et al. Hepatitis network 49. Sources of variability in histological scoring of chronic viral hepatitis. Hepatology 2005;41:257–264.

[25] Poynard T, Halfon P, Castera L, Charlotte F, Bail BL, Munteanu M, et al. Variability of the area under the receiver operating characteristic curves in the diagnostic evaluation of liver fibrosis markers: impact of biopsy length and fragmentation. Aliment Pharmacol Ther 2007;25:733–739.

[26] Poynard T, Munteanu M, Deckmyn O, Ngo Y, Drane F, Messous D, et al. Applicability and precautions of use of liver injury biomarker FibroTest. A reappraisal at 7 years of age. BMC Gastroenterol 2011;11:39.

[27] Coco B, Oliveri F, Maina AM, Ciccorossi P, Sacco R, Colombatto P, et al. Transient elastography: a new surrogate marker of liver fibrosis influenced by major changes of transaminases. J Viral Hepat 2007;14:360–369.

[28] Poynard T, Ngo Y, Munteanu M, Thabut D, Massard J, Moussalli J, et al. Biomarkers of liver injury for hepatitis clinical trials: a meta-analysis of longitudinal studies. Antivir Ther 2010;15:617–631.

[29] Poynard T, Munteanu M, Ngo Y, Castera L, Halfon P, Ratziu V, et al. ActiTest accuracy for the assessment of histological activity grades in patients with chronic hepatitis C, an overview using Obuchowski measure. Gastroenterol Clin Biol 2010;34:388–396.

[30] Poynard T, Ratziu V, Naveau S, Thabut D, Charlotte F, Messous D, et al. The diagnostic value of biomarkers (SteatoTest) for the prediction of liver steatosis. Comp Hepatol 2005;4:10.

[31] Poynard T, Munteanu M, Colombo M, Bruix J, Schiff E, Terg R, et al. FibroTest is an independent predictor of virologic response in chronic hepatitis C patients retreated with pegylated interferon alfa-2b and ribavirin in the EPIC(3) program. J Hepatol 2011;54:227–235.

[32] Ngo Y, Munteanu M, Messous D, Charlotte F, Imbert-Bismut F, Thabut D, et al. A prospective analysis of the prognostic value of biomarkers (FibroTest) in patients with chronic hepatitis C. Clin Chem 2006;52:1887–1896.

[33] Vergniol J, Foucher J, Terrebonne E, Bernard PH, Le Bail B, Merrouche W, et al. Non-invasive tests for fibrosis and liver stiffness predict 5-year outcomes of patients with chronic hepatitis C. Gastroenterology 2011;40:1970–1979.

[34] Ngo Y, Benhamou Y, Thibault V, Ingiliz P, Munteanu M, Lebray P, et al. An accurate definition of the status of inactive hepatitis B virus carrier by a combination of biomarkers (FibroTest–ActiTest) and viral load. PLoS One 2008;2:e2573.

[35] Naveau S, Gaudé G, Asnacios A, Agostini H, Abella A, Barri-Ova N, et al. Diagnostic and prognostic values of noninvasive biomarkers of fibrosis in patients with alcoholic liver disease. Hepatology 2009;49:97–105.

[36] La Haute Autorité de Santé (HAS) in France – The HAS recommendations for the management of the chronic hepatitis C using non-invasive biomarkers. Available from: <http://www.has-sante.fr/portail/display.jsp?id=c_476486> [accessed August 2007].