Systematic review with meta-analysis: direct comparisons of biomarkers for the diagnosis of fibrosis in chronic hepatitis C and B

M Houot, Y Ngo, M Munteanu, S Marque, T Poynard

Systematic review with meta-analysis: direct comparisons of biomarkers for the diagnosis of fibrosis in chronic hepatitis C and B

M. Houot*, Y. Ngo*, M. Munteanu*, S. Marque[†] & T. Poynard[‡,§]

*BioPredictive, Paris, France.
[†]Capionis, Paris, France.
[‡]Hepatology Department, Assistance Publique-Hôpitaux de Paris, Hôpital Pitié-Salpêtrière, Paris, France.
[§]INSERM & Université Pierre et Marie Curie - Univ Paris 06, UMR_S 938, Centre de Recherche Saint-Antoine & Institute of Cardiometabolism and Nutrition (ICAN), Paris, France.

Correspondence to:
Dr T. Poynard, 47-83 Boulevard de l'Hôpital, 75651 Paris Cedex 13, France.
E-mail: tpoynard@teaser.fr

SUMMARY

Background
Blood tests and transient elastography (TE), proposed as alternatives to biopsy for identifying advanced fibrosis (METAVIR-stage-F2 or greater) or cirrhosis, have never been compared using an intention to diagnose approach, with direct comparisons only, and Bayesian approach.

Aim
To permit more appropriate comparisons.

Methods
From articles overview (2002–2014), we selected studies that directly compared the diagnostic accuracy of FibroTest, aspartate aminotransferase–platelet ratio index (APRI), FIB4 or TE, with biopsy as a reference, in patients with chronic hepatitis C (CHC) or B (CHB). Investigators abstracted and checked study details and quality by using pre-defined criteria. Bayesian method in intention to diagnose was the primary outcome.

Results
Of 1321 articles identified, 71 studies including 77 groups according to aetiology (All-CB) were eligible: 37 Only-C, 28 Only-B and 12 Mixed-C-B. There were 185 direct comparisons between the area under the ROC curves (AUROCs), 99 for the diagnosis of advanced fibrosis and 86 for cirrhosis. In All-CB, Bayesian analyses revealed significant AUROCs differences in identifying advanced fibrosis in favour of FibroTest vs. TE [credibility interval: 0.06(0.02–0.09)], FibroTest vs. APRI [0.05 (0.03–0.07)] and for identifying cirrhosis TE vs. APRI [0.07 (0.02–0.13)] and FIB4 vs. APRI [0.04(0.02–0.05)]. No differences were observed between TE and FibroTest, for identifying cirrhosis in All-CB, and in sub-groups (Only-C, Only-B, Mixed-CB) for both cirrhosis and fibrosis.

Conclusions
In CHC and CHB, APRI had lower performances than FIB-4, TE and FibroTest. TE had lower performance than FibroTest for identifying advanced fibrosis in All-CB, without significant difference for identifying cirrhosis in all groups.

Aliment Pharmacol Ther

INTRODUCTION

The global burden of cirrhosis and of primary liver cancer, its main complication, increased by 30% from 1990 to 2010 in US, and Europe reaching the 8th and the 30th rank of leading diseases and injuries contributing to years of life lost (YLLs) due to pre-mature mortality.[1] This increase in YLLs may not only be related to the spread of hepatitis C[2,3] but also to the misdiagnosis of liver fibrosis progression in the four most frequent chronic liver diseases: viral hepatitis chronic hepatitis B (CHB),[4] chronic hepatitis C (CHC), alcoholic liver disease (ALD) and non-alcoholic fatty liver disease (NAFLD).[5]

To be more effective we suggested a 'fibrosis-oriented' direct screening strategy and not a 'liver-disease-oriented' indirect strategy.[6] To improve the access to fibrosis staging, and replace liver function test and biopsy, many non-invasive fibrosis biomarkers have been assessed since 1991,[7] including systematic reviews since 2002.[8] Most of these reviews focused on a specific liver disease such as CHC,[9,10] CHB,[11–13] NAFLD,[14,15] and ALD.[16]

These reviews, including those from our team, have had several methodological limitations that could result in misleading conclusions when tests were compared.[17] The main limitations of previous systematic overviews were the absence of the following five methods: direct comparisons between tests[9]; intention to diagnose (ITD) analysis:[18,19] Bayesian comparisons[20]; adjustment on spectrum effect[18]; individual data meta-analysis[17] and systematic review of prognostic performances.[17,21]

Direct comparisons have been reviewed only once, in CHC, and showed slightly better performance of FibroTest vs. APRI.[9] For CHB, the last overviews of blood tests did not focus on direct comparisons (Table S1) .[11–13]

Bayesian comparisons are based on less assumption about data distribution than standard frequentist methods.[20,22] Frequentist conclusion is sensitive to multiple comparisons and multiple looks. Bayesian method provides a 'credible interval' which is more robust and can be used in lieu of the *P*-value of frequentist tests as a measure of the evidence strength. Contrarily to the frequentist 'confidence interval', a 95% 'credible interval' is one that has a 95% chance of containing the population risk ratio.[23] Using this approach would improve standard methods of meta-analysis, which give an incomplete picture of the relative benefits of diagnostic tests.

Due to these limitations, we aimed to perform an improved systematic review, focusing on direct comparisons between the four most frequently used, discussed and recommended tests,[3,4,24,25] at least for CHC since 2001: FibroTest,[26] aspartate aminotransferase (AST)–platelet (PLT) ratio index (APRI),[27] FIB4 index,[28] and transient elastography (TE) by Fibroscan.[29] We focus also on the most investigated chronic liver diseases, CHC and CHB, as too few direct comparisons have been performed so far in ALD and NAFLD.[14–16, 25] Because of the limited number of direct comparisons between tests as well as their universal use for both CHC and CHB, the primary population was the meta-analysis including studies in CHC and CHB.

METHODS

The protocol was developed using a standardised process available in Data S1. We focused on direct comparisons of the AUROCs between FibroTest, APRI, FIB4 and TE, for the diagnosis of advanced fibrosis (METAVIR F2F3F4) and cirrhosis (F4), in ITD. AUROCs' differences were assessed by four methods. These methods were performed in four groups of patients according to aetiology: studies including only CHC (Only-C), CHB (Only-B), Mixed-CB and the pooling of all (All-CB). The primary method was the Bayesian method in ITD, and the primary population All-CB due to the limited number of studies.

Blood tests combined several components: FibroTest combined in patented algorithm alpha-2 macroglobulin, apolipoprotein A1, haptoglobin, total bilirubin and gamma-glutamyl transpeptidase (adjusted according to age and gender), APRI is the ratio between AST (in upper limit of the normal, ULN units) and platelet counts; FIB4 formula is the following: (age [year] $\times$ AST [U/L])/(PLT [10^9/L] $\times$ ALT [U/L]).

Study selection

The criteria for inclusion of studies were the following: adult patients (mean of age 18 years or older) explored for Only-C, Only-B or Mixed-CB; in Mixed-CB, when other diseases than CHC or CHB (Mixed-CBO) were mixed, Others must represented less than 49% of CHC and CHB total; liver biopsy was taken as a reference; fibrosis must have been staged using METAVIR scoring system or any equivalent score, expressed as 'clinically significant (advanced) fibrosis', defined as bridging fibrosis (stage F2–F3–F4), and cirrhosis (stage F4); at least two tests simultaneously performed of the following four tests: FibroTest, APRI, FIB4 and TE using Fibroscan with M-probe (TE) with performances expressed using AUROC. Discrepancies were resolved through consensus.

Data sources and searches

The objective of our search strategy was to identify studies with the defined inclusion criteria. We searched MEDLINE (January 2002 to February 2014), and Scopus. The literature was reviewed using PubMed with the following keywords: fibrosis, cirrhosis, liver, APRI, FibroTest, FibroSure, FIB-4, liver stiffness measurement, TE and FibroScan.

Study quality rating

One investigator (YN) abstracted details about study design, patient population, setting, interventions, analysis, follow-up and results, and a second investigator reviewed data for accuracy (MH). Two investigators (TP, YN) independently applied pre-defined criteria of Chou *et al.* to assess the quality of each study as good, fair or poor.[9] Discrepancies were resolved through consensus.

Statistical methods

We assess the range of differences using different methods (M1–M4) for the various comparisons, noting those instances in which the differences were generally larger across analytic methods or when the median estimate was really different from the estimates based on pooling. The endpoints were the AUROCs for advanced fibrosis and for cirrhosis. Four methods were used to assess the differences in AUROCs between the four tests.

The primary end point was the Bayesian method (M4) never used before, nonfrequentist and assessing the pooled direct differences. The descriptive direct method (M1) with median and range was the method of reference recommended by Chou *et al.*,[9] which permitted to approve the use of blood tests such as FibroTest for DAA payments in US. Two other methods were used, as used in several overviews, one frequentist indirect (M2) assessing difference between pooled AUROC difference, and one frequentist direct (M3) assessing the pooled direct difference between AUROCs.

M4, a Bayesian approach, enabled the limitations of the frequentist methods to be overcome, especially providing credibility interval instead of confidence interval (CI). As the covariance between two tests was not given in studies, and there was no independent population in direct comparison that would allow the covariance between two tests to be set to 0, the 'frequentist approach' might be not relevant. Therefore, we developed a Bayesian model that allowed us to relax some of these limitations (Data S1). A random effects model of meta-analysis within a Bayesian framework was used to assess the pooled AUROC differences in each pair of tests. M4

used 'R2jags' package in R software 3.0.2 (https://www.r-project.org/). We used 10 000 iterations, a burn-in of 1000, a thin of 2 and 3 Markov chain Monte Carlo. We checked graphically the convergence of the Monte Carlo samples by Gelman–Rubin convergence diagnostic, autocorrelation and successive convergence for each parameter. The difference was considered as 'significant', when the 95% credibility interval did not include the 0 value.

M2, a frequentist indirect approach, assessed the difference between pooled AUROCs. A meta-analysis with random effects was performed for each diagnostic test, stratified by aetiology, endpoint and comparisons. As there were only two tests by head-to-head comparison, two meta-analyses were performed (one by test) and the difference of the two pooled AUROCs was assessed. M3, a frequentist direct approach, assessed the pooled difference between AUROCs. We used standard direct meta-analysis (inverse variance-weighted method) with random effects, which assessed the pooled difference in AUROCs between two tests as stratified by aetiology and endpoint.[30] As the covariance between the two AUROCs was unknown, it was set to 0 although this hypothesis was unlikely. For M2 and M3, the pooled differences were considered 'significant', when the 95% CI did not include 0 value. For frequentist methods, no corrections were used to prevent the risk of multiple testing as the significance of the primary outcome was tested with the credibility interval, which reduces the risk of false-positive conclusions.[21, 22]

Intention to diagnose. Comparison with per-protocol analysis, which excluded tests' failure or unreliable tests, ITD only impacted TE and FibroTest, which have rules for excluding failure and nonreliable results.[18] Indeed, APRI and FIB4 had several risks of nonreliable results due mainly to the impact of necro-inflammatory activity on transaminases ALT or AST, but without recognised standard criteria for reliability or expression of ULN.[31] Two ITD methods were used. The primary ITD method (ITD1) used was the intuitive method suggested by Poynard *et al.*[18] The hypothesis was that among the non-applicable tests, the subject had 50% probability to be correctly classified. The other method (ITD2) used three scenarios to check the significance, was detailed in Data S1.

Sensitivity analyses. A per-protocol analysis, and ITD2 were performed with all studies. The following sensitivity analyses were performed after successive exclusion of: studies with poor quality; studies including special populations (HIV coinfection, transplantation, haemodialysed,

dialysis, associated mixed cryoglobulinemia vasculitis, beta-thalassaemia); studies in CHC with persistent normal ALT; Mixed-CB; Mixed-CB without those which included other disease than CHC or CHB (Mixed-CBO); studies with less than 100 patients, and non-independent studies.

Spectrum effect. The spectrum effect on AUROC variability was assessed, by viewing the association between the AUROCs and fibrosis spectrum (DANA indices), as previously described.[32, 33]

RESULTS

Selection of studies

From 2001 to 2014, of the 1,321 biomarkers' articles identified, 71 studies were found to be eligible (Data S2 and Tables S2, S3), 17 (24%) were from France and 10 (14%) from China; 66 (93%) studies were independent from the authors of this overview, three were performed by the authors and two were mixed (inclusion of patients from authors' team but not involved in the data analyses). Of the 77 groups of patients (this total was named 'All-CB', as one study could included between one to three groups of aetiology such as one Only-C, one Only-B or one Mixed-CB), analysed in the 71 studies, 37 included Only-C, 28 Only-B and 12 Mixed-CB. There were 185 direct comparisons between the four tests' AUROCs, 99 for the diagnosis of advanced fibrosis (12 725 patients) and 86 for cirrhosis (10 929 patients). Of the 185 comparisons, 28 involved FibroTest vs. TE, 35 FibroTest vs. APRI, 11 FibroTest vs. FIB4, 35 TE vs. APRI, 13 TE vs. FIB4 and 63 APRI vs. FIB4 (Figure 1).

Median applicability rates were 0.870 for TE and 0.965 for FibroTest. The ULN for AST was described in 23 of 105 (22%) comparisons including APRI. The overall ULN range was 31–75 IU/mL. No description of the control groups used for ULN definition was given.[31] The methodological quality of studies was good in four (6%), fair in 53 (74%) and poor in 14 (20%).

Comparisons of test performances

Whatever the aetiology, there was a large 'intra-test' variability in the AUROC medians, with overlaps between their ranges both for blood tests and TE, particularly when the number of studies was lower than five, which underlined the interest of direct comparisons (Table S4, Data S3).

Primary outcome using Bayesian methods, all-CB. The primary outcome, median of differences between AUR-

OCS using Bayesian methods in all-CB studies was significant according to credibility interval for identifying advanced fibrosis in two comparisons in favour of FibroTest vs. TE [0.06 (0.02–0.09)], and FibroTest vs. APRI [0.05 (0.03–0.07)] (Table 1, Figure 2a). For identifying cirrhosis, significant differences were observed in two comparisons: TE vs. APRI [0.07 (0.02–0.13)] and FIB4 vs. APRI [0.04 (0.02–0.05)]; no significant difference was observed for the remaining comparisons (Table 1, Figure 3a).

Nonprimary outcomes using Bayesian methods, subgroups (Data S3). In Only-C (Table 2), three significant differences were observed; FibroTest vs. APRI [advanced fibrosis: 0.05 (0.01–0.08)], TE vs. APRI [cirrhosis: 0.07 (0.01–0.13)] and FIB4 vs. APRI [cirrhosis: 0.03 (0.01 to −0.05)]. In Only-B (Table 3) and Mixed-CB (Table 4), no significant differences were observed.

Impact of different ITD analyses among Bayesian comparisons (Table S5). Only comparisons including TE were modified by ITD2. In All-CB, the same credible interval differences were observed in favour of FT vs. TE, with higher difference from 0.17(0.08–0.26) to 0.22 (0.07–0.37) for advanced fibrosis; for cirrhosis, difference were significant from 0.14(0.03–0.23) to 0.22(0.07–0.37). The significant difference (0.07) observed between TE and APRI for cirrhosis disappeared.

Impact of per-protocol method among Bayesian analyses (Table S6). In All-CB, the significant differences between FibroTest and TE for advanced fibrosis disappeared in per-protocol, and appeared for cirrhosis in favour of TE: 0.07 (0.05–0.10).

Sensitivity analyses for Bayesian analysis (Tables S7 and S8). Significant differences in favour of TE vs. APRI observed for cirrhosis in All-CB, and in Only-C, were no more significant after exclusion of the poor quality studies, studies with less than 100 patients, and the studies including special populations.

The difference in favour of FT vs. TE increased for advanced fibrosis from 0.06 (0.02–0.09) to 0.08 (0.05–0.11), when studies with less than 100 patients were excluded. After exclusion of the 12 Mixed-CB, difference in favour of FT vs. TE was no more significant for advanced fibrosis, in All-CB but with same difference [from 0.06(0.02–0.09) to 0.06(0.00–0.11)]. The exclusion of the only study performed in Only-C with persistent normal ALT did not change results.
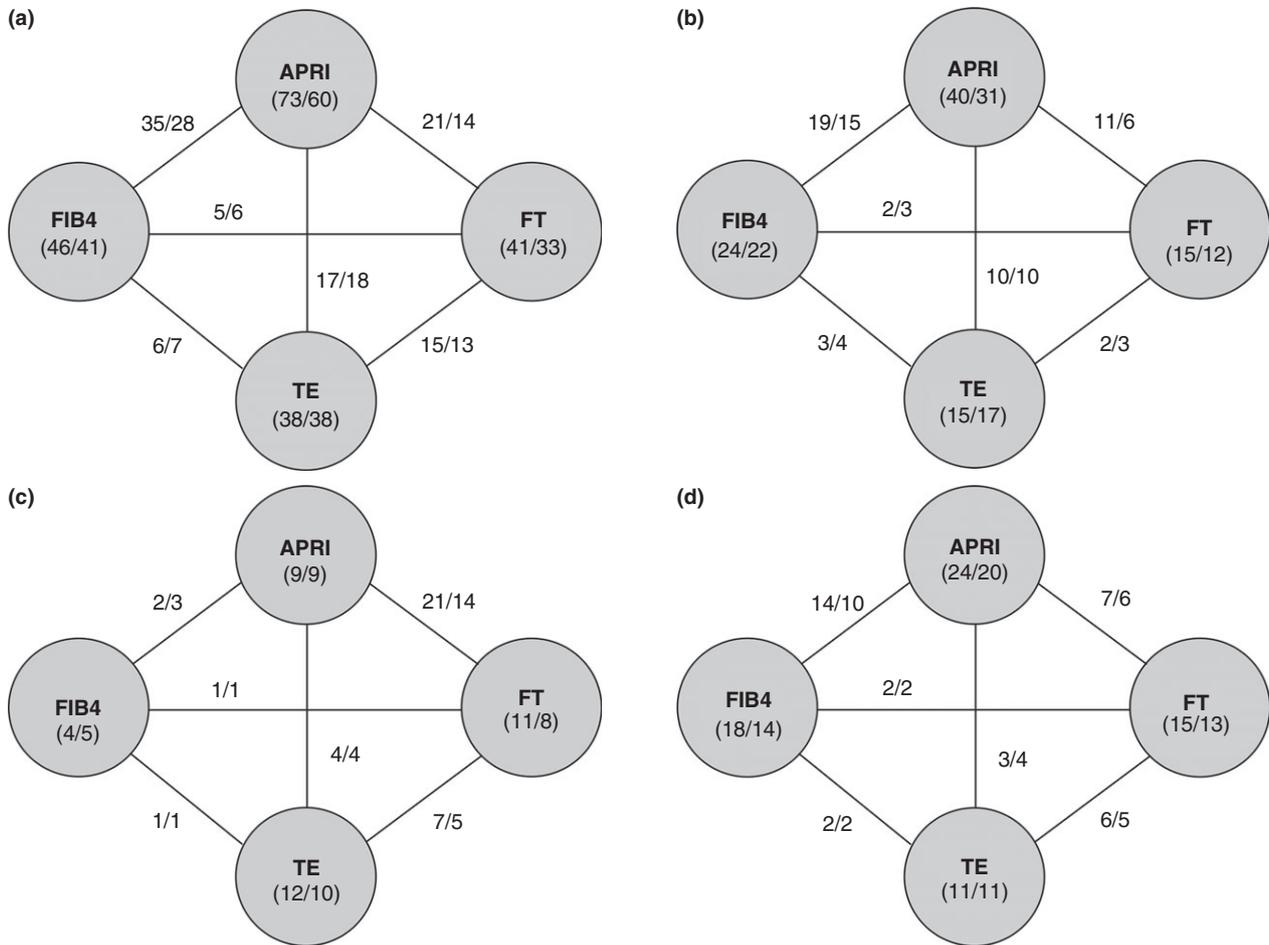
Figure 1 | Direct comparisons between APRI, FIB4, FibroTest and TE. (a) Direct comparisons of biomarkers performance in All-CB patients. Of the 185 comparisons, 28 (15/13) (15 Fibrosis F2F3F4/13 cirrhosis F4) involved FibroTest vs. TE, 35 (21/14) FibroTest vs. APRI, 11 (5/6) FibroTest vs. FIB4, 35 (21/14) TE vs. APRI, 13 (6/7) TE vs. FIB4 and 63 (35/28) APRI vs. FIB4. (b) Direct comparisons of biomarkers performance in Only-C patients. Of the 88 comparisons, 5 (2/3) (2 F2F3F4/ 3 F4) involved FibroTest vs. TE, 17 (11/6) FibroTest vs. APRI, 5 (2/3) FibroTest vs. FIB4, 20 (10/10) TE vs. APRI, 7 (3/4) TE vs. FIB4 and 34 (19/15) APRI vs. FIB4. (c) Direct comparisons of biomarkers performance in Mixed-CB patients. Of the 34 comparisons, 12 (7/5) involved FibroTest vs. TE, 5 (3/2) FibroTest vs. APRI, 2 (1/1) FibroTest vs. FIB4, 8 (4/4) TE vs. APRI, 4 (2/2) TE vs. FIB4 and 5 (2/3) APRI vs. FIB4. (d) Direct comparisons of biomarkers performance in Only-B patients. Of the 63 comparisons, 11 (6/5) involved FibroTest vs. TE, 13 (7/6) FibroTest vs. APRI, 4 (2/2) FibroTest vs. FIB4, 7 (3/4) TE vs. APRI, 4 (2/2) TE vs. FIB4, and 24 (14/10) APRI vs. FIB4.

When the five studies with FibroTest non-independent authors were excluded, for advanced fibrosis, difference between FibroTest and APRI was reduced in Only-C and no more significant, from 0.05(0.01–0.08) to 0.03 (−0.01–0.07). However in All-CB, significant differences persisted in favour of FibroTest vs. TE and vs. APRI, for advanced fibrosis.

When the three studies with mixed-CBO patients were excluded, differences remained identical (Table S7).

Frequentist methods. In all-CB, the following significant differences were also observed by frequentist methods in favour of FibroTest vs. TE for advanced fibrosis by M3 [0.07 (0.05–0.09)], FibroTest vs. APRI by M2 [0.06 (0.02–0.09)] and M3 [0.05 (0.03–0.07)]. No-significant differences were observed for the remaining four comparisons. For cirrhosis, the following five significant differences were also observed by frequentist methods in favour of FibroTest vs. APRI by M2 [0.07 (0.01–0.13)], FibroTest vs. FIB4 by M3 [0.06 (0.01–0.10)], TE vs. APRI by M2 [0.08 (0.01–0.14)] and M3 [0.07 (0.02–0.13)], TE vs. FIB4 by M2 [0.11 (0.03–0.19)] and M3 [0.12 (0.03–0.21)] and FIB4 vs. APRI by M3 [0.04 (0.02–0.05)]. No-significant differences were observed for the remaining comparison (FibroTest vs. TE).

Table 1 | Direct comparisons of biomarkers performance in All-CB patients (*n* = 185)

Fibrosis stage	Test A	Test B	n	M1 Descriptive Median difference (range)	M2 indirect Pooled AUROC difference [CI95%]	M3 Standard Pooled AUROC difference [CI95%]	M4 Bayesian Pooled AUROC difference [CI95%]
F2F3F4	FT	TE	15	0.04 (−0.10–0.14)	0.04 [−0.01 to 0.10]	0.07 [0.05–0.09]*	0.06 [0.02–0.09]†
	FT	APRI	21	0.05 (−0.06 to 0.18)	0.06 [0.02–0.09]*	0.05 [0.03–0.07]*	0.05 [0.03–0.07]†
	FT	FIB4	5	0.03 (−0.01 to 0.08)	0.02 [−0.06 to 0.11]	0.02 [−0.02 to 0.07]	0.02 [−0.05 to 0.10]
	TE	APRI	17	0.01 (−0.14 to 0.25)	0.04 [−0.03 to 0.12]	0.04 [0.00–0.08]	0.04 [0.00–0.09]
	TE	FIB4	6	0.04 (−0.13 to 0.15)	0.03 [−0.15 to 0.21]	0.03 [−0.04 to 0.10]	0.03 [−0.08 to 0.15]
	APRI	FIB4	35	−0.01 (−0.13 to 0.10)	−0.01 [−0.05 to 0.02]	−0.01 [−0.03 to 0.01]	−0.01 [−0.03 to 0.01]
F4	FT	TE	13	0.01 (−0.18 to 0.07)	0.01 [−0.04 to 0.05]	0.01 [−0.03 to 0.04]	0.00 [−0.04 to 0.04]
	FT	APRI	14	0.08 (−0.13 to 0.35)	0.07 [0.01–0.13]*	0.05 [0.00–0.10]	0.05 [0.00–0.11]
	FT	FIB4	6	0.05 (−0.03 to 0.11)	0.06 [−0.01 to 0.12]	0.06 [0.01–0.10]*	0.05 [−0.03 to 0.12]
	TE	APRI	18	0.08 (−0.07 to 0.37)	0.08 [0.01–0.14]*	0.07 [0.03–0.12]*	0.07 [0.02–0.13]†
	TE	FIB4	7	0.13 (−0.01 to 0.25)	0.11 [0.03–0.19]*	0.12 [0.03–0.21]*	0.12 [0.00–0.24]
	APRI	FIB4	28	−0.04 (−0.24 to 0.10)	−0.04 [−0.08 to 0.00]	0.04 [−0.05 to 0.02]*	−0.04 [−0.05 to 0.02]†

FibroTest (FT) had higher performance than TE and APRI for the diagnosis of advanced fibrosis (F2F3F4).

Transient elastography (TE) and FIB4 had higher performance than APRI for the diagnosis of cirrhosis (F4).

* Statistically significant difference for confidence interval (not including 0; *P* < 0.05).

† The primary outcome was the Bayesian pooled AUROC difference, in all-CB patients, considered as statistically significant when credibility interval not included 0.

In subpopulations, Only-C, Only-B and mixed-CB, similar differences or trends were observed among subpopulations in comparisons with the overall comparisons.

In Only-C (Table 2), for four comparisons, significant differences were observed; in favour of FibroTest vs. TE (advanced fibrosis by M3: 0.06), FibroTest vs. APRI (advanced fibrosis by M3: 0.05), TE vs. APRI (cirrhosis by M3: 0.07), and FIB4 vs. APRI (cirrhosis by M3 0.04). No significant differences between tests were observed in eight comparisons.

In Only-B (Table 3), for six comparisons, significant differences were observed; in favour of FibroTest vs. TE (advanced fibrosis by M3: 0.07); FibroTest vs. APRI by M2–M3 [advanced fibrosis 0.07–0.11 and cirrhosis (0.13–0.11)]; TE vs. APRI (advanced fibrosis by M3: 0.13), and TE vs. FIB4 (advanced fibrosis by M2: 0.09 and cirrhosis by M2–M3: 0.18–0.19). No significant differences between tests were observed in six comparisons.

In Mixed-CB (Table 4), for two comparisons, significant differences were observed; in favour of FibroTest vs. TE [advanced fibrosis by M3: 0.05 (0.01–0.09)]; FibroTest vs. APRI (advanced fibrosis by M3: 0.05); and no-significant differences in the remaining 10 comparisons.

Summary of Bayesian vs. frequentist pooled results. Pooled AUROC differences values were similar, especially between the two direct methods Bayesian-M4 and M3; however, the number of differences reaching the significance level was twice lower for Bayesian-M4 (7/48 comparisons), vs. (18/48) M3 (Figures 2 and 3). For All-CB, Bayesian-M4 was discordant with M3 for only two comparisons (out of 12): FibroTest-FIB4 and TE-FIB4 for cirrhosis, with borderline differences and small power (twice less studies) than for the concordant results. For Only-C, there was only one discordance: FibroTest vs. TE for advanced fibrosis (borderline and only two studies). For Only-B, and Mixed-CB all comparisons by Bayesian-M4 were nonsignificant; leading to six and two discordances with M3 respectively.

Spectrum effect (Table S8, Figure S1). There was a significant association between the spectrum of advanced fibrosis estimated by DANA, and the tests' performance assessed by AUROC [185 comparisons, Spearman correlation (SC) = 0.17; *P* = 0.02], the higher associations being observed for FIB4 comparisons: vs. APRI (SC/*P*-value 0.37/0.03), vs. FibroTest (0.75/0.01) and vs. TE (0.57/0.09).

DISCUSSION

This overview of the four most frequently used fibrosis tests in CHC and CHB, is the first to focus only on direct comparisons, performed in ITD, and using a Bayesian approach. These methods permitted statistical comparisons and ranking of these four tests. When compared with previous overviews, this pilot design had major advantages but still limitations.
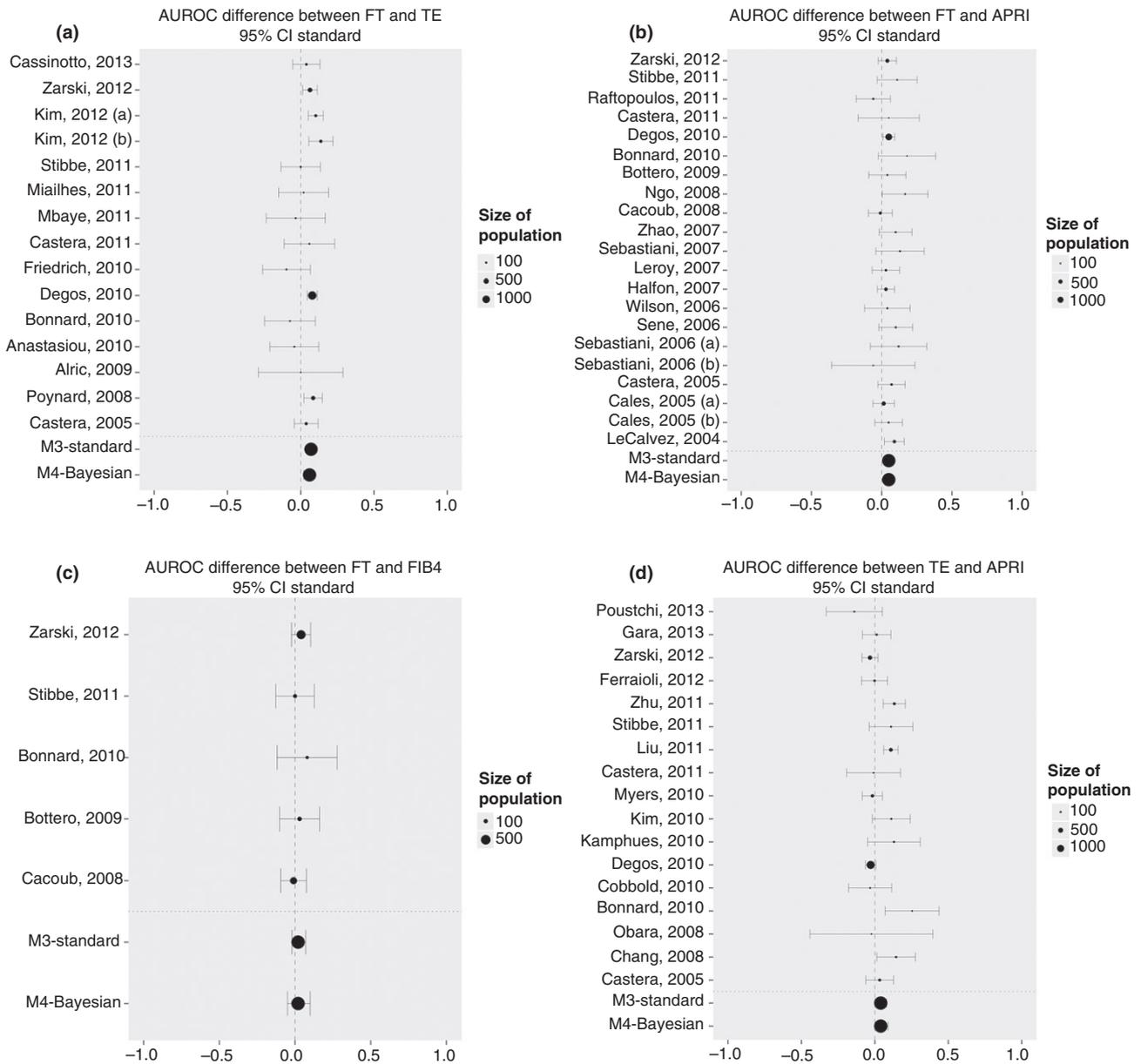
Figure 2 | Meta-analysis of direct comparisons in chronic hepatitis C and B (All-CB) for identifying advanced fibrosis.

Advantages

The first advantage of direct vs. indirect comparison was the reduction in spectrum effects, and of other causes of clinical diversity, as inter-patient variability disappeared. Due to the binary definition of advanced fibrosis[32, 33] and the heterogeneity of cirrhosis stage,[34, 35] the spectrum effect is a major source of misleading interpretation when AUROCs are indirectly compared. The classification of test performance according to category of AUR-OCs (e.g. 0.90–1.0 classified as 'excellent', and less than 0.70 as 'poor') can be misleading when assessed by indirect comparisons.[9] Indeed and as previously described,[32]

we observed that, when the prevalence of each fibrosis stage varied widely as in the FibroTest vs. FIB4 studies, the AUROCs for identifying advanced fibrosis could vary from 0.64 to 0.80 (Table S8, Figure S1). Using direct comparisons, we did not need 'adjusted' AUROCs based on a standardised distribution of fibrosis stages (DANA) as well as Obuchowski measures,[9, 32, 33] as the prevalences of fibrosis stages were the same, each patients being his own control.

The second advantage was to take into account the test applicability. Due to the low applicability of TE (80%) a fair comparison with blood tests, needed ITD.[36]
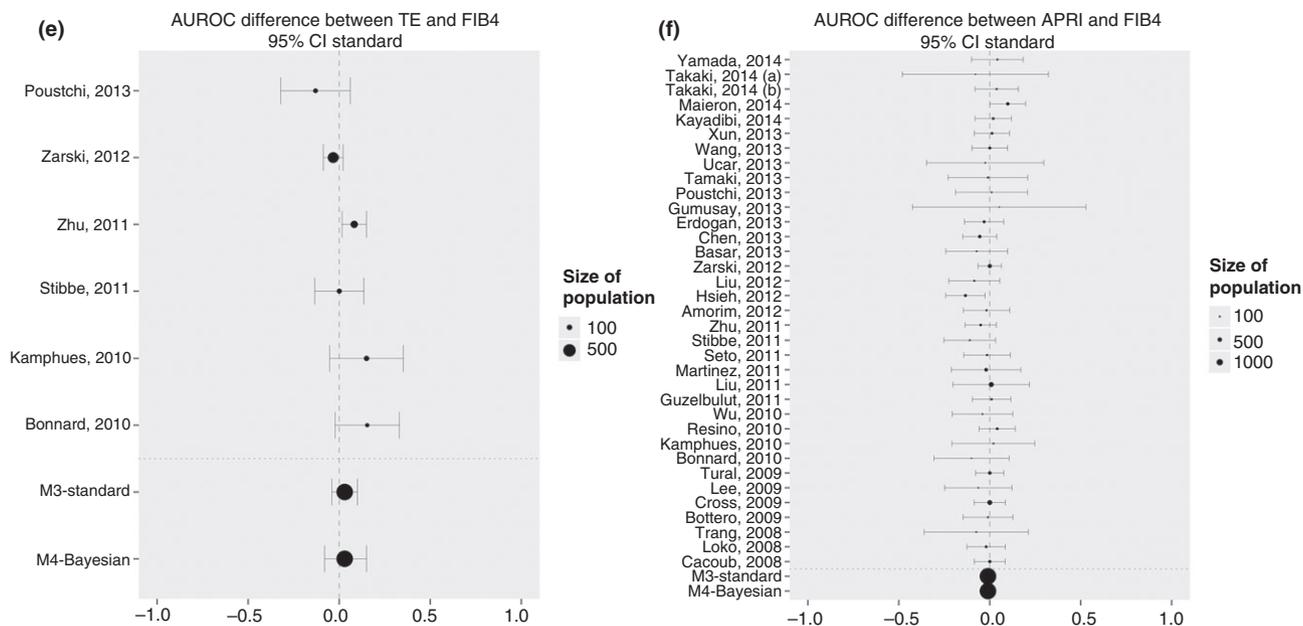
(e) AUROC difference between TE and FIB4
95% CI standard

(f) AUROC difference between APRI and FIB4
95% CI standard

Figure 2 | (Continued).

In the present overview, TE applicability (87%) was in the range observed in CHC and CHB. For the first time TE was directly compared, in ITD, to validated blood tests. We clearly confirmed the absence of significant pooled differences between TE and FibroTest for identifying cirrhosis,[5] contrarily to several isolated studies, reviews or editorials,[36–39] Furthermore for identifying advanced fibrosis, we observed significant differences in favour of FibroTest vs. TE, both the operator effect and activity false positive are other main limitations for TE.[31, 36–41] Despite ITD, TE had higher performances than APRI and FIB4, both for identifying advanced fibrosis or cirrhosis. Pooled analyses 'per-protocol', would have overestimated the performance of TE, whatever the statistical method used. The performances of TE were even worse, when another ITD method was used (Table S5). The applicability of FibroTest was in the range (96.5%) of tertiary centres studies. In large population, the applicability rate of Fibrotest was 99%.[41] Applicability remained to be assessed for APRI and FIB4 as ALT and AST varied according to numerous nonfibrosis injuries, definitions of upper limit of normal and age.[31, 36, 40, 42]

The third advantage was the introduction of the Bayesian method for the first time in an overview of fibrosis tests as performed for CHB treatments.[20] This approach overcame some limitations of the frequentist methods, mainly the risk of false-positive conclusion, and an estimate of a credibility interval more useful for ranking

several tests in lieu of *P*-values of frequentist methods.[22, 23] Bayesian approach also allowed overcoming the lack of information about the unknown relationships (covariance) between tests.

Moreover, informative priors can be used when there is prior knowledge (e.g. expert opinion, or previous studies) leading to have a stronger influence on the posterior distribution and hence on the estimate. According to our works demonstrating the impact of spectrum effect and applicability on AUROCs,[18] as well as the conclusions of Chou *et al.* overview, which was mainly descriptive despite several rankings,[9] (Data S3) we have considered the prior distribution as not informative. The only conclusion for direct comparisons in this overview was a difference in AUROCs favour of FT vs. APRI but without statistical test.

Due to this reduction in variability, for the first time, it was possible to assess the statistical significance of the differences between the AUROCs, observed in direct comparisons between these three blood tests, and vs. TE, with a maximum of power for the primary endpoint in all-CB population. Therefore, it was possible to rank these tests according to their differences in diagnostic performances, contrarily to previous overviews separately performed in CHC or CHB, and using frequentist methods in indirect comparisons.[3, 4, 10, 13, 36]

In a Bayesian framework, comparison between two diagnostic tests, which had never been compared in a publication, can be performed as they had been com-
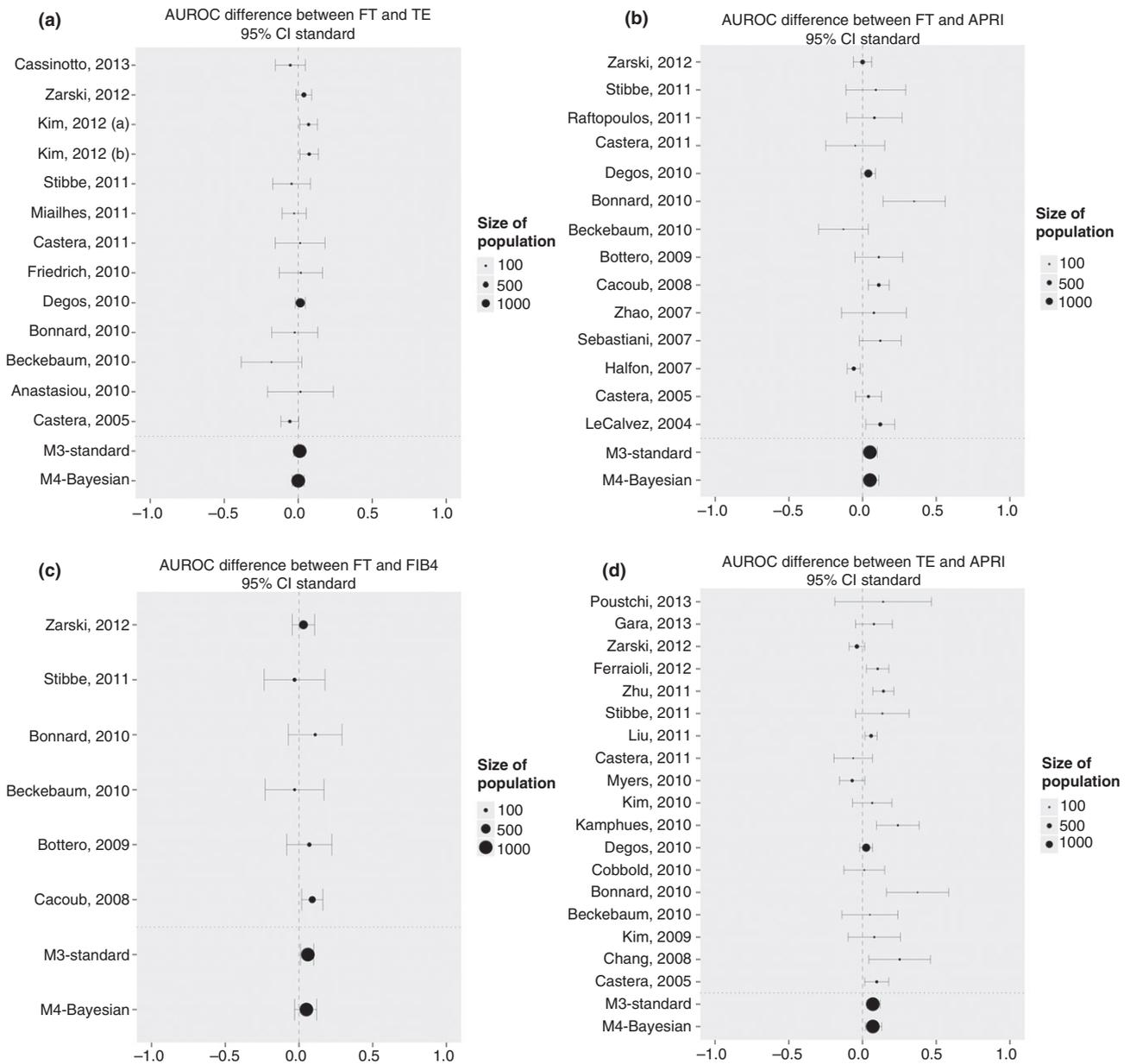
Figure 3 | Meta-analysis of direct comparisons in chronic hepatitis C and B (All-CB) for identifying cirrhosis.

pared to another diagnostic test. This should be performed when more direct comparisons will be available. More than 30 new tests (blood or imaging) for identifying advanced fibrosis in CHC and CHB have become available within the past decade.[4, 9–13] Studies comparing these tests directly have been limited to 2–10 at a time, whereas traditional meta-analytic techniques were limited to comparisons of two tests. This has left clinicians to make their own judgments about the relative performances of tests, for which head-to-head comparisons have not been available.

Our results, using a more appropriate methodology, could challenge the conclusions of recent tests' overviews in viral hepatitis.

In CHC, one extensive overview of blood tests direct comparisons (without ITD), used descriptive methods without statistical tests and concluded 'APRI was associated with a slightly lower AUROC than FibroTest for fibrosis (18 studies; median difference, −0.03; range, −0.10 to 0.07), but there was no difference for cirrhosis (seven studies; median difference, 0.0; range, −0.04 to 0.06)'.[9] Another recent frequentist overview (indirect
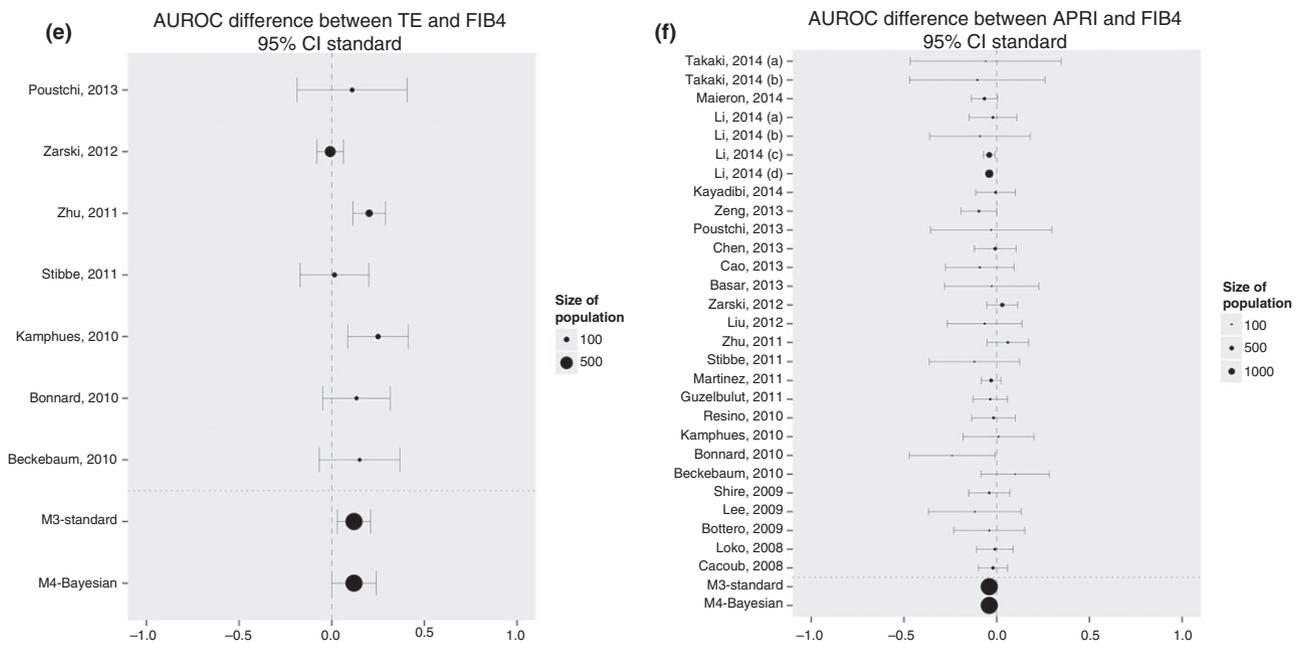
(e)

AUROC difference between TE and FIB4
95% CI standard

(f)

AUROC difference between APRI and FIB4
95% CI standard

Figure 3 | (Continued).

Table 2 | Direct comparisons of biomarkers performance in Only-C patients (*n* = 88)

Fibrosis stage	Test A	Test B	*n*	M1 descriptive Median difference (range)	M2 indirect Pooled AUROC difference [CI$_{95\%}$]	M3 standard Pooled AUROC difference [CI$_{95\%}$]	M4 Bayesian Pooled AUROC difference [CI$_{95\%}$]
F2F3F4	FT	TE	2	0.05 (0.04–0.06)	0.05 [−0.03 to 0.12]	0.06 [0.01–0.10]*	0.05 [−0.90 to 1.00]
	FT	APRI	11	0.04 (−0.06 to 0.12)	0.05 [0.00–0.10]	0.05 [0.02–0.07]*	0.05 [0.01–0.08]†
	FT	FIB4	2	0.02 (−0.01 to 0.04)	0.01 [−0.18 to 0.20]	0.02 [−0.03 to 0.07]	0.02 [0.82–0.90]
	TE	APRI	10	0.01 (−0.14 to 0.13)	0.02 [−0.09 to 0.13]	0.03 [−0.02 to 0.07]	0.03 [−0.04 to 0.09]
	TE	FIB4	3	−0.03 (−0.13 to 0.15)	−0.01 [−0.32 to 0.29]	−0.02 [−0.13 to 0.10]	−0.01 [−0.43 to 0.44]
	APRI	FIB4	19	0.00 (−0.13 to 0.10)	0.00 [−0.04 to 0.05]	0.00 [−0.02 to 0.03]	0.00 [−0.03 to 0.03]
F4	FT	TE	3	−0.06 (−0.18 to 0.04)	−0.03 [−0.11 to 0.06]	−0.04 [−0.14 to 0.06]	−0.05 [−0.53 to 0.36]
	FT	APRI	6	0.02 (−0.13 to 0.12)	0.02 [−0.05 to 0.09]	0.02 [−0.05 to 0.09]	0.02 [−0.10 to 0.13]
	FT	FIB4	3	0.03 (−0.03 to 0.09)	0.05 [−0.04 to 0.14]	0.05 [0.00–0.11]	0.04 [−0.24 to 0.31]
	TE	APRI	10	0.07 (−0.04 to 0.24)	0.07 [0.01–0.13]*	0.07 [0.02–0.11]*	0.07 [0.01–0.13]†
	TE	FIB4	4	0.13 (−0.01 to 0.25)	0.09 [−0.03 to 0.21]	0.11 [−0.03 to 0.25]	0.11 [−0.16 to 0.40]
	APRI	FIB4	15	−0.03 (−0.10 to 0.10)	−0.02 [−0.07 to 0.03]	−0.04 [−0.05 to −0.02]*	−0.03 [−0.05 to −0.01]†

* Statistically significant difference for confidence interval (not including 0; *P* < 0.05).

† Bayesian pooled AUROC difference, considered as statistically significant when credibility interval not included 0.

comparisons without ITD) of fibrosis biomarker concluded, 'APRI and FIB-4, fail to classify a significant proportion of patients who fall into the grey zone of indeterminate values. Proprietary non-invasive tests with the possible exception of FibroTest, are insufficiently validated in independent cohorts. The increasingly used FibroScan does not have validated cutoffs for specific fibrosis stages. Therefore, non-invasive tests need better-quality studies and further validation, particularly for the diagnosis of moderate fibrosis.'[10]

In CHB, a recent frequentist overview of 30 studies (13 with direct comparisons), concluded 'The heterogeneity of APRI for detecting significant fibrosis was affected by median age, and for cirrhosis was affected by aetiology. On the basis of the analysis, we claim that FibroTest has excellent diagnostic accuracy for identification of HBV-related significant fibrosis and cirrhosis. FIB-4 has modest benefits and may be suitable for wider scope implementation.'[13] For CHB, the 2015 WHO report concluded after a frequentist overview of non-invasive tests (indirect com-

Table 3 | Direct comparisons of biomarkers performance in Only-B patients (*n* = 63)

Fibrosis stage	Test A	Test B	*n*	M1 descriptive Median difference (range)	M2 indirect Pooled AUROC difference [CI$_{95\%}$]	M3 Standard Pooled AUROC difference [CI$_{95\%}$]	M4 Bayesian Pooled AUROC difference [CI$_{95\%}$]
F2F3F4	FT	TE	6	0.04 (−0.07 to 0.14)	0.05 [−0.08 to 0.18]	0.07 [0.01–0.13]*	0.06 [−0.05 to 0.14]
	FT	APRI	7	0.10 (−0.06 to 0.18)	0.07 [0.02–0.13]*	0.07 [0.01–0.14]*	0.08 [−0.01 to 0.17]
	FT	FIB4	2	0.06 (0.03–0.08)	0.05 [−0.06 to 0.16]	0.05 [−0.06 to 0.16]	0.05 [−0.86 to 0.99]
	TE	APRI	3	0.13 (−0.01 to 0.25)	0.11 [−0.10 to 0.33]	0.13 [0.02–0.24]*	0.13 [−0.30 to 0.55]
	TE	FIB4	2	0.12 (0.08–0.15)	0.12 [−0.04 to 0.28]	0.09 [0.03 to 0.15]*	0.10 [−0.94 to 1.00]
	APRI	FIB4	14	−0.03 (−0.10 to 0.05)	−0.03 [−0.08 to 0.01]	−0.03 [−0.07 to 0.00]	−0.03 [−0.07 to 0.01]
F4	FT	TE	5	0.01 (−0.03 to 0.07)	0.05 [−0.03 to 0.13]	0.04 [−0.01 to 0.09]	0.03 [−0.06 to 0.11]
	FT	APRI	6	0.10 (−0.05 to 0.35)	0.13 [0.02–0.24]*	0.11 [0.03–0.20]*	0.11 [−0.02 to 0.24]
	FT	FIB4	2	0.09 (0.07–0.11)	0.09 [−0.03 to 0.21]	0.09 [−0.03 to 0.20]	0.09 [−0.92 to 1.00]
	TE	APRI	4	0.11 (−0.06 to 0.37)	0.12 [−0.05 to 0.29]	0.12 [−0.04 to 0.29]	0.13 [−0.21 to 0.48]
	TE	FIB4	2	0.17 (0.13–0.20)	0.18 [0.06 to 0.29]*	0.19 [0.11 to 0.27]*	0.18 [−0.73 to 1.00]
	APRI	FIB4	10	−0.05 (−0.24 to 0.06)	−0.05 [−0.14 to 0.03]	−0.04 [−0.09 to 0.01]	−0.05 [−0.11 to 0.02]

* Statistically significant difference for confidence interval (not including 0; *P* < 0.05).

Table 4 | Direct comparisons of biomarkers performance in Mixed-CB patients (*n* = 34)

Fibrosis Stage	Test A	Test B	*n*	M1 descriptive Median difference (range)	M2 indirect Pooled AUROC difference [CI$_{95\%}$]	M3 standard Pooled AUROC difference [CI$_{95\%}$]	M4 Bayesian Pooled AUROC difference [CI$_{95\%}$]
F2F3F4	FT	TE	7	0.00 (−0.10 to 0.08)	0.03 [−0.02 to 0.09]	0.05 [0.01 to 0.09]*	0.04 [−0.04 to 0.10]
	FT	APRI	3	0.05 (0.01–0.11)	0.04 [−0.02 to 0.11]	0.05 [0.01–0.08]*	0.05 [−0.24 to 0.36]
	FT	FIB4	1	NA	NA	NA	NA
	TE	APRI	4	0.05 (−0.03 to 0.14)	0.04 [−0.08 to 0.17]	0.03 [−0.05 to 0.11]	0.04 [−0.14 to 0.24]
	TE	FIB4	1	NA	NA	NA	NA
	APRI	FIB4	2	−0.09 (−0.11 to −0.06)	−0.09 [−0.22 to 0.03]	−0.09 [−0.21 to 0.02]	−0.09 [−1.00 to 0.92]
F4	FT	TE	5	0.01 (−0.05 to 0.02)	−0.02 [−0.06 to 0.03]	0.00 [−0.03 to 0.04]	−0.01 [−0.10 to 0.07]
	FT	APRI	2	0.06 (0.04 to 0.09	0.04 [−0.01 to 0.09]	0.04 [−0.01 to 0.09]	0.06 [−0.88 to 1.00]
	FT	FIB4	1	NA	NA	NA	NA
	TE	APRI	4	0.08 (−0.07 to 0.25)	0.04 [−0.05 to 0.13]	0.06 [−0.07 to 0.18]	0.06 [−0.20 to 0.36]
	TE	FIB4	1	NA	NA	NA	NA
	APRI	FIB4	3	−0.12 (−0.12 to −0.04)	−0.06 [−0.16 to 0.03]	−0.06 [−0.16 to 0.03]	−0.08 [−0.42 to 0.24]

NA, not applicable, only one study.

* Statistically significant difference for confidence interval (not including 0; *P* < 0.05).

parisons): APRI is recommended as the preferred non-invasive test to assess for the presence of cirrhosis (APRI score >2 in adults) in resource-limited settings. TE (e.g. FibroScan) or FibroTest may be the preferred non-invasive tests in settings, where they are available and cost is not a major constraint.[4]

Therefore, our Bayesian analyses in ITD of direct comparisons in CHC and CHB, the primary outcome, permitted to reinforce the evidence in favour of FibroTest vs. APRI and TE for the diagnosis of advanced fibrosis, already suggested but not claimed by the previous overviews. However, mainly due to the ITD our analysis concluded in the absence of difference between the performances of FibroTest and TE for the diagnosis of cirrhosis (Table 3). The lower performances of biomarkers using ALT or AST (FIB4 and APRI) were rational as they are both limited by the false-positive cases induced by activity and the variability in definitions of transaminase ULN for APRI.[31, 36]

We recognised the absence of significant difference between TE and FT for detecting cirrhosis or fibrosis in subpopulations including only CHC and CHB. However, two recommendations of recent EASL-ALEH guidelines referring to 'viral hepatitis' can be challenged, at least for FT and in ITD. 'TE and serum biomarkers have equivalent performance for detecting significant fibrosis in

patients with viral hepatitis' and 'TE is the most accurate non-invasive method for detecting cirrhosis in patients with viral hepatitis'.

The fourth advantage was the robustness of main results (Table S7). The lower performances of APRI vs. TE and vs. FIB4 for cirrhosis, and the absence of significant difference between TE and FibroTest for cirrhosis (from 0.00 to −0.01), persisted in all sensitivity analyses. The higher performance of FibroTest vs. TE in advanced fibrosis was nonsignificant, only after exclusion of Mixed-CB, which did not decrease the median difference, but enlarged the credibility interval.

Limitations

We acknowledge that our pilot overview was not complete, as we reviewed only the four most frequently used tests. This choice was justified by the limited number of direct comparisons and by the pragmatic need for clinicians to clarify ranking between these tests performances.[3, 4, 24, 25] Update of direct comparisons' overview will be necessary. This will be particularly interesting for the new elastography methods, which could have better applicability than TE.[43]

Variation in cut-off values is the main source of heterogeneity in diagnostic studies. We were not able to assess pooled sensitivity and specificity as recommended,[44] as among the 71 studies included (77 groups), the remaining number of groups giving predetermined cut-offs results was far too small, $n = 15$ (13 F2F3F4 and 2 F4), to permit appropriate bivariate pooled analyses of sensitivity and specificity (Table S2, Figure S2).

As FIB4 cut-offs were initially validated for F3F4 and not for F2F3F4 of F4 diagnosis, specific validations were needed before comparing performance to FT or APRI.[28] We were also not able to use an elegant recent method proposed for ITD analysis which recommended to consider the patients with non-evaluable results (failure or nonreliable results) in the group of false positive if the reference (here biopsy) is positive and in the group of false negative if the reference is negative.[19] Among the 71 studies included, 17 had non-interpretable results and among them, only three studies gave details of fibrosis stages at biopsy (Table S9).

We acknowledge the possible conflict of interest. The sensitivity analyses excluding non-independent gave similar results at one exception. In Only-C, the Bayesian credibility interval between FibroTest and APRI (advanced fibrosis) became no more significant and therefore also justify updates of such overviews.

The relatively small number of studies was a limitation, and results obtained for comparisons with fewer than four studies, frequent for FIB4, should be confirmed in further studies. The weaker results were therefore those obtained for comparisons between FIB4 vs. FibroTest and FIB4 vs. TE. However, it was reassuring to observe homogeneous trends for differences according to aetiology between Only-C and Only-B groups. Ideally, a pooling of studies with individual data could enhance the quality of such overviews. We succeeded in doing that for our first overviews, but it was no longer possible afterwards due to the increasing number of publications.[15, 17]

Here, the reference for AUROCs estimates was biopsy, which is far from a perfect reference, with 25% variability between fibrosis stages even for a 25 mm specimen length. Few studies had obtained greater AUROC than 0.90 only because, there was a spectrum effect or the number of patients was too small with a huge CI. If a study included a high number of patients, and if the analysis takes into account the spectrum effect, the AUROC estimate will converge to the maximum possible AUROC according to the size of the biopsy.[45] In the absence of large surgical biopsy, it is not possible to reach greater AUROCs than 0.90, even for a perfect biomarker.[46] In this context of reference variability, the statistically significant median differences observed between AUROCs, between 0.04 and 0.12, are probably also clinically relevant for the choice of clinicians. As already observed by previous overviews,[9] the majority of included studies were fair and poor quality, and progress in such diagnostic studies are expected with more specialised recommendations (Table S2).[47]

We did not include all indirect comparisons in a mixed Bayesian approach due to the high number of publications, and the risk of heterogeneity. We also acknowledge that we used multiple comparisons and frequentist tests, which increased the risk of statistically significance just by chance. However, we focused on the primary end point, the pooled direct AUROC differences, assessed by the Bayesian method (M4) in ITD and in studies pooling CHC and CHB, with the maximum of power, and used the credibility interval, which reduces the risk of false-positive conclusions.[22, 23]

Another limitation was the absence of pooling the results of prognostic studies. Others and we have already published such meta-analyses, which suggest better prognostic values of FibroTest vs. APRI and FIB4, and values similar to TE in per-protocol, but better in ITD1; updates are needed however.[21, 34, 35] Last, we focused only on the diagnostic performances without comparing the cost benefits or the efficiency of these tests, which will require specific overviews.

In conclusion, pooling AUROCs differences in direct comparisons, analysed in ITD with Bayesian methods, permitted to improve the evidence based comparisons between the three most frequently used fibrosis blood tests and TE. Similar rankings were observed in CHC and CHB. APRI had lower performances than FIB-4, TE and FibroTest. TE had lower performance than FibroTest for identifying advanced fibrosis in viral hepatitis when CHC and CHB were pooled, without significant difference for identifying cirrhosis.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Association between tests' performance (AUROC-ITD) and fibrosis spectrum (DANA index).

Figure S2. Pooled sensitivity and specificity of direct comparisons.

Table S1. Previous systematic reviews on direct comparisons of tests.

Table S2. Characteristics of 71 included studies with direct comparisons between tests.

Table S3. Characteristics of 39 excluded studies after discussion.

Table S4. Median AUROCs' differences between tests directly compared.

Table S5. Pooled AUROCs' difference using Bayesian method for 4 ITD methods.

Table S6. Pooled AUROCs' difference using Bayesian in per protocol vs. primary ITD (ITD1).

Table S7. Sensitivity analyses for Bayesian methods.

Table S8. Association between tests performance (AUROC) and fibrosis spectrum (DANA index) in Only-C.

Table S9. Details of biopsy in non-interpretable tests.

Data S1. Methods: Intention to diagnose, sensitivity analyses.

Data S2. Included and non-included studies.

Data S3. Direct comparisons of biomarkers performance.

Data S4. Other sensitivity analyses.

AUTHORSHIP

Guarantor of the article: None.

Author contributions: TP: study concept and design; analysis and interpretation of data; statistical analysis; drafting; study supervision. MH, YN, MM: acquisition of data; analysis and interpretation of data; statistical analysis; drafting; critical revision of the manuscript. SM: statistical analysis; drafting; critical revision of the manuscript. All authors approved the final version of the article, including the authorship list.

REFERENCES

1. US Burden of Disease Collaborators. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. *JAMA* 2013; **310**: 591–608.
2. Mohd-Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 2013; **57**: 1333–42.
3. WHO Library Cataloguing-in-Publication Data. Guidelines for the screening, care and treatment of persons with hepatitis C infection. WHO April 2014. Available at: http://apps.who.int/iris/bitstream/10665/111747/1/9789241548755_eng.pdf?ua=1 (accessed 15 October 2015).
4. WHO Library Cataloguing-in-Publication Data. Guidelines for the prevention, care and treatment of persons with chronic hepatitis B infection. WHO March 2015. Available at: http://apps.who.int (accessed 15 October 2015).
5. 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**: 586–7.
6. Poynard T, Lebray P, Ingiliz P, *et al.* Prevalence of liver fibrosis and risk factors in a general population using non-invasive biomarkers (FibroTest). *BMC Gastroenterol* 2010; **10**: 40.
7. Poynard T, Aubert A, Bedossa P, *et al.* A simple biological index for detection of alcoholic liver disease in drinkers. *Gastroenterology* 1991; **100**: 1397–402.
8. Gebo KA, Herlong HF, Torbenson MS, *et al.* Role of liver biopsy in management of chronic hepatitis C: a systematic review. *Hepatology* 2002; **36** (5 Suppl. 1): S161–72.
9. Chou R, Wasson N. Blood tests to diagnose fibrosis or cirrhosis in patients with chronic hepatitis C virus infection: a systematic review. *Ann Intern Med* 2013; **158**: 807–20.
10. Tsochatzis EA, Crossan C, Longworth L, *et al.* Cost-effectiveness of noninvasive liver fibrosis tests for treatment decisions in patients with chronic hepatitis C. *Hepatology* 2014; **60**: 832–43.
11. Poynard T, Ngo Y, Munteanu M, *et al.* Noninvasive markers of hepatic fibrosis in chronic hepatitis B. *Curr Hepat Rep* 2011; **10**: 87–97.
12. Salkic NN, Jovanovic P, Hauser G, Brcic M. FibroTest/Fibrosure for significant liver fibrosis and cirrhosis in chronic hepatitis B: a meta-analysis. *Am J Gastroenterol* 2014; **109**: 796–809.
13. Xu XY, Kong H, Song RX, *et al.* The effectiveness of noninvasive biomarkers to predict hepatitis B-related significant fibrosis and cirrhosis: a systematic review and meta-analysis of diagnostic test accuracy. *PLoS ONE* 2014; **9**: e100182.

14. Miller MH, Ferguson MA, Dillon JF. Systematic review of performance of non-invasive biomarkers in the evaluation of non-alcoholic fatty liver disease. *Liver Int* 2011; **31**: 461–73.

15. Poynard T, Lassailly G, Diaz E, *et al.* Performance of biomarkers FibroTest, ActiTest, SteatoTest, and NashTest in patients with severe obesity: meta analysis of individual patient data. *PLoS ONE* 2012; **7**: e30325.

16. Parkes J, Guha IN, Harris S, *et al.* Systematic review of the diagnostic performance of serum markers of liver fibrosis in alcoholic liver disease. *Comp Hepatol* 2012; **11**: 5.

17. Poynard T, Morra R, Ingiliz P, *et al.* Biomarkers of liver fibrosis. *Adv Clin Chem* 2008; **46**: 131–60.

18. Poynard T, de Ledinghen V, Zarski JP, *et al.* FibroTest and Fibroscan performances revisited in patients with chronic hepatitis C. Impact of the spectrum effect and the applicability rate. *Clin Res Hepatol Gastroenterol* 2011; **35**: 720–30.

19. Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. *BMJ* 2012; **345**: e6717.

20. Woo G, Tomlinson G, Nishikawa Y, *et al.* Tenofovir and entecavir are the most effective antiviral agents for chronic hepatitis B: a systematic review and Bayesian meta-analyses. *Gastroenterology* 2010; **139**: 1218–29.

21. Poynard T, Ngo Y, Perazzo H, *et al.* Prognostic value of liver fibrosis biomarkers: a meta-analysis. *Gastroenterol Hepatol (N Y)* 2011; **7**: 445–54.

22. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med* 1999; **130**: 1005–13.

23. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ* 2001; **322**: 226–31.

24. AASLD guidelines. Available at: http://www.hcvguidelines.org/full-report/hcv-testing-and-linkage-care

25. EASL-ALEH Clinical Practice Guidelines. Non-invasive tests for evaluation of liver disease severity and prognosis. *J Hepatol* 2015; **63**: 237–64.

26. Imbert-Bismut F, Ratziu V, Pieroni L, *et al.* Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: a prospective study. *Lancet* 2001; **357**: 1069–75.

27. Wai CT, Greenson JK, Fontana RJ, *et al.* A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology* 2003; **38**: 518–26.

28. Sterling RK, Lissen E, Clumeck N, *et al.* APRICOT Clinical Investigators. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology* 2006; **43**: 1317–25.

29. Castéra L, Vergniol J, Foucher J, *et al.* Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C. *Gastroenterology* 2005; **128**: 343–50.

30. Hartung J, Knapp G, Sinha BK. *Statistical Meta-analysis with Applications.* NewYork: John Wiley & Sons, 2008. ISBN 978-0-470-29089-7.

31. Perazzo H, Pais R, Munteanu M, *et al.* Variability in definitions of transaminase upper limit of the normal impacts the APRI performance as a biomarker of fibrosis in patients with chronic hepatitis C: "APRI c'est fini ?" *Clin Res Hepatol Gastroenterol* 2014; **4**: 432–9. pii: S22107401(14)00101-6.

32. Poynard T, Halfon P, Castera L, *et al.* Standardization of ROC curve areas for diagnostic evaluation of liver fibrosis markers based on prevalences of fibrosis stages. *Clin Chem* 2007; **53**: 1615–22.

33. Lambert J, Halfon P, Penaranda G, *et al.* How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. *Clin Chem* 2008; **54**: 1372–8.

34. Poynard T, Vergniol J, Ngo Y, *et al.* Staging chronic hepatitis C in seven categories using fibrosis biomarker (FibroTest™) and transient elastography (FibroScan®). *J Hepatol* 2014; **60**: 706–14.

35. Poynard T, Vergniol J, Ngo Y, *et al.* Staging chronic hepatitis B into seven categories, defining inactive carriers and assessing treatment impact using a fibrosis biomarker (FibroTest) and elastography (FibroScan). *J Hepatol* 2014; **61**: 994–1003.

36. Castera L. Noninvasive methods to assess liver disease in patients with hepatitis B or C. *Gastroenterology* 2012; **142**: 1293–302.

37. Poynard T, de Ledinghen V, Zarski JP, *et al.* 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. *J Hepatol* 2012; **56**: 541–8.

38. Munteanu M, Ratziu V, Poynard T. FibroStic: a large confirmatory study for non-invasive biomarkers accuracy, if correctly interpreted. *J Hepatol* 2011; **55**: 233; author reply 234–5.

39. Tapper EB, Castera L, Afdhal NH. FibroScan (vibration-controlled transient elastography): where does it stand in the United States practice. *Clin Gastroenterol Hepatol* 2015; **13**: 27–36.

40. Nascimbeni F, Lebray P, Fedchuk L, *et al.* Significant variations in elastometry measurements made within short-term in patients with chronic liver diseases. *Clin Gastroenterol Hepatol* 2015; **13**: 763–71.

41. Poynard T, Munteanu M, Deckmyn O, *et al.* Applicability and precautions of use of liver injury biomarker FibroTest. A reappraisal at 7 years of age. *BMC Gastroenterol* 2011; **11**: 39.

42. Chao DT, Lim JK, Ayoub WS, Nguyen LH, Nguyen MH. Systematic review with meta-analysis: the proportion of chronic hepatitis B patients with normal alanine transaminase ≤ 40 IU/L and significant hepatic fibrosis. *Aliment Pharmacol Ther* 2014; **39**: 349–58.

43. Poynard T, Munteanu M, Luckina E, *et al.* Liver fibrosis evaluation using real-time shear wave elastography: applicability and diagnostic performance using methods without a gold standard. *J Hepatol* 2013; **58**: 928–35.

44. Pavlov CS, Casazza G, Nikolova D, *et al.* Transient elastography for diagnosis of stages of hepatic fibrosis and cirrhosis in people with alcoholic liver disease. *Cochrane Database Syst Rev* 2015; **1**: CD010542.

45. Bedossa P, Dargère D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. *Hepatology* 2003; **38**: 1449–57.

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

47. Boursier J, de Ledinghen V, Poynard T, *et al.* An extension of STARD statements for reporting diagnostic accuracy studies on liver fibrosis tests: the Liver-FibroSTARD standards. *J Hepatol* 2015; **62**: 807–15.