Hepatology International (2019) 13:91–101
https://doi.org/10.1007/s12072-018-9907-5

ORIGINAL ARTICLE

Systematic review: diagnostic accuracy of non-invasive tests for staging liver fibrosis in autoimmune hepatitis

Shanshan Wu[1] · Zhirong Yang[2,3] · Jialing Zhou[1] · Na Zeng[1] · Zhiying He[1] · Siyan Zhan[4] · Jidong Jia[1,5] · Hong You[1,5]

Received: 14 August 2018 / Accepted: 22 October 2018 / Published online: 15 November 2018

Abstract

Background and aims Non-invasive fibrosis assessment has been highly recommended in many liver diseases. However, comparative diagnostic accuracy of laboratory markers, ultrasound and magnetic resonance elastography (MRE) for fibrosis in autoimmune hepatitis (AIH) patients has not been established.

Methods Medline, Embase and Cochrane Library were searched. Primary outcome was significant fibrosis (SF), advanced fibrosis (AF) and cirrhosis, defined as Metavir stage $F \geq 2$, $F \geq 3$ and $F = 4$ according to liver biopsy. Hierarchical summary receiver operating characteristic curve (ROC) model was used to evaluate diagnostic accuracy of non-invasive methods. Summary area under ROC (AUROC) and diagnostic odds ratio (DOR) with 95% confidence interval (CI) were calculated. The Grading of Recommendations Assessment, Development and Evaluation system was used to assess quality of evidence.

Results Overall, 16 studies with 861 patients were included, comparing aspartate aminotransferase to platelet ratio index (APRI), fibrosis-4 index (FIB-4), aspartate aminotransferase/alanine aminotransferase ratio, transient elastography (TE), acoustic radiation force impulse, shear wave elastography and MRE versus liver biopsy. Among all non-invasive markers, TE had good performance for fibrosis staging. Summary AUROCs and DORs of TE were 0.90 (95% CI 0.87, 0.92) and 23.7, 0.91 (95% CI 0.89, 0.93) and 31.6, 0.89 (95% CI 0.86, 0.92) and 80.5 for staging SF, AF and cirrhosis, whereas APRI and FIB-4 showed poor performance for detecting AF (DOR, 4.6 and 4.7) and cirrhosis (DOR, 5.5 and 12.9).

Conclusions TE performs well to stage liver fibrosis in patients with AIH, compared with other laboratory non-invasive indexes. Nevertheless, diagnostic accuracy of APRI and FIB-4 is poor.

Keywords Autoimmune hepatitis · Liver fibrosis · Non-invasive methods · Transient elastography

List of abbreviations

AIH	Autoimmune hepatitis
MRE	Magnetic resonance elastography
SF	Significant fibrosis
AF	Advanced fibrosis

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12072-018-9907-5) contains supplementary material, which is available to authorized users.

✉ Hong You
youhong30@sina.com

[1] National Clinical Research Center of Digestive Diseases, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China

[2] Primary Care Unit, Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge CB18RN, UK

[3] Department of Population Medicine, Harvard Medical School, Harvard Pilgrim Health Care Institute, Boston, MA 02215, USA

[4] Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Centre, Beijing 100191, China

[5] Liver Research Center, Beijing Friendship Hospital, Capital Medical University, 95 Yong-an Road, Xi-Cheng District, Beijing 100050, China

ROC	Receiver operating characteristic curve
AUROC	Summary area under ROC
DOR	Diagnostic odds ratio
CI	Confidence interval
APRI	Aspartate aminotransferase to platelet ratio index
FIB-4	Fibrosis-4 index
AAR	Aspartate aminotransferase/alanine aminotransferase ratio
TE	Transient elastography
ARFI	Acoustic radiation force impulse
SWE	Shear wave elastography
HCC	Hepatocellular carcinoma
AASLD	American Association for the Study of Liver Diseases
EASL	European Association for the Study of the Liver
AST	Aspartate aminotransferase
ALT	Alanine aminotransferase
PC/SD	Platelet count to spleen diameter
NAFLD	Non-alcoholic fatty liver disease
PPV	Positive predictive value
NPV	Negative predictive value
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies-2 scale
GRADE	The Grading of Recommendations Assessment Development and Evaluation
LR+	Positive likelihood ratio
LR−	Negative likelihood ratio
ULN	Upper limit normal

Introduction

Autoimmune hepatitis (AIH) is a non-resolving chronic inflammatory liver disease, classically characterized by interface hepatitis, hypergammaglobulinemia, circulating autoantibodies, elevated transaminase levels and response to immunosuppression [1, 2]. The disease, although treated with immunosuppression therapy, still can result in cirrhosis, hepatocellular carcinoma (HCC), decompensation and death [3–5]. Nearly 3% of treated AIH patients develop cirrhosis every year, and 1–6% of patients with cirrhosis progress to HCC [4, 5]. According to the guidelines of the American Association for the Study of Liver Diseases (AASLD) and the European Association for the Study of the Liver (EASL), assessment of liver fibrosis and cirrhosis is essential to guide treatment strategies in patients with AIH [6, 7].

Liver biopsy is considered as the gold standard for the evaluation of liver fibrosis stage. However, it is an invasive procedure with risk of complications, sampling error and inter-observer variability [6–8]. Moreover, sequential biopsies are unfeasible in clinical practice for purpose of regular dynamic monitoring liver fibrosis stage. Therefore, it is essential to develop accurate non-invasive methods to assess disease progression and guide therapy [9].

Several non-invasive markers, including laboratory and radiological tests, have been proposed. Laboratory panel markers include aspartate aminotransferase to platelet ratio index (APRI) [10] and fibrosis-4 index (FIB-4) [11], the aspartate aminotransferase (AST)/alanine aminotransferase (ALT) ratio (AAR) and other less commonly used markers (such as platelet count to spleen diameter (PC/SD) ratio and non-alcoholic fatty liver disease (NAFLD) fibrosis score). Radiological tests include transient elastography (TE), acoustic radiation force impulse (ARFI), two-dimensional shear wave elastography (2D-SWE) and magnetic resonance elastography (MRE). However, all indices are introduced and validated in chronic viral hepatitis and NAFLD. There are limited data on their performance to stage liver fibrosis in AIH, since AIH differs from other diseases with fluctuating inflammation. Although several studies have investigated one or more non-invasive tests in AIH recently, results are still inconsistent due to the relatively small sample size and heterogeneous AIH conditions [12–16]. Until now, according to the guideline released by EASL, no recommendation can be made with current evidence on the use of non-invasive tests in AIH [7].

In view of this, we aimed to conduct a systematic review to assess the comparative diagnostic accuracy of laboratory markers (FIB-4, APRI, AAR, PC/SD ratio, NAFLD fibrosis score), ultrasound (TE, ARFI, 2D-SWE) and magnetic resonance technology (MRE) for liver fibrosis stage assessment in patients with AIH.

Methods

This study is registered with PROSPERO, number CRD2018090903.

Data sources and searches

Medline, Embase and the Cochrane Central Register of Controlled Trials were searched from inception to January 20th, 2018 using the following keywords: autoimmune hepatitis, liver fibrosis, cirrhosis, APRI, FIB-4, elasticity imaging techniques, transient elastography, shear wave elastography, magnetic resonance elastography and acoustic radiation force impulse (Online appendix 1 for full details about the search strategy). The website of AASLD and EASL was searched for the annual conference abstracts. In addition, we also checked the reference list of all relevant articles to identify additional studies.

Study selection

Studies were included in the systematic review according to the following criteria: (1) participants of the study were patients with AIH. If the original study enrolled patients with different liver diseases, data for patients with AIH should be reported separately. (2) The index tests assessed in the original study for staging liver fibrosis were APRI, FIB-4, AAR, TE, 2D-SWE, MRE, ARFI, MRS (magnetic resonance spectroscopy), PC/SD ratio, and NAFLD fibrosis score. (3) Liver biopsy was used as the reference standard for staging liver fibrosis according to Metavir score or other pathological score systems that can be transformed to Metavir score. (4) Data were available to calculate the value of true positive, true negative, false positive and false negative for each index test. (5) Number of AIH cases in the original study should be greater than 10, according to the prevalence of AIH. (6) Studies should be cross-sectional, and single-gate. Reviews, editorials, letters, guidelines and protocols were excluded. Articles focused on animals or basic research were also excluded.

The eligibility of studies for inclusion criteria was assessed independently by two reviewers (SSW and NZ) in duplicate. Any discrepancies were resolved by consensus between the two independent reviewers or by a senior investigator (JLZ).

Data extraction and quality assessment

Data were extracted using Microsoft Excel 2010 with respect to study information (author, publication year, sample size, study period, region, study design, no. of centers, types of index test), participant characteristics (diagnostic criteria of AIH, age, gender, baseline ALT level, pre- or post-treatment), information on liver biopsy (length of liver biopsy, pathological scoring system, blind, time interval between liver biopsy and index test), performance of different index tests (cut-off values, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under receiver operating curve (AUROC)) and information on methodology. Four investigators (SSW, NZ, ZYH and JLZ) extracted data independently, in duplicate.

For original studies without sufficient data for meta-analysis, we sent emails to the authors for more details. In the case of non-response from the authors after two emails, the studies were excluded.

Risk of bias of included studies was assessed according to the Quality Assessment of Diagnostic Accuracy Studies-2 scale (QUADAS-2), including four domains (patient selection, index test, reference standard, flow and timing) [17]. Additionally, the GRADE (The Grading of Recommendations Assessment, Development, and Evaluation) framework was used to assess the quality of evidence contributing to performance estimate of each index test, which characterizes the quality of a body of evidence on the basis of the risk of bias, indirectness, inconsistency, imprecision, and publication bias for staging liver fibrosis. The estimated minimum sample size needed for each index test detecting liver fibrosis or cirrhosis was calculated according to Flahault method [18]. Two investigators (SSW and SYZ) evaluated the risk of bias for included studies independently, in duplicate. Any discrepancies were resolved by discussion with a senior investigator (ZRY).

Data synthesis and analysis

All analyses were conducted using STATA 13.0 (summary ROC curve, forest plot of DOR, estimation of heterogeneity and Deeks' funnel plot), Reviewer Manager Version 5.3 (risk of bias graph and risk of bias summary graph) and Meta-Disc Version 1.4 (simple pooling of sensitivity and specificity).

Definition of liver fibrosis

Significant fibrosis (SF), advanced fibrosis (AF) and cirrhosis were defined as stage $F \geq 2$, $F \geq 3$ and $F = 4$ according to liver biopsy scoring systems such as Metavir, Desmet & Scheuer, and Batts &Ludwig.

Methods for diagnostic accuracy evaluation

The value of true positive, true negative, false positive and false negative for detecting SF, AF and cirrhosis was calculated based on sensitivity, specificity and sample size of patients in each original study. When there was more than one cut-off value staging liver fibrosis in the same original study, the one closer to the common cut-off value was selected in the meta-analysis. Descriptive analysis was performed to describe the cut-off values.

For index tests staging liver fibrosis with number of original studies ≥ 4, hierarchical models including HSROC (hierarchical summary ROC) model and bivariate model were used to evaluate diagnostic accuracy, which considered the correlation between sensitivity and specificity. Summary AUROC, summary sensitivity, summary specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR−) and DOR (diagnostic odds ratio) with 95% confidence interval (CI) were calculated as effect measures. Additionally, SROC plot and forest plot for DOR were performed for each index test in detecting SF, AF and cirrhosis. For index tests staging liver fibrosis with number of original studies < 4, simple pooling method was used to calculate summary sensitivity, summary specificity, LR+, LR− and DOR with 95% CI, since the number of studies was not enough to perform hierarchical model.

Methods for heterogeneity and publication bias assessment

The I^2 statistic was calculated to assess heterogeneity for the diagnostic accuracy of each non-invasive method, as a measure of the proportion of the overall variation that is attributable to between-study heterogeneity. The Cochrane Q test was used to evaluate heterogeneity statistically. An @@I^2 value > 50% or a p value < 0.05 may be considered to represent substantial heterogeneity. Additionally, subgroup analysis was performed according to treatment status (pre-treatment vs post-treatment). However, the number of original studies for each index test was not enough to explore heterogeneity using meta-regression.

Publication bias assessment for each non-invasive method detecting SF, AF and cirrhosis was performed by Deeks' plot, a linear regression of log DOR against 1/sqrt (effective sample size) to test asymmetry of funnel plot and with $p < 0.1$ for the slope coefficient indicating significant asymmetry.

Results

Study and patient characteristics

Overall, 16 studies met the inclusion criteria (Online appendix 2 for full reference list). Flowchart of study selection is shown in Fig. 1. Ten index tests were analyzed, including laboratory tests (FIB-4, APRI, AAR and NAFLD fibrosis score), ultrasound (TE, 2D-SWE and ARFI), magnetic resonance technology (MRE and MRS), and combination of laboratory test and ultrasound (PC/SD ratio). A total of 10, 6 and 8 studies were focused on TE, FIB-4 and APRI detecting SF, AF and cirrhosis. The characteristics of the included studies are summarized in Table 1. Publication year varied from 2005 to 2017, mainly (75%) from 2016 to 2017. Most studies were from Asia (44%) and Europe (44%). There were eight (50%) prospective studies and five (31%) retrospective studies.

A total of 867 patients with AIH contributed to the analysis of diagnostic performance on staging liver fibrosis, predominantly female (72%, range 58–87%). The mean age of patients was 46.0 years [standard deviation (SD) 11.1 year], and the median baseline ALT level was 56 U/L (range 21–606 U/L). Among all AIH patients, 39% (339/867) were untreated, 35% (300/867) were treated (post-treatment), and other patients' treatment status (26%) was unclear. The non-invasive tests were performed on the same day or within 3 days or 7 days of liver biopsy in most studies.

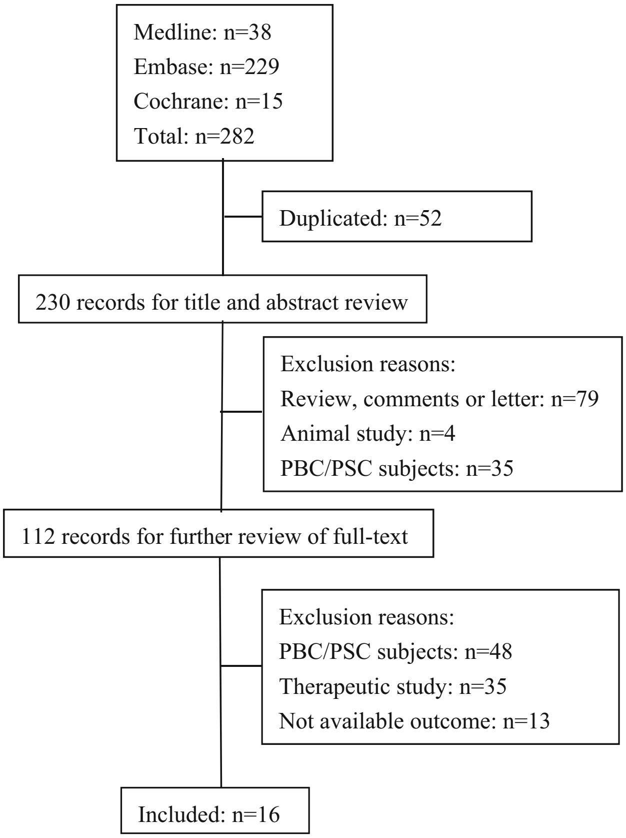


Fig. 1 Flowchart for study selection in the meta-analysis

Methodological quality and risk of bias results

Results of quality assessment for included studies are listed in Online appendix 3. Patient selection, index test and flow and timing were appropriately described in majority of the studies (100.0, 87.5 and 87.5%, respectively). By contrast, reference standard was not clearly reported in 68.7% of the cases. In patient selection, all studies enrolled patients consecutively without inappropriate exclusion. However, two studies were considered as unclear risk for applicability concerns since these only focused on children. Regarding the index test, two studies were considered as unclear risk because no blind information was declared when interpreting the results of TE. In terms of reference standard, five studies were considered as unclear risk because the results of reference standard were interpreted with unclear blind of the results of index test. For flow and timing, two studies were considered as unclear risk since the interval between reference standard and index test was unclear. Overall, the risk of bias across studies was relatively low.

Table 1 Characteristics of studies included in the meta-analysis

No.	Author (year)	Study period	Region	Diagnostic criteria	Model	Study design	Sample size	Mean age	Female (%)	ALT level	Mean BMI	Pre-/post-treatment	Scoring system	Interval	Blind	Length (mm)
1	Xu Q (2017)	2014–2016	China	IAIHG 2008	1, 2, 3	Prospective	100	45.0	81.0	131.5	NA	Pre	Metavir	Same day	Y	>10
2	Anastasiou OE (2016)	2008–2013	Germany	IAIHG 2008	1, 2, 3, 4, 8	Retrospective	53	47.3	58.5	606.4	NA	Pre(35) + Post(18)	Metavir	4 (2–17) days	Y	≥14
3	Guo L (2017)	2012–2017	China	IAIHG 2008	1, 2, 3	Retrospective	108	46.5	81.5	146.5	23.52	NA	Metavir	3 days	NA	≥15
4	Zeng J (2017)	2011–2016	China	IAIHG 2008	6	Prospective	81	45.6	81.6	78.5	21.6	Pre	Metavir	3 days	Y	≥15
5*	Hartl J (2016)	2007–2010	Germany	EASL 2015	1	Prospective	34	53.0	82.0	48.5	NA	Post	Desmet & Scheuer	Within 3 months	Y	≥8
6*	Hartl J (2016)	2008–2015	Germany	EASL 2015	1	Retrospective	60	52.0	83.0	35.0	NA	Post	Desmet & Scheuer	Within 4 months	Y	≥8
7	Wang J (2017)	2007–2015	China	IAIHG 1999	1, 2, 3, 4, 9	Retrospective	36	51.6	NA	217.4	27.7	Pre(17) + Post(19)	Metavir	Within 3 months	Y	NA
8	Sheptulina A (2016)	2008–2014	Germany	IAIHG 1999	2, 3, 4, 7	Prospective	76	40.0	85.5	54.4	25.0	Pre(22) + Post(54)	Metavir	7 days	Y	≥14
9	Efe C (2015)	2004–2010	Turkey	IAIHG 2008	5	Retrospective	15	40.9	86.7	32.0	NA	Post	Metavir	Within 14 days	NA	NA
10	Kim JK (2014)	2008–2014	Korea	IAIHG 1999	1	Retrospective	47	NA	87.2	NA	NA	NA	Metavir	NA	NA	NA
11	Youssef A (2013)	NA	Egypt	NA	1	Retrospective	16	NA	NA	NA	NA	NA	Metavir	NA	NA	NA
12	Piwczynska K (2016)	NA	Poland	NA	3	Prospective	46	14.5	71.8	NA	NA	NA	Batts & Ludwig	NA	NA	NA
13	Paranagua VD (2017)	2012–2015	Brazil	NA	1, 5	Prospective	33	NA	84.8	NA	28.6	Post	Metavir	Same day	NA	NA
14	Harrison L (2016)	2013–2015	UK	IAIHG 1999	1	Prospective	27	56.0	NA	21.0	NA	Post	Ishak	Same day	NA	NA
15	Nishikawa H (2016)	2005–2015	Japan	IAIHG 1999	2, 3, 4	Prospective	84	64.0	82.1	57.5	NA	Pre	Metavir	NA	NA	NA
16	Abdo AA (2005)	1996–2004	Arabic	IAIHG 1999	3, 4	Retrospective	39	45.4	65.0	268.0	NA	Post	Metavir	NA	NA	NA
17	Puustinen L (2017)	NA	Finland	IAIHG 2008	10	Prospective	12	42.8	83.3	28.5	NA	NA	Metavir	Within 1 month	NA	NA

*No. 5 and no. 6 are the same publication; models are represented by the following numbers: 1, transient elastography (TE); 2, fibrosis-4 index (FIB-4); 3, aspartate aminotransferase to platelet ratio index (APRI); 4, aspartate aminotransferase to alanine transaminase ratio (AAR); 5, acoustic radiation force impulse (ARFI); 6, two-dimensional shear wave elastography (2D-SWE); 7, platelet count to spleen diameter (PC/SD) ratio; 8, NAFLD (non-alcoholic fatty liver disease) fibrosis score; 9, magnetic resonance elastography (MRE); 10, magnetic resonance spectroscopy (MRS)

IAIHG International Autoimmune Hepatitis Group, *EASL* European Association for the Study of the Liver, *AASLD* American Association for the study of liver diseases, *ALT* alanine transaminase, *BMI* body mass index, *NA* not available

Diagnostic accuracy for SF (*F* ≥ 2)

Eight out of 16 studies assessed the performance of 7 non-invasive methods for detecting SF (stage F2–F4). Overall, 7 (429 patients), 2 (161 patients) and 2 (161 patients) studies investigated the diagnostic accuracy of TE, FIB-4 and APRI, respectively. Besides, there was only one study that focused on AAR, PC/SD ratio, NAFLD fibrosis score and 2D-SWE, separately. The original data on sensitivity, specificity and AUROC of each non-invasive method are listed in Online appendix 4 (Supplementary Table 1).

The summary sensitivity, summary specificity, LR+, LR− and cut-off values of the non-invasive methods are listed in Table 2. With the cut-off ranging from 5.8 to 7.0 Kpa with a median of 6.3 Kpa, the summary sensitivity and specificity of TE were much greater than other six non-invasive methods, with 0.82 (95% CI 0.69, 0.90) and 0.79 (95% CI 0.69, 0.86), respectively. The FIB-4 has a relatively good summary specificity (0.85, 95% CI 0.70, 0.94) with poor summary sensitivity (0.50, 95% CI 0.41, 0.59), whereas APRI had a relatively moderate summary sensitivity (0.70, 95% CI 0.61, 0.78) with poor summary specificity (0.56, 95% CI 0.40, 0.72).

Regarding the summary AUROC values (Fig. 2), TE showed the best performance (0.90, 95% CI 0.87, 0.92) for staging SF (heterogeneity $I^2 = 44\%$, $p = 0.085$), followed by 2D-SWE (0.85), whereas FIB-4 (range 0.66–0.66), APRI (range 0.60–0.64) and AAR (0.60) showed poor performance for detecting SF. In terms of DOR (Fig. 3), TE also demonstrated the best performance (23.7, 95% CI 8.7, 64.6) for staging SF.

Diagnostic accuracy for AF (*F* ≥ 3)

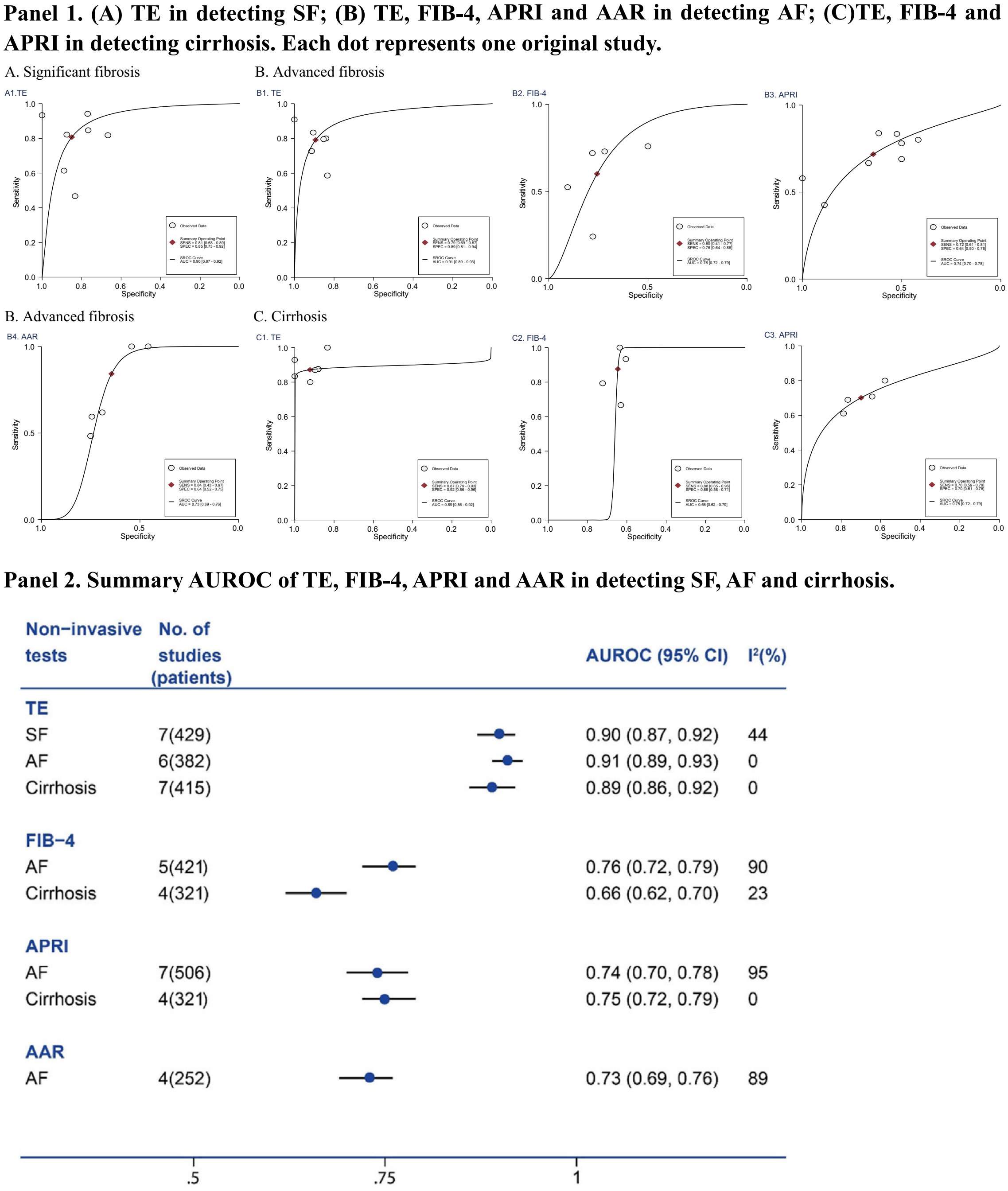
Overall, 12 of 16 studies investigated the performance of nine non-invasive methods for staging AF (stage F3–F4).

Table 2 Summary sensitivities, specificities, LR+, and LR− of TE, FIB-4, APRI, AAR and ARFI at various diagnostic thresholds for prediction of SF, AF, and cirrhosis

	Cut-off values	No. of studies (no. of patients)	Summary sensitivity (95% CI)	Summary specificity (95% CI)	Summary LR+ (95% CI)	Summary LR− (95% CI)
TE						
SF	5.80–7.00	5 (329)	0.82 (0.69, 0.90)	0.79 (0.69, 0.86)	3.80 (2.50, 5.70)	0.23 (0.13, 0.42)
	9.10–10.05	2 (100)	0.70 (0.56, 0.81)	0.98 (0.87, 0.99)	14.63 (1.41, 151.60)	0.22 (0.02, 2.61)
AF	8.18–8.75	2 (208)	0.80 (0.71, 0.87)	0.85 (0.76, 0.91)	5.18 (3.27, 8.22)	0.24 (0.16, 0.35)
	10.40–12.10	4 (174)	0.74 (0.61, 0.84)	0.93 (0.86, 0.97)	7.68 (2.88, 20.48)	0.27 (0.11, 0.63)
Cirrhosis	11.00–12.67	4 (268)	0.88 (0.77, 0.94)	0.88 (0.83, 0.92)	7.40 (5.10,10.80)	0.14 (0.07, 0.27)
	16.00–19.00	3 (147)	0.86 (0.70, 0.95)	0.97 (0.92, 0.99)	21.66 (5.08,92.30)	0.18 (0.09, 0.38)
FIB-4						
SF	2.90–3.20	2 (161)	0.50 (0.41, 0.59)	0.85 (0.70, 0.94)	3.25 (1.52, 6.94)	0.59 (0.47, 0.73)
AF	1.75–2.37	3 (229)	0.73 (0.64, 0.81)	0.70 (0.61, 0.78)	2.29 (1.42, 3.71)	0.39 (0.28, 0.53)
	3.21–5.10	2 (192)	0.37 (0.27, 0.47)	0.83 (0.74, 0.90)	2.35 (0.46, 11.97)	0.73 (0.39, 1.36)
Cirrhosis	2.59–3.40	4 (321)	0.88 (0.65, 0.96)	0.65 (0.58, 0.71)	2.50 (2.00, 3.10)	0.19 (0.06, 0.62)
APRI						
SF	0.88–1.45	2 (161)	0.70 (0.61, 0.78)	0.56 (0.40, 0.72)	1.64 (1.13, 2.38)	0.52 (0.35, 0.77)
AF	0.50–0.90	4 (299)	0.81 (0.74, 0.87)	0.52 (0.44, 0.60)	1.70 (1.40, 2.00)	0.36 (0.25, 0.52)
	1.24–2.13	4 (246)	0.58 (0.45, 0.70)	0.76 (0.55, 0.89)	2.40 (1.30, 4.50)	0.55 (0.43, 0.70)
Cirrhosis	1.50–2.00	4 (321)	0.70 (0.59, 0.79)	0.70 (0.61, 0.78)	2.30 (1.80, 3.10)	0.43 (0.30, 0.60)
AAR						
SF	0.72	1 (53)	0.52	0.78	2.36	0.62
AF	0.70–1.10	4 (252)	0.84 (0.43, 0.97)	0.64 (0.52, 0.75)	2.40 (1.80, 3.10)	0.24 (0.05, 1.15)
Cirrhosis	0.94–1.40	3 (213)	0.61 (0.48, 0.73)	0.82 (0.75, 0.88)	3.28 (1.99, 5.43)	0.49 (0.35, 0.68)
ARFI[a]						
Cirrhosis	1.65	2 (48)	0.80 (0.52, 0.96)	0.82 (0.65, 0.93)	4.01 (1.86, 8.66)	0.28 (0.11, 0.69)

LR+ positive likelihood ratio, *LR−* negative likelihood ratio, *TE* transient elastography, *FIB-4* fibrosis-4 index, *APRI* aspartate aminotransferase to platelet ratio index, *AAR* aspartate aminotransferase to alanine transaminase ratio, *ARFI* acoustic radiation force impulse, *SF* significant fibrosis, *AF* advanced fibrosis

[a] No available data on diagnostic accuracy of ARFI in detecting significant fibrosis and advanced fibrosis



Non−invasive tests	No. of studies (patients)	AUROC (95% CI)	I²(%)
TE			
SF	7(429)	0.90 (0.87, 0.92)	44
AF	6(382)	0.91 (0.89, 0.93)	0
Cirrhosis	7(415)	0.89 (0.86, 0.92)	0
FIB−4			
AF	5(421)	0.76 (0.72, 0.79)	90
Cirrhosis	4(321)	0.66 (0.62, 0.70)	23
APRI			
AF	7(506)	0.74 (0.70, 0.78)	95
Cirrhosis	4(321)	0.75 (0.72, 0.79)	0
AAR			
AF	4(252)	0.73 (0.69, 0.76)	89

Fig. 2 The summary ROC plots of non-invasive tests with no. of original studies ≥ 4. Panel 1. **a** TE in detecting SF, **b** TE, FIB-4, APRI and AAR in detecting AF, **c** TE, FIB-4 and APRI in detecting cirrhosis. Panel 2. Summary AUROC of TE, FIB-4, APRI and AAR in detecting SF, AF and cirrhosis. *AUROC* area under the receiver operator curve, *TE* transient elastography, *FIB-4* fibrosis-4 index, *APRI* aspartate aminotransferase to platelet ratio index, *AAR* aspartate aminotransferase to alanine transaminase ratio, *SF* significant fibrosis, *AF* advanced fibrosis

Approximately, six (382 patients), five (421 patients), seven (506 patients) and 4 (252 patients) studies evaluated the diagnostic accuracy of TE, FIB-4, APRI and AAR, respectively. However, there was only one study that focused on PC/SD ratio, NAFLD fibrosis score, MRE, MRS and 2D-SWE, separately.

As shown in Table 2, TE also showed good summary sensitivity [0.80 (95% CI 0.71, 0.87)] and specificity [0.85

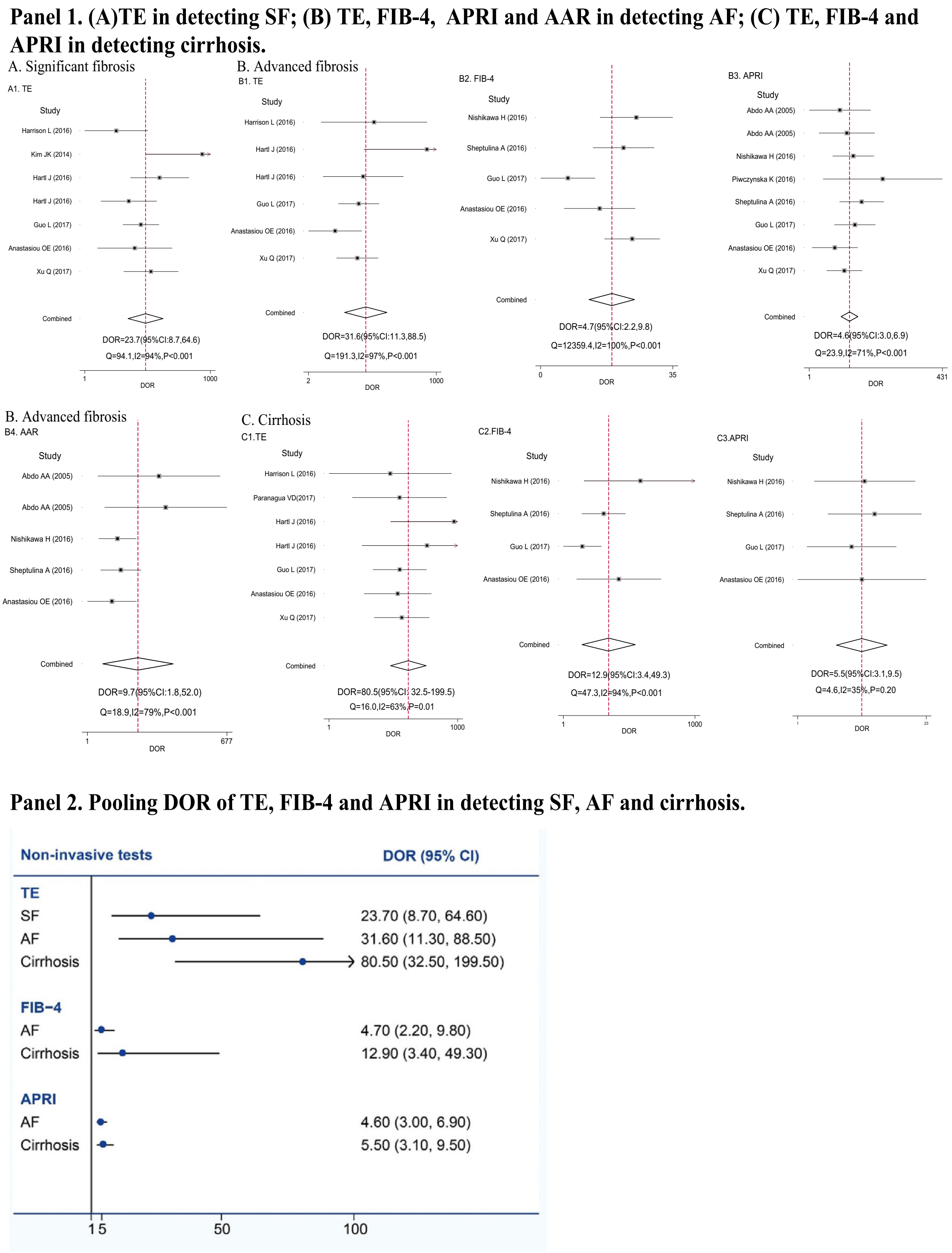


Fig. 3 Forest plots of diagnostic odds ratio (DOR). Panel 1. **a** TE in detecting SF, **b** TE, FIB-4, APRI and AAR in detecting AF, **c** TE, FIB-4 and APRI in detecting cirrhosis. Panel 2. Pooling DOR of TE, FIB-4 and APRI in detecting SF, AF and cirrhosis. *DOR* diagnostic odds ratio, *TE* transient elastography, *FIB-4* fibrosis-4 index, *APRI* aspartate aminotransferase to platelet ratio index, *SF* significant fibrosis, *AF* advanced fibrosis

Table 3 Summary sensitivities and specificities of TE for prediction of SF, AF, and cirrhosis in post-treatment AIH patients

	Cut-off values	No. of studies (patients)	Summary sensitivity (95% CI)	Summary specificity (95% CI)
SF	5.80–7.0	3 (121)	0.80 (0.69, 0.89)	0.76 (0.62, 0.87)
AF	10.4–11.0	3 (121)	0.85 (0.69, 0.94)	0.95 (0.88, 0.99)
Cirrhosis	11.0–16.0	4 (154)	0.90 (0.73, 0.97)	0.97 (0.75, 1.00)

TE transient elastography, *SF* significant fibrosis, *AF* advanced fibrosis, *AIH* autoimmune hepatitis

(95% CI 0.76, 0.91)], with the cut-off ranging from 8.2 to 8.8 Kpa with a median of 8.5 Kpa. When the cut-off varied from 10.4 to 12.1 Kpa with a median of 10.7 Kpa, the summary sensitivity decreased to 0.74 (95% CI 0.61, 0.84) as specificity increased to 0.93 (95% CI 0.86, 0.97). Additionally, MRE appeared with good performance (0.90 for sensitivity and 1.00 for specificity) for detecting AF; however, there was only one study (Online appendix 4, Supplementary Table 2).

According to Fig. 2, the summary AUROC values for TE, FIB-4, APRI and AAR were 0.91 (95% CI 0.89, 0.93), 0.76 (95% CI 0.72, 0.79), 0.74 (95% CI 0.70, 0.78) and 0.73 (95% CI 0.69, 0.76). No heterogeneity was detected among TE studies ($I^2 = 0\%$, $p = 0.495$). In descending order, DOR values of these non-invasive methods (Fig. 3) were 31.6 (TE), 9.7 (AAR), 4.7 (FIB-4) and 4.6 (APRI).

Diagnostic accuracy for cirrhosis (*F* = 4)

Eleven of 16 studies evaluated the diagnostic accuracy of nine non-invasive methods for staging cirrhosis (stage F4). Approximately seven (415 patients), four (321 patients), four (321 patients), three (213 patients) and two (48 patients) studies explored the performance of TE, FIB-4, APRI, AAR and ARFI, respectively. Only one study focused on PC/SD ratio, NAFLD fibrosis score, MRE and 2D-SWE, separately.

As listed in Table 2, with cut-off from 11.0 to 12.7 Kpa with a median of 12.4 Kpa, TE had the largest summary sensitivity (0.88, 95% CI 0.77, 0.94) and summary specificity (0.88, 95% CI 0.83, 0.92) for detecting cirrhosis compared with FIB-4, APRI, AAR and ARFI. When the cut-off ranged from 16.0 to 19.0, the summary sensitivity slightly decreased (0.86, 95% CI 0.70, 0.95) whereas specificity dramatically increased to 0.97 (95% CI 0.92, 0.99). The summary AUROC for TE indicated best performance with 0.89 (95% CI 0.86, 0.92), and no heterogeneity was detected among studies ($I^2 = 0\%$, $p = 0.367$). Worthy to be mentioned, one study found good performance of MRE (0.92 for sensitivity, 0.96 for specificity and 0.98 for AUROC) to diagnose cirrhosis (Online appendix 4, Supplementary Table 3). In terms of DOR (Fig. 3), TE had the largest value (80.5, $I^2 = 63\%$, $p = 0.01$), followed by FIB-4 (12.9, $I^2 = 94\%$, $p < 0.001$) and APRI (5.5, $I^2 = 35\%$, $p = 0.20$).

Subgroup analysis

Results of subgroup analyses by treatment status indicated good diagnostic accuracy for staging liver fibrosis in post-treatment patients (Table 3), whereas analysis for pre-treatment could not be performed due to limited data. However, we sorted diagnostic data of TE according to treatment status in Online appendix 5, which showed good performance for detecting SF, AF and cirrhosis in treatment-naïve patients.

Publication bias

Although number of original studies was fewer, we still performed the linear regression test and Deeks' funnel plot to evaluate publication bias. The results were demonstrated in Online appendix 6. No evidence of publication bias was detected for these non-invasive methods staging SF, AF and cirrhosis (all p value > 0.10), except for AAR for detecting AF (p value = 0.091).

GRADE framework for evidence quality

According to GRADE, the quality of evidence generally ranged between low and high, but was rated as moderate for most comparisons due to imprecision. However, the quality was moderate for TE versus liver biopsy staging SF, AF and cirrhosis due to limited data on treatment-naïve patients (Online appendix 7 for the estimation of the minimum sample size needed and Online appendix 8 for quality of evidence according to GRADE framework).

Discussion

Accurately staging liver fibrosis, particularly through non-invasive methods, is required for disease progression evaluation during therapy in patients with AIH. Our systematic review and meta-analysis with 16 studies and 867 patients suggested that transient elastography performs well to stage liver fibrosis in patients with AIH, compared with other available laboratory non-invasive indexes. However, diagnostic accuracy of APRI and FIB-4 is poor. In addition, MRE may appear to have excellent diagnostic accuracy for staging liver fibrosis.

As reported, TE has been widely used as a reliable non-invasive tool to stage liver fibrosis in chronic hepatitis B or hepatitis C [19–21]. Our results demonstrated good diagnostic performance of TE in patients with AIH for detecting SF, AF and cirrhosis, although TE may be affected by hepatic inflammation and AIH was always accompanied by elevated transaminase levels [7]. In the present study, the median baseline ALT level was 56 U/L (range 21–606 U/L), 39% of patients were untreated and 35% were post-treatment. Since transaminase levels would decrease significantly after treatment, we conducted subgroup analyses by treatment status, indicating good performance for detecting SF, AF and cirrhosis in both treatment-naïve and post-treatment patients, but the data were limited. Further large studies are needed to validate this conclusion.

Consistent with other studies about patients with diverse chronic liver diseases [22–25], MRE seems to have promising performance for detecting advanced fibrosis and cirrhosis in patients with AIH, and is superior to laboratory assessments. It seems that diagnostic accuracy of MRE is not influenced by hepatic inflammation, no matter in untreated (mean ALT level of 298 U/L) or treated patients (mean ALT level of 145 U/L). Nevertheless, the sample size is small and the findings need to be confirmed in studies with a large number of participants.

Additionally, our findings indicated that APRI and FIB-4 were poor for staging liver fibrosis in patients with AIH. Unlike with studies in chronic viral hepatitis or NAFLD, APRI and FIB-4 were proved to offer reliable accuracy for detecting cirrhosis or advanced fibrosis in those patients [26, 27]. The different results might be due to relatively higher ALT levels in AIH patients, since patients with ALT more than two times ULN (upper limit normal) were excluded in studies with other chronic liver diseases [28].

A major strength of our study is the comprehensive and substantial analysis of diagnostic accuracy of these non-invasive tests compared with liver biopsy as the gold standard in patients with AIH for the first time. In addition to simple pooling of sensitivity and specificity, we conducted detailed hierarchical summary ROC analyses to address the threshold effect of studies. Furthermore, we assessed quality of evidence and incorporate it into explaining the results by the GRADE framework, which is the latest evidence and more useful for further guideline and clinical practice.

Several limitations should be mentioned and taken into account when interpreting the data from this study. First, only trials in English were included, which may lead to potential publication bias. However, we did a comprehensive and thorough search, and the Deeks' funnel plots were almost symmetrical except for AAR for detecting AF (p value = 0.091), which was marginally significant. Second, we did not have access to original studies' data, so we could not perform an individual patient data meta-analysis to properly assess in our analyses potentially relevant effect modifiers such as different baseline levels of ALT, disease duration, diagnostic criteria of AIH and length of liver biopsy. Moreover, many studies did not mention the length of liver biopsy, or only a short length of biopsy, which may lead to a sampling error of biopsy diagnosis. Finally, due to insufficient data and limited number of studies, we cannot compare the diagnostic performance of these non-invasive tests between pre-treatment and post-treatment patients with AIH. Five out of six studies did not report available outcome information, although we have sent emails twice. Further studies are needed since the diagnostic accuracy in treatment-naïve patients is more important in clinical practice.

In conclusion, transient elastography performs well to stage liver fibrosis in patients with AIH, compared with other laboratory non-invasive indexes. Nevertheless, diagnostic accuracy of APRI, FIB-4 and AAR is poor. Further studies are needed to evaluate the diagnostic performance of these non-invasive indexes in patients with AIH.

Acknowledgements We are grateful to all cooperating organizations and their staff whose hard work made this study possible. ZRY is supported by the Cambridge Trust and the China Scholarship Council. Thanks to Prof. Jing Hua for supplying cut-off value, sensitivity and specificity of APRI and FIB-4 detecting advanced fibrosis.

Authorship SSW and HY designed the study and drafted the manuscript. SSW, NZ, ZYH and JLZ extracted the data, SSW, ZRY, and SYZ evaluated the quality. SSW, ZRY, and SYZ assessed the quality of evidence by GRADE framework. SSW and ZRY analyzed the data. HY and JDJ interpreted the results, incorporated comments for the co-authors and finalized the manuscript. All authors approved the final version of the paper.

Financial support This study is funded by Beijing Municipal Administration of Hospitals' Youth Program (QML20170107) and Beijing Talents Fund (2016000021469G226). The sponsor had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Compliance with ethical standards

Conflict of interest Shanshan Wu, Zhirong Yang, Jialing Zhou, Na Zeng, Zhiying He, Siyan Zhan, Jidong Jia, and Hong You have no conflict of interests.

References

1. Michael PM, Ansgar WL, Diego V. Autoimmune hepatitis-update 2015. J Hepatol 2015;62:100–111
2. Wang Q, Yang F, Miao Q, et al. The clinical phenotypes of autoimmune hepatitis: a comprehensive review. J Autoimmun 2016;66:98–107
3. Czaja AJ, Carpenter HA. Progressive fibrosis during corticosteroid therapy of autoimmune hepatitis. Hepatology 2004;39:1631–1638
4. Montano-Loza AJ, Carpenter HA, Czaja AJ. Predictive factors for hepatocellular carcinoma in type 1 autoimmune hepatitis. Am J Gastroenterol 2008;103:1944–1951
5. Yeoman AD, Al-Chalabi T, Karani JB, et al. Evaluation of risk factors in the development of hepatocellular carcinoma in autoimmune hepatitis: implications for follow-up and screening. Hepatology 2008;48:863–870
6. Manns MP, Czaja AJ, Gorham JD, et al. Diagnosis and management of autoimmune hepatitis. Hepatology 2010;51:2193–2213
7. European Association for the Study of the Liver. EASL clinical practice guidelines: autoimmune hepatitis. J Hepatol 2015;63:971–1004
8. Rockey DC, Caldwell SH, Goodman ZD, et al. Liver biopsy. Hepatology 2009;49:1017–1044
9. Czaja AJ. Review article: the prevention and reversal of hepatic fibrosis in autoimmune hepatitis. Aliment Pharmacol Ther 2014;39:385–406
10. 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;38:518–526
11. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317–1325
12. Xu Q, Sheng L, Bao H, et al. Evaluation of transient elastography in assessing liver fibrosis in patients with autoimmune hepatitis. J Gastroenterol Hepatol 2017;32(3):639–644
13. Guo L, Zheng L, Hu L, et al. Transient elastography (FibroScan) performs better than non-invasive markers in assessing liver fibrosis and cirrhosis in autoimmune hepatitis patients. Med Sci Monit 2017;23:5106–5112
14. Hartl J, Denzer U, Ehlken H, et al. Transient elastography in autoimmune hepatitis: timing determines the impact of inflammation and fibrosis. J Hepatol 2016;65(4):769–775
15. Sheptulina A, Shirokova E, Nekrasova T, et al. Platelet count to spleen diameter ratio non-invasively identifies severe fibrosis and cirrhosis in patients with autoimmune hepatitis. J Gastroenterol Hepatol 2016;31(12):1956–1962
16. Wang J, Malik N, Yin M, et al. Magnetic resonance elastography is accurate in detecting advanced fibrosis in autoimmune hepatitis. World J Gastroenterol 2017;23(5):859–868
17. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529–536
18. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol 2005;58(8):859–862
19. Kim JH, Kim MN, Han KH, et al. Clinical application of transient elastography in patients with chronic viral hepatitis receiving antiviral treatment. Liver Int 2015;35:1103–1115
20. Lee HW, Yoo EJ, Kim BK, et al. Prediction of development of liver-related events by transient elastography in hepatitis B patients with complete virological response on antiviral therapy. Am J Gastroenterol 2014;109:1241–1249
21. 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;128:343–450
22. Loomba R, Wolfson T, Ang B, et al. Magnetic resonance elastography predicts advanced fibrosis in patients with nonalcoholic fatty liver disease: a prospective study. Hepatology 2014;60:1920–1928
23. Cui J, Heba E, Hernandez C, et al. Magnetic resonance elastography is superior to acoustic radiation force impulse for the diagnosis of fibrosis in patients with biopsy-proven nonalcoholic fatty liver disease: a prospective study. Hepatology 2016;63:453–461
24. Venkatesh SK, Wang G, Lim SG, et al. Magnetic resonance elastography for the detection and staging of liver fibrosis in chronic hepatitis B. Eur Radiol 2014;24:70–78
25. Huwart L, Sempoux C, Vicaut E, et al. Magnetic resonance elastography for the noninvasive staging of liver fibrosis. Gastroenterology 2008;135:32–40
26. Xiao G, Yang J, Yan L. Comparison of diagnostic accuracy of aspartate aminotransferase to platelet ratio index and fibrosis-4 index for detecting liver fibrosis in adult patients with chronic hepatitis B virus infection: a systemic review and meta-analysis. Hepatology 2015;61:292–302
27. Xiao G, Zhu S, Xiao X, et al. Comparison of laboratory tests, ultrasound, or magnetic resonance elastography to detect fibrosis in patients with nonalcoholic fatty liver disease: a meta-analysis. Hepatology 2017;66(5):1486–1501
28. de Oliveira AC, El-Bacha I, Vianna MV, et al. Utility and limitations of APRI and FIB4 to predict staging in a cohort of nonselected outpatients with hepatitis C. Ann Hepatol 2016;15:326–332