Original research

Diagnostic accuracy of non-invasive tests for advanced fibrosis in patients with NAFLD: an individual patient data meta-analysis


ABSTRACT

Objective Liver biopsy is still needed for fibrosis staging in many patients with non-alcoholic fatty liver disease. The aims of this study were to evaluate the individual diagnostic performance of liver stiffness measurement by vibration controlled transient elastography (LSM-VCTE), Fibrosis-4 Index (Fib-4) and NAFLD (non-alcoholic fatty liver disease) Fibrosis Score (NFS) and to derive diagnostic strategies that could reduce the need for liver biopsies.

Design Individual patient data meta-analysis of studies evaluating LSM-VCTE against liver histology was conducted. Fib-4 and NFS were computed where possible. Sensitivity, specificity and area under the receiver operating curve (AUROC) were calculated. Biomarkers were assessed individually and in sequential combinations.

Results Data were included from 37 primary studies (n=5735; 45% women; median age: 54 years; median body mass index: 30 kg/m²; 33% had type 2 diabetes; 30% had advanced fibrosis). AUROCs of individual LSM-VCTE, Fib-4 and NFS for advanced fibrosis were 0.85, 0.76 and 0.73. Sequential combination of Fib-4 cut-offs (<1.3; ≥2.67) followed by LSM-VCTE cut-offs (<8.0; ≥20.0 kPa) to rule-in or rule-out advanced fibrosis had sensitivity and specificity (95% CI) of 66% (63–68) and 86% (84–87) with 33% needing a biopsy to establish a final diagnosis. Fib-4 cut-offs (<1.3; ≥3.48) followed by LSM cut-offs (<8.0; ≥20.0 kPa) to rule out advanced fibrosis or rule in cirrhosis had a sensitivity of 38% (37–39) and specificity of 90% (89–91) with 19% needing biopsy.

Conclusion Sequential combinations of markers with a lower cut-off to rule-out advanced fibrosis and a higher cut-off to rule-in cirrhosis can reduce the need for liver biopsies.

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is the hepatic manifestation of the metabolic syndrome with high prevalence worldwide. Most patients remain asymptomatic for long periods of time (years/decades) with slowly progressive disease, but
What are the new findings?

- Existing non-invasive test cut-offs are validated for their use as screening biomarkers to rule out advanced fibrosis in a study group of 5735 patients.
- The sequential combination of Fibrosis-4 Index (FIB-4) (<1.3; ≥2.67) and liver stiffness measurement by vibration controlled transient elastography (LSM-VCTE) (<8.0 kPa; ≥10.0 kPa) which is increasingly used in routine practice has a false negative rate of 9% for advanced fibrosis.
- The diagnostic performance of LSM-VCTE for advanced fibrosis is influenced by biopsy quality, body mass index and presence of type 2 diabetes.
- An algorithm combining FIB-4 and LSM-VCTE sequentially with lower cut-offs to rule out advanced fibrosis (FIB-4 <1.3; LSM-VCTE <8.0 kPa) and with upper cut-offs to rule-in and positively diagnose cirrhosis without the need for liver biopsy with specificity of 95% (FIB-4 ≥3.48; LSM-VCTE ≥20.0 kPa) or 98% (FIB-4 ≥4.63; LSM-VCTE ≥28.0 kPa) can reduce the need for liver biopsies from 33% to 19% or 24%, respectively.

How might it impact on clinical practice in the foreseeable future?

- The non-invasive test cut-offs for the diagnosis of cirrhosis can be incorporated into clinical practice as they have been validated in a large group of patients.
- Application of these cut-offs can lead to a decrease in the need for liver biopsies in secondary care.

The diagnostic performance of LSM-VCTE for advanced liver fibrosis can be accompanied by varying degrees of fibrosis. The currently accepted reference standard for diagnosing NAFLD is liver biopsy with specificity of 95% (FIB-4 ≥3.48; LSM-VCTE ≥20.0 kPa) or 98% (FIB-4 ≥4.63; LSM-VCTE ≥28.0 kPa) can reduce the need for liver biopsies from 33% to 19% or 24%, respectively.

A large body of evidence also exists on how non-invasive tests (NITs) could be used to risk-stratify patients for the presence of advanced fibrosis. These approaches usually involve sequential application of two NITs, with the first tier of a simple, inexpensive, serum-based test performed in the community (eg, Fibrosis-4 Index (FIB-4) or NAFLD Fibrosis Score (NFS)), followed by a second tier of liver stiffness measurement (LSM) (eg, vibration controlled transient elastography: VCTE), or a proprietary serum-based test (eg, enhanced liver fibrosis test; ELF). A lower and an upper threshold are usually used in each tier of testing to rule out (those with a NIT result less than the lower threshold) or rule in (those with a NIT result more than the upper threshold) patients at high risk of advanced fibrosis. Patients with indeterminate results in both tiers of testing would need a liver biopsy for risk stratification. The main value of these approaches lies in their high negative predictive value to rule out patients with low risk of advanced fibrosis who can be safely managed in primary care.

Despite the increasing evidence to support these approaches, some aspects of their application require further clarifications. First, there is no consensus on which NIT thresholds to use for this purpose. For example, FIB-4 upper cut-offs of 3.25 and 2.67 have been described, while other investigators omit the FIB-4 upper cut-off altogether. There is also some uncertainty about the performance of NITs in specific patient subgroups, such as those with type 2 diabetes or obesity. Furthermore, for patients who are ruled in as being at high risk of advanced fibrosis (F3–4), liver biopsy is often needed to identify those with cirrhosis who would need surveillance for HCC. Developing approaches that can minimise the need for liver biopsy in secondary care is therefore an area of untapped need.

To address these problems, we conducted an individual patient data meta-analysis (IPDMA) with three main aims: (1) to evaluate the performance of LSM-VCTE and compare it to the performance of FIB-4 and NFS as screening tests to rule out advanced fibrosis; (2) to evaluate NIT combination strategies to minimise the number of cases that would need a liver biopsy in secondary care; (3) to explore factors that influence diagnostic accuracy.

METHODS

This IPDMA was reported in accordance with the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-IPD Statement and was registered as PROSPERO CRD42019157661.

Criteria for considering studies for the IPD meta-analysis

Patients

Studies reporting data on adults (≥18 years) with NAFLD and paired liver histology and LSM-VCTE were eligible. When studies reported study groups of participants with unselected aetiologies, only IPD of those with NAFLD were sought.

Index tests

The index test of main interest was LSM-VCTE performed with FibroScan (Echosens, France). Results for serum-based biomarkers NFS, FIB-4, aspartate aminotransferase (AST) to alanine aminotransferase (ALT) ratio and AST-to-platelet ratio index (APRI) were also computed where data were available. Online supplemental table 1 summarises the definition of NITs considered in this IPDMA.

Universally accepted cut-offs for diagnosing different groups of fibrosis stages do not exist (several suggested cut-offs are presented in online supplemental table 2). For LSM-VCTE, <7.9 kPa and ≥9.6 kPa are the most used for respectively ruling out and in, advanced fibrosis.

Reference standard

Only studies reporting histological classification of liver fibrosis based on the non-alcoholic steatohepatitis Clinical Research Network (NASH CRN) staging system were considered.

Target conditions

Advanced fibrosis (F3–4) and cirrhosis (F4) were the target conditions of interest. To fulfil the aims of the study, cut-offs...
were selected to rule out or rule in advanced fibrosis, and to rule out advanced fibrosis or rule in cirrhosis.

Study design
All study designs were considered if they were reporting on patients with NAFLD undergoing both liver biopsy and LSM-VCTE within 6 months. No language restrictions were applied.

Establishing collaborations
Authors of eligible studies were contacted by email and reminders were sent if a response was not received within 2 weeks. Only data from studies that received ethical approval were used. Additional ethical approval was not sought for the meta-analysis as only anonymised data were provided.

Data verification
Range checks of measurement values provided for individual patients were carried out and authors were asked to provide clarifications where necessary. Missing data were queried until received or confirmed as unavailable. Missing data were handled in the analysis by pairwise deletion.

LSM-VCTE with median stiffness ≥7.1 kPa and IQR-to-median LSM ratio >30% were considered unreliable. These were included in the main analysis and were later compared in a subgroup analysis to reliable measurements, to assess whether they can be reliably used to diagnose advanced fibrosis.

Authors were provided with a template table of required data (online supplemental table 3) and were asked to deduplicate data were possible. We also checked for duplicate entries and where identified these were removed.

Data analysis
Quality and bias assessment
The quality of studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2).

IPD meta-analysis
The original data sets were merged, a study identification variable was added, and descriptive statistical analysis of the data sets was conducted. Dichotomous variables are displayed as percentages. Continuous variables are reported as means with SD, or medians with IQRs according to the distribution of the data.

Analyses were done per protocol, as we did not have information on failed LSM-VCTE. To express the diagnostic performance of NITs, non-parametric, empirical receiver operating characteristic (ROC) curves were constructed for the target conditions of interest. Diagnostic performance was expressed as the area under the ROC curve (AUROC) with 95% CI, based on De Long’s method. AUROCs were compared using De Long’s test statistic.

Thresholds to maximise the Youden index (ie, sensitivity+specificity−1), for 90% sensitivity, and for 90% specificity were reported. The diagnostic performance of previously published cut-offs was also evaluated. Sequential combinations of serum biomarkers and LSM-VCTE were evaluated, by computing sensitivity, specificity and proportions of misclassified and indeterminate patients.

Positive and negative predictive values (PPV and NPV) were estimated for prevalences within the range of those reported in the original studies. The number of false positive and false negative results for 100 theoretical cases was also reported.

The main analysis was conducted to maximise data for each NIT. For a valid comparison of the performance of NITs, a separate analysis was conducted in the subgroup of patients where all three of VCTE, FIB-4 and NFS were available in each participant.

To fulfil the aim of developing testing strategies that reduce the number of patients in need of a liver biopsy, lower cut-offs for ruling out advanced fibrosis and upper cut-offs for ruling in cirrhosis were used. The rationale for this approach is illustrated in the online supplemental figure 1. The upper cut-offs for identifying cirrhosis were chosen at 95% and 98% specificity in a derivation set and tested in a validation set. Derivation and validation sets were obtained by random sampling from the IPD study group in a 3:2 ratio. These upper cut-offs were combined with lower cut-offs from the literature for ruling out advanced fibrosis and the algorithm was tested in the whole IPD study group. For ease of reference, we also examined the cut-offs of 8 kPa and 10 kPa (corresponding to the most common VCTE cut-offs in the literature of 7.9 kPa and 9.6 kPa rounded to the nearest integer) and also rounded our cirrhosis cut-offs to the nearest integer to facilitate application in clinical practice.

Only test-positive and test-negative patients were included in the calculation of diagnostic performance indices, and patients in the indeterminate group were excluded from calculations.

Subgroup analysis was performed according to biopsy length (<20 mm, ≥20 mm), number of portal tracts in biopsy samples (<11, ≥11), biopsy quality (intermediate: 10 mm ≤length <20 mm; high: length ≥20 mm and ≥11 tracts), age (four quartiles), sex, body mass index (BMI; BMI <25 kg/m², 25 kg/m² ≤BMI <30 kg/m², BMI ≥30 kg/m²), presence of type 2 diabetes mellitus (T2DM), continent of provenance (Europe, Asia), probes used (M, XL), reliability criteria for LSM-VCTE (reliable (median LSM <7.1 kPa or median LSM ≥7.1 kPa and IQR/median LSM <0.30) vs unreliable (median LSM ≥7.1 kPa and IQR/median LSM ≥0.30)), and aminotransferase levels (ALT or AST<40, 40≤ALT or AST<100, ALT or AST≥100; ALT<40 and AST<40, ALT≥40 or AST≥40).

All statistical analyses were performed using R (V.1.2.1335, R Foundation for Statistical Computing, Vienna, Austria) with the pROC package. V95% CIs were calculated using 500 stratified bootstrap replicates using the boot package.

VCTE probe types
The analysis to account for probe type is described in the online supplemental materials.

Patient and public involvement
Patients and the public were not involved in the conduct of this study as there was no direct patient participation in the study.

RESULTS Search process and data collection
Ten thousand three hundred ninety-two articles were identified in a search performed for a larger systematic review evaluating the diagnostic performance of LSM-VCTE and other index tests for the staging of fibrosis and diagnosis of NASH in adult patients with NAFLD. After removing duplicates, and screening titles, abstracts, and full texts, 59 studies examining VCTE were identified.

Authors of 37 studies shared useable data (figure 2). Authors of more than one study supplied data in a single dataset and, overall, we received 30 data sets including data from 6371 patients. After removing duplicates (n=628) and patients with

Hepatology

65% of studies, as these either excluded technical failures from domain were judged to have high risk or unclear risk of bias in 2 and 3. Only one study had low risk of bias or low applicability concerns.

**Study quality**
The methodological quality of the studies assessed with the QUADAS-2 tool is summarised in online supplemental figures 2 and 3. Only one study had low risk of bias or low applicability concerns in all QUADAS-2 domains.28 The flow and timing domain were judged to have high risk or unclear risk of bias in 65% of studies, as these either excluded technical failures from their final diagnostic performance analysis or did not report them.

**Validating the diagnostic performance of LSM by VCTE and serum-based tests for detecting advanced fibrosis**
LSM-VCTE, FIB-4, NFS, APRI and AST/ALT ratio had corresponding AUROCs of 0.85, 0.76, 0.73, 0.70, 0.64 for identifying advanced fibrosis (table 3), and 0.90, 0.80, 0.78, 0.72, 0.69 for the identification of cirrhosis (online supplemental table 6). LSM-VCTE performed significantly better (p<10⁻¹⁵) in detecting both advanced fibrosis and cirrhosis than all serum-based tests. This relationship was preserved when performing a head-to-head comparison of LSM-VCTE, FIB-4 and NFS in the same group of patients (online supplemental tables 7 and 8).

When considering cut-offs from the literature, we evaluated lower and higher cut-offs separately. For any given test, as would be expected, low thresholds yielded higher sensitivity and high thresholds were associated with higher specificity (online supplemental table 9). Indicative PPV and NPV are also provided for the range of prevalences (5%–50%) reported in the primary study (online supplemental tables 10–14).

APRI and AST/ALT ratio had only modest diagnostic performance for advanced fibrosis (AUROC ≤0.70, table 3), and were therefore not considered further.

None of the thresholds regarded in isolation resulted in both a high sensitivity (≥80%) and high specificity (≥80%) (figure 2, table 3, online supplemental tables 9 and 15, and online supplemental figure 4). Therefore, we explored the use of a lower and an upper cut-off. LSM-VCTE literature cut-offs performed well in only two cases (<7.1 kPa and ≥14.1 kPa: 83% sensitivity, 90% specificity; and <7.9 kPa and ≥9.6 kPa: 84% sensitivity, 78% specificity), while for other LSM-VCTE, NFS and FIB-4 thresholds a high specificity was observed (FIB-4: 91% for <1.3 and ≥2.67, 95% for <1.3, ≥3.25) but sensitivity was <60% (table 4). In addition, the proportion of indeterminate cases was >30% for serum-based NTIs. Threshold pairs derived from the IPD study group did not reduce the proportion of misclassified and indeterminate patients seen with literature-based threshold pairs (table 4).

We further evaluated the performance of LSM-VCTE, FIB-4 and NFS to diagnose advanced fibrosis in sequential combinations of serum-based NTIs and LSM-VCTE. When selecting threshold combinations for FIB-4 and NFS available in the literature (<1.3 & ≥2.67, <1.3 & ≥3.25 for FIB-4; <−1.455 & ≥0.676 for NFS) and pairing them with the best threshold pair for LSM-VCTE (<7.9 kPa & ≥9.6 kPa, identified as the one with highest sensitivity and lowest indeterminate proportion), the proportion of patients in the indeterminate group was 5%. While both the FIB-4+LSM VCTE and NFS+LSM VCTE sequential combinations had specificity >80%, their sensitivity was ≤80% (table 5). A better sensitivity was reached by using thresholds derived from the IPD study group (<0.88 & ≥2.31 for FIB-4; <−2.55 & ≥0.28 for NFS), but the proportion of indeterminate cases was near 20% in those cases and the proportions of patients needing LSM-VCTE was also larger than when using literature cut-offs (table 5).

**Algorithms to minimise the need for liver biopsy**
In the derivation set, the cut-offs for 95% and 98% specificity for the diagnosis of cirrhosis were respectively 20.4 kPa and 27.6 kPa for LSM-VCTE, 3.48 and 4.63 for FIB-4 and 1.01 and 1.57 for NFS. These cut-offs performed similarly in the validation set (online supplemental tables 16 and 17).

---

**Study and population characteristics**
The characteristics of the 30 data sets are summarised in table 1. Studies were conducted in Europe (67%), Asia (40%) and Australia (3%). Data availability is shown in online supplemental table 3. FIB-4 and NFS were determined in 5393 (94%) and 3248 (57%) patients, respectively. Median age was 54 years, 2570 (45%) patients were women, 33% had diabetes and 43% had BMI ≥30 kg/m². Overall, 30% had advanced fibrosis and 11% had cirrhosis. Details of the IPD study group are included in table 2, and online supplemental tables 4 and 5.

**Missing biopsy (n=14) or LSM-VCTE (n=194) data, the final dataset consisted of 5735 unique patients.**

**Figure 1** Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart illustrating the identification and selection process for studies finally included in this individual patient data meta-analysis. IPD, individual patient data; LSM, liver stiffness measurement; VCTE, vibration controlled transient elastography.
<table>
<thead>
<tr>
<th>Data set ID</th>
<th>Country</th>
<th>Study design</th>
<th>Number of participants (n)</th>
<th>Age (year)</th>
<th>BMI (kg/m²)</th>
<th>WC (cm)</th>
<th>M/F</th>
<th>Recruitment interval</th>
<th>Hardware used</th>
<th>Probe used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agrawal et al</td>
<td>UK</td>
<td>MC, P, CS</td>
<td>25</td>
<td>47.8 (19–70)</td>
<td>27.7 (15.8–35.7)</td>
<td>95.4 (39–120)</td>
<td>18/7</td>
<td>2009–2012</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Akyut et al</td>
<td>Turkey</td>
<td>SC, P, CS</td>
<td>88</td>
<td>46.0 (24–62)</td>
<td>30.3 (18.3–41.8)</td>
<td>101.5 (70–143)</td>
<td>50/38</td>
<td>–</td>
<td>FibroScan 502 Touch</td>
<td>M</td>
</tr>
<tr>
<td>Boursier et al</td>
<td>France</td>
<td>MC, P, CS</td>
<td>1063</td>
<td>56.1 (18–83)</td>
<td>31.6 (16.3–55.5)</td>
<td>108.3 (58–174)</td>
<td>613/450</td>
<td>–</td>
<td>–</td>
<td>M or XL</td>
</tr>
<tr>
<td>Cassinotto et al</td>
<td>France</td>
<td>SC, P, CS</td>
<td>61</td>
<td>55.9 (22–81)</td>
<td>30.1 (16.2–46.6)</td>
<td>103.6 (72–125)</td>
<td>40/21</td>
<td>2010–2012</td>
<td>–</td>
<td>M and XL</td>
</tr>
<tr>
<td>Cassinotto et al</td>
<td>France</td>
<td>MC, P, CS</td>
<td>286</td>
<td>56.6 (18–80)</td>
<td>32.2 (20.3–57.4)</td>
<td>109.8 (68–168)</td>
<td>171/115</td>
<td>2011–2015</td>
<td>–</td>
<td>M and XL</td>
</tr>
<tr>
<td>Chan et al</td>
<td>Malaysia</td>
<td>SC, P, CS</td>
<td>146</td>
<td>50.4 (18–73)</td>
<td>29.4 (6.9–41.2)</td>
<td>98.3 (79–127)</td>
<td>80/66</td>
<td>2012–2013</td>
<td>–</td>
<td>FibroScan 502 Touch</td>
</tr>
<tr>
<td>Chan et al</td>
<td>Malaysia, Hong Kong</td>
<td>MC, P, CC</td>
<td>153</td>
<td>54.0 (24–76)</td>
<td>29.9 (20.1–44.8)</td>
<td>98.4 (69–141)</td>
<td>68/85</td>
<td>–</td>
<td>FibroScan 502 Touch</td>
<td>M and XL</td>
</tr>
<tr>
<td>Eddowes et al</td>
<td>UK</td>
<td>MC, P, CS</td>
<td>358</td>
<td>53.3 (19–77)</td>
<td>34.2 (19.5–53.2)</td>
<td>117.2 (65–158)</td>
<td>206/152</td>
<td>–</td>
<td>–</td>
<td>M or XL</td>
</tr>
<tr>
<td>Eddowes et al</td>
<td>UK</td>
<td>MC, P, CS</td>
<td>50</td>
<td>50.2 (18–73)</td>
<td>33.6 (23.6–47.8)</td>
<td>109.4 (89–132)</td>
<td>28/22</td>
<td>2014–2015</td>
<td>–</td>
<td>M or XL</td>
</tr>
<tr>
<td>Gaia et al</td>
<td>Italy</td>
<td>SC, P, CS</td>
<td>68</td>
<td>48.6 (29–63)</td>
<td>28.0 (21.2–40.2)</td>
<td>100.2 (56–124)</td>
<td>48/20</td>
<td>2007–2009</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Garg et al</td>
<td>India</td>
<td>SC, P, CS</td>
<td>76</td>
<td>38.2 (20–65)</td>
<td>45.2 (32.7–73.8)</td>
<td>90.6 (43–128)</td>
<td>16/60</td>
<td>2014–2016</td>
<td>FibroScan 502 XL Touch</td>
<td></td>
</tr>
<tr>
<td>Karls et al</td>
<td>Germany</td>
<td>SC, P, CS</td>
<td>41</td>
<td>45.7 (28–64)</td>
<td>47.7 (33.7–60.1)</td>
<td>112.1 (65–158)</td>
<td>–</td>
<td>13/28</td>
<td>–</td>
<td>FibroScan 502 XL</td>
</tr>
<tr>
<td>Labenz et al</td>
<td>Germany</td>
<td>SC, P, CS</td>
<td>126</td>
<td>47.4 (20–73)</td>
<td>31.6 (23.5–50.4)</td>
<td>109.4 (69–132)</td>
<td>28/22</td>
<td>2014–2015</td>
<td>–</td>
<td>M or XL</td>
</tr>
<tr>
<td>Lee et al</td>
<td>Korea</td>
<td>SC, P, CS</td>
<td>94</td>
<td>55.5 (19–82)</td>
<td>27.2 (19.1–36.3)</td>
<td>100.2 (56–124)</td>
<td>48/20</td>
<td>2007–2009</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Lupstor et al</td>
<td>Romania</td>
<td>SC, P, CS</td>
<td>72</td>
<td>42.4 (20–69)</td>
<td>28.7 (21.0–41.5)</td>
<td>102.4 (60–124)</td>
<td>51/1</td>
<td>2007–2009</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Mahadeva et al</td>
<td>Malaysia</td>
<td>SC, P, CS</td>
<td>131</td>
<td>49.9 (23–73)</td>
<td>28.7 (18.5–43.1)</td>
<td>93.5 (43–128)</td>
<td>66/65</td>
<td>2009–2010</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Okajima et al</td>
<td>Japan</td>
<td>SC, P, CS</td>
<td>173</td>
<td>56.3 (18–81)</td>
<td>27.2 (16.5–40.3)</td>
<td>113.5 (69–126)</td>
<td>84/89</td>
<td>2013–2015</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Ooi et al</td>
<td>Australia</td>
<td>MC, P, CS</td>
<td>82</td>
<td>44.5 (18–67)</td>
<td>46.2 (29.1–74.0)</td>
<td>136.5 (101–192)</td>
<td>23/59</td>
<td>2015–2016</td>
<td>–</td>
<td>M or XL</td>
</tr>
<tr>
<td>Pavlides et al</td>
<td>UK</td>
<td>SC, P, CS</td>
<td>70</td>
<td>53.5 (25–77)</td>
<td>34.5 (23.0–57.3)</td>
<td>112.5 (80–149)</td>
<td>42/28</td>
<td>2011–2015</td>
<td>–</td>
<td>M or XL</td>
</tr>
<tr>
<td>Petta et al</td>
<td>Italy</td>
<td>MC, P&amp;R, CS</td>
<td>234</td>
<td>45.5 (15–78)</td>
<td>28.2 (15.7–40.7)</td>
<td>99.4 (69–126)</td>
<td>16/95</td>
<td>2008–2013</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Petta et al</td>
<td>France, Hong Kong, Italy</td>
<td>MC, P, CS</td>
<td>260</td>
<td>54.6 (15–87)</td>
<td>29.4 (16.5–46.6)</td>
<td>100.9 (74–148)</td>
<td>122/138</td>
<td>–</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Petta et al</td>
<td>Italy</td>
<td>MC, P, CS</td>
<td>474</td>
<td>45.5 (19–77)</td>
<td>29.2 (15.2–49.5)</td>
<td>99.6 (47–164)</td>
<td>275/199</td>
<td>–</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Shen et al</td>
<td>China</td>
<td>MC, P, CS</td>
<td>101</td>
<td>59.0 (16–67)</td>
<td>27.0 (20.1–37.3)</td>
<td>92.9 (75–120)</td>
<td>74/27</td>
<td>2012–2014</td>
<td>FibroScan 502 M</td>
<td>M or XL</td>
</tr>
<tr>
<td>Staufner</td>
<td>Austria</td>
<td>MC, P, CS</td>
<td>186</td>
<td>49.6 (19–83)</td>
<td>32.5 (19.0–56.9)</td>
<td>102.0 (71–148)</td>
<td>106/80</td>
<td>2011–2016</td>
<td>FibroScan 502 Touch</td>
<td>M or XL</td>
</tr>
<tr>
<td>Wong et al</td>
<td>Hong Kong, France</td>
<td>MC, P, CS</td>
<td>464</td>
<td>53.8 (20–83)</td>
<td>30.5 (17.3–48.0)</td>
<td>102.0 (71–148)</td>
<td>201/263</td>
<td>2009–2017</td>
<td>–</td>
<td>M and XL</td>
</tr>
<tr>
<td>Wong et al</td>
<td>Hong Kong, France</td>
<td>MC, P, CS</td>
<td>273</td>
<td>51.6 (21–77)</td>
<td>28.8 (16.5–54.0)</td>
<td>96.2 (65–144)</td>
<td>147/126</td>
<td>2003–2009</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Yoneda et al</td>
<td>Japan</td>
<td>MC, P, CS</td>
<td>97</td>
<td>52.1 (19–76)</td>
<td>26.5 (17.3–38.5)</td>
<td>98.9 (47–128)</td>
<td>41/56</td>
<td>&lt;2008</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Younes et al</td>
<td>Italy</td>
<td>SC, P, CS</td>
<td>289</td>
<td>44.8 (15–78)</td>
<td>28.8 (17.5–41.7)</td>
<td>98.9 (47–128)</td>
<td>199/90</td>
<td>–</td>
<td>–</td>
<td>M</td>
</tr>
<tr>
<td>Ziol et al</td>
<td>France</td>
<td>SC, P, CS</td>
<td>13</td>
<td>49.3 (39–60)</td>
<td>29.4 (23.8–34.6)</td>
<td>10/3</td>
<td>2003–2005</td>
<td>–</td>
<td>–</td>
<td>M</td>
</tr>
</tbody>
</table>

– Data not available; BMI, body mass index; CC, case-control; CS, cross-sectional; f, females; M, males; MC, multicentre; P, prospective; R, retrospective; SC, single-centre; WC, waist circumference.

Algorithms combining FIB-4 (lower cut-off of 1.3 as described in the literature and upper cut-offs of 3.48 and 4.63 as described above) and LSM by VCTE (lower cut-off rounded to 8.0 kPa and upper cut-offs rounded to 20.0 kPa and 28.0 kPa, as described above) were then compared with the traditional way of applying these tests, also with rounded cut-offs for LSM by VCTE (8 kPa and 10 kPa) (figure 3). This approach increased the number of patients requiring a LSM.
Subgroup and sensitivity analyses

In subgroup analysis for the diagnosis of advanced fibrosis (online supplemental table 19), NITs performed better in patients with lower BMI (AUROCs LSM-VCTE: 0.91, p<0.005; FIB-4: 0.81, p<0.001; NFS: 0.76, p<0.025), without T2DM (LSM-VCTE: 0.87, p<10^{-6}; FIB-4: 0.77, p<0.01), and with biopsies shorter than 20 mm (LSM-VCTE: 0.87, p<0.005; FIB-4: 0.80, p<0.001; NFS: 0.79, p<0.05), or with fewer than 11 portal tracts (LSM-VCTE: 0.86, p=0.01; FIB-4: 0.79, p=0.04; NFS: 0.78, p<0.005).

Diagnostic performance was also lower in patients in the youngest age quartile (<43 years, AUROC: 0.58, p<0.001) and in women (AUROC: 0.71, p<0.005) for NFS, while continent of provenance did not have a significant effect for any NITs. In patients with normal levels of ALT (ALT<40) FIB-4 performed worse (AUROC: 0.73) than in patients with ALT≥40 and ALT<100 (AUROC: 0.77, p<0.01).

Table 3

| Diagnostic performance of non-invasive tests for advanced fibrosis (F3-F4) |
|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| LSM by VCTE (n=5489)     | FIB-4 (n=5393)            | NFS (n=3248)             | APRI (n=5477)            | AST/ALT (n=5434)          |
| Advanced fibrosis, %     | 30                        | 30                       | 29                       | 30                       |
| AUROC                    | 0.85 (0.84-0.86)          | 0.76 (0.74-0.77)         | 0.73 (0.71-0.75)         | 0.70 (0.69-0.72)          | 0.64 (0.62-0.65)          |
| Sensitivity, %           | 90% Se                    | 90% Sp                   | 90% Se                   | 90% Se                   | 90% Sp                   |
| Threshold                | YI                        | 90% Se                   | 90% Sp                   | YI                       | 90% Se                   | 90% Sp                   |
|                      | 9.1 (7.5-79)              | 12.1 (52-57)             | 1.44 (36-41)             | -1.6 (78-62)             | 1.1 (64-69)              | 1.7 (-73-77)             |
| Sensitivity, %           | 77 (89-91)                | 55 (89-91)               | 69 (67-72)               | 90 (88-91)               | 75 (78-72)               | 90 (89-92)               |
| Specificity, %           | 78 (59-61)                | 90 (89-91)               | 70 (69-72)               | 39 (34-40)               | 63 (61-65)               | 90 (89-91)               |
| Misclassified, %         | 22 (22-33)                | 31 (20-21)               | 30 (30-31)               | 46 (46-47)               | 34 (34-46)               | 48 (48-50)               |

For each non-invasive test thresholds were selected according to Youden’s index (YI), and fixed at 90% sensitivity (90% Se) and 90% specificity (90% Sp). 95% CIs were estimated with 500 bootstrap replicates.

ALT, alanine aminotransferase; APRI, AST-to-platelet ratio index; AST, aspartate aminotransferase; FIB-4, Fibrosis-4 Index; LSM, liver stiffness measurement; NFS, NAFLD (non-alcoholic fatty liver disease) Fibrosis Score; VCTE, vibration controlled transient elastography.
Hepatology

performed better in patients with AST<40 (AUROC: 0.76), than in patients with AST≥100 (AUROC: 0.65, p<0.01). FIB-4 performed better in patients with at least one abnormal aminotransferase measurement (AUROC: 0.72, p=0.014). For cirrhosis, the trends were similar, except that for the diagnosis of cirrhosis, LSM by VCTE performed better in the youngest age group (AUROC: 0.97, p<10^{-4}) and NIT diagnostic performance was independent of aminotransferase levels (online supplemental table 20).

The diagnostic performance of LSM-VCTE was significantly lower in patients with unreliable LSms (p<10^{-4}; both for advanced fibrosis and cirrhosis) when applying the Boursier-criteria, but not when only considering IQR/median LSM <0.30. The proportion of unreliable results was 12% both in the advanced fibrosis and cirrhosis groups (online supplemental table 21).

There was no difference in the diagnostic performance of LSM-VCTE between the M and XL probes in the subgroup of patients who had undergone LSM by both probes (online supplemental table 22).

In a sensitivity analysis of patients with LSM matched to BMI (only M probe measurements if BMI <30 kg/m^2 and only XL probe measurements if BMI ≥30 kg/m^2), there was no significant difference between the diagnostic performance of LSM-VCTE when comparing to the entire IPD study group (online supplemental table 23).

Figure 2  Distribution of sensitivities and specificities over the possible threshold ranges for liver stiffness measurement (LSM) by vibration controlled transient elastography (VCTE) (A), Fibrosis-4 Index (FIB-4) (B) and NAFLD (non-alcoholic fatty liver disease) Fibrosis Score (NFS) (C) when considering the diagnosis of advanced fibrosis. Insets show the distribution of cut-offs identified from the literature. Horizontal dashed lines are representing the minimum acceptable criteria for considering a test as having high sensitivity (≥80%) and high specificity (≥80%).
DISCUSSION

Through an extensive collaboration network with authors of primary studies we were able to collect the largest dataset of its kind ever to be reported on. This includes a diverse set of study groups from Europe, Asia, and Australia, 30% of whom had advanced fibrosis. We believe that our findings are therefore relevant for patients typical of secondary care in these territories and may be applied in the development of new strategies or in the consolidation of existing practices in evaluating patients for referral to secondary care.

A few studies evaluated the diagnostic performance of LSM-VCTE and other NITs, but most report on fewer than 500 patients. One similarly large study reported on patients screened for inclusion in clinical trials, where the prevalence of advanced fibrosis was 71%, making it difficult to make generalisations about its applicability in routine practice or compare its results to ours. A smaller study with 1073 patients with NAFLD of whom 29% had advanced fibrosis examined the diagnostic performance of LSM by VCTE. The authors of that study reported AUC and specificity values similar to our findings, however they reported increased sensitivity. Other smaller studies reported similar prevalence of advanced fibrosis and similar AUROCs for LSM-VCTE.

Overall, the diagnostic performance of LSM-VCTE for advanced fibrosis was good (AUROC=0.85), while that of FIB-4 and NFS in the same group was moderate (AUROC=0.76 for FIB-4, AUROC=0.73 for NFS). None of the studied NITs had both sufficiently high sensitivity and specificity (≥80%) when used with single cut-offs. Diagnostic performance was higher for detecting cirrhosis, as reported in previous studies.29,31,33,34 LSM-VCTE had the highest sensitivity and specificity, both in the case of a single cut-off (9.1 kPa obtained by maximising the Youden index; 77% and 78%) and for two cut-offs (<7.4 kPa & ≥12.1 kPa; 83% and specificity of 90%), but with a proportion of 39% of patients ending up with an indeterminate result, similar to 41% indeterminate patients reported in the original paper.31

LSM-VCTE thresholds identified in our study group (<9.1 kPa; <7.4 kPa & ≥12.1 kPa) were similar to thresholds reported in the literature (<9.9 kPa; <7.1 kPa & ≥14.1 kPa, <7.9 kPa & ≥9.6 kPa). However, thresholds for FIB-4 (<1.44; <0.88 & ≥2.31) and NFS (<−1.39; <−2.55 & ≥0.28) defined in our IPD study group spanned a wider range than those reported in the literature (<1.3 & ≥2.67 or <1.3 & ≥3.25 for FIB-4; <−1.455 & ≥0.676 for NFS).

Our findings are in line with the existing literature suggesting that sequential combinations of NITs increase sensitivity and specificity.29 Additionally, we have found NFS+LSM VCTE and FIB-4+LSM VCTE combinations to have similar sensitivity and specificity as recently reported by Boursier et al.35 Such combined testing strategies can reduce the number of indeterminate cases and reduce the costs associated with liver biopsies.

Furthermore, we propose an approach that could minimise the need for liver biopsies further, by using upper cut-offs with 95% and 98% specificity for the identification of cirrhosis. The rationale for this approach is explained in the online supplemental discussion. When using the 95% specificity cut-off, the proportion of patients needing liver biopsy decreases from 33% to 19% (figure 3). However, in this approach, 345 of 656 patients ‘ruled-in’ as having cirrhosis do not have histologically diagnosed cirrhosis. While this may seem like a high proportion...
of patients with false positive results, this must be interpreted in the light of two factors. First, the limitations of liver biopsy could mean that these patients are falsely classified as not having cirrhosis histologically. Furthermore, patients without cirrhosis on histology and with high NIT values could have equivalent risks as patients with cirrhosis on histology. For example, it is known from the hepatitis C literature that patients without cirrhosis on liver biopsy but with a high FIB-4 (>3.25) still had a significant risk of developing HCC after hepatitis C treatment, demonstrating that NITs can have added benefit beyond the histological diagnosis of cirrhosis alone. The rate of false positive results for cirrhosis can be decreased by choosing cut-offs with higher specificity, but this will come at the expense of doing more biopsies. Despite this encouraging result, this is an area where more information is needed, particularly longitudinal data comparing the prognostic value of LSM-VCTE and other NITs against histology, and ultimately, the cost effectiveness of the various cut-offs would need to be evaluated.

Surprisingly, subgroup analyses showed that the diagnostic accuracy of NITs was better in cases with poor biopsy quality. This finding is difficult to explain but a similar observation was reported previously in a large group of patients screened for clinical trials. The use of local biopsy reports as reference standard reported previously in a large group of patients screened for clinical trials. The use of local biopsy reports as reference standard. Furthermore, some heterogeneity in the performance of LSM-VCTE application underwent significant changes, initially with the introduction of the XL probe, followed by the advice to measure skin-to-capsule distance (SCD) and the introduction of the Automatic Probe Selection tool. There was therefore some heterogeneity in the performance of LSM-VCTE, with early studies using only the M probe to assess all patients, while only a subset of studies assessed SCD to guide probe selection. Furthermore, one third of the included studies was carried out in France, as the technology used for LSM by VCTE originates from there. Lastly, our data confirm that LSM-VCTE had superior accuracy to serum-based tests, and this is independent of probe type, sex, ALT, AST, and participants’ continent of origin. There was, however, some dependence on the presence of T2DM, BMI and for the detection of cirrhosis, and we did not check for subgroup-specific cut-offs, but these should be explored in future studies.

Our study has several strengths, including the large size of the IPD study group and composition with prevalence of advanced fibrosis of 30%, which makes it relevant to routine practice. Furthermore, the proportion of unreliable VCTE measurements in our study was 12%, in keeping with the literature. However, we acknowledge some limitations. We did not have any data from the USA and very few studies from Australia, so the results could not be globally applicable, due to differences in BMI across study populations. In addition, due to the nature of our study, we had to use the locally provided histology results possibly introducing bias. Furthermore, we covered a large chronological period, during which LSM-VCTE application underwent significant changes, initially with the introduction of the XL probe, followed by the advice to measure skin-to-capsule distance (SCD) and the introduction of the Automatic Probe Selection tool. There was therefore some heterogeneity in the performance of LSM-VCTE, with early studies using only the M probe to assess all patients, while only a subset of studies assessed SCD to guide probe selection. Furthermore, one third of the included studies was carried out in France, as the technology used for LSM by VCTE originates from there. Lastly, our data confirm that LSM-VCTE had superior accuracy to serum-based tests, and this is independent of probe type, sex, ALT, AST, and participants’ continent of origin. There was, however, some dependence on the presence of T2DM, BMI and for the detection of cirrhosis, and we did not check for subgroup-specific cut-offs, but these should be explored in future studies.

Our study examined some of the most widely available NITs. While it cannot be considered exhaustive, it can be regarded as the benchmark against which newer NITs can be tested. This is particularly important as new tests are continuously being developed (FibroTest-FibroSURE, ActiTest, ELF55). Furthermore, newer tests are also needed for patients with ‘at risk’ NASH (NASH-F2–3) who would be candidates for clinical trials or treatments, once approved therapies become available (FAST score, NIS4, cTAG46).

In conclusion, our study provides further validation of the use of sequential combination of FIB-4 and LSM-VCTE to rule out patients with NAFLD and advanced fibrosis who can be managed in primary care. We have shown how the use of lower cut-offs to rule in cirrhosis in combination with lower cut-offs to

### Table 5 Diagnostic performance of combinations of NAFLD (non-alcoholic fatty liver disease) Fibrosis Score (NFS) and liver stiffness measurement (LSM) by vibration controlled transient elastography (VCTE), and Fibrosis-4 Index (FIB-4) and LSM by VCTE tests to diagnose patients with advanced fibrosis

<table>
<thead>
<tr>
<th>Advanced fibrosis, %</th>
<th>FIB-4 &amp; LSM by VCTE</th>
<th>NFS &amp; LSM by VCTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=5159)</td>
<td>(n=3094)</td>
</tr>
<tr>
<td>Thresholds for blood-based NIT</td>
<td>&lt;0.88, ≥2.31*</td>
<td>&lt;2.55, ≥0.28*</td>
</tr>
<tr>
<td>Sensitivity, %</td>
<td>87 (77–88)</td>
<td>77 (74–81)</td>
</tr>
<tr>
<td>Specificity, %</td>
<td>81 (79–83)</td>
<td>83 (81–85)</td>
</tr>
<tr>
<td>PPV, %</td>
<td>91 (90–92)</td>
<td>91 (89–93)</td>
</tr>
<tr>
<td>Indeterminate, %</td>
<td>16 (14–17)</td>
<td>15 (13–17)</td>
</tr>
<tr>
<td>Misclassification, %</td>
<td>18 (17–19)</td>
<td>20 (18–21)</td>
</tr>
<tr>
<td>Patients undergoing LSM by VCTE, %</td>
<td>51 (50–53)</td>
<td>56 (54–59)</td>
</tr>
</tbody>
</table>

95% CIs were estimated with 500 bootstrap replicates.

*Thresholds were determined from the literature. For LSM by VCTE, a threshold pair yielding the highest sensitivity and specificity while having the smallest proportion of indeterminate cases in diagnosing advanced fibrosis was chosen.

NIT, non-invasive test; PPV, positive predictive value; NFS, Fibrosis Score; LSM, liver stiffness measurement; BMI, body mass index; T2DM, type 2 diabetes mellitus; ALT, alanine aminotransferase; AST, aspartate aminotransferase; NASH, non-alcoholic steatohepatitis; SCD, skin-to-capsule distance; M probe, middle transducer; XL probe, extra-large transducer.
rule out advanced fibrosis can lead to a reduction in the number of patients who would need to undergo liver biopsy.

**Author affiliations**

1. Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK
2. Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
3. Translational Gastroenterology Unit, University of Oxford, Oxford, UK
4. NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust and the University of Oxford, Oxford, UK
5. Division of Gastroenterology and Hepatology, Department of Medicine III, Medical University of Vienna, Vienna, Austria
6. Laboratoire HIFIH, UPRES EA 3859, SFR ICAT 4208, Université d’Angers, Angers, Pays de la Loire, France
7. Service d’Hépato-Gastroenterologie et Oncologie Digestive, Centre Hospitalier Universitaire d’Angers, Angers, Pays de la Loire, France
8. Echosens SA, Paris, Ile-de-France, France

---

**Figure 3** Sankey diagrams showing the distribution of patients in true positive, true negative, false positive, false negative and indeterminate groups for a sequential combination of Fibrosis-4 Index (FIB-4) and liver stiffness measurement (LSM) by vibration controlled transient elastography (VCTE) when using different thresholds for each testing tier. A lower threshold was used to rule out patients without advanced fibrosis and an upper threshold ruled in patients with advanced fibrosis when applying both tests (A). In an alternative model, a lower threshold was used to rule out patients without advanced fibrosis, but the upper threshold ruled in only patients with cirrhosis (B, C). Two different pairs of thresholds were chosen for this hybrid strategy: the lower cut-off for both FIB-4 and LSM by VCTE were determined from the literature; upper cut-offs were both determined as corresponding to 95% specificity in detecting cirrhosis (B) or both corresponding to 98% specificity in detecting cirrhosis (C). In the application of the algorithm described in (A) 33% of patients would need to have a liver biopsy for the diagnosis of cirrhosis (those in the indeterminate group to rule out advanced fibrosis and those in the rule in group to identify cirrhosis). With the application of an upper cut-off to rule in cirrhosis without the need of biopsy, only patients in the indeterminate group need to have a biopsy. The latter strategy results in fewer patients undergoing biopsy (18% and 24% depending on the threshold used). Tables next to each panel contain the number and proportion of patients in each of the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) groups for FIB-4 and LSM by VCTE.
Atsushi Nakajima http://orcid.org/0000-0002-6263-1436
Christophe Cassinotto http://orcid.org/0000-0001-5136-4742
Johannes Wiegand http://orcid.org/0000-0001-9233-4064

Hepatology
 cardio-

management of NAFLD is associated with improvement in markers of liver and

analyze and compare ROC curves

stiffness evaluation by transient elastography.

European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO).


4. Spilker TS, Sterling RK, Davison BA, et al. European practice guidelines for the management of non-

2. Wyllie LM, Frisch DR, Segura M, et al. Performance of the aspartate aminotransferase-AST (fast) score for the non-


10. Ekstedt M, Fränzén LE, Mathiesen UL, et al. The NAFLD fibrosis score: a noninvasive system for the diagnosis of liver fibrosis in patients with biopsy-


17. Lin Z-H, Xin Y-N, Dong Q-J, et al. Performance of the aspartate aminotransferase-to-platelet ratio index for the staging of hepatitis C-related fibrosis: an updated meta-


27. Ivanou NV, Feld JJ. What are the benefits of a sustained virologic response to direct-acting antiviral therapy for hepatitis C virus infection? Gastroenterology 2019;156:46–60.


30. Alkayali T, Qutairi L, Kaya E, et al. Clinical utility of noninvasive scores in assessing advanced hepatic fibrosis in patients with type 2 diabetes mellitus: a study in biopsy-


31. Eren F, Kaya E, Yilmaz Y. Accuracy of Fibrosis-4 index and non-alcoholic fatty liver disease fibrosis scores in metabolic (dysfunction) associated fatty liver disease, according to body mass index: failure in the prediction of advanced fibrosis in lean and morbidly obese individuals. Eur J Gastroenterol Hepatol.


36. Harrison SA, Ratuvi Z, Bourjoll J, et al. A blood-based biomarker panel (NiSi4) for non-


Gut fibrosis tests and liver stiffness measurement by FibroScan in non-alcoholic fatty liver disease. *J Hepatol* 2016;65:570–8.


53 Boursier J, Lannes A, Shili S. The new fibrometervcte outperforms recommended liver fibrosis tests in NAFLD. *Hepatology* 2018;68.


