Biomarkers for staging fibrosis and non-alcoholic steatohepatitis in non-alcoholic fatty liver disease (the LITMUS project): a comparative diagnostic accuracy study

Biomarkers for staging fibrosis and non-alcoholic steatohepatitis in non-alcoholic fatty liver disease (the LITMUS project): a comparative diagnostic accuracy study

Yasaman Vali*, Jenny Lee*, Jerome Boursier, Salvatore Petta, Kristy Wonders, Dina Tiniakos, Pierre Bedossa, Andreas Geier, Sven Francque, Mike Allison, Georgios Papatheodoridis, Helena Cortez-Pinto, Raluca Pais, Jean-Francois Dufour, Diana Julie Leeming, Stephen A Harrison, Yu Chen, Jeremy F Cobbold, Michael Pavlides, Adriaan G Holleboom, Hannele Yki-Jarvinen, Javier Crespo, Morten Karsdal, Rachel Ostroff, Mohammad Hadi Zafarmand, Richard Torstenson, Kevin Duffin, Carla Yunis, Clifford Brass, Mattias Ekstedt, Guruprasad P Aithal, Jörn M Schattenberg, Elisabetta Bugianesi, Manuel Romero-Gomez, Vlad Ratziu, Quentin M Anstee†, Patrick M Bossuyt†, on behalf of the Liver Investigation: Testing Marker Utility in Steatohepatitis (LITMUS) consortium investigators‡

Summary

Background. The reference standard for detecting non-alcoholic steatohepatitis (NASH) and staging fibrosis—liver biopsy—is invasive and resource intensive. Non-invasive biomarkers are urgently needed, but few studies have compared these biomarkers in a single cohort. As part of the Liver Investigation: Testing Marker Utility in Steatohepatitis (LITMUS) project, we aimed to evaluate the diagnostic accuracy of 17 biomarkers and multimarker scores in detecting NASH and clinically significant fibrosis in patients with non-alcoholic fatty liver disease (NAFLD) and identify their optimal cutoffs as screening tests in clinical trial recruitment.

Methods. This was a comparative diagnostic accuracy study in people with biopsy-confirmed NAFLD from 13 countries across Europe, recruited between Jan 6, 2010, and Dec 29, 2017, from the LITMUS metacohort of the prospective European NAFLD Registry. Adults (aged ≥18 years) with paired liver biopsy and serum samples were eligible; those with excessive alcohol consumption or evidence of other chronic liver diseases were excluded. The diagnostic accuracy of the biomarkers was expressed as the area under the receiver operating characteristic curve (AUC) with liver histology as the reference standard and compared with the fibrosis-4 index for liver fibrosis (FIB-4) in the same subgroup. Target conditions were the presence of NASH with clinically significant fibrosis (ie, at-risk NASH; NAFLD Activity Score ≥4 and F≥2) or the presence of advanced fibrosis (F≥3), analysed in all participants with complete data. We identified thresholds for each biomarker for reducing the number of biopsybased screen failures when recruiting people with both NASH and clinically significant fibrosis for future trials.

Findings. Of 1430 participants with NAFLD in the LITMUS metacohort with serum samples, 966 (403 women and 563 men) were included after all exclusion criteria had been applied. 335 (35%) of 966 participants had biopsy-confirmed NASH and clinically significant fibrosis and 271 (28%) had advanced fibrosis. For people with NASH and clinically significant fibrosis, no single biomarker or multimarker score significantly reached the predefined AUC 0·80 acceptability threshold (AUCs ranging from 0·61 [95% CI 0·54–0·67] for FibroScan controlled attenuation parameter to 0·81 [0·75–0·86] for SomaSignal), with accuracy mostly similar to FIB-4. Regarding detection of advanced fibrosis, SomaSignal (AUC 0·90 [95% CI 0·86–0·94]), ADAPT (0·85 [0·81–0·89]), and FibroScan liver stiffness measurement (0·83 [0·80–0·86]) reached acceptable accuracy. With 11 of 17 markers, histological screen failure rates could be reduced to 33% in trials if only people who were marker positive had a biopsy for evaluating eligibility. The best screening performance for NASH and clinically significant fibrosis was observed for SomaSignal (number needed to test [NNT] to find one true positive was four [95% CI 4–5]), then ADAPT (six [5–7]), MACK-3 (seven [6–8]), and PRO-C3 (nine [7–11]).

Interpretation. None of the single markers or multimarker scores achieved the predefined acceptable AUC for replacing biopsy in detecting people with both NASH and clinically significant fibrosis. However, several biomarkers could be applied in a prescreening strategy in clinical trial recruitment. The performance of promising markers will be further evaluated in the ongoing prospective LITMUS study cohort.

Introduction Non-alcoholic fatty liver disease (NAFLD), a leading cause of chronic liver disease, spans a histological spectrum from steatosis to non-alcoholic steatohepatitis (NASH) with progressive hepatic fibrosis, leading to cirrhosis or hepatocellular carcinoma in a subset of people.1 This progressive disorder is predicted to become more prevalent in the next decade, consuming substantial health-care resources and becoming an increasing public health challenge.1 Despite its high prevalence, accurate diagnosis and provision of effective management of NAFLD remain challenging, with generally poor public health readiness internationally. This failure is partly due to a lack of clarity on biomarker performance in identifying patients with NASH at an increased risk of disease progression and poor clinical outcomes (ie, at-risk NASH, defined as those patients with both NASH and clinically significant fibrosis [NAS≥4 and F≥2]), deterring their implementation by clinicians.2

People with NAFLD and advanced fibrosis (F≥3) are at increased risk of adverse liver-related outcomes, liver transplantation, and death.3 International guidelines recommend assessment of NAFLD for early identification of high stages of liver fibrosis (F≥3).4 In the absence of approved pharmacological therapies, identification of people with NASH and clinically significant fibrosis (F≥2) is essential to support recruitment into therapeutic clinical trials. The current reference standard for detecting NASH and staging fibrosis is liver histology. Liver biopsy sampling is invasive, resource-intensive, prone to sampling error, and has a small but appreciable risk of complications.5 Despite debates regarding its limitations, participants recruited to NAFLD trials require biopsy to qualify for enrolment. There is, therefore, an urgent need for non-invasive biomarkers to support clinical care and facilitate the evaluation of new therapies. Several non-invasive biomarkers have been proposed, with variable performance in detecting fibrosis and NASH. Few studies have compared these biomarkers in a single cohort to evaluate their relative performances. The aim of this comparative diagnostic accuracy study in individuals with biopsy-confirmed NAFLD, part of the Liver Investigation: Testing Marker Utility in Steatohepatitis (LITMUS) consortium,6 is to evaluate the performance of 17 non-invasive single biomarkers, multimarker scores, and vibration-controlled transient elastography (VCTE) in identifying people with both NASH and clinically significant fibrosis and those with advanced fibrosis, with liver biopsy as the reference standard. We aimed to increase the efficiency of future drug trial enrolment by selecting thresholds for each marker that lead to an acceptable screen failure rate in identifying eligible participants.

Methods

Study design and participants Data were collected in the LITMUS metacohort section of the European NAFLD Registry, a cohort of individuals with NAFLD prospectively recruited via standardised procedures and monitoring (details of which have been published previously), by multiple collaborators.7 Briefly, the European NAFLD Registry recruited patients with suspected NAFLD from secondary and tertiary care centres in Austria, Belgium, Finland, France, Germany, Greece, Italy, the Netherlands, Portugal, Spain, Sweden, Switzerland, and the UK between Jan 6, 2010, and Dec 29, 2017. Participants were required to provide written informed consent before inclusion. Studies contributing to the metacohort were approved by the relevant ethical committees in the participating countries and conform to the guidelines of the Declaration of Helsinki.7 Adults aged 18 years or older with suspected NAFLD and paired liver biopsy and serum samples (obtained within 6 months of one another) were eligible for inclusion in the study group; patients with biopsy-confirmed NAFLD were included in the analysis. All participants had to meet the predefined eligibility criteria of the European NAFLD Registry7 and had a liver biopsy as part of the routine

diagnostic tests for presumed NAFLD, confirming the presence of NAFLD. Most participants had originally been referred for investigation due to abnormal biochemical tests (eg, alanine aminotransferase [ALT] or γ-glutamyltransferase [GGT]) or an ultrasonographically detected bright liver associated with features of the metabolic syndrome. Participants with excessive alcohol consumption (ie, >20–30 g per day) or evidence of other chronic liver diseases, such as hepatitis B or C, were excluded.

Clinical assessment

Detailed clinical data were collected from all participants by a trained investigator and entered directly into a central registry. Sex data were self-reported, with the options "Female" or "Male" provided. BMI was calculated by dividing weight (kg) by height (m) squared. Clinical laboratory blood assays were done in laboratories of the recruitment centres. Lipid (ie, LDL, HDL, cholesterol, and triglyceride) and liver (ie, platelet count, ALT, aspartate aminotransferase [AST], and GGT) profiles were collected. Comorbidities, such as dyslipidaemia (fasting triglyceride concentration ≥150 mg/dL [1·7 mmol/L] or fasting HDL 7·0 mmol/L, or those on treatment for diabetes) were also recorded.

Liver biopsy

Liver biopsy samples were considered to be of adequate size and technical quality to be suitable for clinical diagnosis by the reporting pathologists and were histologically examined locally in each centre by expert liver pathologists prospectively after routine clinical tests.8 NAFLD activity was assessed according to the NASH Clinical Research Network (NASH CRN) scoring system; steatosis and lobular inflammation were scored on four-point scales (0–3) and ballooning was scored on a three-point scale (0–2).9 Liver fibrosis was graded on a five-point scale (0–4), according to Kleiner and colleagues' scoring system.9

Biomarker measurements

All serum samples were collected with standard collection kits and processed locally before storage at –80°C, according to prespecified biobanking standard operating procedures. These were prospectively collected samples that were stored and subsequently analysed. Samples were shipped in batches on dry ice from recruitment sites to the LITMUS Central Biobank, where serum samples were catalogued and subsequently sent for central analysis at Nordic Biosciences (Herlev, Denmark), a laboratory accredited by the College of American Pathologists. Only serum samples collected within 6 months of liver biopsy were eligible for this analysis. All measurements were done by investigators masked to all clinical data associated with the samples. Because of differences in available sample volumes, not all biomarkers could be measured in every participant. The biomarkers measured in the central LITMUS laboratory were CK-18 M30 (M30 Apoptosense ELISA 10011, VLVbio [Nacka, Sweden]), CK-18 M65 (M65 EpiDeath ELISA 10040, VLVbio), PRO-C3 (ELISA-based),10 PRO-C4 (ELISA-based),11 and PRO-C6 (ELISA-based).12 Liver stiffness measurement (LSM) and controlled attenuation parameter (CAP) determined by VCTE (FibroScan, Echosens, Paris, France) within 6 months of liver biopsy were also evaluated. Probe sizes were selected as advised by device guidelines.

Multimarker scores

We used available clinical laboratory data to calculate Fibrosis-4 index for liver fibrosis (FIB-4) using its originally published formula.13 The following previously reported multimarker scores were also calculated: MACK-3 (HOMA-IR, AST, and CK-18 M30);14 a scoring system proposed by Cao and colleagues in 2013 (ALT, platelet count, CK-18 M30, and triglycerides), hereafter referred to as Cao 2013;15 ADAPT (age, platelet count, diabetes, and PRO-C3);16 FIBC3 (age, BMI, diabetes, platelet count, and PRO-C3);17 ABC3D (age, BMI, diabetes, platelet count, and PRO-C3);17 NAFLD Fibrosis Score (NFS; age, BMI, impaired fasting glycaemia or diabetes, ratio of AST to ALT, platelet count, and albumin);18 and APRI (AST to platelet count ratio).19 Enhanced Liver Fibrosis (ELF) test scores, based on hyaluronic acid, tissue inhibitor of matrix

metalloproteinase-1, and aminoterminal propeptide of procollagen type III,20 were measured in the central laboratory with the ADVIA Centaur CP system (Siemens, Munich, Germany). The SomaSignal serum tests for fibrosis, steatosis, inflammation, and ballooning were assayed at the UK SomaLogic facility (Oxford, UK). The SomaSignal NASH tests are modified, aptamer-based, elastic net logistic regression models trained and validated against biopsy results for each component (ie, steatosis [containing 12 protein analytes], lobular inflammation [containing 14 protein analytes], hepatocellular ballooning [containing five protein analytes], and fibrosis [containing eight protein analytes]). The tests were developed as dichotomised protein-phenotype models for clinically relevant severity of steatosis (NAFLD Activity Score [NAS] 0 vs 1–3), hepatocellular ballooning (0 vs 1–2), lobular inflammation (0–1 vs 2–3), and fibrosis (stages 0–1 vs 2–4).21 We multiplied the model probabilities for steatosis, inflammation, and ballooning as a SomaSignal marker for NASH. These 17 markers were selected because their performance in detecting either fibrosis or NASH had been previously reported in the literature, including systematic reviews. Most biomarkers had been developed for their ability to detect advanced fibrosis but, considering the strong pathophysiological links and clinically recognised coexistence of steatohepatitis (the biological driver of liver injury) and fibrogenesis (the chronic wound healing response), we considered it relevant to study a wide range of potential biomarkers in the analysis.

Target conditions

Clinically significant fibrosis was defined as F≥2 and advanced fibrosis was defined as F≥3. The NAS, the sum of the steatosis, lobular inflammation, and ballooning— scored according to the NASH CRN system—ranged from 0 to 8. NASH was defined as the presence of steatosis, lobular inflammation, and hepatocellular ballooning. This definition was operationalised in accordance with standard clinical trial practice as a NAS of 4 or more, with at least one point in each component.22,23 The main target condition was the combination of clinically significant fibrosis and NASH (ie, at-riskNASH). This combination has been defined by health authorities (eg, the US Food and Drug Administration [FDA], the European Medicines Agency [EMA], and the Center for Drug Evaluation and Research) as the crucial inclusion criterion in phase 3 drug development for treatment of non-cirrhotic NASH with liver fibrosis. Furthermore, we evaluated the performance of the biomarkers in detecting presence of advanced fibrosis (ie, F≥3 independent of the presence of histologically defined NASH).

Statistical analysis

Non-parametric, empirical receiver operating characteristic (ROC) curves were constructed for each biomarker and multimarker score. Diagnostic accuracy was expressed through the area under the ROC curve (AUC) with its 95% CI, calculated with the DeLong method.24 To be considered as a diagnostic marker of acceptable accuracy, an AUC of at least 0·80 in detecting both NASH and clinically significant fibrosis was expected. Recruiting 966 participants, of whom 35% were anticipated to have the target condition, provided at least 80% power to reject the null hypothesis that the AUC does not exceed the minimally acceptable value of 0·80 if the actual AUC is 0·85 or more, and at least 99% power if the actual AUC is 0·87 or more, at a 5% type I error rate.25 The performance of the FIB-4 score was used for comparison. Because of the absence of any existing validated non-invasive test for NASH and the high collinearity between NASH and fibrosis stage, performance of FIB-4, a widely used simple fibrosis test, was used as a comparator in all target conditions. To account for differences in the subgroups of patients for which marker results were available, we recalculated the AUC for FIB-4 in each of the corresponding marker subgroups. In an additional head-to-head direct comparison, we evaluated the performance of the most clinically available biomarkers and scores (ie, PRO-C3, CK-18 M30 and M65, ELF, NFS, APRI, ADAPT, FIBC3, ABC3D, and FIB-4) in a subgroup in which results for all ten biomarkers and scores were available. We also evaluated the subgroupspecific performance of all markers according to diabetes status (R-package ROCnReg). We aimed to identify an optimal cutoff level for each biomarker as a screening test to identify people with both NASH and fibrosis. The optimal cutoff would allow a hypothetical trial enrolment screen failure rate (based on

liver biopsy) not exceeding 33%. This value was selected on the basis of a survey conducted among clinicians and drug developers in the LITMUS consortium. The biopsy screen failure rate corresponds to 1 minus the positive predictive value (PPV). Because of the differences between the subgroups in which biomarker results were available, we calculated the minimally required likelihood ratio for a positive biomarker result to provide the corresponding screen failure rate and PPV on the basis of a single measure of the prevalence using the formula

LR = × 1–prev PPV 1–PPV

in which prev denotes the prevalence of NASH and clinically significant fibrosis in the population being tested. On the basis of the proportion of people with NASH and clinically significant fibrosis in the LITMUS metacohort study group, we estimated the prevalence to be 35%. To achieve a screen failure rate not exceeding 33%, the likelihood ratio of a positive biomarker result would then have to be at least 3·77. Of all positivity thresholds with a likelihood ratio exceeding 3·77, we selected the one with the highest sensitivity, thereby maximising efficiency at the preselected acceptable screen failure rate. We report sensitivity, specificity, proportion of positive biomarker results (at 35% prevalence), true positive fraction (ie, proportion of potential study participants with a biomarker-positive result who have fibrosis and NASH at biopsy), and number needed to test to find one eligible trial participant after liver biopsy in individuals with a biomarker-positive result (ie, the inverse of the true positive fraction). 95% CIs were based on 10000 bootstrap samples. All statistical analyses were done with R version 4.2.1. This Article is reported according to the Standards for the Reporting of Diagnostic Accuracy studies guidelines.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Of the 1430 participants with serum samples in the LITMUS metacohort, 1279 had complete biopsy data and were eligible for our study. Of these, 1018 had serum and biopsy samples taken within 6 months of one another (mean time interval between biopsy and blood sampling was 2 days [SD 41·8]). Data from 966 participants in the LITMUS metacohort were included in this analysis after all exclusion criteria had been applied (figure 1). 335 (35%) of 996 participants had NASH and clinically significant fibrosis and 271 (28%) of 966 had biopsy-confirmed advanced fibrosis. Mean age was 51·2 years (SD 13·0), 563 (58%) were men and 403 (42%) were women, and 875 (91%) were White. In the group of participants with NASH and clinically significant fibrosis, we observed higher liver enzymes and a higher proportion of participants with hypertension (69%) and diabetes (64%) at baseline compared with the 631 people without NASH and clinically significant fibrosis (table 1). Participants with missing data were not included in the analysis. 23 (2%) of 966 participants had a NAS score of 0, 68 (7%) had a NAS score of 1, 116 (12%) had a NAS score of 2, 187 (19%) had a NAS score of 3, 224 (23%) had a NAS score of 4, 186 (19%) had a NAS score of 5, 119 (12%) had a NAS score of 6, 40 (4%) had a NAS score of 7, and 3 (We also evaluated each biomarker as a screening test for recruiting trial participants with NASH and clinically significant fibrosis (ie, at-risk NASH). The optimal threshold for each marker corresponded to a screen failure rate not exceeding 33% while maximising sensitivity (table 3). We could define such a threshold for most evaluated biomarkers and scores, but the corresponding proportion of people testing positive for the biomarker varied widely as a result of the differences in the underlying distributions of marker results in people with and people without NASH and clinically significant fibrosis. The single biomarker for which the highest proportion of participants tested positive and, consequently, found the highest proportion of participants with a true positive result was PRO-C3. At a threshold of 24·05 ng/mL, 17 of 100 people with NAFLD would test positive for PRO-C3 and qualify for a screening biopsy, of which 11 would have a true positive diagnosis of NASH and clinically significant fibrosis. Therefore, for every nine individuals being tested for PRO-C3, one eligible

participant with biopsy-proven NASH and clinically significant fibrosis would be identified as eligible for trial recruitment. By contrast, the optimal diagnostic screening threshold for PRO-C4 was 433·35 ng/mL. At that threshold, six of 100 people with NAFLD being tested would have a positive test result and could be selected for biopsy to evaluate trial eligibility. Of these six, four would have NASH and clinically significant fibrosis and two would not. This result corresponds to a 33% failure rate, but at a lower efficiency than PRO-C3; 23 individuals would have to be tested for PRO-C4 to find one eligible trial participant with biopsy-confirmed NASH and clinically significant fibrosis. The screening tools with the best performance were the SomaSignal test, ADAPT, and MACK-3. 35 of 100 people with NAFLD tested with SomaSignal, 24 of 100 tested with ADAPT, and 21 of 100 tested with MACK-3 would test positive for the marker at the selected thresholds and would have a screening biopsy at trial entry. Subsequently, 24 of the people who tested positive with SomaSignal, 16 of those who tested positive with ADAPT, and 14 of those who tested positive with MACK-3 would then be histologically eligible and have a true positive result for NASH and clinically significant fibrosis. The highest sensitivity was observed for the SomaSignal test (table 3). At a different prevalence of NASH and clinically significant fibrosis, the optimal thresholds and proportions will be different. The number of participants with a positive marker test and the number of those with a true positive diagnosis after biopsy for, as an example, four of the markers and scores included in the head-tohead comparison at various levels of prevalence are shown in figure 3. For detecting advanced fibrosis, of the single biomarkers, only LSM VCTE reached the predefined 0·80 threshold, with an AUC of 0·83 (0·80–0·86) compared with 0·73 (0·70–0·78) for FIB-4 (table 2). PRO-C3 and PRO-C6 were the single biomarkers with the next best performance (table 2). PRO-C4 and both forms of CK-18 performed less well than FIB-4 (table 2; figure 2B). Five different multimarker scores exceeded the 0·80 AUC threshold in detecting advanced fibrosis, but only two were statistically significant (table 2). The SomaSignal test had an AUC of 0·90 (0·86–0·94) versus 0·72 (0·66–0·79) for FIB-4; the ADAPT score had an AUC of 0·85 (0·81–0·89) versus 0·76 (0·71–0·81) for FIB-4. FIBC3 had an AUC of 0·82 (0·78–0·87), ABC3D had an AUC of 0·81 (0·76–0·85), and ELF had an AUC of 0·80 (0·76–0·83; table 2; figure 2B) A direct head-to-head comparison of ten circulating biomarker and multimarker scores (ie, PRO-C3, CK-18 M30, CK-18 M65, ELF, NFS, APRI, ADAPT, FIBC3, ABC3D, and FIB-4) was done in 335 participants (the number of participants who had test results for all ten tests listed). In this subgroup, 121 (36%) had both NASH and clinically significant fibrosis and 91 (27%) had advanced fibrosis confirmed by liver biopsy (appendix p 11). We observed that the AUCs for detecting NASH and clinically significant fibrosis and advanced fibrosis were similar in this subgroup to the AUCs in the main analysis (appendix pp 6,7,12). In an additional comparison, the performance of each marker was evaluated separately in people with and without diabetes. In our study group, 406 (42%) of 966 participants had diabetes. For detecting NASH and clinically significant fibrosis, performance was marginally lower in participants with diabetes, although only significantly lower for the multimarker score ABC3D, with an AUC of 0·74 (0·68–0·79) for people without diabetes versus 0·56 (0·48–0·64) for people with diabetes (appendix p 13). There were no significant differences in detecting advanced fibrosis between the two subgroups, with similar AUCs (appendix p 13).

Discussion

In this comparative diagnostic accuracy study, we used data and samples collected from the LITMUS metacohort to evaluate the performance of several markers in identifying people with NASH and clinically significant fibrosis (ie, NASH and F≥2; at-risk NASH) or people with advanced fibrosis (F≥3) using liver histology as the reference standard. Based on the ROC analyses, none of the evaluated single biomarkers met our prespecified 0·80 threshold in detecting NASH and clinically significant fibrosis. Of the multimarker scores, best performance was observed for the SomaSignal test, composed of 35 different proteins. AUC values were higher for detecting advanced fibrosis than for detecting NASH and clinically significant fibrosis. The SomaSignal test, the ADAPT score, and LSM VCTE exceeded our prespecified 0·80 AUC threshold. Recruitment for clinical trials is, at present, based on liver biopsy. Screening for individuals with NASH and clinically significant fibrosis is limited by high screen failure rates for histological assessment. A successful

screening biomarker would be expected to identify most people NASH and clinically significant fibrosis while substantially reducing the number requiring biopsy. We proposed a strategy for preselecting participants for liver biopsy by targeting a screen failure rate not exceeding 33%. With this strategy, we observed that some tests would substantially reduce the number of potential participants who would need to have a liver biopsy with acceptable sensitivity, as only those positive for the marker would require further evaluation. Without a screening biomarker, all 966 participants would require biopsy to identify the 335 people with NASH and clinically significant fibrosis. Therefore, participant selection efficiency would be 335 (35%) of 966. The biomarker assessed in this study with the best performance, SomaSignal, would reduce the number of patients requiring biopsy by 65%, from 966 to 338, resulting in 232 identified people with NASH and clinically significant fibrosis (participant selection efficiency of 69%). Many of the other biomarkers measured in this study would similarly increase participant selection efficiency but with lower sensitivity, resulting in a lower identification rate compared with the SomaSignal test. We note that the 33% screen failure rate we applied was defined on the basis of expert opinion, and other researchers and clinicians might arrive at a different acceptable proportion based on factors such as feasibility and costs. The thresholds identified here should be externally validated as several factors, such as disease spectrum, might affect the performance of the tests in diagnostic screening for trial recruitment.

A major strength of the study was the centralised measurement of all novel biomarkers instead of the use of local, historical measurements, although measured in batches. The analysis was done by an independent group of expert epidemiologists with no vested interest in showing superior performance of any test. We provide comparative accuracy data for many staple fibrosis tests and newer developments proposed for NAFLD, which can supplement guideline development for their suggested use in the future. The limitations of this study included the small serum sample volume that impeded measuring all biomarkers in all participants. Groups with biomarker results were only partly overlapping, which is why we present pairwise comparisons with FIB-4 as a readily available comparator. Stability of these markers is not also well understood, which is why we did not include samples collected before 2010. Histological scoring was not centralised and variability in recognition of elementary lesions or composite diagnoses might have occurred.26 Histologybased semiquantitative scoring is an imperfect reference standard, restricting the accuracy of grading necroinflammatory activity and staging fibrosis.5 Another limitation was the limited quantity and subsequent exhaustion of sample material, restricting the measurement of most biomarkers in different subsets of participants. Some of the biomarkers we assessed were not originally proposed for identifying patients with NASH and clinically significant fibrosis. As many are available to clinicians, and considering the high collinearity between NASH and fibrosis stage, we pragmatically decided to assess their diagnostic performance for this key histological aspect as well. Various markers27 and multimarker scores23,28 have been specifically developed for the diagnosis of NASH and clinically significant fibrosis and will need to be compared with the biomarkers with the best performance from the current study. The analysis presented in this Article focuses on serum-based biomarkers and multimarker scores. Although we included additional analysis on VCTE, a non-invasive technology proposed to evaluate liver aetiologies, other non-invasive imaging technologies have been proposed and should be further studied. We note the influence of recruiting patients from mostly tertiary care centres in multiple countries. Factors such as differences in prevalence, epidemiology, referral patterns, and clinical tests leading up to biopsy might affect the generalisability of our findings to other settings. The performance of many markers in this study was comparable to findings published elsewhere. Collagenbased markers and scores analysed from participants enrolled in the CENTAUR phase 2b trial showed that the single marker PRO-C3 had a marginally worse performance than FIB-4 in detecting advanced fibrosis, whereas the ADAPT score had a higher AUC.29 Incorporating a direct marker of fibrosis into the algorithm resulted in an improvement from simple scores, such as APRI. The ELF test had a performance consistent with that presented in a meta-analysis, which reported a summary AUC of 0·83 (0·71–0·90) for identifying advanced fibrosis.30 For NASH and clinically significant fibrosis, Chuah and colleagues31 concluded MACK-3 had similar performance to FIB-4

and outperformed single markers, such as CK-18 antigens. Compared with another large meta-analysis, the CK-18 M30 antigens showed consistent AUCs for NASH and clinically significant fibrosis (AUC 0·73, 0·57–0·85).32 When interpreting contrasting results between studies or subgroups, spectrum effects should be considered.33 Test performance often varies across population subgroups, as can be seen in the varying AUC estimates for FIB-4 in the partly overlapping subgroups in our analysis. The performance of NAFLD markers to correctly identify people with advanced fibrosis will vary with the relative proportions of individuals with F0 fibrosis and F4 fibrosis in the study group. Having more people with F0 fibrosis or F4 fibrosis in the study group will make it easier to differentiate between those with and without advanced fibrosis. There is a clear difference between the left-skewed fibrosis distribution in our study group, which represents the profile typically seen in secondary and tertiary care clinical practice, and the NIMBLE stage 1 NASH CRN study, which had approximately equal numbers in the five fibrosis stage subgroups.34 The limited performance of biomarkers in detecting NASH and clinically significant fibrosis provides a mandate for further study of novel biomarker algorithms, implementing both hypothesis-driven approaches from pathophysiology and machine learning approaches. Such approaches should be adapted to a specific target condition and context of use. The ultimate utility of these or any other biomarkers would be their ability to predict clinical outcomes. The longitudinal outcome data currently being generated by LITMUS within the Europe NAFLD Registry will be an important asset for evaluating their prognostic value. We have done one of the largest comparative diagnostic accuracy studies, with 17 different non-invasive markers for NAFLD. The results from this study showed that none of the single biomarkers achieved the desired performance to replace liver histology in detecting NASH and clinically significant fibrosis. However, some multimarker scores, such as the SomaSignal test and ADAPT, are promising tools for identifying advanced fibrosis. No biomarkers have been approved by the US FDA or EMA, which further highlights the urgency of the LITMUS consortium's aim to validate and advance towards regulatory qualification markers for NAFLD and NASH. The LITMUS project will continue to collect data in the prospective LITMUS study cohort and will analyse blood-based and imaging biomarkers to further facilitate the evaluation of new and existing interventions in trials and to improve the clinical care and outcomes of people with NAFLD.

Contributors. QMA and PMB conceptualised and designed the study. JB, SP, KW, DT, PB, AG, SF, MA, GP, HC-P, J-FD, DJL, SAH, JC, MP, AGH, HY-J, JC, MK, RO, CY, CB, ME, GPA, JMS, EB, MR-G, VR, and QMA assisted in data collection. YV, JL, PMB, KW, and QMA accessed and verified the raw data. JL and YV did data analyses and drafted the manuscript. PMB supervised the data analyses and JB, YC, KD, RO, QMA, and VR commented on the data analyses. All authors critically revised the manuscript and approved of the final version. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Acknowledgments. This study was done as part of the LITMUS project, which received funding from the Innovative Medicines Initiative 2 Joint Undertaking (grant agreement number 777377). This Joint Undertaking receives support from the EU Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations.

References

1 Riazi K, Azhari H, Charette JH, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022; 7: 851–61.

2 Lazarus JV, Anstee QM, Hagström H, et al. Defining comprehensive models of care for NAFLD. Nat Rev Gastroenterol Hepatol 2021; 18: 717–29.

3 Taylor RS, Taylor RJ, Bayliss S, et al. Association between fibrosis stage and outcomes of patients with nonalcoholic fatty liver disease: a systematic review and meta-analysis. Gastroenterology 2020; 158: 1611–25.

4 Francque SM, Marchesini G, Kautz A, et al. Non-alcoholic fatty liver disease: a patient guideline. JHEP Rep Innov Hepatol 2021; 3: 100322.

5 Davison BA, Harrison SA, Cotter G, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. J Hepatol 2020; 73: 1322–32.

6 Innovative Medicines Initiative. LITMUS: liver investigation: testing marker utility in steatohepatitis. 2017. https://www.imi. europa.eu/projects-results/project-factsheets/litmus (accessed Aug 20, 2022).

7 Hardy T, Wonders K, Younes R, et al. The European NAFLD Registry: a real-world longitudinal cohort study of nonalcoholic fatty liver disease. Contemp Clin Trials 2020; 98: 106175.

8 Bedossa P, FLIP Pathology Consortium. Utility and appropriateness of the fatty liver inhibition of progression (FLIP) algorithm and steatosis, activity, and fibrosis (SAF) score in the evaluation of biopsies of nonalcoholic fatty liver disease. Hepatology 2014; 60: 565–75.

9 Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005; 41: 1313–21.

10 Nielsen MJ, Nedergaard AF, Sun S, et al. The neo-epitope specific PRO-C3 ELISA measures true formation of type III collagen associated with liver and muscle parameters. Am J Transl Res 2013; 5: 303–15.

11 Leeming DJ, Nielsen MJ, Dai Y, et al. Enzyme-linked immunosorbent serum assay specific for the 7S domain of collagen type IV (P4NP 7S): a marker related to the extracellular matrix remodeling during liver fibrogenesis. Hepatol Res 2012; 42: 482–93.

12 Sun S, Henriksen K, Karsdal MA, et al. Collagen type III and VI turnover in response to long-term immobilization. PLoS One 2015; 10: e0144525.

13 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: 1317–25.

14 Boursier J, Anty R, Vonghia L, et al. Screening for therapeutic trials and treatment indication in clinical practice: MACK-3, a new blood test for the diagnosis of fibrotic NASH. Aliment Pharmacol Ther 2018; 47: 1387–96.

15 Cao W, Zhao C, Shen C, Wang Y. Cytokeratin 18, alanine aminotransferase, platelets and triglycerides predict the presence of nonalcoholic steatohepatitis. PLoS One 2013; 8: e82092.

16 Daniels SJ, Leeming DJ, Eslam M, et al. ADAPT: an algorithm incorporating PRO-C3 accurately identifies patients with NAFLD and advanced fibrosis. Hepatology 2019; 69: 1075–86.

17 Boyle M, Tiniakos D, Schattenberg JM, et al. Performance of the PRO-C3 collagen neo-epitope biomarker in non-alcoholic fatty liver disease. JHEP Rep Innov Hepatol 2019; 1: 188–98.

18 Angulo P, Hui JM, Marchesini G, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007; 45: 846–54.

19 Wai C-T, 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.

20 Guha IN, Parkes J, Roderick P, et al. Noninvasive markers of fibrosis in nonalcoholic fatty liver disease: Validating the European Liver Fibrosis Panel and exploring simple markers. Hepatology 2008; 47: 455–60.

21 Sanyal AJ, Williams SA, Lavine JE, et al. Defining the serum proteomic signature of hepatic steatosis, inflammation, ballooning and fibrosis in non-alcoholic fatty liver disease. J Hepatol 2022; published online Dec 14. https://doi.org/10.1016/j.jhep.2022.11.029

22 Corey KE, Pitts R, Lai M, et al. ADAMTSL2 protein and a soluble biomarker signature identify at-risk non-alcoholic steatohepatitis and fibrosis in adults with NAFLD. J Hepatol 2022; 76: 25–33.

23 Newsome PN, Sasso M, Deeks JJ, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with nonalcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020; 5: 362–73.

24 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837–45.

25 Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 2004; 50: 1118–25.

26 Brunt EM, Clouston AD, Goodman Z, et al. Complexity of ballooned hepatocyte feature recognition: defining a training atlas for artificial intelligence-based imaging in NAFLD. J Hepatol 2022; 76: 1030–41.

27 Harrison SA, Ratziu V, Boursier J, et al. A blood-based biomarker panel (NIS4) for non-invasive diagnosis of non-alcoholic steatohepatitis and liver fibrosis: a prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020; 5: 970–85.

28 Noureddin M, Truong E, Gornbein JA, et al. MRI-based (MAST) score accurately identifies patients with NASH and significant fibrosis. J Hepatol 2022; 76: 781–87.

29 Nielsen MJ, Leeming DJ, Goodman Z, et al. Comparison of ADAPT, FIB-4 and APRI as non-invasive predictors of liver fibrosis and NASH within the CENTAUR screening population. J Hepatol 2021; 75: 1292–300.

30 Vali Y, Lee J, Boursier J, et al. Enhanced liver fibrosis test for the non-invasive diagnosis of fibrosis in patients with NAFLD: a systematic review and meta-analysis. J Hepatol 2020; 73: 252–62.

31 Chuah KH, Wan Yusoff WNI, Sthaneshwar P, Nik Mustapha NR, Mahadeva S, Chan WK. MACK-3 (combination of hoMa, Ast and CK18): a promising novel biomarker for fibrotic non-alcoholic steatohepatitis. Liver Int 2019; 39: 1315–24.

32 Lee J, Vali Y, Boursier J, et al. Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: a systematic review and meta-analysis. PLoS One 2020; 15: e0238717.

33 Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002; 137: 598–602.

34 Sanyal AJ, Shankar SS, Yates K, et al. Primary results of the nimble stage 1-NASH CRN study of circulating biomarkers for nonalcoholic steatohepatitis and its activity and fibrosis stage. AASLD Meeting; Nov 12–15, 2021 (suppl 1).

Tables

Figures