Supplementary File S2. Methods.

Patients. Description of FibroFrance cohort, FLIP cohorts and EPIC-3 trial.

FibroFrance, is a program organized in 1997 to assess the burden of chronic liver diseases in France (Clinical trial French registry no: DRCD-2013-1 and ClinicalTrials.org no: NCT01927133), and STROBE statements were followed. The protocol was approved by the institutional review board, regulatory agency and performed in accordance with principles of Good Clinical Practice. All patients provided written informed consent before entry. All authors had access to the study data and reviewed and approved the final manuscript.

The FLIP cohort included adult patients at risk of NASH, i.e. presenting with ultrasound defined steatosis, and/or increased liver function test values, and/or metabolic risk factors (overweight, visceral adiposity, T2-diabetes, arterial hypertension, dyslipidemia) without well identified chronic liver diseases including alcohol consumption of < 50 g/day. All patients had to have had a liver biopsy. The nine participating centers were located in Paris, Seville, Newcastle, Bologna, Turin, Modena, Sao Paulo, Bern and Vienna. A central e-CRF was created, and data quality was enforced by a central data manager. The inclusion criteria in this retrospective analysis were the presence of reliable FibroTest, ActiTest and SteatoTest (FibroMax package), as well as the reading of the biopsy using the SAF scoring system by one of the experts from the FLIP Pathology Consortium. The validation of the patented FibroMax package was pre-determined in the FLIP protocol aims before patient inclusion (http://www.flip-fp7.eu/).

For hepatitis-C, the individual data from FibroFrance and from a large, international, multicenter, randomized trial of peginterferon alfa-2b (PEGIntron; Schering
Plough Corp. Kenilworth, NJ) or interferon alfa-2b (Intron A; Schering Plough Corp) in combination with ribavirin (Rebetol; Schering Plough Corp) described in detail elsewhere were obtained.\textsuperscript{12,13} The protocol was approved by the institutional review board, regulatory agency and performed in accordance with principles of Good Clinical Practice. All patients provided written informed consent before entry. All authors had access to the study data and reviewed and approved the final manuscript.\textsuperscript{19,20}

**Blood tests**

The FibroTest, the updated NashTest-2, the updated SteatoTest-2, are patented tests (BioPredictive, Paris, France) that have been validated extensively to assess the stages of fibrosis, the necro-inflammatory activity and steatosis grades of steatosis, using the SAF and CRN scoring systems for NAFLD,\textsuperscript{2,3,4,5,12,13,15,18,21} and the METAVIR scoring system for hepatitis-C.\textsuperscript{22} FibroTest and first NashTest and SteatoTest are already used in patients at risk of NASH in 40 countries including USA branded as NASH-FibroSure (LabCorp, Burlington, North Carolina, USA). The FibroTest includes serum \(\alpha_2\)-macroglobulin, apolipoprotein-A1, haptoglobin, total bilirubin, and \(\gamma\)-glutamyl transpeptidase. NashTest-2 combined 11 components with different weights, including the seven components of FibroTest, plus cholesterol and triglycerides.\textsuperscript{5} SteatoTest-2 combined with different weights, 10 out of the 11 components of NashTest-2, without bilirubin, but with fasting glucose.\textsuperscript{4} As the presence of at least 5% of steatosis is mandatory for the definition of NASH in CRN and FLIP scoring systems, but also recognized the possibility of burning steatosis (NASH without steatosis at biopsy, in cases without any cause of ballooning or lobular inflammation) two NashTest-2 values were assessed by the algorithm. The standard NashTest-2 followed the histological definition. and cases with SteatoTest-2
grade S0 (< to the 0.40 grade S1 cutoff) were graded as non-NASH. The second value was the NashTest-2-raw value even if the SteatoTest-2 concluded as grade S0 (< to the 0.40 grade S1 cutoff).

The scores of these biomarkers range from 0 to 1.00, the highest scores being attributed to the most severe lesions.

The preanalytical analytical procedures were those recommended by BioPredictive. For FibroFrance, tests were analyzed on fresh serum, for FLIP patients the serum were stored at −80°C and centralized to the Paris reference center and for hepatitis-C patients in a US centralized lab. Exclusion criteria were nonreliable results identified using security control algorithms.

**Histological references**

The SAF scoring system, specific for NAFLD features and permitting simplified construction of blood tests, has been described elsewhere. The goal of the SAF test was to find a compromise between the development of a simple, easily applied system for making a firm diagnosis in individual patients, even when applied by nonspecialists, and of a more reliable and discriminating system for therapeutic trials or for the assessment of biomarker diagnostic performance. A FLIP histopathology consortium of eight members developed the FLIP algorithm, a diagnostic tool for the diagnosis and staging of severe forms of NAFLD. According to the combination of each semi-quantification of the three elementary features of NAFLD using the SAF score for steatosis, inflammatory activity and fibrosis respectively. The steatosis score (S) assesses the quantities of large or medium-sized lipid droplets, with the exception of foamy microvesicles, and rates them from 0 to 3 (S0: < 5%; S1: 5–33%, mild; S2: 34–66%, moderate; S3: > 66%, marked). Activity grade (A, from 0 to 4) is the unweighted addition of hepatocyte ballooning (0–2) and lobular
inflammation (0–2). Cases with A0 (A=0) had no activity; A1 (A =1) had mild activity; A2 (A=2) moderate activity; A3 (A=3) severe activity and A4 (A=4) had very severe activity. Fibrosis stage (F) was assessed using the score described as follows: stage 0 (F0) =none; stage 1 (F1) =1a or 1b perisinusoidal zone 3 or 1c portal fibrosis; stage 2 (F2) = perisinusoidal and periportal fibrosis without bridging; stage 3 (F3) = bridging fibrosis and stage 4 (F4) = cirrhosis.21

To reduce interobserver variability and homogenize the reading using the new SAF-FLIP histological classification, we used only reports reviewed by members of the FLIP Pathology Consortium (DT and PB for the FLIP subpopulation and FC for the FibroFrance subpopulation). In patients with hepatitis-C, the METAVIR scoring system was used.22

Statistical methods.

The method for assessing an index permitting to adjust a binary-AUROCs according to the different prevalences of fibrosis stages.

The full method is described elsewhere (Poynard T, Halfon P, Castera L, et al. Standardization of ROC curve areas for diagnostic evaluation of liver fibrosis markers based on prevalences of fibrosis stages. Clin Chem. 2007;53;1615–1622.

In summary AUROCs are impacted by the prevalence of each fibrosis stage used to define a binary feature i.e. clinically significant fibrosis defined as F2, F3, F4 and non-significant fibrosis defined as F0, F1. For chronic hepatitis C the first definition of “advanced” fibrosis included F2 with F3 and F4, and non-advanced fibrosis only stage F0 and F1. The acronym “DANA index” came from the difference between the mean fibrosis stage of advanced fibrosis minus the mean fibrosis stage of non-advanced fibrosis).
Nowadays “advanced fibrosis” in NASH patients is defined by F3-F4 stages and no more by F2-F3-F4, now named “clinically significant fibrosis” by SAF scoring system. Therefore, it would be now less confusing to use the acronym “F34vsF012 index” as the “difference between the mean of F3F4 in fibrosis stage unit and the mean of F0F1F2”, and not to use the “DANA-Index”. To decrease the complexity for the publication’s readers we finally used the name “SpectrumF3F4-Index”, who remained clear for the definition of the stages F3F4.

This spectrum index assesses the relationship between the prevalence of each stage and the AUROC of a fibrosis marker for F3F4 and permits standardizing comparisons. The standardization can be performed to transform any different prevalence profile to 2 standard distributions of fibrosis stages: a uniform distribution and a naturally observed distribution. The uniform distribution of fibrosis stages was defined by a prevalence of 0.20 for each of the 5 stages (uniform prevalence). In the case of a uniform prevalence, the mean fibrosis stage in METAVIR units is 3.5 for F3F4 fibrosis: [mean of (F3+F4)/2 = (3+4)/2=3.5] vs 1.0 for non-significant fibrosis [mean of (F2+F1+F0)/3 = (2+1+0)/3= 1.0]. In this uniform prevalence distribution, the difference between the mean fibrosis stage of F3F4 minus the mean fibrosis stage of F012 was 2.5. The naturally observed distribution of fibrosis stages was defined by the prevalence of stages observed among patients in the context of use of the tests. The regression between multiple observed binary AUROCs such as F3F4vsF0F1F2 and the F34vsF012-Index enables to predict thereafter an adjusted-AUROC from the F34vsF012-Index of a study even without individual data.

If the prevalence of each stage is given in the publication of a diagnostic test or in a meta-analysis, the F34vsF012-Index (“SpectrumF3F4-Index”) is estimated by the following
formula: \( F34vsF012-\text{Index} = \text{mean F3F4 estimated by } \frac{\left(\text{prevalence F3 } \times 3 + \text{prevalence F4 } \times 4\right)}{\left(\text{prevalence F3 } + \text{prevalence F4}\right)} \) minus the mean non-advanced fibrosis as estimated by \( \frac{\text{prevalence F1 + Prevalence F2} \times 2}{\text{prevalence F0 } + \text{prevalence F1} + \text{prevalence F2}} \).

We defined the uniform SpectrumF3F4-Index at 2.5, each stage prevalence =0.20, giving a mean of 3.5 CRN fibrosis in the F3F4 fibrosis group and a mean of 1 in the F0F1F2 group. An adjusted uniform AUROC, relating the observed index value to that seen when fibrosis stages are uniformly distributed (SpectrumF3F4-Index=2.5), can be calculated from the regression formula linking the observed F3F4 AUROC (binary-AUROC) to SpectrumF3F4-Index.

The naturally observed SpectrumF3F4-Index in NASH here was 2.34, in 501 consecutive patients of a liver clinic in a tertiary center. The AUROC, which was standardized at the DANA value of 2.34, is the value for a given test for the diagnosis of F3F4 when fibrosis stages are distributed as observed in the reference population.

*Comparison of Obuchowski measures in patients with and without T2-diabetes*

Our main hypothesis was that there was no significant difference between the Obuchowski measure of the following three blood tests, FibroTest, NashTest-2 (with and without mandatory steatosis) and SteatoTest-2 for the prediction of fibrosis stages, NASH grades and steatosis grades respectively, in T2-diabetes vs. matched non-TD2M controls.

The primary endpoint was the Obuchowski measure, FibroTest with five SAF ordinal stages of fibrosis (F0–F4), NashTest-2 with five ordinal SAF grades (A0–A4) and SteatoTest, with four ordinal grades (S0–S3). The penalty function was related to the number of classes difference: for five classes, 0.25 for one class difference (adjacent classes), 0.50 for two for and 0.75 for three.6
As a sensitivity analysis, the Obuchowski measure was also assessed in patients with chronic hepatitis C, including between T2-diabetes vs matched non-T2-diabetes, for FibroTest and SteatoTest-2. NashTest-2 were not applicable in hepatitis-C, as the histological reference is not based on ballooning and lobular inflammation.

*Case-control studies to compare performances in patients with and without T2-diabetes*

The first secondary endpoints were two case-control studies. Due to the variability of the histological reference in chronic liver diseases, retrospective case-control studies are the simplest design for testing the hypothesis of differences between liver tests performances due to the presence or absence of T2-diabetes. Nowadays it seemed complicated to perform a prospective study including at least 100 subjects with T2-diabetes and 100 paired matched controls, with biopsy and simultaneous biomarkers.

The indirect comparisons of binary AUROCS are erroneous, without matching on the spectrum of histological features and the prevalences of their main confounding factors.\(^6\,^7\) Therefore we chose to match the controls according to the four main factors associated, besides the presence of T2-diabetes, with the severity of NAFLD (male gender, age $\geq$50 years, liver fibrosis stages $\geq$ stage 2) and the severity of overweight (BMI $\geq$ 30 kg/m\(^2\)).\(^{23,24,25,26}\)

In order to construct a case-controlled population with the maximum of power and the minimum of differences between the confounding factors, the following strategy was constructed. First, to maximize the power, no T2-diabetes patients were excluded. Second, we classified T2-diabetes and controls in 16 groups according to the 16 possible combinations of the 4 confounding (2-classes) factors. To obtain a non-significant difference of the confounding factors prevalences between cases and controls, controls
were excluded step by step, blindly to NashTest-2 and SteatoTest-2 results. Only the binary result of FibroTest was known as it was used as an adjustment factor for advanced fibrosis at the standard validated 0.48 cutoff. A non-significant difference was defined as a Fisher exact test $\geq 0.05$.

As a sensitivity analysis, the same case-control study was constructed in patients with chronic hepatitis C.

Construction and comparisons of adjusted-AUROCs in patients with and without diabetes

The subsequent secondary endpoint was to use an adjusted-AUROC for comparing biomarkers performances. In the absence of individual data, it is impossible to assess the Obuchowski measure, but it is possible to assess an adjusted-AUROCs which adjusts the binary-AUROCs (assessed by the empirical method) according to the prevalences of all stages, between patients, with and without T2-diabetes. This method was previously used in patients with chronic hepatitis C and described elsewhere.\textsuperscript{7}

Construction of “dF34vsF012”, (also named “SpectrumF3F4-Index” an index of spectrum effect

In patients at risk of NASH, most of studies on biomarkers used the stages F3F4 to define “advanced” fibrosis, which became the standard endpoint for binary-AUROCs. The adjusted-AUROC was predicted by the linear regression analysis linking the observed binary-AUROC to the “SpectrumF3F4-Index”, in the integrated base of 600 subjects at risk of NASH, with and without diabetes and in the only study which gave the FibroTest results and the prevalence of fibrosis stages in cases with T2-diabetes.\textsuperscript{2} In T2-diabetes, and contrarily to hepatitis-C, very few studies have estimated the natural prevalence of fibrosis
stages. Indeed, here we used a prospective cross-sectional study with 501 patients enrolled from two hepatology outpatient clinics in Cleveland, USA which had an almost uniform distribution of stages (F0, 16.8%/F1, 23.1%/F2, 19.7%/F3, 19.7%/F4, 20.6%) with a dF34 vs F012 = 2.59. Therefore, the “natural” prevalence in the context of use of a hepatology outpatient clinic, was close to an “uniform” prevalence with dF34 vs F012 = 2.50, each stage prevalence = 0.20, giving a mean of 3 METAVIR fibrosis units in the F2F3F4 fibrosis group and a mean of 1 in the F0F1 group.

Comparisons of FibroTest adjusted-AUROCs for F3F4 between T2-diabetes and non T2-diabetes

The regression curves linking the observed binary-AUROCs to the dF34 vs F012 index were compared for the 42 different combinations of prevalences in subsets of patients at risk of NASH with biopsy, in patients with and without T2-diabetes.

Furthermore, in the absence of individual data or AUROCs stratified according to the presence of T2-diabetes, it was possible from published studies giving the prevalence of diabetes, to assess a possible impact of diabetes. For this purpose, we compared the regression lines of the subset of studies with a prevalence of T2-diabetes equal or above the median vs. the subset with prevalence lower than the median. These data were those of a recent published overview of the performances of transient elastography and magnetic resonance elastography, completed with the studies of FibroTest.

Pilot study for comparing the impact of spectrum effect on biomarkers performances

To prevent a spectrum effect, the recommended method is the direct face to face comparison of biomarkers in the same patients using Obuchowski measure, in the context...
of use population. However, in the absence of such study, the impact of spectrum effect on biomarkers can be compared using the regression lines between binary-AUROCs and an index of spectrum such as the SpectrumF3F4-Index.

Sample size

We seek for a minimum of 100 cases of T2-diabetes to reach the median of the number of subjects included in diagnostic studies evaluating blood tests performances with liver biopsy in a recent overview and 400 controls to find appropriate matching. Comparison between Obuchowski measures used Z-test and equivalence tests of means, (H0: μ1 - μ2 ≤ -2.000 or μ1 - μ2 ≥ 3.000 versus Ha, the equivalence being concluded at an alpha risk of 0.01). Means comparison between several groups used multiple comparison Tukey-Kramer's test.

NCSS12 and numROC-software were used for statistical analyses.