Prospective external validation of a new non-invasive test for the diagnosis of non-alcoholic steatohepatitis in patients with type 2 diabetes

Thierry Poynard, Valérie Paradis, Jimmy Mullaert, Olivier Deckmyn, Nathalie Gault, Estelle Marcault, Pauline Manchon, Nassima Si Mohammed, Beatrice Parfait, Mark Ibberson, et al.

Received: 10 June 2021 | First decision: 14 June 2021 | Accepted: 11 July 2021

DOI: 10.1111/apt.16543

AP&T Alimentary Pharmacology & Therapeutics WILEY

Prospective external validation of a new non-invasive test for the diagnosis of non-alcoholic steatohepatitis in patients with type 2 diabetes

Thierry Poynard[1] | Valérie Paradis[1] | Jimmy Mullaert[1] | Olivier Deckmyn[1] | Nathalie Gault[1] | Estelle Marcault[1] | Pauline Manchon[1] | Nassima Si Mohammed[1] | Beatrice Parfait[1] | Mark Ibberson[2] | Jean-Francois Gautier[1] | Christian Boitard[1] | Sébastien Czernichow[1] | Etienne Larger[1] | Fabienne Drane[1] | Jean Marie Castille[1] | Valentina Peta[1] | Angélique Brzustowski[3] | Benoit Terris[1] | Anais Vallet-Pichard[1] | Dominique Roulot[4] | Cédric Laouénan[1] | Pierre Bedossa[1] | Laurent Castera[1] | Stanislas Pol[1] | Dominique Valla[1,3] | the Quid-Nash consortium

[1]Paris, France

[2]Lausanne, Switzerland

[3]Clichy, France

[4]Bobigny, France

Correspondence
Thierry Poynard, Hepato-Gastroenterology Department, Groupe Hospitalier Pitié Salpêtrière, 75013 Paris, France.
Email: thierry@poynard.com

Funding information
The RHU QUID-NASH Project, funded by Agence Nationale de la Recherche programe Investissements d'Avenir, (Reference ANR-17-T171105J-RHUS-0009 to DV), Agence Nationale de la Recherche, is carried out by Institut National de la Recherche Medicale, Paris Descartes University, Paris Diderot University, Centre National de la Recherche Scientifique, Centre de l'Energie Atomique, Servier, Biopredictive, and Assistance Publique-Hôpitaux de Paris under the coordination of Prof. Dominique Valla and project leader Angélique Brzustowski.

Summary

Background: One of the unmet needs in patients with type 2 diabetes mellitus (T2DM) is the prediction of non-alcoholic liver disease by non-invasive blood tests, for each of the three main histological features, fibrosis, non-alcoholic steatohepatitis (NASH) and steatosis.

Aims: To validate externally the performances of a recent panel, Nash-FibroTest, for the assessment of the severity of fibrosis stages, NASH grades and steatosis grades.

Methods: We prospectively analysed 272 patients with T2DM. Standard definitions of stages and grades were used, and analyses were centralised and blinded. The performances of the FibroTest, NashTest-2 and SteatoTest-2 were assessed using the Obuchowski measure (OM), the main outcome recommended as a summary measure of accuracy includeing all pairwise stages and grades comparisons, which is not provided par the extensively used binary area under the ROC curve.

Results: The diagnostic performance of each component of the panel was significant. OM (SE; significance) of the FibroTest, the NashTest-2 and the SteatoTest-2 was 0.862 (0.012; $P < 0.001$), 0.827 (0.015; $P < 0.001$) and 0.794 (0.020; $P < 0.01$), respectively. For ballooning and lobular inflammation, OM was 0.794 (0.021; $P < 0.001$) and 0.821 (0.017; $P < 0.001$), respectively. In a *post hoc* analysis the FibroTest outperformed VCTE by 4.1% (2.5-6.5; $P < 0.001$) for reliability, with a

The members of the Quid-Nash consortium are listed in File S1.

The complete list of authors' affiliations are listed in Appendix 1.

The Handling Editor for this article was Dr Colin Howden, and it was accepted for publication after full peer-review.

non-significant difference for OM for fibrosis staging, 0.859 (0.012) for FibroTest vs 0.870 (0.009) for VCTE.
Conclusions: From a single blood sample, the panel provides non-invasive diagnosis of the stages of fibrosis, and the grades of NASH and steatosis in patients with T2DM.
Trial registration number: NCT03634098.

1 | INTRODUCTION

One of the unmet needs in patients with components of the metabolic syndrome such as android obesity, type 2 diabetes mellitus (type 2 diabetes), hyperlipidaemia and hypertension is easy access to non-invasive tests (NITs) to assess the severity of non-alcoholic steato-hepatitis (NASH), including its three main histological features, steatosis, activity and fibrosis. Although the courses and clinical relevance of these features differ, they are intercorrelated.[1,2] Steatosis is the early feature of disease, the progression of fibrosis is the most accurate predictor of mortality and severe liver events, and inflammatory 'activity' is the biological driver of the progression of fibrosis.[3,4]

Thus, the availability of one NIT for each feature would provide a simple alternative to liver biopsy for the surveillance and treatment strategy in patients at risk of NASH. Among the numerous available NITS for the diagnosis of NASH, one panel, called the 'Nash FibroTest' panel, provides a specific test for each of the three features including the FibroTest (FibroSure in the USA)[5]; NashTest-2 for NASH[6]; and SteatoTest-2 for steatosis.[7] This panel was constructed and internally validated in 600 patients at risk of NAFLD in a multicentre cohort, using the steatosis activity fibrosis score (SAF-score), which defines the grades of NASH and steatosis and their associated clinical outcomes independently.[8-10]

These tests were then used in a prospective phase-2 trial of selonsertib,[7] and in an ongoing phase-3 trial of obeticholic acid,[11] and validated in a retrospective analysis of 220 patients, all with type 2 diabetes.[12] The latter study is the only existing trial evaluating type 2 diabetes patients. In a review of 25 NITs of NAFLD, the prevalence of type 2 diabetes ranged between 14% and 50%.[13] Of all the components of the metabolic syndrome, type 2 diabetes is the most important risk factor for NAFLD and non-alcoholic steatohepatitis (NASH), and the most important clinical predictor of adverse outcomes such as advanced liver fibrosis and mortality.[14] Indeed, a meta-analysis of population-based observational studies found that type 2 diabetes is associated with a more than twofold increase in the risk of developing severe liver disease.[15] The performance of NITs for the diagnosis of the features of liver disease in type 2 diabetes patients is controversial,[16,17] because of the absence of prospective evaluations as well as methodological limitations such as biopsy sampling variability,[18] intra- and interobserver variability for scoring the features,[18] the impact of the spectrum effect on binary AUROCs,[19-21] and inappropriate histological references to assess the performance of NITS such as the NAS score.[8,9,22] Thus, the first aim of the prospective QUID-NASH research program focusing on type 2 diabetes (https://rhu-quidnash.com), was to externally validate the performance of the 'Nash FibroTest' panel in the specific context of diabetology outpatient clinics using a centralised biopsy review, and appropriate methods.

2 | RESEARCH DESIGN AND METHODS

2.1 | Study participants and design

The primary outcome of this prospective cross-sectional multicentre study in patients with type 2 diabetes is to assess the diagnostic accuracy of the FibroTest, NashTest-2 and SteatoTest-2 using liver histology as the reference to evaluate liver fibrosis, activity and steatosis. NAFLD was suspected based on the presence of abnormal liver enzymes as well as an ultrasound scan showing a bright liver echo pattern, in patients with type 2 diabetes diagnosed in a diabetology outpatient clinic.

The STARD and FibroSTARD guidelines were followed (File S2) particularly for items 13.7 and 13.8.[23] Consecutive patients were prospectively recruited between October 2018 and 2020 in four outpatient diabetology clinics in the Assistance-Publique-Hôpitaux-de-Paris (File S1). The study (NCT03634098) was approved by the Research Ethics Committee #18.021-2018-A00311-54. All patients gave written informed consent. The study was performed in accordance with the declaration of Helsinki. All authors had access to the study data and reviewed and approved the final manuscript.

2.2 | Main analyses

The primary objective of the study was to evaluate the diagnostic accuracy of each component of the NashFibroTest, the FibroTest, the NashTest and the SteatoTest, in relation to the histological evaluation of fibrosis, Nash activity and steatosis. The primary endpoint was the validation of the FibroTest because the stage of fibrosis is the main prognostic criterion compared the grades of Nash or Steatosis.[1-4]

2.3 | Inclusion and exclusion criteria in the validation population

Inclusion criteria were as follows: patients were ≥18 years of age, able to give written informed consent, with type 2 diabetes defined according to American Diabetes Association (ADA) or World Health Organization (WHO) criteria,[24] and were scheduled, independently from this study, to undergo a liver biopsy for investigation of suspected NAFLD within 4 weeks after ultrasonography and alanine aminotransferase (ALT) assessment. These patients had abnormal transaminases had to be negative for standard tests for liver diseases (File S3). Exclusion criteria were as follows: patients with HBV, HCV and autoimmune diseases, pregnant women, patients without national health insurance, with a history of chronic liver disease, patients with serum haemoglobin <7g/L or <10g/L in the presence of cardiovascular or pulmonary disease, patients who refused liver biopsy or tests, patients with significant alcohol consumption (≥30 g/day for males and ≥20 g/day for females) and by serum carbohydrate deficient transferrin per cent >2%, and patients with a terminal disease.

2.4 | Patient characteristics

The following characteristics were recorded in all patients: age, gender, body mass index (BMI), temperature, the presence of diabetes and arterial pressures. A 12-hour fasting blood test was performed locally on fresh samples for assessment of the following parameters: platelet count, aspartate transaminase (AST), ALT, gamma-glutamyltransferase (GGT), alkaline phosphatase, albumin, bilirubin, fasting glucose, total cholesterol, high-density and low-density cholesterol lipoproteins, triglyceride, ferritin, urea, creatinine, alpha-2-macroglobulin (A2M), A1C-haemoglobin, insulinaemia, HOMA score, urea, creatinine, sodium, potassium, calcium and C-reactive protein.

2.5 | Histopathologic evaluation in the validation group

Liver biopsy (intercostal or transvenous) was performed in all patients according to the standard local procedure. Biopsy specimens were fixed in formalin, embedded in paraffin and stained with haematoxylin and eosin and Sirius Red. Slides were analysed in each centre by an experienced pathologist (VP, BT) and then centrally reviewed by a single experienced pathologist (PB) for the read-outs, blinded to all patient characteristics. The length and the number of fragments were assessed, and the quality scored according to a three-class classification (adequate, marginal and inadequate). The cause of any inadequate liver biopsy was specified: length, fragmentation or technical issues, that is, inadequate staining, or granuloma. NASH was diagnosed according to the presence of steatosis, hepatocyte ballooning (three grades 0-2) and lobular inflammation (three grades 0-2) with at least 1 point for each category. NAFLD activity (Nash score) was scored using both the SAF (main outcome in four classes),[8-10] and NASH-CRN scoring systems, which are different for several feature scores (File S3).[21]

Fibrosis was scored using the same SAF and CRN definition in five stages from 0 to 4.[21] Steatosis was scored in four grades (from 0, 'less than 5%', 1 '5%-30%', 2 '33%-66%', to 3 'more than 66% of hepatocyte with steatosis'). Portal inflammation and Mallory bodies were also recorded by grade into three classes. Liver biopsies were categorised by pathologists as a normal liver (no liver pathology), NAFL (steatosis but no NASH), NASH or other diagnosis when no NAFLD but other histological features suggesting another diagnosis were observed.

2.6 | Nash FibroTest panel

The FibroTest is called the NASH-FibroSure® (LabCorp) in the USA. The FibroTest includes A2M, apolipoprotein-A1, haptoglobin, total bilirubin and GGT.[5,25] The comparative components of the FibroTest, the new NashTest-2,[6] the SteatoTest-2[7] and the original NashTest, and SteatoTest are described in Table 1. Compared to the original NashTest,[26] NashTest-2 was developed for a quantitative diagnosis of NASH (SAF score as reference) with no need for the body mass index (BMI). Compared to the original SteatoTest,[27] SteatoTest-2 was constructed without total bilirubin and BMI.[7] The tests were all adjusted for age and gender. All components were assessed on fresh samples. The pre-analytical and analytical procedures were those recommended by BioPredictive. Exclusion criteria were the non-reliable results identified using security control algorithms.[28] Using both the FibroTest and the NashTest-2, it was possible to predict the presence or the absence of clinically significant NAFLD as defined by the histological SAF score: fibrosis stage ≥F2 (FibroTest >0.48), the standard cutoff for stage F2-F3-F4,[5,25] and/or activity grade ≥grade A2, (NashTest-2 ≥0.50).[6]

2.7 | Effect of the uncertainty of biopsy on tests performances

Biopsy is an imperfect gold standard.[18,29-33] We used the method recently suggested by McHugh et al,[33] and for the first time we assess the effect of uncertainty in the patient classification (Files S3 and S4). The performance of any test must be evaluated with reference to a comparator. The presence of classification uncertainty in the comparator (here biopsy) is therefore an important confounding factor when interpreting the diagnostic performance of the test (ie FibroTest). We report the comparator uncertainty together with the estimated performance of the test. As increasing amounts of noise are introduced into the biopsy, such as the biopsy length,[32,33] the apparent performance of the diagnostic test compared to biopsy the comparator, will decrease accordingly. Each amount of uncertainty (here the false positive/negative of biopsy according to the specimen length vs large surgical biopsy) was randomly introduced into

100 iterations and the aggregate results are shown. This simulation was implemented using the online simulation tool, https://imperfect-gold-standard.shinyapps.io/classification-noise/.[33] The performances of liver biopsy where those assessed using large surgical biopsy as the ground truth for staging.[34] The percentage of 25 mm liver biopsy that was correctly classified for fibrosis by the METAVIR score was 75%. Thus, a 25 mm biopsy was considered to have a sensitivity of 82.5% and a specificity of 82.5% for the percentage of correct classifications into the five stages of fibrosis. The same method was applied for the NashTest-2. There was no large surgical biopsy for ground truth in NASH, thus we used the repeated biopsies results as ground truth as recommended.[18,33-35]

2.8 | Discordance analysis

A major discordance was defined as a difference >2 stages for fibrosis, or >2 grades for activity according to the SAF score, which could influence clinical decision-making. For steatosis, as NAFLD and NASH required the presence of steatosis no major discordances could be observed. To attribute these major discordances to biopsy or to the Nash-FibroTest panel, reliable VCTE and FIB4 were used for the staging of fibrosis, ALT, AST and GGT levels were used to grade significant NASH. All cases with such major discordances were independently adjudicated by two clinicians DV and TP.[32]

2.9 | Comparisons between NashFibroTest, VCTE and FIB-4

A prospective, direct comparison between the FibroTest, VCTE and FIB4 in intention to diagnose and per-protocol analyses would have required 600 cases, based on the multiple comparisons between the Obuchowski measure and reliability.[36] These comparisons have been scheduled in other work packages of the Quid-Nash consortium (https://rhu-quidnash.com). In this study, we performed a post hoc analysis to compare the reliabilities and diagnostic performances of FibroTest, VCTE and FIB4 for fibrosis and SteatoTest-2 and controlled attenuation parameter (CAP) for steatosis. The VCTE FibroScan (FibroScan 502Touch model Echosens, Paris, France) examination was performed by nurses or physicians trained and certified by the manufacturer and blinded to the patient's histological evaluation and NashFibroTest. Only examinations with at least 10 valid liver stiffness measurements (LSMs) as well as those with LSMs median/IQR ratio ≥30%, both for LSMs, and CAP were considered to be valid.[30,37,38] FIB-4 was assessed with the original formula: age ([yr] × AST [IU/L])/((PLT [10^9/L]) × (ALT [IU/L])$^{1/2}$).[39]

2.10 | Statistical analysis

The chosen same sample size of n = 300 for the primary aim of the study was the same as that used for the internal validation of SteatoTest-2, and for validation of the original SteatoTest.[7] Evidence of differences in variables between the stages of fibrosis and the grades of NASH or of steatosis was evaluated with the Kruskal-Wallis test followed by Dunnett's tests with a post hoc comparison. *P* values <0.05 were considered to be statistically significant.

The overall diagnostic accuracy of tests (main outcome) and VCTE and FIB4 (post hoc analysis) was estimated by the Obuchowski measure together with the standard error, to take into account the spectrum effect.[19,20,23] The performances of the FibroTest, NashTest-2 and SteatoTest-2 were assessed using the Obuchowski measure, the main outcome recommended as a summary measure of accuracy which

TABLE 1 Comparison of the components of FibroTest, NashTest-2 and SteatoTest-2, the three tests available in the Nash FibroTest panel

	NASH FibroTest			First generation	
Components	**FibroTest**	**NashTest-2**	**SteatoTest-2**	**NashTest**	**SteatoTest**
Reference	**Imbert-Bismut[25]**	**Poynard[6]**	**Poynard[7]**	**Poynard**	**Poynard**
Apha-2 macroglobulin	X	X	X	X	X
Apolipoprotein-A1	X	X	X	X	X
Haptoglobin	X	X	X	X	X
GGT	X	X	X	X	X
Total Bilirubin	X	X	-	X	X
ALT		X	X	X	X
AST		X	X	X	-
Cholesterol		X	X	X	X
Triglyceride		X	X	X	X
Glucose		-	X	X	X
Weight and height (BMI)		-	-	X	X
Age and sex	XX	XX	XX	XX	XX
Total number	7	11	11	14	11

includes all pairwise stages and grades comparisons, which is not provided par the extensively used binary area under the ROC curve.[19,20] The Obuchowski measure can be interpreted as the probability that the non-invasive index will correctly rank 2 randomly chosen patient samples from different fibrosis stages according to the weighting scheme, with a penalty for misclassifying patients. The binary under the ROC curve only measure the probability to be lower or higher than the cutoff, that is, 0.48 for the FibroTest for stages F0F1 vs F2F3F4 (significant fibrosis) that is one comparison. The Obuchowski measure summarises the performance of the all pairwise comparisons, that is 10 comparisons for the five stages of fibrosis (F0 to F4).[19]

'To compare the performance of FibroTest between the original Construction and the Validation subsets, we assess the binary-AUROC "spectrum adjusted" (binaryAUROCsa), together with the associated the difference between the mean fibrosis stages of (F2 + F3 + F4) and the mean fibrosis stages of (F0 + F1) as previously described.[20] This permitted to estimate the spectrum effect without computing the individual data. The binaryAUROCsa is calculated by its linear regression curve with binary-AUROC. The maximum is 4 when all patients are F0 or F4. The minimum is 1 when all patients are F1 or F2. When there is an uniform prevalence of stages, 20% for each five stages, the binaryAUROCsa is 2.5.[20,]

Due to the absence of patients grade S0 and with only two S1 (Table 2), we could only validate the SteatoTest-2 vs the original population, and performed a binary AUROC for the diagnosis of S3 vs S2. Data were reported for standard predetermined thresholds of the stages of fibrosis for the Fibrotest (0.27, 0.48, 0.58 and 0.74 for F1, F2, F3 and F4 respectively), grades of activity for the NashTest-2 (0.25, 0.50 and 0.75 for A1, A2 and A3 respectively) and of steatosis for the SteaoTest-2 (0.40, 0.55 and 0.62 for S1, S2 and S3 respectively). We reported the sensitivity (Se), specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio and negative likelihood ratio together with 95% CI for each cutoff value. We also investigated the performance of the tests in settings with different prevalences using Bayes' equation to estimate post-test probabilities. In this case we used the F2 threshold for fibrosis, and A2 for NASH activity which correspond to clinically significant liver disease.[1,2] The post hoc analysis was performed in intention to diagnose, the reliability and the diagnostic performances being compared by the paired binary test. For FIB-4 there was no definition of reliability in the literature. FibroTest reliability definition followed the manufacturer recommendation.[28] TE reliability was assessed among the participants of the core group, as not prospectively scheduled in the eligible participants.

To assess possible variability due to the length of biopsy 2 subsets was also analysed, one with biopsy length of 15 mm or longer, and one with length lower than 15 mm. The original cut-offs for F2 were used 7.1 kPa for VCTE,[38] and 1.45 for FIB-4.[39] All analyses were performed using the software R, version 3.3.0.32 and NCSS 2020, and in duplicate by two independent teams of statisticians, one independent from the inventor (JM, PM); and one including the inventor (TP). Continuous variables were expressed as medians (interquartile range [IQR]) and categorical variables as absolute figures with percentages. CIs were reported

TABLE 2 Characteristics of included patients

Characteristic	n	Distribution % or median (IQR)	Range
Total	272	100	
Centre			
Lariboisiere-Beaujon	184	66.7	
Cochin-HEGP	79	29.0	
Avicenne	9	3.3	
Demographic and clinical data			
Age	272	59 (52-66)	23-55
BMI	272	32 (28-35)	
Female gender	168	61.8	
Diabetes mellitus	272	100	
Geographical origin			
Europe	230	84.6	
Other	42	15.4	
Hypertension	182	66.9	
Stroke	11	4.0	
Myocardial infarction	9	3.0	
Arteritis	10	3.7	
Retinopathy	50	18.6	
Neuropathy	49	18.4	
Smoker	127	46.7	
Previous alcohol consumption at risk	10	3.7	
Alcohol consumption at inclusion			
None	137	50.4	
Not at risk	135	49.6	
At risk	0	0	
Treatment first year after diagnosis			
Diet or exercise	78	28.7	
Oral	242	90.0	
Insulin	21	7.7	
Blood tests			
Alpha-2 macroglobulin	272	2.10 (1.54-2.76)	1.22-4.32
Apolipoprotein A1	272	1.33 (1.21-1.48)	0.7-2.43
Haptoglobin	272	1.44 (1.06-1.86)	0.8-2.76
GGT	272	56 (36-86)	12-454
Bilirubin	272	9 (6-12)	6-30
ALT	272	49 (36-70)	16-335
AST	272	35 (28-47)	13-163
Fasting glucose	272	8.4 (6.9-10.5)	3.3-19.3
Haemoglobin-glycate	269	7.5 (6.8-8.4)	5.3-13.1
Total cholesterol	272	1.56 (1.33-1.89)	0.71-2.87

(Continues)

TABLE 2 (Continued)

Characteristic	n	Distribution % or median (IQR)	Range
HDL cholesterol	272	0.40 (0.34-0.47)	0.20-1.64
LDL cholesterol	260	0.82 (0.61-108)	0.09-3.12
Triglyceride	272	1.57 (1.10-2.16)	0.50-7.84
Platelets count	272	245 (200-292)	87-478
Prothrombin time	266	100 (93-100)	42-128
Creatinine	270	71 (59-84)	32-551
C-reactive protein	267	2.6 (1.1-5)	0-27
FibroTest range 0-1	272	0.33 (0.16-0.54)	0.04-0.92
NashTest-2 range 0-1	272	0.72 (0.58-0.82)	0-0.96
SteatoTest-2 range 0-1	272	0.76 (0.63-0.86)	0.15-0.96
Time blood test and biopsy (days)	272	55 (31-82)	0-343
Imaging			
Ultrasonography brightness	233	85.7	
Ultrasonography segment hypertrophy	100	36.8	
VCTE			
Performed	269	98.9	
Not performed	3	1.1	
XL probe	168	61.8	
Time VCTE and biopsy (d)	269	27 (0-75)	0-174
Reliability VCTE			
Reliable 10 measures IQR/m <30%	258	94.8	
Missing	6	2.2	
Not reliable	8	2.9	
LSM reliable (kPa)		7.8 (6.1-11.6)	3.6-70.5
10 measures	179	65.8	
>10 measures	86	31.6	
CAP	258		
CAP (dB/m), range 100-400 dB/m	258	338 (304-370)	
CAP IQR	258	27 (18-38)	0-89
CAP reliable IQR/median <30%	142	55.0	
Liver biopsy			
Transparietal	184	67.7	
Transjugular	88	32.3	

(Continues)

TABLE 2 (Continued)

Characteristic	n	Distribution % or median (IQR)	Range
Length of liver biopsy specimen (mm)	272	17.0 (16-18)	5-38
Number of fragments	272	1	1-20
1	124	45.6	
2	83	30.5	
≥3	65	14.9	
Quality biopsy			
Adequate	233	85.7	
Marginal	39	14.3	
Inadequate	0	0	
Fibrosis stage SAF-CRN			
F0	54	19.9	
F1	65	23.9	
F2	50	18.4	
F3	74	27.2	
F4	29	10.7	
Activity grade according to SAF			
A0	57	21.0	
A1	51	18.8	
A2	73	26.8	
A3	91	33.4	
Ballooning grade (SAF)			
B0 no	106	39.0	
B1 round and clear hepatocyte	95	34.9	
B2 ballooned hepatocytes	71	26.1	
Ballooning grade (CRN)			
B0 no	107	39.3	
B1 few	98	36.1	
B2 many	67	24.6	
Lobular inflammation grade (SAF)			
I0 absent	59	21.7	
I1 less than 3 foci	138	50.7	
I2 3 foci or more	75	27.6	
Lobular inflammation grade (CRN)			
I0 no	60	22.0	
I1 less than 2 foci	136	50.0	
I2 2-4 foci	66	24.3	
I3 more than 4 foci	10	3.7	

(Continues)

TABLE 2 (Continued)

Characteristic	n	Distribution % or median (IQR)	Range
Portal inflammation			
I0	92	33.8	
I1	142	52.2	
I2	38	14.0	
Mallory bodies			
M0	239	87.9	
M1	26	9.5	
M2	7	2.6	
Steatosis grade			
S0 <5%	0	0	
S1 5-33	2	0.7	
S2 23-66	59	21.7	
S3 66-100	211	77.6	
Micro steatosis	3	1.1	
NAS score			
0	1	0.4	
1	27	9.9	
2	41	15.1	
3	37	13.6	
4	58	21.3	
5	48	17.6	
6	37	13.6	
7	18	6.6	
8	5	1.8	
Pathologist diagnosis			
Not NAFLD	0	0	
Non-Alcoholic Fatty Liver	108	39.7	
Non-Alcoholic Steato Hepatitis	162	59.6	
Burned-out fibrosis (without inflammation)	2	0.7	

at the 95% level. Details are given in File S3. An explanation of the impact of spectrum effect and of the uncertainty of biopsy is given in File S4.

3 | RESULTS

3.1 | Patient characteristics

The study flow chart of patients included in the validation population with biopsy is presented in Figure 1. Table 2 presents the clinical, serological, histological characteristics and NashFibroTest data for all the 272 included patients.

3.1.1 | Liver biopsies

A total of 325 patients underwent liver biopsy (see Figure 1). The median (IQR) length was 17 mm (8 mm), with 55 (31-82) days between the blood test and biopsy (days) (Table 2). Biopsies were not available in 50 patients. The reading was not centralised in 22 and 18 biopsies were inadequate (Table S1). Only one significant side effect was observed in the 325 patients, one case with an accidental intestinal biopsy, without symptoms.

3.1.2 | Main outcomes

The Obuchowski measure (SE; significance) for the FibroTest was 0.862 (0.012; $P < 0.001$), for NashTest-2 was 0.827 (0.015; $P < 0.001$), and for SteatoTest was 0.794 (0.020; $P < 0.01$) with the corresponding medians and IQR by stages and grades in Figure 2A-C respectively.

The comparisons of test performances in diabetes vs the original population are provided in Table 3, for the Obuchowski measures, standard binary AUROCs, and including the adjusted AUROC for the FibroTest, and Figure 3 for the spectrum of stages and grades.

The influence of the prevalence of advanced fibrosis on the PPV and the NPV (95% CI) is provided in Table 5 and in File S3 for references. The PPV for the standard predetermined cutoff used, was 80% (71-89) and the NPV was 56% (48-63) in type 2 diabetes with a high (56%) prevalence of advanced fibrosis. In a large cohort of 30 761 NAFLD patients with type 2 diabetes and a prevalence of 32% of advanced fibrosis the PPV was 61% (52-69) and the NPV was 78% (69-86). In a group representative of the French general population with the lowest prevalence of advanced fibrosis (2.8%), the PPV was 9% (8-10) and the NPV was 98% (97-99).

The influence of the prevalence of significant NASH on PPV and NPV is also presented in Table 5. In these cases the PPV at the standard predetermined NashTest-2 cutoff (0.50), was 64% (57-70) and the NPV was 63% (45-79) in type 2 diabetes, with a high (56%) prevalence of significant NASH. In a large cohort of 89 427 NAFLD patients with type 2 diabetes and a prevalence of significant NASH of 60% the PPV was 67% (60-73) and the NPV was 59% (49-74). In a group representative of the French general population with the lowest prevalence of NASH (1.1%), the PPV was 1.3% (0.8-18) and the NPV was 99% (98-100).

3.1.3 | Secondary outcomes

The diagnostic performance of the NashTest-2 was also significant for each of the elementary features of NASH activity, according to both the SAF and CRN scoring systems. All w-AUROCs were above 0.790 ($P < 0.001$). Results are presented in Figure S1A for CRN ballooning, Figure S1B for SAF ballooning 0.0794 (0.021), Figure S1C for CRN lobular inflammation, Figure S1D for SAF lobular inflammation 0.821 (0.017), Figure S1E for portal inflammation and Figure S1F for Mallory bodies.

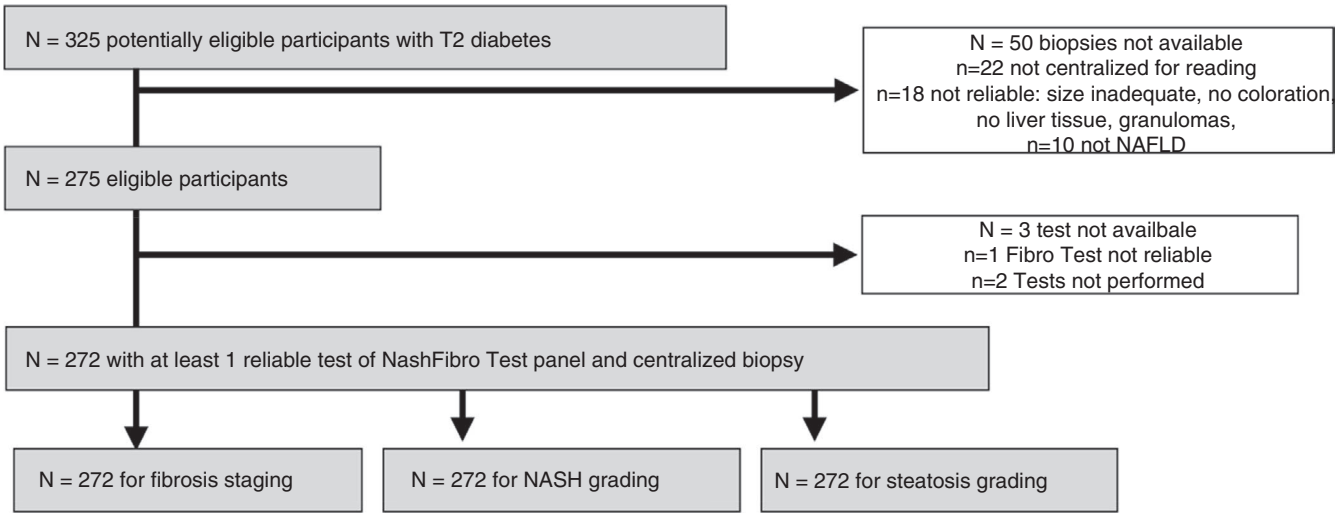

FIGURE 1 Study flow chart of core population with biopsy. Of 325 patients enrolled, 272 were eligible. Eventually among 275 patients with an interpretable biopsy, 272 had reliable FibroTest, NashTest-2 and SteatoTest-2. Only one patient with a non-reliable FibroTest has been excluded

3.1.4 | Post hoc comparisons between NITs, VCTE, FIB-4 and CAP

FibroTest was performed in 273 of the eligible patients and 272 were reliable, for a reliability of 99.6% (98.0-100). A total of 260 of the 272 included patients had a reliable VCTE, for a reliability of 95.6% (92.5-97.4). For the 272 cases with paired NITs the FibroTest reliability outperformed VCTE by 4.1% (2.5-6.5; $P < 0.001$). In an intention to diagnose analysis, the Obuchowski measure (se) for fibrosis stage was 0.859 (0.012) for the FibroTest, 0.870 (0.009) for VCTE, a non-significant difference ($P = 0.10$). If the analysis included only reliable stiffness measurements Obuchowski measures were higher for VCTE 0.910 (0.009), than for FibroTest, 0.862 (0.012; $P = 0.009$). For FIB-4, analysis cannot be performed in intention to diagnose, and the standard Obuchowski measure was 0.828 (0.011) which was lower than the FibroTest ($P = 0.02$) and VCTE ($P = 0.001$).

The diagnostic performances by other endpoints are described in Table S2. The overall results were similar for cases with biopsy ≥15 mm, and for cases with a biopsy <15 mm, the Obuchowski measure was only a higher for VCTE vs FIB-4. Comparison of CAP with SteatoTest-2 cannot be performed in intention to diagnose in the absence of a recognised cutoff for the CAP reliability. The binary AUROC for S3 vs S2 was 0.60 (0.52-0.67) and 0.69 (0.60-0.77), a not significant difference ($Z = 1.71$; $P = 0.09$) between SteatoTest-2 and CAP respectively.

3.1.5 | Effect of the uncertainty of biopsy on tests performances

The effect of biopsy uncertainty on the diagnostic performance of the FibroTest was significantly associated with the length of the specimen (Figure 4A). The maximum expected binary AUROC of an ideal NIT for fibrosis using 25 mm biopsies,[33] as a comparator in a study of 272 patients, would be 0.83. In the present study the median biopsy specimen was 17 mm and the maximum expected AUROC for an ideal NIT decreased to 0.70 due to the 30% misclassification rate of the biopsy.[33]

The effect of biopsy uncertainty on the diagnostic performance of NashTest-2 was even higher than for fibrosis, and also significantly associated with the length of the specimen, with AUROC decreasing from 0.69 with 25 mm to 0.60 with 17 mm (Figure 4B).

3.1.6 | Analysis of severe discordances

A major discordance was found between the biopsy and the FibroTest in 28 patients (10.3%; 7.0-14.5). After adjudication 10 (3.7%; 1.8-6.7) were considered to be a FibroTest error, 5 (1.8%; 0.6-4.2) a biopsy error and 14 (5.1%; 2.8-8.5) indeterminate (Table S3A). A major discordance was found in nine patients (3.3%; 1.5-6.2) between the biopsy and the NashTest-2. After adjudication 7 (3.7%; 1.8-6.7) were considered as an error of biopsy, and two (2.6%; 1.0-5.2) to be indeterminate (Table S3B).

4 | DISCUSSION

This prospective study examined the association of the NashFibroTest panel and liver histology in a cohort of type 2 diabetes patients undergoing biopsy for investigation of suspected NAFLD. The results validated the diagnostic performances with the Obuchowski measure, the primary endpoint. To our knowledge, this is the first prospective study focusing on patients with type 2 diabetes in a context of use of diabetology clinics. The present findings confirm the results of several retrospective studies in patients with type 2 diabetes,[12,16,17] and in subsets of patients at risk of Nash including type 2 diabetes.[5-7]

The results of the comparisons between NashFibroTest and NITs confirmed the similar performance already observed in NAFLD and

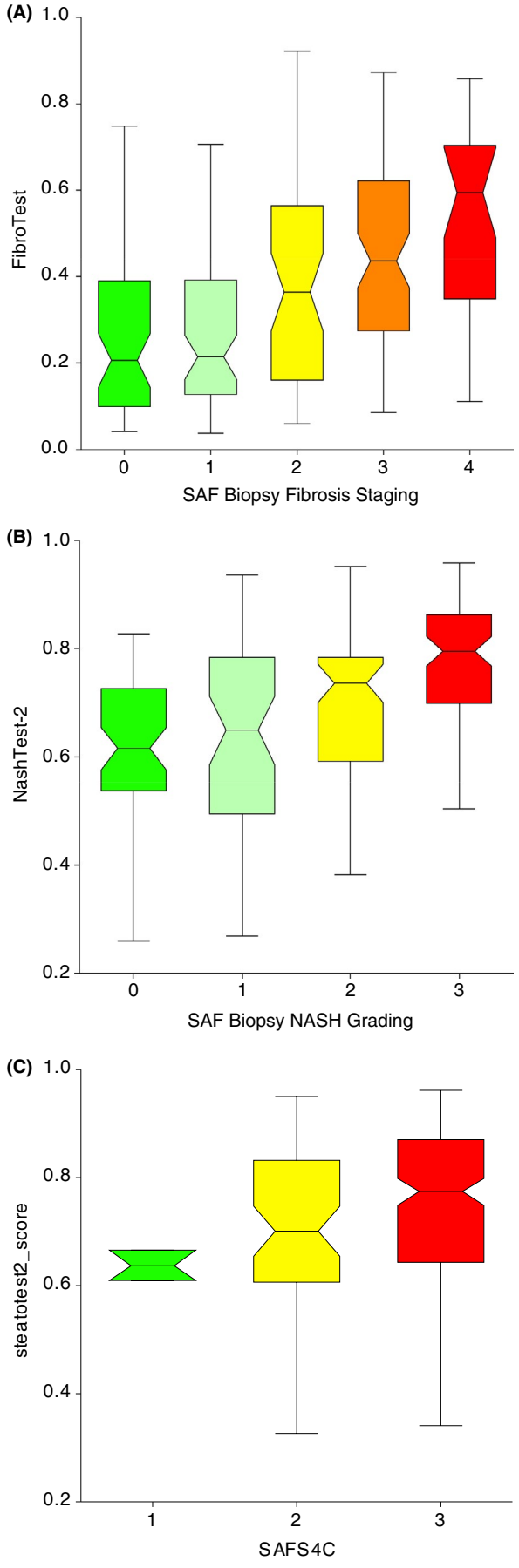

FIGURE 2 FibroTest performance in 272 type 2 diabetes patients for Fibrosis staging. (A) FibroTest was significantly different between Stage F0 (n = 54) vs F2, F3 and F4; Stage F1 (n = 65) vs F2, F3, and F4; Stage F2 (n = 50) vs F0, F1 and F4; Stage F3 (n = 74) vs F0 and F1; Stage F4 (n = 29) vs F0, F1 and F2. All 272 patients had reliable tests and centralised biopsies. The corresponding Obuchowski measure (SE; significance) was 0.862 (0.012; $P < 0.001$). (B) NashTest-2 performance in 272 T2M patients for NASH grading. NashTest-2 was significantly different between Grade A0 (n = 57) vs A2 and A3; Grade A1 (n = 51) vs A3; Grade A2 (n = 73) vs A0 and A3; Grade A3 (n = 91) vs F0, F1 and F2. The corresponding Obuchowski measure (SE; significance) was 0.827 (0.015; $P < 0.001$). (C) SteatoTest-2 performance in 272 T2M patients for Steatosis grading. By definition there was no S0, and only 2 S1. SteatoTest-2 was significantly different between grade S3 (n=207) vs S2 (n=58; P=0.03).

viral hepatitis, for FibroTest vs VCTE in intention to diagnose, with a higher reliability of Fibrotest vs VCTE, as well as the higher performance of FibroTest vs FIB-4, especially when the biopsy was more than 15 mm long.[6,31,36,40]

4.1 | Strengths

The study has several advantages compared to others evaluating the performance of NITs in NAFLD.

All of the elements in the liver-FIBROSTARD checklist were assessed except for cost-effectiveness (File S2).[23] Although the main limitations have been well known since 2003,[34] few studies have used appropriate methods.

This study takes into account a possible spectrum effect using the Obuchowski measure as the main endpoint, as recommended.[19,20,23] This was particularly important because our population had a high prevalence of minimal fibrosis, and for the grading of steatosis because of the very low prevalence of grade 0. In patients with type 2 diabetes, the influence of the spectrum effect on binary AUROCs explains the misleading interpretation of binary AUROCs in the absence of face-to-face comparisons between NITs,[12,16,17] even when using the C-statistic,[17,19] that is, only one comparison for fibrosis staging instead of 10 pairwise comparisons by Obuchowski measure, or the Harrel-C statistic which has a risk of overestimation.[19]

This really means for practice that a clinician can prefer a test with a 0.70 binary AUROC predicting significant fibrosis or significant Nash, because of the methodological quality of the validation of this test. He can also reject a test with 0.90 binary AUROC because the validation studies had not eliminated a spectrum effect or a risk of overestimation due to the uncertainty of biopsy.

The NashFibroTest panel was constructed using the SAF scoring system, which has several advantages compared to the standard CRN scoring system.[6-10,22] A simpler definition of activity was used as a reference: hepatocyte ballooning, and lobular inflammation with at least 1 point for each category. Indeed, this definition does not require the

TABLE 3 Test performance according to statistical methods, by features and population

Features, test and population	N	Method assessing test performance as determined in the protocol		
		Primary endpoint **Obuchowski measure weighted AUROC summarises all pairwise stages or grades performances. Mean (standard error)**	**Not taken as an endpoint. Standard binary AUROC cannot test the pairwise comparisons. Mean (95% CI)**	**Not taken as an endpoint. Binary AUROC spectrum adjusted on the different stages proportion among F2F3F4 and among F0F1**
Fibrosis and FibroTest		5 stages 10 comparisons	F2F3F4 vs F0F1	F2F3F4 vs F0F1
Original construction	541	0.904 (0.007)	0.80 (0.76-0.84)	0.84 (2.11)[a]
Diabetes validation	272	0.862 (0.012)	0.74 (0.65-0.83)	0.76 (2.32)[a]
NASH and NashTest		4 grades 6 comparisons	A3A2 vs A0A1	Not published
Original construction	541	0.827 (0.015)	0.77 (0.73-0.81)	
Diabetes validation	272	0.849 (0.013)	0.70 (0.62-0.75)	
Steatosis and SteatoTest		3 grades 3 comparisons	S2S3 vs S0S1	Not published
Original construction	541	0.801 (0.013)	0.74 (0.65-0.80)	
Diabetes validation	272	0.794 (0.020)	0.74 (0.60-0.83)	

[a]The binary AUROC spectrum adjusted is given with the difference between the mean fibrosis stages of (F2 + F3 + F4) and the mean fibrosis stages of (F0 + F1) in parenthesis.

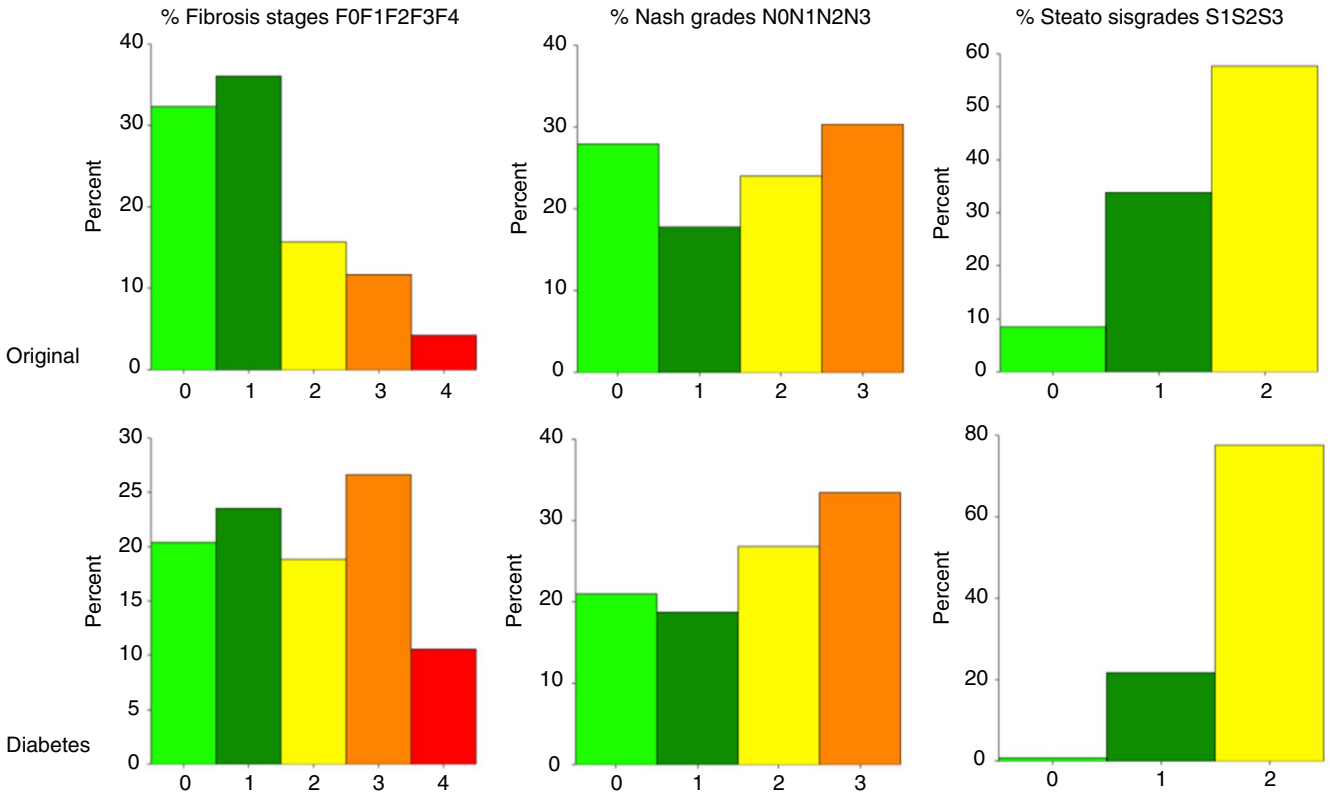

FIGURE 3 Spectrum of stages and grades in the original (upper row) and diabetes (lower row) subset. The spectrum of the stages of Fibrosis was not uniform in the original subset and almost uniform in diabetes. The prevalence of F3 and F4 was twice as high in diabetes as in the original subset. The difference between the mean advanced fibrosis stage *and* non-advanced stage was 2.32 in diabetes and 2.11 in the original population resulting in a slight underestimation of binary AUROC for both subsets vs a perfect uniform distribution. Binary AUROCs were 0.76 and 0.84 after standardisation vs 0.72 and 0.80 before, for the diabetes and original subsets respectively

presence of steatosis or the presence of both lobular inflammation and ballooning. This independence among features reduces the risk of false positive/negatives.[21] The NAS score is not appropriate for the construction of a NIT because it adds the grade of steatosis to the grades of ballooning and lobular inflammation.[7-9,22] A NAS score of 4 can correspond to a patient with grade 3 steatosis and grade 1 lobular

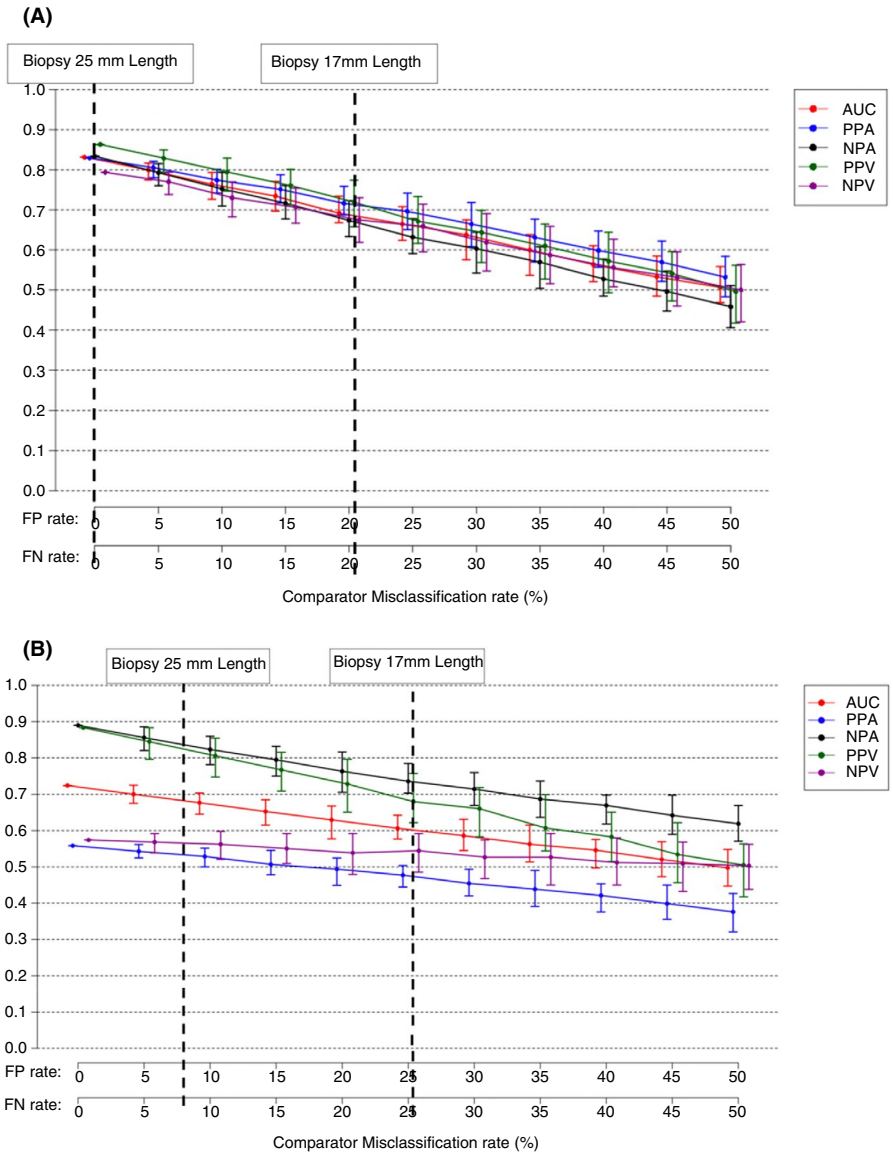

FIGURE 4 The effect of biopsy uncertainty on patient classification, due to specimen length in relation to the diagnostic performance of FibroTest (Panel A) and NashTest-2 (Panel B). The ground truth is a large surgical liver specimen. (FP, false positive, FN, false negative). In this study with 272 patients and a median 17 mm long biopsy the expected area under the ROC curve (AUROC) of the FibroTest (or NashTest-2) as a comparator cannot be more than 0.70 whatever its real performance due to the 30% misclassification rate of the biopsy as comparator. PPV, positive predictive value, and NPV, negative predictive value. The terms positive per cent agreement (PPA) and negative per cent agreement (NPA) are used instead of sensitivity and specificity, respectively, when the comparator is known to contain uncertainty

inflammation as well as to a patient with higher histological activity, grade 2 lobular inflammation, grade 1 ballooning and grade 1 steatosis. Furthermore, the grades of inflammation with the SAF score are more detailed with less inter-pathologist variability, and ballooning is differentiated from round and clear hepatocytes,[9] but not in the CRN score.[22] Our study used a centralised reading by a single expert, reducing the inter-observer variability.

Our results externally validated that a single blood sample provided an independent assessment of the severity of three histological features of NAFLD, the stage of fibrosis, the SAF grades of Nash by NashTest-2 and steatosis by SteatoTest-2, including elementary features of activity. This is an improvement in comparison to our first generation of tests. The sample was analysed in a biochemistry unit and results were obtained within few hours. These results (Obuchowski measures and NPV) confirmed previous studies (Tables 3 and 4).[6] In cases of lower prevalence, NIT of this type with high NPV would be an excellent 'rule out' test, particularly as in the context of use with a relatively low prevalence of NASH (Table 5). The risks of false positive/negative are well known, lower than 2%.[20,28] Another advantage is the numerous studies of FibroTest and SteatoTest whose diagnostic and prognostic performances have been extensively validated in chronic viral hepatitis and alcoholic liver disease.[6,7,38] which are frequently associated with type 2 diabetes. Furthermore, in comparison to the first generation test the NashTest-2 did not include fasting glucose or BMI in its components,[6] which simplifies its use. The SteatoTest-2 has also the advantage of increased reliability, as total bilirubin is no more included.[7]

4.2 | Limitations

The main limitation for the validation of NITs in NAFLD, including ours, is sampling variability which is directly associated with specimen length.[18] In our study the median (IQR) biopsy length of 17 (8) mm does not correspond to the recommended ideal of 25 mm.[34] However, 17 mm is also the mean length of the only retrospective

TABLE 4 Diagnostic performance of FibroTest, NashTest-2 and SteatoTest-2 for the diagnosis of fibrosis stage (SAF-CRN scoring system), Nash grade and steatosis grade, using predetermined cutoffs and SAF-scoring system

	FibroTest	NashTest-2	SteatoTest-2
N	**272**	**272**	**272**
Primary endpoint			
Obuchowski measure (SE)	0.862 (0.012) <0.001	0.849 (0.013) <0.001	0.794 (0.02) <0.01
Secondary endpoints			
Binary AUROC	F2F3F4 vs F0F1	A2A3 vs A0A1	S2S3 vs S0S1
AUROC (95% CI)	0.74 (0.65-0.83)	0.70 (0.62-0.75)	0.74 (0.60-0.83)
Prevalence (n)	0.56 (n = 153/272)	0.60 (n = 164/272)	0.99 (n = 270/272)
Cutoff (range 0-1)	0.48	0.50	0.55
Youden index (Se + Sp − 1)	0.33	0.13	0.27
Sensitivity (95% CI)	0.47 (0.39-0.55)	0.92 (0.87-0.96)	0.85 (0.80-0.89)
Specificity (95% CI)	0.86 (0.78-0.92)	0.20 (0.13-0.29)	Not applicable[a]
Positive predictive value (95% CI)	0.81 (0.71-0.89)	0.62 (0.55-0.69)	0.99 (0.97-0.93)
Negative predictive value (95% CI)	0.56 (0.48-0.63)	0.66 (0.47-0.81)	Not applicable[a]
Positive likelihood ratio LR+	3.29	1.17	Not applicable[a]
Negative likelihood ratio LR−	0.62	0.373	Not applicable[a]
DOR diagnostic odds ratio LR+/LR−	5.33	3.13	Not applicable[a]

[a]Not applicable as only two patients with steatosis S0-S1 to assess specificity and predictive values.

TABLE 5 Impact of the prevalence of significant stage or grade, on PPV and NPV (95% CI), according to FibroTest and NashTest-2 cutoffs

Diagnostic method of advanced fibrosis F2 or significant activity A2	Prevalence, n/total (%), of disease using predetermined cutoff	Context of use	Predictive value (95% CI)	Predetermined NIT cutoff	More sensitive cutoff (1 stage/ grade less)	More specific cutoff (1 stage /grade more)
SAF fibrosis stage ≥F2				>0.48 (F2)	>0.27 (F1)	>0.58 (F3)
Biopsy stage: F2F3F4	153/272 (56%)	Our type 2 diabetes clinic	PPV	80% (71-89)	70% (62-77)	82% (71-89)
			NPV	56% (48-63)	65% (55-74)	51% (45-58)
FibroTest F2 >0.48	209/7463 (2.8%)	General population France	PPV	9% (8-10)	5% (4-6)	9% (8-9)
			NPV	98% (97-99)	99% (98-100)	98% (97-99)
FibroTest F2 >0.48	9896/30,761 (32%)	Type 2 diabetes NAFLD USA	PPV	61% (52-69)	46% (37-55)	63% (54-71)
			NPV	78% (69-86)	83% (65-91)	74% (64-84)
FibroTest F2 >0.48	19797/105,255 (19%)	No type 2 diabetes NAFLD USA	PPV	44% (38-50)	30% (21-39)	46% (38-54)
			NPV	88% (80-96)	91% (90-92)	85% (77-93)
VCTE ≥7.1 kPa	89/1190 (7%)	General Population France	PPV	20% (15-25)	12% (10-14)	21% (16-26)
			NPV	96% (94-98)	97% (95-99)	95% (93-97)
SAF NASH grade ≥A2				>0.50 (A2)	>0.25 (A1)	>0.75 (A3)
Biopsy stage: A2-A3	164/272 (60%)	Our type 2 diabetes clinic	PPV	64% (57-70)	60% (54-66)	76% (67-83)
			NPV	63% (45-79)	33% (1-91)	50% (42-58)
NashTest-1: Nash	80/7463 (1.1%)	General population France	PPV	1.3% (0.8-1.8)	1.1% (0.7-1.5)	2.2% (1.7-2.7)
			NPV	99% (98-100)	99% (98-100)	99% (98-100)
NashTest-2: N2-N3	5416/89,427 (64%)	NAFLD USA	PPV	67% (60-73)	64% (57-70)	79% (70-80)
			NPV	59% (49-74)	30% (1-91)	46% (38-50)

Note: References are detailed in File S3.

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

study in patients with type 2 diabetes,[17] and is within the range of lengths found in 64 studies in NAFLD.[13] This study confirms the effect of this uncertainty using the appropriate definitions as well as its associated simulation tool.[33] The median biopsy specimen was 17 mm, thus the maximum expected binary AUROC for an ideal NIT decreased to 0.70 due to the 30% misclassification rate of the biopsy (Figure 4). Therefore, binary AUROCs of more than 0.80 using biopsies of around 17 mm as a reference, may have been overestimated in past studies.[12,13,17] A minor limitation is also that the NashFibroTest requires fasted samples.

There are several other limitations to the present study. Our study only provides an external validation of the NashFibroTest panel in diabetology clinics and not in a general population. Like in the construction subset, we also acknowledge that all patients preincluded for biopsy required abnormal transaminases, which is usually recommended by ethics committees in France. However, despite patient selection based on increased liver enzymes, the spectrum of stages was uniform up to stage F3 with a lower prevalence of cirrhosis than in the original study. The performances of NashTest-2 were also similar to those of the original study, with an uniform spectrum of grades.

We found the same high sensitivity of SteatoTest-2 (0.85; 0.80-0.89) as in the T2DM subset of the original validation (0.85) and the same limited specificity vs the original SteatoTest.

A cost-effectiveness analyses should be performed like in Hepatitis C.[41] Face-to-face comparisons between the main NITs in intention to diagnose, with appropriate sample size, are mandatory for an objective ranking. Fibroscan can measure two of the three features here and MRE could do all of them. However, it not yet clear for CAP what are the criteria of reliability, and for MRE the performances for staging NASH severity are not yet fully validated. Even if our results confirm the performance observed in United States,[16,17] other validation outside France is needed.

Finally, our results support a simplification of the standard definitions of NAFLD without the mandatory concomitant (temporal) presence of steatosis and inflammation. Most transversal studies assessing NITs in type 2 diabetes, only included patients with at least 5% steatosis at MRI-PDFF, and therefore excluded by definition burnt-out fibrosis (fibrosis without inflammation) or burnt-out NASH (inflammation without steatosis).[42] Only very large cohort studies using NITs such as the NashFibroTest panel, without selection on transaminases values, could estimate the true prevalence of burnt-out fibrosis, limited here to 0.7%.

In summary, despite the limitations of biopsy, this study confirms the significant performances of the NashFibroTest panel for the diagnostic of fibrosis stages, NASH grades and steatosis grades in a cohort of patients with type 2 diabetes.

ACKNOWLEDGEMENTS

RHU QUID-NASH is funded by Agence Nationale de la Recherche (ANR); implemented by Inserm, Université Paris Descartes, Université Paris Diderot, CNRS, CEA, Laboratoires Servier, Biopredictive, and AP-HP; and coordinated by Prof. Dominique Valla and Angélique Brzustowski.

Declaration of personal interests: TP is the inventor of FibroTest/SteatoTest and the founder of BioPredictive, the company that markets these tests. Patents belong to the French Public Organization Assistance Publique-Hôpitaux de Paris. TP plays a role in the study design, analysis and preparation of the manuscript. TP plays also a role in the statistical analysis but to preserve an independence in this validation study the statistical analysis was duplicate by independent academic authors CL and JM, supervised by DV and SP. OD, JMC, FD and VP are BioPredictive employees organised the anonymised assessments of the BioPredictive NITs and did not play a role in the study design, data collection, decision to publish or preparation of the manuscript and only provided financial support in the form of these authors' salaries.

AUTHORSHIP

Guarantors of the article: DV and SP.

Author contributions: TP, DV and SP were involved in experiment conception and design. TP, VP, DV, VP, PB, CB, LC, SP, JFG, FD, JMC, AVP, CB, EL, BT and DR were involved in experiment performance. TP, OD, JM, PM, CL, EM, NSM, BP, MI and AB were involved in data analysis. OD, JMC, FD and DV organised the anonymised assessments of the BioPredictive NITs. TP, OD, VP, DV, SP, LC and JM were involved in drafting of the paper. All authors had access to the study data and reviewed and approved the final manuscript. DV and SP have full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Thierry Poynard https://orcid.org/0000-0002-3726-7230
Sébastien Czernichow https://orcid.org/0000-0001-7353-2532
Stanislas Pol https://orcid.org/0000-0001-9772-9591

REFERENCES

1. European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. *J Hepatol*. 2016;64:1388-1402.
2. Rinella ME, Tacke F, Sanyal AJ, Anstee QM. Report on the AASLD/EASL joint workshop on clinical trial endpoints in NAFLD. *Hepatology*. 2019;70:1424-1436.
3. Hagström H, Nasr P, Ekstedt M, et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. *J Hepatol*. 2017;67:1265-1273.
4. Ratziu V. Back to Byzance: Querelles byzantines over NASH and fibrosis. *J Hepatol*. 2017;67:1134-1136.
5. Ratziu V, Massard J, Charlotte F, et al. Diagnostic value of biochemical markers (FibroTest-FibroSURE) for the prediction of liver

fibrosis in patients with non-alcoholic fatty liver disease. *BMC Gastroenterol*. 2006;6:6.
6. Poynard T, Munteanu M, Charlotte F, et al. Diagnostic performance of a new noninvasive test for nonalcoholic steatohepatitis using a simplified histological reference. *Eur J Gastroenterol Hepatol*. 2018;30:569-577.
7. Poynard T, Peta V, Munteanu M, et al. The diagnostic performance of a simplified blood test (SteatoTest-2) for the prediction of liver steatosis. *Eur J Gastroenterol Hepatol*. 2018;31:393-402.
8. Bedossa P, Poitou C, Veyrie N, et al. Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. *Hepatology*. 2012;56:1751-1759.
9. 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;60:565-575.
10. Nascimbeni F, Bedossa P, Fedchuk L, et al. Clinical validation of the FLIP algorithm and the SAF score in patients with non-alcoholic fatty liver disease. *J Hepatol*. 2020;72:828-838.
11. Younossi ZM, Ratziu V, Loomba R, et al. Obeticholic acid for the treatment of non-alcoholic steatohepatitis: interim analysis from a multicentre, randomised, placebo-controlled phase 3 trial. *Lancet*. 2019;394:2184-2196.
12. Bril F, McPhaul MJ, Caulfield MP, et al. Performance of the SteatoTest, ActiTest, NashTest and FibroTest in a multiethnic cohort of patients with type 2 diabetes mellitus. *J Investig Med*. 2019;67:303-311.
13. Xiao G, Zhu S, Xiao X, Yan L, Yang J, Wu G. Comparison of laboratory tests, ultrasound, or magnetic resonance elastography to detect fibrosis in patients with nonalcoholic fatty liver disease: a meta-analysis. *Hepatology*. 2017;66:1486-1501.
14. Younossi ZM, Golabi P, de Avila L, et al. The global epidemiology of NAFLD and NASH in patients with type 2 diabetes: a systematic review and meta-analysis. *J Hepatol*. 2019;71:793-801.
15. Jarvis H, Craig D, Barker R, et al. Metabolic risk factors and incident advanced liver disease in non-alcoholic fatty liver disease (NAFLD): a systematic review and meta-analysis of population-based observational studies. *PLoS Medicine*. 2020;17:e1003100.
16. Poynard T, Peta V, Deckmyn O, et al. Performance of liver biomarkers, in patients at risk of nonalcoholic steato-hepatitis, according to presence of type-2 diabetes. *Eur J Gastroenterol Hepatol*. 2019;32:998-1007.
17. Bril F, McPhaul MJ, Caulfield MP, et al. Performance of plasma biomarkers and diagnostic panels for nonalcoholic steatohepatitis and advanced fibrosis in patients with type 2 diabetes. *Diabetes Care*. 2020;43:290-297.
18. Ratziu V, Charlotte F, Heurtier A, et al. Sampling variability of liver biopsy in nonalcoholic fatty liver disease. *Gastroenterology*. 2005;128:1898-1906.
19. Lambert J, Halfon P, Penaranda G, et al. How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. *Clin Chem*. 2008;54:1372-1378.
20. Poynard T, Halfon P, Castera L, et al; FibroPaca Group. Standardization of ROC curve areas for diagnostic evaluation of liver fibrosis markers based on prevalences of fibrosis stages. *Clin Chem*. 2007;53:1615-1622.
21. Poynard T, Munteanu M, Charlotte F, et al. Impact of steatosis and inflammation definitions on the performance of NASH tests. *Eur J Gastroenterol Hepatol*. 2018;30:384-391.
22. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*. 2005;41:1313-1321.
23. Boursier J, de Ledinghen V, Poynard T, et al. An extension of STARD statements for reporting diagnostic accuracy studies on liver fibrosis tests: the Liver-FibroSTARD standards. *J Hepatol*. 2015;62:807-815.
24. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2010;33:S62-S69.
25. Imbert-Bismut F, Ratziu V, Pieroni L, et al. Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: a prospective study. *Lancet*. 2001;357:1069-1075.
26. Poynard T, Ratziu V, Charlotte F, et al. Diagnostic value of biochemical markers (NashTest) for the prediction of non alcoholo steato hepatitis in patients with non-alcoholic fatty liver disease. *BMC Gastroenterol*. 2006;6:34.
27. Poynard T, Ratziu V, Naveau S, et al. The diagnostic value of biomarkers (SteatoTest) for the prediction of liver steatosis. *Comp Hepatol*. 2005;4:10.
28. Poynard T, Munteanu M, Deckmyn O, et al. Applicability and precautions of use of liver injury biomarker FibroTest. A reappraisal at 7 years of age. *BMC Gastroenterol*. 2011;11:39.
29. Poynard T, Munteanu M, Luckina E, et al. Liver fibrosis evaluation using real-time shear wave elastography: applicability and diagnostic performance using methods without a gold standard. *J Hepatol*. 2012;58:928-935.
30. Friedrich-Rust M, Poynard T, Castera L. Critical comparison of elastography methods to assess chronic liver disease. *Nat Rev Gastroenterol Hepatol*. 2016;13:402-411.
31. Poynard T, de Ledinghen V, Zarski JP, et al. Relative performances of FibroTest, Fibroscan, and biopsy for the assessment of the stage of liver fibrosis in patients with chronic hepatitis C: a step toward the truth in the absence of a gold standard. *J Hepatol*. 2012;56:541-548.
32. Poynard T, Munteanu M, Imbert-Bismut F, et al. Prospective analysis of discordant results between biochemical markers and biopsy in patients with chronic hepatitis C. *Clin Chem*. 2004;50:1344-1355.
33. McHugh LC, Snyder K, Yager TD. The effect of uncertainty in patient classification on diagnostic performance estimations. *PLoS One*. 2019;14:e0217146.
34. Bedossa P, Dargère D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. *Hepatology*. 2003;38:1449-1457.
35. Vuppalanchi R, Ünalp A, Van Natta ML, et al. Effects of liver biopsy sample length and number of readings on sampling variability in nonalcoholic Fatty liver disease. *Clin Gastroenterol Hepatol*. 2009;7:481-486.
36. Munteanu M, Tiniakos D, Anstee Q, et al. Diagnostic performance of FibroTest, SteatoTest and ActiTest in patients with NAFLD using the SAF score as histological reference. *Aliment Pharmacol Ther*. 2016;44:877-889.
37. Eddowes PJ, Sasso M, Allison M, et al. Accuracy of Fibroscan controlled attenuation parameter and liver stiffness measurement in assessing steatosis and fibrosis in patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2019;156:1717-1730.
38. Castéra L, Vergniol J, Foucher J, et al. Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C. *Gastroenterology*. 2005;128:343-350.
39. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology*. 2006;43:1317-1325.
40. Houot M, Ngo Y, Munteanu M, Marque S, Poynard T. Systematic review with meta-analysis: direct comparisons of biomarkers for the diagnosis of fibrosis in chronic hepatitis C and B. *Aliment Pharmacol Ther*. 2016;43:16-29.
41. Liu S, Schwarzinger M, Carrat F, Goldhaber-Fiebert JD. Cost effectiveness of fibrosis assessment prior to treatment for chronic hepatitis C patients. *PLoS One*. 2011;6:e26783.
42. Powell EE, Cooksley WGE, Hanson R, Searle J, Halliday JW, Powell W. The natural history of nonalcoholic steatohepatitis: a follow-up

study of forty-two patients for up to 21 years. *Hepatology*. 1990;11:74-80.

SUPPORTING INFORMATION

Additional supporting information will be found online in the Supporting Information section.

How to cite this article: Poynard T, Paradis V, Mullaert J, et al. Prospective external validation of a new non-invasive test for the diagnosis of non-alcoholic steatohepatitis, Nash-FibroTest, in patients with type 2 diabetes. *Aliment Pharmacol Ther*. 2021;54:952–966. https://doi.org/10.1111/apt.16543

APPENDIX 1

THE COMPLETE LIST OF AUTHORS' AFFILIATIONS

Thierry Poynard: Groupe Hospitalier Pitié Salpêtrière APHP, Sorbonne Université, UMRs938 & Institute of Cardiometabolism and Nutrition (ICAN), INSERM, Paris, France. Valérie Paradis and Pierre Bedossa: Department of Pathology, Physiology and Imaging, Beaujon Hospital APHP Diderot University, Paris, France. Jimmy Mullaert, Nathalie Gault, Estelle Marcault, Pauline Manchon, Nassima Si Mohammed and Cédric Laouénan: Department of Epidemiology, Biostatistics and Clinical Research, Hôpital Bichat, APHP, Paris, France. Olivier Deckmyn, Fabienne Drane, Jean Marie Castille and Valentina Peta: BioPredictive, Paris, France. Beatrice Parfait: Centre de Resources Biologiques, Hôpital Cochin APHP, Paris, France. Mark Ibberson: Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. Jean-Francois Gautier: Department of Diabetes and Endocrinology, Lariboisière Hospital APHP, Université de Paris, Paris, France. Christian Boitard: INSERM U1016, Cochin Institute, University Paris Descartes, Faculty of Medicine, Sorbonne Paris Cité, Paris, France. Sébastien Czernichow: Department of Diabetology, Cochin Hospital APHP, Inserm U1016, Institute Cochin, Paris, France; and Department of Nutrition and Diabetes, Hôpital Européen George Pompidou APHP, Paris, France. Etienne Larger: Department of Diabetology, Cochin Hospital APHP, Inserm U1016, Institute Cochin, Paris, France. Angélique Brzustowski: Inserm U1149, Centre de Recherche sur l'Inflammation CRI, Clichy, France. Benoit Terris and Stanislas Pol: Department of Anatomic Pathology, Cochin Hospital APHP, INSERM U1016, Paris, France. Anais Vallet-Pichard: Department of Hepatology, Cochin Hospital APHP, Paris, France. Dominique Roulot: Department of Hepatology, Avicenne Hospital APHP, Bobigny, France. Laurent Castera: Department of Hepatology, Beaujon Hospital APHP, Paris, France. Dominique Valla: Inserm U1149, Centre de Recherche sur l'Inflammation CRI, Clichy, France; and Department of Hepatology, Beaujon Hospital APHP, Paris, France.