Long-term outcomes and predictive ability of non-invasive scoring systems in patients with non-alcoholic fatty liver disease

Long-term outcomes and predictive ability of non-invasive scoring systems in patients with non-alcoholic fatty liver disease

Ramy Younes [1 2 †],
Gian Paolo Caviglia [3 †],
Olivier Govaere [1],
Chiara Rosso[3],
Angelo Armandi [3],
Tiziana Sanavia [3],
Grazia Pennisi [4],
Antonio Liguori [5],
Paolo Francione [6],
Rocío Gallego-Durán [7],
Javier Ampuero [7],
Maria J. Garcia Blanco[8],
Rocio Aller [9],
Dina Tiniakos [1 10],
Alastair Burt [1],
Ezio David [11],
Fabio M. Vecchio [5 12],
Marco Maggioni [13],
Daniela Cabibi [14],
María Jesús Pareja [15],
Marco Y.W. Zaki [1 16],
Antonio Grieco [5 17],
Anna L. Fracanzani [6],
Luca Valenti [18],
Luca Miele [5 17],
Piero Fariselli [3],
Salvatore Petta [4],
Manuel Romero-Gomez [7],
Quentin M. Anstee [1 19],
Elisabetta Bugianesi [3].

[1]The Newcastle Liver Research Group, Translational & Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom
[2]Boehringer Ingelheim International, GmbH, Ingelheim, Germany
[3]Department of Medical Sciences, Division of Gastroenterology and Hepatology, A.O. Città della Salute e della Scienza di Torino, University of Turin, Turin, Italy
[4]Sezione di Gastroenterologia, PROMISE, Università di Palermo, Palermo, Italy
[5]Dipartimento Universitario Medicina e Chirurgia Traslazionale, Università Cattolica del Sacro Cuore, Rome, Italy
[6]Unit of Medicine and Metabolic Disease Ca' Granda IRCCS Foundation, Policlinico Hospital, Department of Pathophysiology and Transplantation, University of Milan, Milan Italy
[7]UCM Digestive Diseases and SeLiver Group, Virgen del Rocio University Hospital, Institute of Biomedicine of Seville, University of Seville, Spain
[8]Hospital Universitario de La Princesa, Medicina Interna, Madrid, Spain
[9]Hospital Clínico de Valladolid, Valladolid, Spain
[10]Dept of Pathology, Aretaieion Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece
[11]Department of Pathology, Azienda Ospedaliero-Universitaria Città della Salute e della Scienza, University of Turin, Turin, Italy
[12]Area Anatomia Patologica. Fondazione Policlinico Gemelli IRCCS, Rome, Italy
[13]Department of Pathology, Ca' Granda IRCCS Foundation, Milan, Italy
[14]Pathology Institute, PROMISE, University of Palermo, Palermo, Italy

[15]Pathology Unit, Valme University Hospital, Seville, Spain
[16]Biochemistry Department, Faculty of Pharmacy, Minia University, Egypt
[17]Area Medicina Interna, Gastroenterologia e Oncologia Medica, Fondazione Policlinico A. Gemelli IRCCS, Rome, Italy
[18]Translational Medicine, Department of Transfusion Medicine and Hematology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[19]Newcastle NIHR Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

Correspondence quentin.anstee@ncl.ac.uk (Q.M. Anstee), elisabetta.bugianesi@ unito.it (E. Bugianesi).

Highlights

- Different non-invasive scoring systems (NSS) have been proposed to stratify patients according to the risk of advanced fibrosis.
- In the cross-sectional analysis, HFS showed the best performance for the identification of advanced fibrosis.
- NFS and FIB-4 showed the best performance for the detection of histological cirrhosis.
- After a median follow-up of ~7 years, NFS, HFS and FIB-4 performed similarly well for the prediction of HCC and overall mortality.
- All NSS had limited performance for extrahepatic events, although those incorporating diabetes performed slightly better.

Lay summary

Non-invasive scoring systems are increasingly being used in patients with non-alcoholic fatty liver disease to identify those at risk of advanced fibrosis and hence clinical complications. Herein, we compared various non-invasive scoring systems and identified those that were best at identifying risk, as well as those that were best for the prediction of long-term outcomes, such as liver-related events, liver cancer and death.

Keywords: NASH**,** NSS**,** APRI**,** BARD**,** FIB-4**,** NFS**,** HFS

Abbreviations

ALT, alanine aminotransferase; APRI, AST to platelet ratio index; AST, aspartate aminotransferase; BARD, BMI AST/ALT ratio diabetes; FIB-4, fibrosis-4; GGT, gamma-glutamyltransferase; HCC, hepatocellular carcinoma; HFS, Hepamet fibrosis score; HR, hazard ratio; MCC, Matthew's correlation coefficient; NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic steatohepatitis; NFS, NAFLD fibrosis score; NSS, non-invasive scoring systems.

Financial support

This study has been supported by the EPoS (Elucidating Pathways of Steatohepatitis) consortium funded by the Horizon 2020 Framework Program of the European Union under Grant Agreement 634413 and the Newcastle NIHR Biomedical Research Centre. The authors are contributing members of *The European NAFLD Registry*. The study was also supported by the Italian Ministry of Health, grant RF-2016-02364358 (*Ricerca Finalizzata, Ministero della Salute*), and the Italian Ministry for Education, University and Research (Ministero dell'Istruzione, dell'Università e della Ricerca - MIUR) under the programme "Dipartimenti di Eccellenza 2018 – 2022" Project code D15D18000410001.

Background & Aims
Non-invasive scoring systems (NSS) are used to identify patients with non-alcoholic fatty liver disease (NAFLD) who are at risk of advanced fibrosis, but their reliability in predicting long-term outcomes for hepatic/extrahepatic complications or death and their concordance in cross-sectional and longitudinal risk stratification remain uncertain.
Methods
The most common NSS (NFS, FIB-4, BARD, APRI) and the Hepamet fibrosis score (HFS) were assessed in 1,173 European patients with NAFLD from tertiary centres. Performance for fibrosis risk stratification and for the prediction of long-term hepatic/extrahepatic events, hepatocarcinoma (HCC) and overall mortality were evaluated in terms of AUC and Harrell's c-index. For longitudinal data, NSS-based Cox proportional hazard models were trained on the whole cohort with repeated 5-fold cross-validation, sampling for testing from the 607 patients with all NSS available.
Results
Cross-sectional analysis revealed HFS as the best performer for the identification of significant (F0-1 *vs.* F2-4, AUC = 0.758) and advanced (F0-2 *vs.* F3-4, AUC = 0.805) fibrosis, while NFS and FIB-4 showed the best performance for detecting histological cirrhosis (range AUCs 0.85-0.88). Considering longitudinal data (follow-up between 62 and 110 months), NFS and FIB-4 were the best at predicting liver-related events (c-indices>0.7), NFS for HCC (c-index = 0.9 on average), and FIB-4 and HFS for overall mortality (c-indices >0.8). All NSS showed limited performance (c-indices <0.7) for extrahepatic events.
Conclusions
Overall, NFS, HFS and FIB-4 outperformed APRI and BARD for both cross-sectional identification of fibrosis and prediction of long-term outcomes, confirming that they are useful tools for the clinical management of patients with NAFLD at increased risk of fibrosis and liver-related complications or death.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is the most common aetiology of chronic liver disease, affecting approximately 25% of the adult population.[1,2] Multiple factors contribute in the progression of the disease to its active necro-inflammatory form (non-alcoholic steatohepatitis [NASH]) and eventually to advanced fibrosis and cirrhosis; fibrosis is recognised as the most important long-term prognostic factor for overall and liver-related mortality.[3-5] Currently, both diagnosis of NASH and staging of fibrosis rely on liver biopsy,[6] but the massive number of potential patients with NASH, estimated between 15% and 20% of those with NAFLD, precludes liver biopsy for case finding. Therefore, different non-invasive scoring systems (NSS) have been proposed as simple, first-line tools to stratify patients according to the risk of advanced fibrosis and to help primary care physicians decide whether to refer a patient to a hepatologist.[7-11] However, the accuracy of these scores in predicting long-term outcomes in NAFLD remains uncertain. In 2013, a seminal multicentre study identified NAFLD fibrosis score (NFS) as the best indicator of long-term liver-related events and mortality in 320 patients with biopsy-proven NAFLD.[12] More recently, a single-centre Swedish study including 646 patients with biopsy-proven NAFLD confirmed NFS along with Fibrosis-4 (FIB-4) as the best predictors of overall mortality and severe liver disease.[13] The largest general population study performed so far evaluated NSS' long-term prediction of fatal and non-fatal liver disease in thousands of adults in Sweden.[14] The scores examined (APRI [aspartate aminotransferase (AST) to platelet ratio index], FIB-4, BARD [BMI, AST/alanine aminotransferase (ALT) ratio, diabetes], Forns and NFS) had a suboptimal performance for the 10-year prognostication

of cirrhosis and its complications in the whole population, although performance was satisfactory in people with risk factors for NAFLD at baseline. Notably, the use of non-invasive tests to estimate fibrosis in liver disease should be applied to proper population groups; indeed, according to the current guidelines, screening for NAFLD and NASH is advised in high-risk groups, but not in the general population.
In this study we aimed to investigate the long-term prognostic value of the most common NSS (NFS, FIB-4, BARD, APRI) and of the recently proposed Hepamet fibrosis score (HFS)[11] in a large, multicentre, European population of Caucasian ethnicity with biopsy-proven NAFLD.

Patients and methods

This is a multicentre cohort study of well-characterized Caucasian patients with biopsy-confirmed NAFLD who had been enrolled and prospectively followed up in tertiary centres in Italy (Turin, Milan, Rome, Palermo), the United Kingdom (Newcastle Upon Tyne) and Spain (Seville). Inclusion criteria were the diagnosis of NAFLD confirmed by liver biopsy, and age ≥18 years. From 1995 to 2015, 1,704 study participants had been selected after they had a liver biopsy performed for the suspicion of NASH based on the presence of metabolic risk factors (components of the metabolic syndrome), absence of other causes of liver disease (drug-induced liver disease, viral hepatitis, autoimmune, cholestatic and metabolic/genetic liver disease, alcohol-related) and chronically elevated LFTs or evidence of NAFLD on ultrasound. Alcohol-induced liver disease had been excluded by selecting patients with a negative history of alcohol abuse, as indicated by a weekly ethanol consumption <140 g in women and <210 g in men. The history of alcohol consumption was investigated by interviewing the patients and in many cases by also interviewing close relatives during both the first and subsequent visits. All patients underwent specific clinical, laboratory, radiographic, and/or histological evaluations. At the time of liver biopsy, clinical and laboratory data were collected, including full blood count, routine liver biochemistry and metabolic profile. Patients were followed by gastrointestinal specialists or hepatologists every year or 6 months as appropriate. At each visit, medical history was reviewed along with a routine laboratory work-up to follow the liver disease and other medical conditions. This included the assessment of liver events (end-stage cirrhosis, cirrhosis decompensation including ascites, hepatic encephalopathy and oesophageal bleeding), hepatocellular carcinoma occurrence (defined by imaging/histology criteria following current clinical guidelines),[15] cardiovascular events (acute coronary syndrome [myocardial infarction, unstable angina, need for coronary revascularisation], peripheral arterial ischaemia, acute cerebrovascular event [transient ischaemic attack, acute ischaemic or haemorrhagic stroke]), non-liver-related cancers (including breast, colorectal, lung, prostatic, haematologic, melanoma, pancreatic and urinary tract cancers) and patient deaths. For the purpose of this study, we recruited patients who underwent a liver biopsy before 2016, in order to have a minimum of 3 years of follow-up. None of the enrolled patients had clinical signs/symptoms of cirrhosis, whose diagnosis was solely based on liver biopsy (histological cirrhosis). Treatment for NAFLD consisted of standard recommendations for lifestyle change, in order to achieve and maintain appropriate body weight with increased physical activity and dietary changes.[6] No patients underwent bariatric surgery. Participants whose health status was unknown for more than 12 months after reviewing their medical records were considered lost to follow-up (n = 531, Fig. 1). Patients with advanced fibrosis or histological cirrhosis underwent endoscopic screening for gastroesophageal varices and screening for hepatocellular carcinoma (HCC) at regular intervals following standard care recommendations or guidelines in place at specific times as proposed by the liver societies. The retrospective database of each unit has been prospectively collected according to common criteria, that have been integrated in the protocol of the European NAFLD Registry,[16] established in 2010. As such, the retrospective cohort is based on a common protocol for anthropometric, clinical and biochemical variables. The study was approved by appropriate

regulatory bodies at all the participating centres, and all the patients gave written informed consent for participation in medical research.

Liver histology

All liver biopsies were judged adequate for analysis by the reporting pathologist (average 25 mm) and had an appropriate number of portal tracts for a confident grading and staging of the histological features. They were stained with haematoxylin and eosin, Masson's trichrome and special stains for iron and copper. A total of 7 experienced liver pathologists scored the liver biopsy features (A.B. and D.T. in the Newcastle centre, E.D, F.M.V., M.M. and D.C. in the Italian centres, and M.J.P. in the Seville centre) using the NASH Clinical Research Network scoring system to grade NAFLD diagnostic histological features and stage fibrosis.[17] Pathologists involved in this study participated in previous pathology consortiums where the strength of their overall agreement was above 75%.[18] Histological features included steatosis grade (0-3), lobular inflammation grade (0-3), ballooning grade (0-2), and fibrosis stage (0-4) as recommended.[17] The presence of NASH was recorded and categorized as NASH or non-NASH based exclusively on the pathologists' opinion of whether NASH was present (based on the pattern of injury and the combined presence of steatosis, ballooning and lobular inflammation).[19] A threshold of 5% of hepatocytes showing steatosis was required for the diagnosis of NAFLD. An exception was made for 7 patients who underwent liver biopsy for NAFLD suspicion and revealed cirrhosis without steatosis, as it is well demonstrated that steatosis may disappear at advanced fibrosis stages. All the aforementioned patients had previous risk factors for NAFLD.

Non-invasive scores

We assessed the long-term predictive value of the following 5 NSS, calculated at the time of liver biopsy according to their originally reported formula:

1. APRI[8]: (AST/AST upper limit normal)/(platelet count [10^9/L]) × 100;
2. BARD score[10]: 0–4 scale, BMI ≥28 kg/m^2 = 1 point, AST/ALT ratio ≥0.8 = 2 points, type 2 diabetes mellitus = 1 point;
3. FIB-4 score[20]: (age [years] × AST [U/L])/([platelets (10^9/L)] × √ALT [U/L]);
4. NFS[7]: 1.675 + 0.037 × age (years) + 0.094 × BMI (kg/m^2) + 1.13 × IFG/diabetes (yes = 1, no = 0) + (0.99 × AST/ALT ratio) (0.013 × platelet [×10^9/L]) (0.66 × albumin [g/dl]);
5. HFS[11]: 1/(1 + e [5.390 – 0.986 × age [45-64 years of age] – 1.719 × age [≥65 years of age] + 0.785 × male sex – 0.896 × AST [35-69 IU/L] – 2.126 × AST [≥70 IU/L] – 0.027 × albumin [4-4.49 g/dl] -0.897 × albumin [<4 g/dl] – 0.899 × HOMA [2-3.99 with no diabetes mellitus] – 1.497 × HOMA [≥4 with no diabetes mellitus] – 2.184 × diabetes mellitus – 0.882 × platelets × 1,000/µl [155-219] – 2.233 × platelets × 1,000/µl [<155]]).

Statistical analysis

The ability of each NSS to predict different fibrosis stages was assessed first. The distributions of the scores between different fibrosis stages were compared using Kruskal-Wallis and Dunn's tests for multiple and pairwise comparisons, respectively. Since each score is characterized by different ranges, NSS were normalized by scaling the values. The diagnostic accuracy in discriminating between different stages was evaluated using the following measures: ROC curve analysis calculating the AUC, the specificities and sensitivities estimated at different NSS cut-offs, as well as those maximizing the Youden's Index to represent the best performance achieved by each ROC curve,[21] and finally the prediction accuracy in terms of Matthew's correlation coefficient (MCC). This latter metric was chosen since it has been shown to be robust when there are unbalanced binary outcomes, *i.e.* few events occurred.[20] MCC ranges in the interval (−1 to +1), with extreme values –1

and +1 reached in case of perfect misclassification and perfect classification, respectively, while MCC = 0 for the random classifier. Three comparisons were considered between the following grouped fibrosis stages: F0-1 *vs.* F2-4 to highlight events of significant fibrosis, F0-2 *vs.* F3-4 for advanced fibrosis and F0-3 *vs.* F4 for cirrhosis outcomes. Overall accuracy was estimated using the Obuchowski index, which considers the ROC curves of all possible fibrosis score pairs.[22] The cross-sectional evaluations were performed using the maximum set of patients (608, see Fig. 1) with the fibrosis stage and all 5 NSS available, excluding the Seville cohort since it was used to define the predictive ability of the HFS score for fibrosis risk. In order to statistically assess the best performers for each fibrosis classification from the cross-sectional analysis, a paired De Long test was applied to the resulting AUCs comparing each pair of scores.
Longitudinal analysis was then performed to evaluate the ability of the NSS to discriminate the long-term incidence of the following outcomes, recorded at the end of the follow-up of each patient: liver-related events, HCC, cardiovascular events, extrahepatic cancer and overall mortality. The predictive performance of the scores was evaluated through univariate Cox proportional hazard regression analysis. For each NSS, the prognostic model was estimated using a repeated k-fold cross-validation approach, in order to evaluate the NSS on test sets of patients not used for estimating the parameters of the model, thus avoiding overfitting issues. Specifically, the dataset was split k = 5 times into pairs of training and test sets, sampling for the test data from the maximum group of patients (782–788 patients according to the type of long-term event, see Fig. 1) with the long-term outcomes and all 5 NSS available, but sampling from all the data available for each score to build the training sets. Although each NSS was assessed independently, the same training/test folds were used to evaluate all the scoring systems. This procedure was then repeated 10 times in order to provide a median value and its corresponding median absolute deviation describing a robust statistic for each performance. The median *p* values from the log-rank test and the median hazard ratios (HRs) with the corresponding median confidence intervals were estimated on the training sets, whereas median Harrell's c-indices and median cumulative AUCs integrated across time (iAUC) were estimated on the test sets.[23] We also checked whether the observed longitudinal data in the test sets were consistent with the expected data from the model built for each NSS through calibration and using the Brier score, a metric like the mean squared error, which was integrated along the time grid from the longitudinal study. Finally, in order to statistically assess the best performers for each long-term outcome, we exploited the multiple results obtained for each performance across the splits of the repeated 5-fold cross-validation and paired Wilcoxon Rank Sum test was applied to the resulting c-indices and iAUCs to compare each pair of scores. The same longitudinal analysis was also repeated to investigate the association of fibrosis stages, NASH and diabetes at baseline with long-term outcomes, using data from the whole cohort. For all the statistical tests, *p* values less than 0.05 were considered statistically significant. The analyses were performed in the R environment (v. 4.0.3), using the following main R packages: ggstatsplot (v. 0.6.5), pROC (v. 1.16.2), mltools (v. 0.3.5), pec (v. 2019.11.03), survival (v. 3.2-7), survAUC (v. 1.0.5), survcomp (v. 1.38.0).

Results

Baseline characteristics of the NAFLD cohort

From an initial cohort of 1,704 patients with biopsy-proven NAFLD, after excluding those whose health status was unknown for more than 12 months of reviewing their medical records, the study included 1,173 patients with NAFLD and complete follow-up data. The flow of enrolment is reported in Fig. 1. Baseline characteristics of the cohort are summarised in Table 1. Patients lost to follow-up had similar clinical/biochemical features except for a slightly higher BMI, platelet count and slightly lower ferritin levels, while histological features were comparable. Median age was 49 years old; 65% of the total cohort were male and 28.2% of the population had diabetes at baseline. At histology, 65.7% of patients had features of NASH, while 24.1% presented severe fibrosis (F3/F4).

Of note, in this cohort, cirrhosis was a histological finding at liver biopsy and no patient had clinical signs or symptoms of advanced liver disease. Patients from Italian centres represented 61.6% of the study-cohort, while patients from the United Kingdom and Spain represented 14.7% and 23.8% of the population, respectively (Fig. 1). Considering available biochemical and anthropometric data, APRI, FIB-4, NFS, BARD, HFS have been calculated for 1,162, 1,160, 1,103, 1,140 and 791 patients, respectively.

Cross-sectional analysis: staging of fibrosis

In the whole study population, all NSS showed a significant, stepwise increase across fibrosis stages (Fig. 2, Kruskal-Wallis test *p* values <1e-11). Similar distributions were observed also for the group of 608 patients used to evaluate the NSS (Fig. S1, Kruskal-Wallis test *p* values <1e-11). Considering Dunn's tests, none of the scores in this latter group of patients were able to significantly discriminate all the pairwise combinations of fibrosis stages. As reported in Table 2, Fig. 3A-B and Table S2, among the NSS analysed, HFS showed the highest performance for the identification of significant (F0-1 *vs.* F2-4, AUC = 0.758) and advanced fibrosis (F0-2 *vs.* F3-4, AUC = 0.805), while for the detection of histological cirrhosis (F0-3 *vs.* F4) HFS, NFS and FIB4 showed comparable performance in terms of AUC and statistically higher than the AUCs obtained for APRI and BARD, according to the paired De Long test comparisons and the ROC curves (Table 2 and Fig. 3C). Obuchowski index confirmed that, overall, APRI and BARD performed worse than the other 3 NSS (indices <0.71, Table 2). Conversely, HFS and NFS were the best performing NSS to assess the fibrosis levels (Obuchowski indices 0.802 and 0.79, respectively). In order to have a graphical view of the performance in terms of specificity and sensitivity to understand the best usage of these scores in assessing the fibrosis risk, we also reported in Fig. S2 how these 2 metrics change at different cut-offs of each score.
Finally, if we restrict our evaluation to the 126 patients of the cohort with type 2 diabetes and all the 5 NSS available, an overall decrease in the performance of HFS and NFS was observed, as expected (Tables S3 and S4). However, it is worth noting that, in this population, APRI and FIB-4 were the best performers for the identification of significant fibrosis, while both FIB-4 and NFS showed significantly higher predictive ability in identifying cirrhotic events with respect to APRI and HFS, but not to BARD. Considering all the possible pairwise comparisons between fibrosis stages, FIB-4 showed the highest Obuchowski index (0.768).

Longitudinal analysis: prediction of long-term outcomes from fibrosis status or diagnosis of either NASH or diabetes

Firstly, we investigated the long-term outcomes recorded at the end of the follow-up (*i.e.* liver-related events, HCC, cardiovascular events, extrahepatic cancer events and mortality) in our cohort with respect to the fibrosis status, the histological diagnosis of NASH and the presence of diabetes at baseline. After a median follow-up of 81 months (IQR 62–110 months), 82 (7 %) patients developed liver-related events (end-stage cirrhosis, cirrhosis decompensation including ascites, hepatic encephalopathy and oesophageal bleeding), 17 (1.5 %) patients developed HCC, 103 (8.8 %) patients reported cardiovascular events and 96 (8.2 %) extrahepatic cancers. Cumulative mortality from all causes was 2.6% (31 deaths).
Univariate Cox models were trained (Tables S5 and S6 for log-rank test *p* values, HRs and their confidence intervals) and tested with 5-fold cross-validation, repeating the procedure 10 times. Median Harrell's c-indices obtained from the test data are reported in Table 3 (with corresponding iAUCs and integrated Brier scores in Table S7). For the liver-related events, all the classifications of fibrosis stages showed better associations on both training and test sets with respect to NASH and diabetes (c-indices >0.62 and log-rank test *p* values <1e-6, iAUCs >0.62). For HCC events, advanced fibrosis showed the best Harrell's c-index (0.853 ± 0.0179) and iAUC (0.799 ± 0.0346). Due to the low

number of HCC events (n = 17/1,168), it was not feasible to estimate the predictive performance of stratification of fibrosis stages into F0 *vs.* F1-2 *vs.* F3-4. Cox models estimated for cardiovascular and extrahepatic cancer events showed, overall, the worst performance in terms of c-index (median <0.62, Table 3), iAUC (median <0.59) and Brier score (>0.13, see Table S7) on test data. The only clinical factor showing log-rank test *p* values on training sets <0.01 for cardiovascular and extrahepatic cancer events was, as expected, the diagnosis of diabetes at baseline (Table S5), which also led to the best Harrell's c-indices on test data (Table 3). Considering overall mortality, advanced fibrosis and the stratification of fibrosis stages into F0 *vs.* F1-2 *vs.* F3-4 showed the strongest associations in both training and test data (median Harrell's c-indices >0.8 and median iAUC >0.7 with median integrated Brier scores <0.05).

Longitudinal analysis: prediction of long-term outcomes using NSS

The results obtained from the longitudinal analysis applied to the NSS showed that, overall, NFS, HFS and FIB-4 performed better than APRI and BARD scores (see Table 4 and Tables S8-9). In terms of Harrell's c-index on test data (Table 4), NFS showed the best performance for HCC (0.901 ± 0.0302, with best median iAUC = 0.889 ± 0.0486) and, together with FIB-4, the best results for liver-related events (c-indices and iAUCs >0.75), while HFS and FIB-4 showed the best performance for mortality (c-indices >0.8 and iAUCs >0.85). Performance of NSS was less satisfactory on non-hepatic events: NFS and BARD were the only scores that showed a significant log-rank test *p* value (Table S8) for cardiovascular events, but both with Harrell's c-indices <0.7 and iAUCs <0.6; similarly, only Cox models on NFS, BARD and HFS had significant log-rank test *p* values for extrahepatic cancer events (Table S8), but with both c-indices and iAUCs <0.7. In terms of calibration, all the Cox models for non-hepatic events showed the worst median integrated Brier scores (>0.11, Table S10), indicating that these complex and multi-factorial events require the integration of other types of variables in order to achieve a better prediction.

Discussion

To date, this is the largest study in a European cohort of patients with biopsy-proven NAFLD and considerable follow-up. This study provides meaningful insights both on the natural history of Caucasian patients with NAFLD according to histological degree of liver damage and on the reliability of currently recommended NSS for risk stratification and long-term event prediction.
After ~7 years on average, the most common events in our cohort were, in order of frequency, cardiovascular, non-liver cancer, and cirrhosis complications. However, if we include HCC in liver-related events, the latter would be the second leading cause of morbidity, reported by 8.4% of patients. These results are in line with previous findings summarised in a recent meta-analysis,[24] where liver-related events developed in approximately 11% of patients with NAFLD across 8 studies. In our study, mortality was low (31 deaths), in agreement with the exclusion of clinically overt cirrhosis at first diagnosis, emphasizing once again that an early diagnosis would reduce the burden of end-stage liver disease deaths. In our cohort, stage of biopsy-confirmed liver fibrosis, but not NASH, was the strongest predictor of future hepatic morbidity and, considering the classification of patients in terms of advanced/severe fibrosis, HCC and all-cause mortality regardless of age, sex, BMI and diabetes (Table 3). This is consistent with several previously published studies and a recent systematic review.[25]
In the cross-sectional analysis, all the investigated NSS showed a significant, stepwise increase according to fibrosis stages (Fig. 2). NFS, FIB-4 and HFS discriminated cirrhosis significantly better with respect to APRI and BARD, showing AUCs >0.8 (Table 2), but HFS also showed the best performance in the identification of significant (AUC = 0.758) and advanced (AUC = 0.805) fibrosis in this first validation, thus proving to be a reliable tool in detecting different stages of fibrosis, as confirmed by the highest Obuchowski score observed (0.802). An important issue is the reliability of

NSS in diabetic patients. Not unexpectedly, all NSS including diabetes at baseline as a covariate showed a decreased performance in cross-sectional identification of fibrosis when assessed only on diabetic patients. Despite the lower statistical power due to the number of patients considered, APRI and FIB-4, including low values of albumin and platelets and high AST/ALT ratio or AST/platelet ratio, were able to perform significantly better than the other scores in the identification of diabetic patients with a higher risk of significant fibrosis, reporting AUCs >0.7, with FIB-4 reporting the highest Obuchowski score for all the pairwise comparisons (0.768, Table S3).

In the longitudinal analysis, NFS and FIB-4 showed the best performance for predicting liver-related events, while for HCC events, despite high c-indices (>0.8) for these scores and HFS, NFS performed significantly better than any other NSS (c-index 0.901 ± 0.0302). However, NFS is less performant compared to HFS and FIB-4, which performed best compared to the other NSS (c-indices >0.8). It is worth highlighting that the number of events reported for both HCC and overall mortality is low (only 8 and 15 events in the group of about 785 patients, respectively), therefore trying to establish statistically robust best performers for these 2 long-term events is challenging. We can state that, overall, NSS which performed best in the cross-sectional identification of fibrosis risk were also better predictors of the occurrence of liver-related events, HCC and overall mortality.

On the other hand, the low performance of all the scores for cardiovascular and non-hepatic cancer events suggests that these NSS are not flexible enough to address the prediction of long-term extrahepatic events, which requires the integration of additional covariates. As expected, we found that diabetes at baseline, as well as the scores including diabetes as a covariate (NFS and HFS), performed slightly better in the prediction of these events (Table 3, Table 4). Specifically, considering only those scores reporting significant log-rank test *p* values in the training data (Table S8), NFS and BARD were the only significantly robust models in training able to perform best on cardiovascular events and on extrahepatic cancer events (Table 4 – together with HFS, which was able to achieve a significant log-rank test *p* value [<0.001] only for this latter type of event [Table S8]). Quite interestingly, in our cohort, a significant difference in the cardiovascular event rate was found in patients with significant fibrosis (12.2% in F2/3/4 *vs.* 5.9% in F0/1 respectively, $p = 0.011$), but not in patients with severe fibrosis (F3/4), in agreement with another study that identified a higher incidence of cardiovascular events in individuals with stage 3 fibrosis compared to stage 4.[26] This is consistent with the fact that relevant comorbidities do already exist before the development of fibrosis and represent the major causes of mortality in the early stages of fibrotic NASH.

As repeatedly reported, baseline diagnosis of NASH was not associated with any event in both the training and test sets (log-rank test *p* values >0.05 for all events, Table S5); nevertheless, NASH is the driving mechanism ultimately leading to liver scarring and disease progression,[27] but this cannot be demonstrated on the basis of the available data. On the other hand, we might speculate that the current paradigm of excluding patients without NASH from clinical trials might not be prudent since such patients might in fact develop NASH and, as this study shows, have a similar risk of liver-related outcomes.[28] The natural history of NASH is much less predictable than that of other chronic liver diseases. Most likely, liver damage does not follow a simple path from steatosis to steatohepatitis to cirrhosis, but progression results from repetitive bouts of inflammation alternating with a reparative immune response.[27,29] This waxing and waning of hepatic injury is highly related to lifestyle changes or to the onset of diabetes that can quickly affect the course of inflammation/fibrosis,[28] which are not captured by the scores at baseline. In fact, in our cohort, incident diabetes led to a higher cumulative probability of cardiovascular events, extrahepatic cancers, overall mortality and liver-related outcomes, the latter occurring in 10.7% of patients who developed diabetes compared with 4.1% in those who did not.

Our study presents several strengths. To the best of our knowledge, this is the largest multicentre study evaluating non-invasive fibrosis algorithms in patients with biopsy-proven NAFLD and a follow-up averaging almost a decade. In addition, it provides the first validation of the long-term prognostic value of the recently published HFS.

We acknowledge our study has some limitations, such as the enrolment bias at tertiary centres or the absence of a shared treatment protocol across all centres. Further, we were not able to adjust for specific treatment modalities. The lack of a central pathologist scoring the liver biopsies is another limitation of the study. However, all recruiting centres are well-known for their interest in NAFLD and involved in several scientific collaborations with homogeneous protocols for patient enrolment and NAFLD histological scoring through the years. We also recognise that our study was carried out in Caucasians only, thus these results cannot be generalised to other ethnicities. In addition, data on modern imaging tools or novel blood-based non-invasive biomarkers were not available, so that a direct comparison was not possible. Finally, as already reported above, the number of events available for HCC and overall mortality are limited (<3% with respect to the number of patients). These events were further limited in the test data used for assessing the long-term outcome predictions, but it is worth highlighting that all the available events in the cohort were used for training the Cox models using the repeated 5-fold cross-validation schema. Nevertheless, this is the largest cohort specific for patients with biopsy-proven NAFLD (without overt cirrhosis) reported so far; we hope that future studies will provide bigger datasets in order to define meta-analyses investigating long-term outcomes with higher statistical robustness.
In conclusion, we believe that this study provides a comprehensive critical description of the ability of NSS to identify patients with NAFLD at higher risk of liver-related complications and overall mortality. Among the 5 scores analysed, NFS, FIB-4 and HFS were shown to be reliable tools for prediction of cirrhosis and, alternately, long-term liver-related events (NFS and FIB-4), HCC (NFS) and overall mortality (FIB-4 and HFS), with HFS showing statistically higher performance in the prediction of significant and severe fibrosis with respect to the other NSS scores (Table 2, Table 4). In diabetic patients with NAFLD, scores including diabetes in the algorithm have inferior performance, whereas APRI or FIB-4 perform better for significant fibrosis and FIB-4 for overall fibrosis risk predictions (Table S3). However, the scores including diabetes are more effective compared to the other scores in predicting cardiovascular and extrahepatic cancer events, despite overall performance for these events being low (c-indices always <0.7, Table 4). Lastly, beyond the main result represented by the overall higher performance observed in NFS, HFS and FIB-4 with respect to APRI and BARD, it is worth noting that none of the scores were able to achieve very high accuracy in terms of prediction of fibrosis levels (MCCs above 0 but all <0.5, Tables S2 and S4) In the longitudinal analysis, we only observed Harrell's c-indices above 0.8 (Table 4) in a few cases (HCC and overall mortality). These results suggest that further solutions need to be developed, with consideration given to machine learning-based approaches.

References

[1] Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016;64:73–84.
[2] Marengo A, Jouness RI, Bugianesi E. Progression and natural history of nonalcoholic fatty liver disease in adults. Clin Liver Dis 2016;20:313–324.
[3] Eslam M, Sanyal AJ, George J, International Consensus Panel. MAFLD: a consensus-driven proposed nomenclature for metabolic associated fatty liver disease. Gastroenterology 2020;158:1999–2014.
[4] Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, et al. Liver fibrosis, but No other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology 2015;149:389–397.
[5] Hagstrom H, Nasr P, Ekstedt M, Hammar U, Stal P, Hultcrantz R, 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.

[6] European Association for the Study of the Liver, European Association for the Study of Diabetes, European Association for the Study of Obesity. EASL-EASD-EASO Clinical Practice Guidelines for the management of nonalcoholic fatty liver disease. J Hepatol 2016;64:1388–1402.
[7] Angulo P, Hui JM, Marchesini G, Bugianesi E, George J, Farrell GC, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007;45:846–854.
[8] Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology 2003;38:518–526.
[9] Shah AG, Lydecker A, Murray K, Tetri BN, Contos MJ, Sanyal AJ, et al. Comparison of noninvasive markers of fibrosis in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2009;7:1104–1112.
[10] Harrison SA, Oliver D, Arnold HL, Gogia S, Neuschwander-Tetri BA. Development and validation of a simple NAFLD clinical scoring system for identifying patients without advanced disease. Gut 2008;57:1441–1447.
[11] Ampuero J, Pais R, Aller R, Gallego-Durán R, Crespo J, García-Monzón C, et al. Development and validation of Hepamet fibrosis scoring system-A simple, noninvasive test to identify patients with nonalcoholic fatty liver disease with advanced fibrosis. Clin Gastroenterol Hepatol 2020;18:216–225.
[12] Angulo P, Bugianesi E, Bjornsson ES, Charatcharoenwitthaya P, Mills PR, Barrera F, et al. Simple noninvasive systems predict long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology 2013;145:782–789.
[13] Hagstrom H, Nasr P, Ekstedt M, Stal P, Hultcrantz R, Kechagias S. Accuracy of noninvasive scoring systems in assessing risk of death and liver-related endpoints in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2019;17:1148–1156.
[14] Hagstrom H, Talback M, Andreasson A, Walldius G, Hammar N. Ability of noninvasive scoring systems to identify individuals in the population at risk for severe liver disease. Gastroenterology 2020;158:200–214.
[15] European Association for the Study of the Liver. EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2018;69:182–236.
[16] Hardy T, Wonders K, Younes R, Aithal GP, Aller R, Allison M, et al. The European NAFLD Registry: a real-world longitudinal cohort study of nonalcoholic fatty liver disease. Contemp Clin Trials 2020;98:106175.
[17] Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005;41:1313–1321.
[18] 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.
[19] Bedossa P, Poitou C, Veyrie N, Bouillot JL, Basdevant A, Paradis V, et al. Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. Hepatology 2012;56:1751–1759.
[20] Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317–1325.
[21] Perkins NJ, Schisterman EF. The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006;163(7):670–675.
[22] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21:6.
[23] Lambert J, Halfon P, Penaranda G, Bedossa P, Cacoub P, Carrat F. How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. Clin Chem 2008;54:1372–1378.

[24] Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. JASA 2007;102:527–537.
[25] Taylor RS, Taylor RJ, Bayliss S, Hagstrom H, Nasr P, Schattenberg JM, et al. Association between fibrosis stage and outcomes of patients with nonalcoholic fatty liver disease: a systematic review and meta-analysis. Gastroenterology 2020;158:1611–1625.
[26] Sutti S, Albano E. Adaptive immunity: an emerging player in the progression of NAFLD. Nat Rev Gastroenterol Hepatol 2020;17(2):81–92.
[27] Vilar-Gomez E, Calzadilla-Bertot L, Wai-Sun Wong V, Castellanos M, Aller-de la Fuente R, Metwally M, et al. Fibrosis severity as a determinant of cause-specific mortality in patients with advanced nonalcoholic fatty liver disease: a multi-national cohort study. Gastroenterology 2018;155:443–457.
[28] McPherson S, Hardy T, Henderson E, Burt AD, Day CP, Anstee QM. Evidence of NAFLD progression from steatosis to fibrosing-steatohepatitis using paired biopsies: implications for prognosis and clinical management. J Hepatol 2015;62:1148–1155.
[29] Schuppan D, Surabattula R, Wang XY. Determinants of fibrosis progression and regression in NASH. J Hepatol 2018;68:238–250.

Table 1. Clinical and demographic characteristics of the total patient population.

Variable [n]	**Total (n = 1,173)**
Age (years) [1,173]	49 (38–57)
Sex	
Female	413
Male	760
BMI (kg/m^2) [1,136]	29.4 (26.3–33.8)
Waist circumference (cm) [937]	101 (93–110)
ALT (IU) [1,169]	59 (41–88)
AST (IU) [1,170]	37 (28–54)
Total bilirubin (mg/dl) [1,138]	0.6 (0.5–0.9)
Albumin (g/dl) [1,047]	4.5 (4.3–4.8)
Alkaline phosphatase (IU/L) [1,117]	82 (64–113)
Platelet (x10^9) [1,164]	225 (186–269)
Glucose (mg/dl) [1,122]	96 (86–114)
Triglycerides (mg/dl) [1,125]	136 (96–190)
Total cholesterol (mg/dl) [1,133]	194 (167–227)
HDL-cholesterol (mg/dl) [1,039]	48 (41–58)
LDL-cholesterol (mg/dl) [1,007]	118 (91–146)
Ferritin (ng/ml) [983]	175 (90–314)
Diabetes at baseline [1,173]	331
Steatosis grade [1,173]	
0	7*
1	470
2	426
3	270
Fibrosis stage [1,173]	
0/1/2	890
3/4	283
NASH [1,173]	771

Data are presented as median (IQR), or number of patients with a condition. Number in square

brackets after each variable indicates the number of patients who had that variable measured.
ALT, alanine aminotransferase; AST, aspartate aminotransferase; NASH, non-alcoholic steatohepatitis.
∗Patients who underwent liver biopsy for suspicion of NASH and showed F4 fibrosis at histology, with steatosis less than 5%.

Table 2. Diagnostic accuracy of NSS to predict fibrosis.

Empty Cell	**F0-1 (n = 331) *vs.* F2-4 (n = 277)**	**F0-2 (n = 477) *vs.* F3-4 (n = 131)**	**F0-3 (n = 573) *vs.* F4 (n = 35)**	**Obuchowski**
APRI	0.669 (0.626–0.712)	0.72 (0.67–0.771)	0.709 (0.612–0.805)	0.704 (0.622–0.786)
FIB-4	0.697 (0.656–0.739)	0.733 (0.682–0.783)	0.856∗ (0.801–0.911)	0.774 (0.713–0.836)
NFS	0.7 (0.658–0.742)	0.761 (0.715–0.808)	0.876∗ (0.819–0.933)	0.79 (0.729–0.851)
BARD	0.651 (0.609–0.692)	0.677 (0.627–0.726)	0.736 (0.655–0.818)	0.698 (0.624–0.772)
HFS	0.758∗ (0.719–0.796)	0.805∗ (0.764–0.845)	0.82∗ (0.768–0.872)	0.802 (0.746–0.858)

AUC values calculated on the 608 patients with available data for all 5 scores and excluding the Seville cohort. Obuchowski index was computed as in Lambert *et al.* (Clin Chem, 2008). Lower and upper confidence intervals are in brackets. For each fibrosis group comparison displayed on the top of the table, the number of patients corresponding to each group is reported in brackets.
APRI, aspartate aminotransferase to platelet ratio index; BARD, BMI aspartate aminotransferase/alanine aminotransferase ratio diabetes; FIB-4, fibrosis-4; HFS, Hepamet fibrosis score; NFS, NAFLD fibrosis score; NSS, non-invasive scoring systems.
∗Reporting statistically higher AUCs with respect to the NSS without asterisk for the same comparison, according to the paired 2-sided De Long test (p value <0.05).

Table 3. Testing associations between the long-term outcomes and fibrosis stage, diagnosis of NASH and diabetes using univariate cox proportional hazard models.

Empty Cell	**Liver-related events (82/1,170)**	**HCC (17/1,168)**	**Cardiovascular events (103/1,170)**	**Extrahepatic cancer (96/1,165)**	**Mortality (31/1,173)**
F0 *vs.* F1-2 *vs.* F3-4	0.685 ± 0.0187∗	n.c.	0.518 ± 0.0392	0.462 ± 0.0209	0.821 ± 0.0355∗

Empty Cell	**Liver-related events (82/1,170)**	**HCC (17/1,168)**	**Cardiovascular events (103/1,170)**	**Extrahepatic cancer (96/1,165)**	**Mortality (31/1,173)**
Significant fibrosis (F0-1 *vs.* F2-4)	0.626 ± 0.0326∗	0.755 ± 0.00665	0.543 ± 0.0193	0.491 ± 0.0198	0.727 ± 0.0219
Advanced fibrosis (F0-2 *vs.* F3-4)	0.685 ± 0.0241∗	0.853 ± 0.0179∗	0.535 ± 0.0196	0.492 ± 0.0307	0.814 ± 0.031∗
Cirrhosis (F0-3 *vs.* F4)	0.693 ± 0.00925∗	0.665 ± 0.095	0.519 ± 0.0128	0.529 ± 0.0113	0.732 ± 0.0382
NAFL *vs.* NASH	0.455 ± 0.0269	0.647 ± 0.017	0.49 ± 0.0386	0.479 ± 0.0153	0.416 ± 0.0391
Diabetes	0.58 ± 0.0228	0.545 ± 0.117	0.611 ± 0.0188∗	0.581 ± 0.0208∗	0.652 ± 0.027

Median Harrell's c-indices with corresponding median absolute deviations estimated on the test sets from the cross-validation. For each long-term outcome displayed on the top of the table, number of events/total number of outcomes are reported in brackets.
n.c., not calculable, due to overfitting issues; HCC, hepatocellular carcinoma; NAFL, non-alcoholic fatty liver; NASH, non-alcoholic steatohepatitis.
∗Reporting statistically higher c-indices with respect to the clinical characteristics without asterisk for the same comparison, according to the paired Wilcoxon Rank Sum test (*p* value <0.05).

Table 4. Testing associations between NSS and long-term outcomes using univariate cox proportional hazard models.

Empty Cell	Liver-related events (42/787)	HCC (8/785)	Cardiovascular events (65/787)	Extrahepatic cancer (68/782)	Mortality (15/788)
APRI	0.600 ± 0.0351	0.788 ± 0.0362	0.467 ± 0.0293	0.514 ± 0.0184	0.703 ± 0.0669
FIB-4	0.783 ± 0.0288∗	0.853 ± 0.0516	0.6 ± 0.0253∗	0.614 ± 0.0153	0.850 ± 0.0135∗
NFS	0.796 ± 0.0231∗	0.901 ± 0.0302∗	0.648 ± 0.0394∗	0.661 ± 0.0209∗	0.789 ± 0.0991
BARD	0.728 ± 0.0181	0.772 ± 0.0345	0.644 ± 0.0442∗	0.624 ± 0.0105∗	0.571 ± 0.0205
HFS	0.729 ± 0.0175	0.824 ± 0.0578	0.633 ± 0.0202∗	0.641 ± 0.0381∗	0.849 ± 0.0187∗

Median Harrell's c-indices with corresponding median absolute deviations were estimated on the test sets from the cross-validation from the patients with available data for the 5 scores. For each long-term outcome displayed on the top of the table, number of events/total number of outcomes are reported in brackets. APRI, aspartate aminotransferase to platelet ratio index; BARD, BMI aspartate aminotransferase/alanine aminotransferase ratio diabetes; FIB-4, fibrosis-4; HCC, hepatocellular carcinoma; HFS, Hepamet fibrosis score; NFS, NAFLD fibrosis score; NSS, non-invasive scoring systems.
∗Reporting statistically higher c-indices with respect to the NSS without asterisk for the same comparison, according to the paired Wilcoxon Rank Sum test (*p* value <0.05).

Graphical abstract

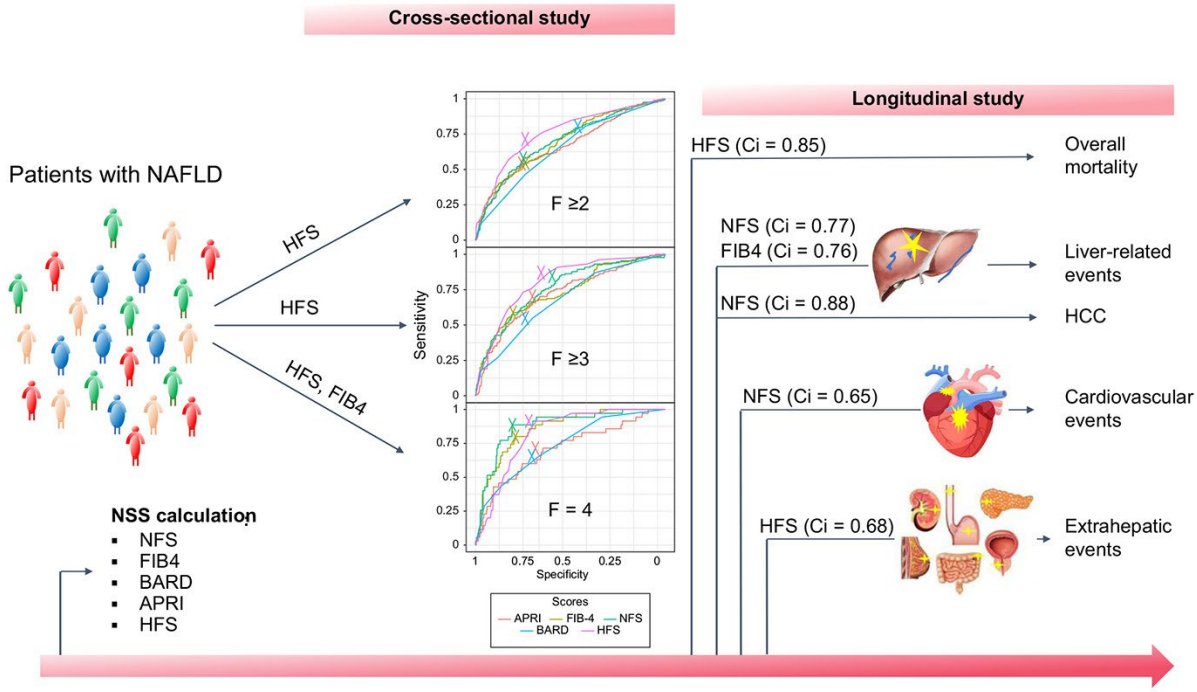

Fig. 1. Flow chart of patient enrolment in the study cohort.
∗Patients whose health status was unknown for more than 12 months of reviewing their medical records. †Sevilla cohort was excluded. ∗∗Patients with available data for the calculation of all the five scores (APRI, FIB-4, NFS, BARD and HFS). HCC, hepatocellular carcinoma; NSS, non-invasive scoring systems.

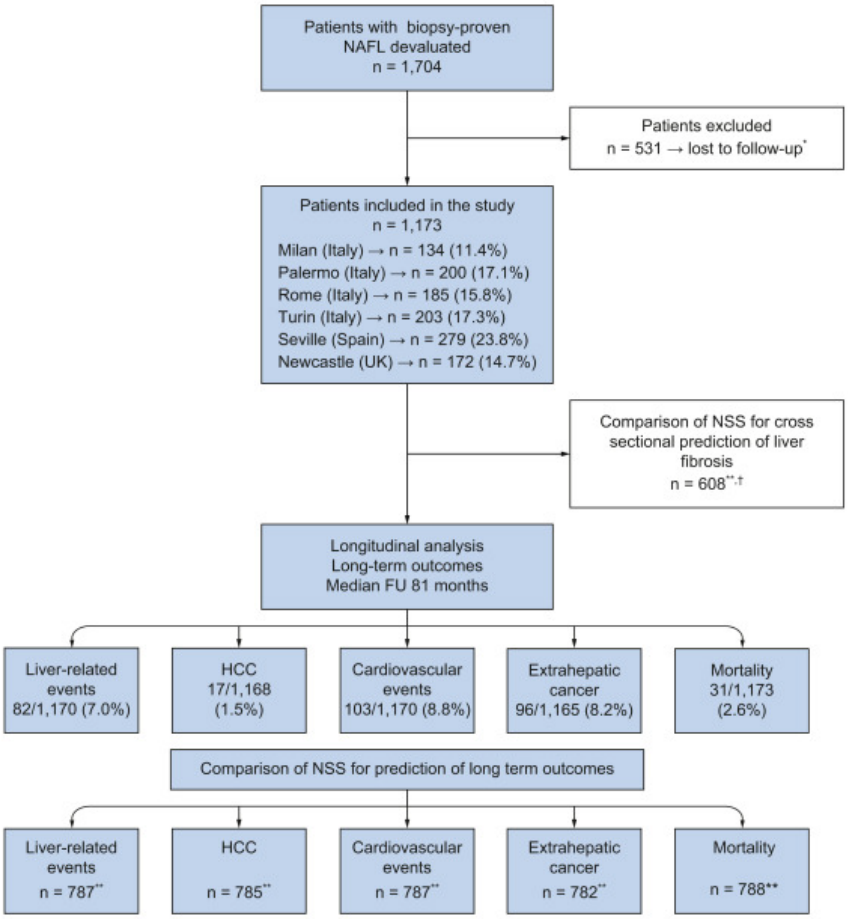

Fig. 2. Box-Violin plot of score values according to fibrosis stage on the 608 patients with all 5 NSS available.
Kruskal-Wallis and Dunn tests for multiple and pairwise comparisons, respectively. ∗: showing Dunn test *p* value <0.05. ∗∗: showing Dunn test *p* value <0.01. ∗∗∗: showing Dunn test *p* value <0.001. APRI, aspartate aminotransferase to platelet ratio index; BARD, BMI aspartate aminotransferase/alanine aminotransferase ratio diabetes; FIB-4, fibrosis-4; HFS, Hepamet fibrosis score; NFS, NAFLD fibrosis score; NSS, non-invasive scoring systems. (This figure appears in color on the web.)

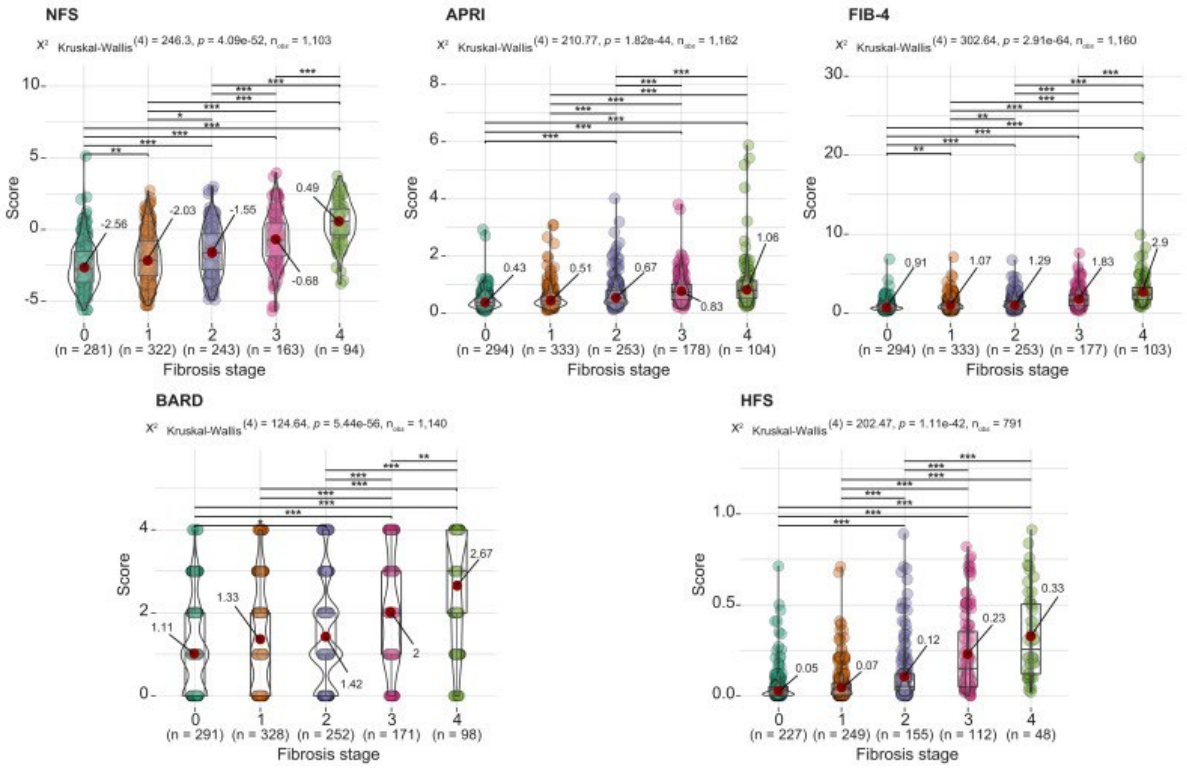

Fig. 3. ROC curves of the 5 NSS.

(A) Prediction of significant fibrosis, (B) severe fibrosis and (C) cirrhosis. The "X" symbols correspond to the estimated Youden index points used to calculate MCC, specificity and sensitivity. APRI, aspartate aminotransferase to platelet ratio index; BARD, BMI aspartate aminotransferase/alanine aminotransferase ratio diabetes; FIB-4, fibrosis-4; HFS, Hepamet fibrosis score; MCC, Matthew's correlation coefficient; NFS, NAFLD fibrosis score; NSS, non-invasive scoring systems. (This figure appears in color on the web.)