International Journal of Medical Informatics 185 (2024) 105373

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

NAIF: A novel artificial intelligence-based tool for accurate diagnosis of stage F3/F4 liver fibrosis in the general adult population, validated with three external datasets

Samir Hassoun [a,1,*], Chiara Bruckmann [a,1,*], Stefano Ciardullo [b,c], Gianluca Perseghin [b,c], Fabio Marra [d], Armando Curto [d], Umberto Arena [d], Francesco Broccolo [e,*], Francesca Di Gaudio [a,f]

[a] *Unità Operativa Centro Controllo Qualità e Rischio Chimico (CQRC), Azienda Ospedaliera Villa Sofia Cervello, viale Strasburgo 233, 90146 Palermo, Italy*
[b] *Department of Medicine and Surgery, University of Milano-Bicocca, via Modigliani 10, 20900 Monza, Italy*
[c] *Department of Medicine and Rehabilitation, Policlinico di Monza, Monza, via Modigliani 10, 20900 Monza, Italy*
[d] *Dipartimento di Medicina Sperimentale e Clinica, University of Florence, Largo Giovanni Alessandro Brambilla, 3, 50134 Firenze Italy*
[e] *Department of Experimental Medicine, University of Salento, 73100 Lecce, Italy*
[f] *PROMISE-Promotion of Health, Maternal-Childhood, Internal and Specialized Medicine of Excellence G. D'Alessandro, Piazza delle Cliniche, 2, 90127 Palermo, Italy*

ARTICLE INFO

Keywords:
AI diagnostic tools
Clinical decision support
Liver fibrosis prediction
Machine learning
Noninvasive diagnostic methods
External validation

ABSTRACT

Objective: The purpose of this study was to determine the effectiveness of a new AI-based tool called NAIF (NAFLD-AI-Fibrosis) in identifying individuals from the general population with advanced liver fibrosis (stage F3/F4). We compared NAIF's performance to two existing risk score calculators, aspartate aminotransferase-to-platelet ratio index (APRI) and fibrosis-4 (Fib4).
Methods: To set up the algorithm for diagnosing severe liver fibrosis (defined as Fibroscan® values E ≥ 9.7 KPa), we used 19 blood biochemistry parameters and two demographic parameters in a group of 5,962 individuals from the NHANES population (2017–2020 pre-pandemic, public database). We then assessed the algorithm's performance by comparing its accuracy, precision, sensitivity, specificity, and F1 score values to those of APRI and Fib4 scoring systems.
Results: In a kept-out sub dataset of the NHANES population, NAIF achieved a predictive precision of 72 %, a sensitivity of 61 %, and a specificity of 77 % in correctly identifying adults (aged 18–79 years) with severe liver fibrosis. Additionally, NAIF performed well when tested with two external datasets of Italian patients with a Fibroscan® score E ≥ 9.7 kPa, and with an external dataset of patients with diagnosis of severe liver fibrosis through biopsy.
Conclusions: The results of our study suggest that NAIF, using routinely available parameters, outperforms in sensitivity existing scoring methods (Fib4 and APRI) in diagnosing severe liver fibrosis, even when tested with external validation datasets. NAIF uses routinely available parameters, making it a promising tool for identifying individuals with advanced liver fibrosis from the general population.
Word count abstract: 236.

1. Introduction

Liver fibrosis (LF) is caused by an excess extracellular matrix (mainly collagens) associated with chronic liver injury and nodular regeneration. LF can eventually lead to cirrhosis and/or hepatocellular carcinoma [1]. Alcohol abuse, nonalcoholic fatty liver disease (NAFLD), nonalcoholic steatohepatitis (NASH), and chronic HBV or HCV infection are leading causes of LF. The classic method for evaluating LF is liver biopsy [2,3]. Unfortunately, liver biopsy is an invasive and expensive procedure with a small but significant burden of complications [4].

New guidelines recommend noninvasive methods for assessing liver fibrosis, reducing the need for liver biopsy. The most widely used

* Corresponding authors.
E-mail addresses: samir.dipatre.hassoun@gmail.com (S. Hassoun), bruckmannchiara@gmail.com (C. Bruckmann), francesco.broccolo@unisalento.it (F. Broccolo).

[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.ijmedinf.2024.105373
Received 5 October 2023; Received in revised form 5 February 2024; Accepted 7 February 2024
Available online 10 February 2024

method is liver stiffness measurement using Fibroscan® [5,6]. Blood-based biomarkers and routine laboratory tests have also been proposed [7]. The four-factor fibrosis index (Fib4) [8,9] and aspartate aminotransferase platelet ratio (APRI) [10,11] scores are commonly used scores. Fib4 and APRI were originally developed to identify liver scarring in patients with HCV and HBV, to determine whether a biopsy is necessary. While Fib4 and APRI tests have adequate specificity for LF, their sensitivity is scarce. Additionally, Fib4 has an "indeterminate" category that 30–40 % of patients fall into, making it difficult to use as a screening tool. [12]. Lastly, Fib4 is only applicable to individuals between the ages of 35 and 65.

Machine learning (ML) algorithms can identify non-linear correlations in laboratory, demographic, and clinical parameters that linear methods like Fib4 and APRI scores cannot detect [13–19]. This can lead to improved diagnostic ability. Several algorithms optimized to identify LF [15,20–22] use anthropometric and clinical parameters (such as body mass index, abdominal circumference, or blood pressure values). However, relying on these parameters limits the algorithm's use to patients who attend medical appointments. A risk assessment tool that predicts the presence of LF based on routine blood values, readily available to everyone, could enable massive population screening and be clinically beneficial. Such a tool would not depend on prior suspicion of liver disease or other conditions like diabetes, dyslipidemia, or hypertension.

Our research aimed to develop a tool that detects advanced fibrosis in the general population using blood biomarkers. This allows early treatment before complications arise.

Previously, we developed an ML model incorporating medical records as a clinical variable to classify LF and we tested it on a group of 5962 individuals, finding that the SMOTE-NC oversampling SVM model had the highest predictive power [22]. In the present study, we maintained the same pipeline as reported previously but removed biometric parameters and physical exams from our features, evaluating the performance of our simplified AI-based tool, NAIF (NAFLD-AI-Fibrosis) exclusively based on 19 laboratory blood parameters (alaninaminotranspherase (ALT), blood albumin, alkaline phosphatase (ALP), aspartataminotranspherase (AST), glucose, gammagliutamyltranspherase (GGT), total bilirubin, triglycerides, uric acid, percentage of lymphocytes, percentage of neutrophils, total count of red blood cells, hemoglobin, hematocrit, platelet count, glycated hemoglobin, C-reactive protein, ferritin, HDL cholesterol), plus the age and sex of the subjects. Then, we compared the results of NAIF, APRI and Fib4 to identify patients with severe LF. Finally, we performed external validation with three different datasets to ensure proper evaluation of our model [23].

2. Methods

The study population of the NHANES dataset is described in **Supplementary information**. **Supplementary Table S1** summarizes the demographic and laboratory characteristics of all subjects included in the study (5962 individuals), and the statistical difference between patients with non-fibrosis and patients with LF for all parameters, compared with a paired *t*-test (Excel Microsoft), while statistical details of the two subgroups (patients with non-fibrosis and patients with LF) were previously reported in [22].

2.1. Construction of predictive model

2.1.1. Model architecture

For the construction and validation of the learning algorithm, the software Orange v3.34.0 [24], developed by Bioinformatics Lab at the University of Ljubljana, Slovenia, in open source, was used. Details of the data preparation and model calibration are reported in Supplementary Information, **Paragraphs S3-S4**. Performance statistics (average, standard deviation, and confidence interval) were calculated with Excel Microsoft. The selected features are reported in Supplementary information, **Paragraph S2**.

2.1.2. Features chi-squared ranking

The attributes were ranked using Orange's "Rank" widget [22,24]. Based on the chi-square internal scorer, the "Rank" widget scores variables according to their correlation with the numeric target variable.

2.2. External validation datasets

In accordance with the guidelines of the 6th revision of the 1975 Declaration of Helsinki, three external validation datasets were gathered. The data was provided did not require patient consent as determined through the Research Ethics Boards review process. All the subsets of subjects of the external validation datasets were always run in parallel for the NAIF, FIB4, and APRI scores. In this evaluation we mapped liver stiffness ≤9.7 kPa (first and second datasets), and liver biopsy stage F3-F4 (together with liver stiffness measured by elastography, third dataset). The patients enrolled for the external validations did not have their race/ethnicity data recorded.

To ensure reliable *meta*-validation, we utilized the Degree of Correspondence Ψ (psi), as suggested in [23]. This non-parametric and distribution-free technique is a metric for quantifying the degree of similarity between two datasets. We tested against the NHANES dataset represented by 5962 individuals, three validation datasets to see how similar were to the training dataset: i) the internal kept-out dataset (corresponding to the 5 % kept-out dataset randomly extracted from the total study population); ii) the first external dataset (n = 52 patients, see below), which was the only one without any missing values, and iii) the second external dataset (n = 55 subjects, see below), after populating the missing values (16,6% with the average values. To do this, we used an online tool to calculate the value of Ψ [25,26].

2.2.1. First external validation dataset (Fibroscan® evaluation of liver fibrosis)

n = 52 patients of the Endocrinology Department of the Policlinico Hospital of Palermo, Italy, with LF's medical records, were considered the first external validation dataset. The subjects included in the second dataset were 26 males and 25 females between the ages of 28 and 81. The results of transient elastography were used to identify LF. The average stiffness value in the population was 14,6 KPa, with a minimum value of 9,7 KPa and a maximum value of 37,5 KPa. Non-HCV-related liver disease, co-infection with hepatitis B (HBV), HIV, HCC, and decompensated liver cirrhosis were the exclusion criteria. **Supplementary Table S2** summarizes the statistics of the first external validation dataset.

2.2.2. Second external validation dataset (Fibroscan®evaluation of liver fibrosis)

n = 55 subjects followed at the Liver Unit of Azienda Ospedaliero-Universitaria Careggi (Florence, Italy) (26 controls and 29 with LF), 27 males and 28 females aged between 19 and 83 years were considered as the second external validation dataset. As for the first dataset, the results of transient elastography were used to identify LF. The average Fibroscan® value of the control population was 5 KPa, with a minimum value of 2,5 KPa and a maximum value of 8,1 KPa. The average value of the target population (with LF) was 18 KPa, with a minimum value of 10, 2 KPa and a maximum value of 49,7 KPa. Non-HCV-related liver disease, co-infection with hepatitis B (HBV), HIV, HCC, and decompensated liver cirrhosis were the exclusion criteria. **Supplementary Table S3** and **Supplementary Tables S4 and S5** summarize the statistics of the second external validation dataset.

2.2.3. Third external validation dataset (evaluation of liver fibrosis through liver biopsy)

n = 13 patients (5 males, 8 females), aged between 38 and 73 years (**Supplementary Table S6**), with a biopsy sample taken between January 2016 and December 2022). These patients underwent percutaneous liver biopsy, which revealed the LF Kleiner stage (F3 or F4).

Patients with decompensated cirrhosis were excluded. The fibrosis stage F4 biopsies were collected from patients who had undergone a biopsy procedure for diagnostic purposes and had no clinical signs of decompensation, such as ascites, hepatic encephalopathy, jaundice, or varices. Each liver biopsy that was included had a length of more than 1.0 cm and six or more complete portal tracts, indicating that it was of acceptable quality. We further evaluated the ability of APRI, FIB4 and NIAF models to assess significant LF (defined as F3-F4 Kleiner stages) using the biopsy-assessed fibrosis stage as the reference standard.

3. Results

3.1. Performance evaluation of NAIF

In the pipeline described in this study, the NHANES (2017–2020 pre pandemic public database) cohort was divided into a training&testing set and two validation sets. The first kept-out validation dataset is composed of randomly selected 150 individuals with Fibroscan® values E ≥ 9.7 KPa and 150 individuals with Fibroscan® values < 9.7 KPa. The second kept-out dataset consists of a randomly extracted 5 % sample of the whole population (i.e. 298 individuals), where the prevalence of the target condition reflects the general population percentages. The study flow is illustrated in Fig. 1. Compared to a recently described study [20], separating the validation sets before any data processing or manipulation, allows an unbiased evaluation of algorithm‘s performance, excluding any overfitting issue. Briefly, our previous model [22] was simplified using 21 features and learned through stratified sampling to distinguish individuals assigned with label 0 (without LF) from patients assigned with label 1 (with LF). The most predictive feature was GGT, followed by glycohemoglobin, according to the chi-square ranking (Table 1). The training&testing average values obtained for five runs are: AUC of 87 ± 0,2% accuracy of 78 ± 0,7% F1 of 78 ± 1 %, precision of 77 ± 0,9% and sensitivity of 80 ± 2 %.

Table 1
Chi-square scoring to identify target parameters for the algorithm, using 21 variables. We calculated the Chi-square between each feature and the target and selected rank features with top Chi-square scores. According to the Chi-square ranking, the most predictive feature is GGT, followed by glycohemoglobin and C-reactive protein.

Ranks	Parameter	Abbreviation	Chi-square (X^2)
1	Gammagliutamyltranspherase (GGT)	lbxsgtsi	109.101
2	glycohemoglobin	lbxgh	108.539
3	C-reactive protein	lbxhscrp	100.122
4	glucose	lbxsgl	85.518
5	uric acid	lbxsua	71.986
6	HDL cholesterol	lbdhdd	66.344
7	alaninaminotranspherase (ALT)	lbxsatsi	62.735
8	triglycerides	lbxstr	43.208
9	age	ridageyr	42.546
10	aspartataminotranspherase (AST)	lbxsassi	26.953
11	ferritin	lbxfer	23.776
12	blood albumin	lbxsal	23.634
13	alkaline phosphatase (ALP)	lbxsapsi	18.940
14	percentage of lymphocytes	lbxlypct	13.191
15	total count of white cells	lbxrbcsi	7.933
16	percentage of segmented neutrophils	lbxnepct	7.151
17	platelet count	lbxpltsi	4.997
18	total bilirubin	lbxstb	4.245
19	sex	riagendr	3.516
20	hematocrit	lbxhct	2.571
21	hemoglobin	lbxhgb	1.383

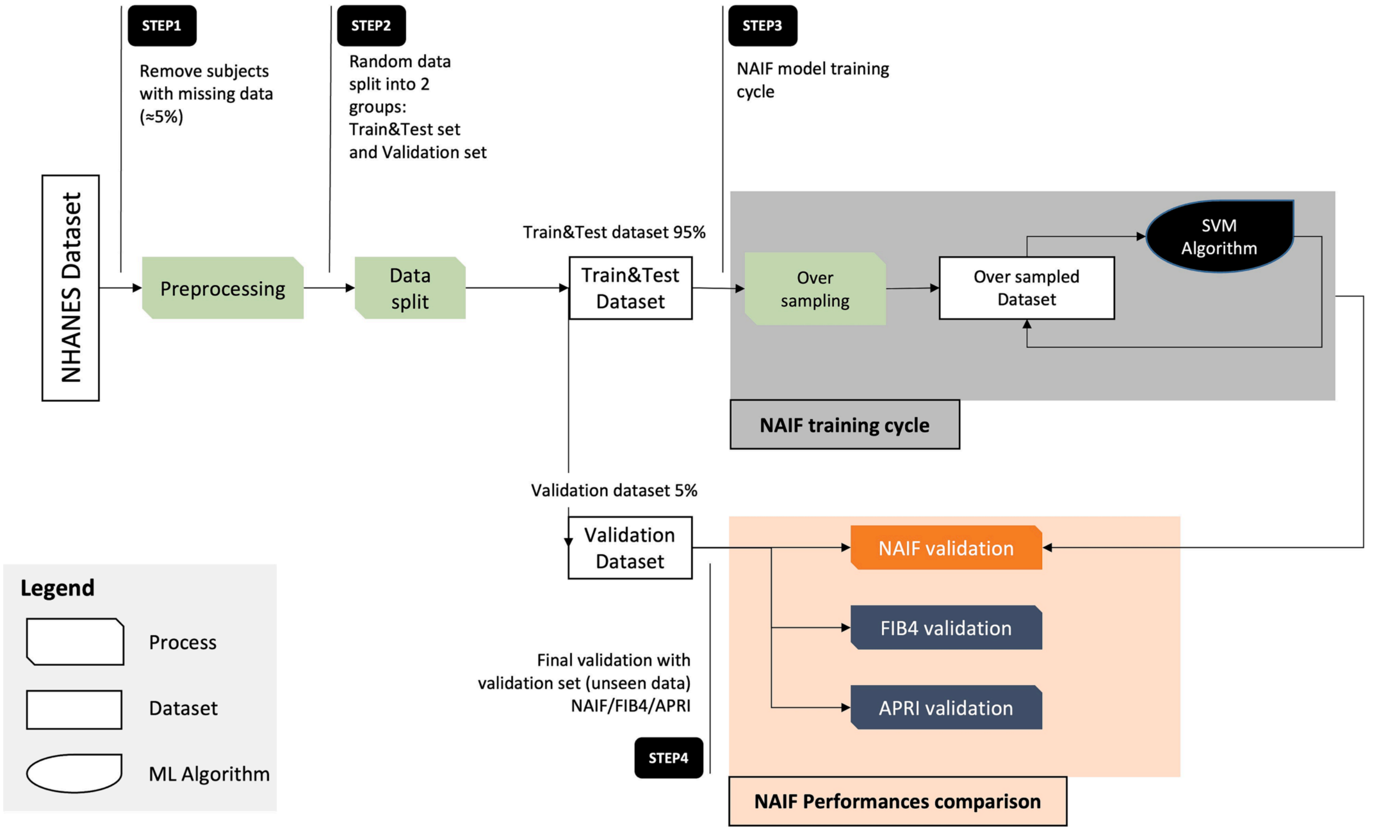


Fig. 1. Architecture of the pipeline. Firstly, subjects with missing data are removed from the NHANES Dataset in step 1. Then, the data is split into training and testing data (95 %) for NAIF algorithm learning, which undergoes oversampling and training cycles. The hold-out dataset (5 %) is used to validate NAIF with data that was not seen during the training and testing. Additionally, the validation dataset is used to test the Fib4 and APRI scores against NAIF and compare their performances.

Then, we determined the performances of NAIF in correctly classifying individuals with or without LF. Therefore, we challenged the algorithm with the two kept-out validation subsets and measured its performance. We performed five independent runs, each one with differently randomized evaluation subsets. With the first validation subset, NAIF had an accuracy of 69 ± 1,8%, a sensitivity of 61 ± 4 %, an AUC of 77 ± 1 %, a precision of 72 ± 1,8%, a specificity of 77 ± 2 %, and an F1 score of 66 ± 2 % (Table 2**)**. With the second kept-out dataset NAIF resulted to have an accuracy of 74 ± 3 %, a sensitivity of 80 ± 8 %, an AUC of 83 ± 5 %, a precision of 20 ± 2 %, a specificity of 74 ± 3 % and an F1 score of 32 ± 3 % (Table 3). Notably, the similarity between the training dataset and second kept-out validation dataset was Ψ = 0.77, indicating, as expected, a significant similarity in their sources. As part of our experimental design, in a separate branch of the pipeline, the same validation kept-out subsets were evaluated in parallel with Fib4 and APRI (Tables 2 and 3). Of note, individuals < 35 years were excluded from validation subsets for Fib4 and APRI evaluation. The confusion matrix was used to relate the actual target values with those predicted by models (Fig. 2). All correct predictions are in the purple diagonal of the confusion matrix, all the wrong prediction are in the pink diagonal. By summing up the two rows of the confusion matrix, it is possible to deduce the total number of positive (labelled = 1) and negative (labelled = 0) samples in the kept-out subsets, reported in bold in Fig. 2. In **Supplementary Figs. S1 and S2** more examples of confusion matrices are reported while in Table 4 are summarized the averages and standard deviations of negative and positive predicted values, over five independent runs. Notably, the Fib4 score works by classifying the subjects into three categories: i) healthy, where advanced fibrosis is excluded; ii) patients likely to have an advanced state of fibrosis; and iii) patients not classified, where further investigation is advised; of note, in the statistics of the Fib4 performance, in this study, we considered people with indetermined classification of Fib4 as with fibrotic liver, as for them further investigations are recommended.

A good diagnostic model should have high true positive and true negative rates, which means correct classifications and accuracy, while keeping low false positive and false negative rates, which refers to wrong classifications. However, an initial test applied to the general population for screening purposes should prioritize maximizing sensitivity to identify individuals affected by the target disease. The NAIF tool has been found to be more effective than Fib4 and APRI in identifying individuals with severe fibrosis. This is because it drastically reduces the false negative discovery rate, meaning it has a higher sensitivity in detecting individuals with liver fibrosis (sensitivity is defined as TP/(TP + FN)). We aimed to develop a tool that could screen the general population for liver fibrosis without any prior suspicion or clinical signs. Our primary focus was to minimize the false discovery rate and maintain high sensitivity. The first kept-out validation dataset showed that NAIF could effectively differentiate healthy individuals from those with fibrosis, identifying 65 % of liver fibrosis patients with a 35 % false negative rate (Fig. 2**A**). On the other hand, Fib4 identified only 30 % of liver fibrosis patients with a higher false negative rate of 70 % (Fig. 2**B**). APRI, however, was only able to identify 5 % of liver fibrotic subjects but with a very high false negative rate of 95 % (Fig. 2**C**).

Table 2

Prediction of significant liver stiffness defined as measured liver stiffness > 9.7 kPa. The table shows the evaluation using the kept-out (completely unseen) dataset, for 150 individuals with liver fibrosis and 150 controls randomly selected from the study population. Parallel comparison on the same validation subsets of the three scores, applied to an adult population >18 <79 years. For Fib4 and ARI the population <35 years was excluded. Sensitivity (true positive rate) refers to the probability of a positive test, conditioned on truly being positive. Specificity (true negative rate) refers to the probability of a negative test, conditioned on truly being negative. Precision is the ratio of system generated results that correctly predicted positive observations (True Positives) to the system's total predicted positive observations, both correct (True Positives) and incorrect (False Positives). The F1 Score is the weighted average (or harmonic mean) of Precision and sensitivity. Mean and standard deviations are reported for each value, referred to five different validation runs, along with the 95 % confidence interval (CI).

	Accuracy	Precision	Sensitivity	Specificity	F1 score
NAIF	69 ± 1,8 %	72 ± 1,8 %	61 ± 4,0 %	77 ± 2,0 %	66 ± 2,7 %
CI (95 %)	67.4–70.5 %	70.9–74 %	57.7–64.7 %	75–78.5 %	63.9–68.6 %
Fib4	53 ± 3,0 %	67 ± 4,9%	31 ± 3,0	81 ± 3,9%	42 ± 3,5 %
CI (95 %)	50.4–55.6 %	62.7–71.3 %	28.1–33.3 %	77.5–84.3 %	39–45.1 %
APRI	49 ± 2,0%	100 ± 0 %	9 ± 2,3	100 ± 0 %	16 ± 3,9 %
CI (95 %)	47.5–51 %	100 %	6.8–10.9 %	100 %	12.8–19.6 %

Table 3

The table shows the evaluation using the kept-out (completely unseen) dataset, for the 5 % of the subjects randomly selected from the general population. Parallel comparison on the same validation subsets of the three scores, applied to an adult population >18 <79 years. For Fib4 and ARI the population <35 years was excluded. Mean and standard deviations are reported for each value, referred to five different validation runs, along with the 95 % confidence interval (CI).

	Accuracy	Precision	Sensitivity	Specificity	F1 score
NAIF	74 ± 3 %	20 ± 2 %	80 ± 8 %	74 ± 3 %	32 ± 3 %
CI (95 %)	72–77 %	18–22 %	73–87 %	71–76 %	29–34 %
Fib4	76 ± 2 %	14 ± 1 %	32 ± 4 %	80 ± 2 %	20 ± 2 %
CI (95 %)	74–77 %	14–15 %	28–35 %	78–82 %	18–21 %
APRI	91 ± 1 %	65 ± 10 %	14 ± 3 %	99 ± 0 %	24 ± 4 %
CI (95 %)	90–92 %	56–73 %	12–17 %	99–99 %	20–27 %

With the second kept-out validation dataset, NAIF correctly classified 74 % of individuals with liver fibrosis with a lower false negative rate of 26 % (Fig. 2**D**). Fib4 identified only 33 % of subjects with liver fibrosis with a higher false negative rate of 67 % (Fig. 2**E**). Similarly, APRI identified only 19 % of the target disease classification, with an 81 % false negative rate (Fig. 2**F**).

It is interesting to note that Fib4 could only confidently assign 10 % of subjects with liver fibrosis to their correct classification, while 20 % were placed in the indeterminate classification in the first validation dataset. In comparison, in the second validation dataset, Fib4 could confidently recognize only 14 % of the subjects with liver fibrosis.

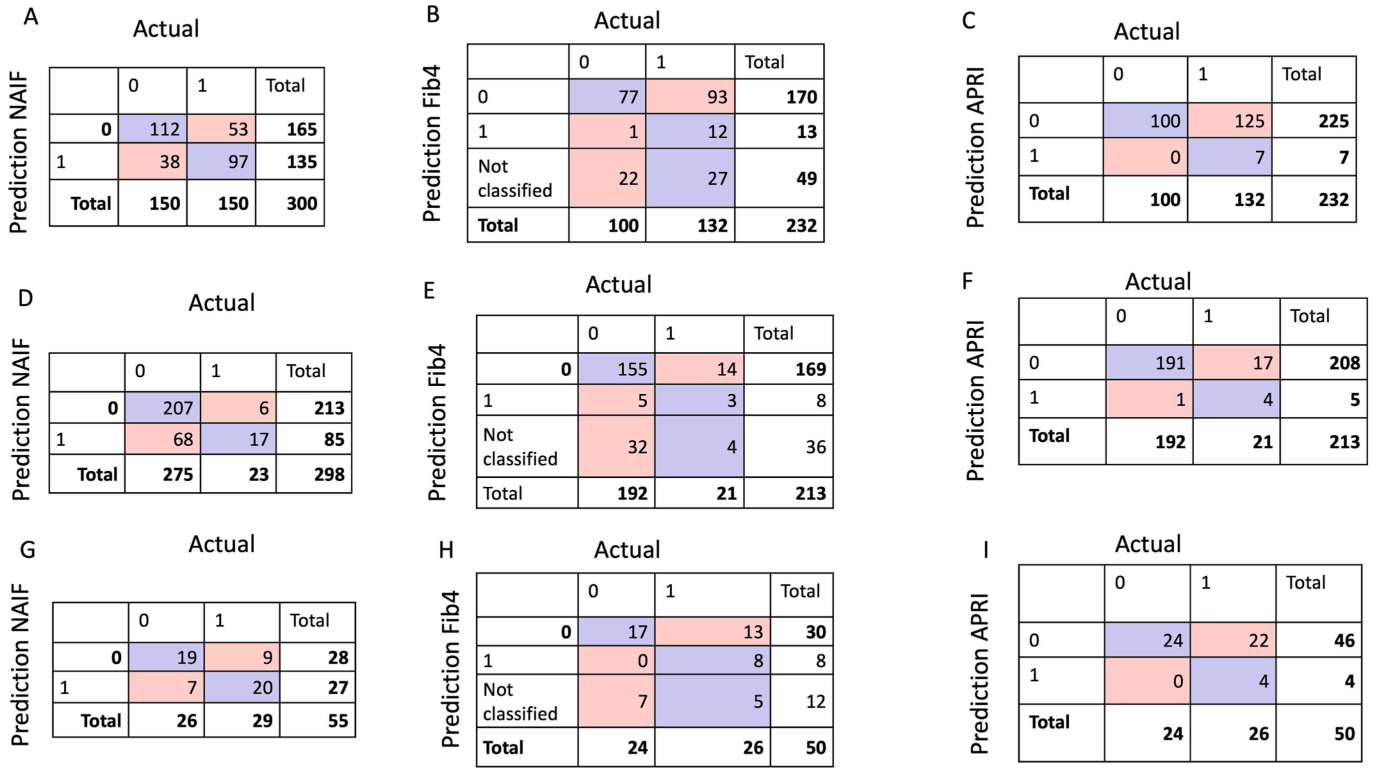
Extracting 5 % of the entire dataset for validation reflects a more realistic situation, but it reduces the number of true positives, thereby decreasing the precision of NAIF and Fib4 (precision is defined as TP/(TP + FP)). APRI, with an FP rate of 0, consistently outperforms, and its precision remains unaffected. Finally, the high F1 scores of NAIF with both the kept-out validation datasets indicate the solid overall performance of NAIF binary classification model. It signifies that NAIF can effectively identify positive cases while minimizing false positives and negatives.

Average values and standard deviation of accuracy, precision, sensitivity, specificity, and F1 score values for NAIF, Fib4 and APRI scores are reported in Tables 2 and 3, along with the 95 % Confidence interval (CI) for five independent runs.

To further validate these findings, we compared NAIF, Fib4, and APRI on three independent datasets.

3.2. *Validation of NAIF with three external datasets*

The external validation of NAIF was performed based on three datasets collected at two different hospitals in Italy, encompassing in a total of 94 LF-positive cases (including 13 diagnoses through liver biopsy) and 26 controls (without LF). External validation uses new participant-level data, independent from those used for model development, to examine whether the model's predictions are reliable (accurate enough) in individuals from the potential population(s) for clinical use.



A

Actual

Prediction NAIF

	0	1	Total
0	112	53	165
1	38	97	135
Total	150	150	300

B

Actual

Prediction Fib4

	0	1	Total
0	77	93	170
1	1	12	13
Not classified	22	27	49
Total	100	132	232

C

Actual

Prediction APRI

	0	1	Total
0	100	125	225
1	0	7	7
Total	100	132	232

D

Actual

Prediction NAIF

	0	1	Total
0	207	6	213
1	68	17	85
Total	275	23	298

E

Actual

Prediction Fib4

	0	1	Total
0	155	14	169
1	5	3	8
Not classified	32	4	36
Total	192	21	213

F

Actual

Prediction APRI

	0	1	Total
0	191	17	208
1	1	4	5
Total	192	21	213

G

Actual

Prediction NAIF

	0	1	Total
0	19	9	28
1	7	20	27
Total	26	29	55

H

Actual

Prediction Fib4

	0	1	Total
0	17	13	30
1	0	8	8
Not classified	7	5	12
Total	24	26	50

I

Actual

Prediction APRI

	0	1	Total
0	24	22	46
1	0	4	4
Total	24	26	50

Fig. 2. Confusion matrices of NAIF, Fib4 and APRI. All correct predictions are in the purple diagonal of the confusion matrices, all the wrong prediction are in the pink diagonal. In A, B and C are shown the results of the first kept-out dataset of the NHANES database; the pipeline automatically selects 150 individuals without severe liver fibrosis and 150 individuals with severe liver fibrosis. In D, E and F are shown the results of the second kept-out NHANES dataset, where the pipeline randomly selects 5 % of the total study population (i.e. 298 individuals). In G, H and I are shown the confusion matrices of the second external dataset, that comprises 29 individuals with severe liver fibrosis, defined as Fibroscan®values ≥ 10,2 KPa and 26 individuals without severe liver fibrosis, defined as Fibroscan®values ≤ 8,1 KPa. For Fib4 and APRI, individuals > 35 years are not included in the analysis. In purple the correct classifications, in pink the wrong classifications. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Negative and positive predictive values based on the confusion matrices of five independent runs. Average and standard deviation are reported. Color code of the rows is matching the color code of the confusion matrices: all correct predictions are in purple; all the wrong prediction are in pink. True positives (TP): patients with liver fibrosis correctly classified (labelled as 1 and classified as 1); True negatives (TN): individuals without liver fibrosis correctly classified (labelled as 0 and classified as 0); False positives (FP): individuals without liver fibrosis not correctly classified (labelled as 0 and classified as 1); False negatives (FN): individuals with liver fibrosis not correctly classified (labelled as 1 and classified as 0).

1st kept-out dataset of the NHANES database
(150 individuals with severe liver fibrosis and 150 controls)

	NAIF	Fib4	APRI
True positives (TP)	30,5 ± 2,0%	17,0 ± 1,6%	4,8 ± 1,2 %
True negatives (TN)	38,3 ± 1,0%	35,8 ± 2,5 %	44,3 ± 1,1 %
False positives (FP)	11,5 ± 1,0%	8,3 ± 1,5 %	0
False negatives (FN)	18,9 ± 2,1%	38,4 ± 1,7%	50,6 ± 2,0%

2nd kept-out dataset of the NHANES database
(randomly selected 5 % of the total study population)

	NAIF	Fib4	APRI
True positives (TP)	5,9 ± 0.9 %	2,9 ± 0,3 %	1,3 ± 0,3 %
True negatives (TN)	68,2 ± 3,7%	70,9 ± 5,2 %	89,6 ± 1,3 %
False positives (FP)	24,1 ± 2,9 %	17,2 ± 1,0 %	0,6 ± 0,2%
False negatives (FN)	1,4 ± 0,5%	6,4 ± 1,1 %	8,0 ± 0,9%

The first external dataset was derived from a hospital in Palermo, Italy, and comprised a population of 52 individuals with Fibroscan® values ≥ 9.7 KPa (**Supplementary Table S2**). NAIF correctly classified 32/52 patients with severe LF (61.5 %), Fib4 2/51 (3.9 % of correct classification), but assigning 6 of them to indeterminate classification, which if considered as true positives leads to 8/51 correct classification (16 %); APRI correctly identified 2/51 individuals with LF (3.9 %), missing 49/51 diagnosis (96,1%).

Notably, the similarity between the training dataset and first external dataset was Ψ = 0.04, indicating significant differences in their sources. Fig. 3**A** shows a scatter plot of the two most relevant features identified by the tool (GTT and C-reactive protein) for the NHANES dataset (in blue) and the external dataset (in orange), obtained through the SelectKBest method, and the Mutual_Info_Classif method. Each datapoint is a case/instance in each dataset. For more details on this score refer to [25]. Fig. 3**B** displays the density and cumulative distribution function diagrams of the first principal components of the combined datasets. Despite the differences, NAIF can accurately classify most patients, demonstrating the algorithm's reliability. Indeed, when a validation set has limited resemblance to the training&testing set but the predictive model performs well, it can be considered dependable and robust.

The second external dataset was derived from a hospital in Firenze, Italy, and included 29 individuals with Fibroscan® values ≥ 10,2 KPa and 26 individuals with Fibroscan® values < 8.1 KPa (**Supplementary Table S3** and **Supplementary Tables S4 and S5**). The similarity between the training dataset and second external dataset was Ψ = 0, indicating a striking difference in their sources. Fig. 2**G-I** reports the confusion matrix of the classifications, and Table 5 shows the performances of NAIF, Fib4 and APRI. NAIF correctly identified 20/29

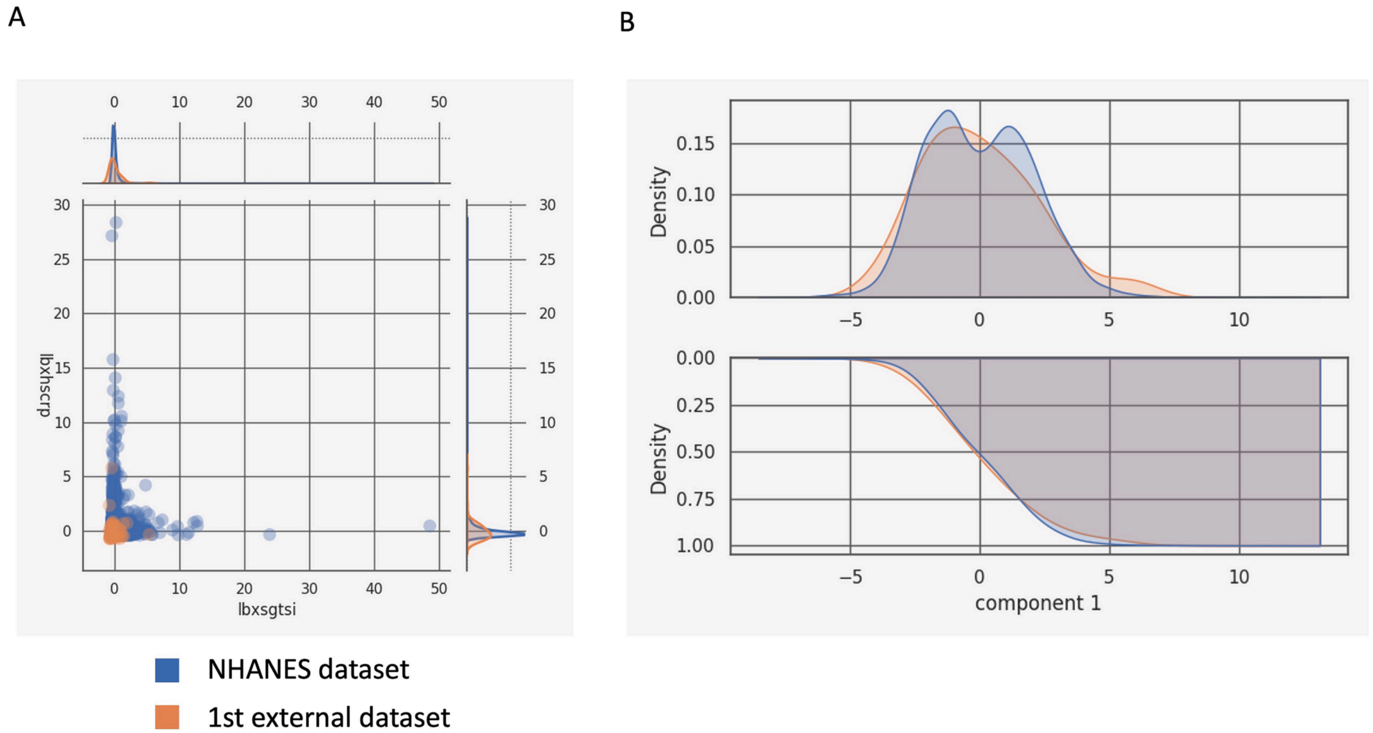


Fig. 3. Degree of correspondence. Panel A: scatter plots depict the primary features of both the NHANES dataset (in blue) and the external dataset (in orange) for the features selected by the tool C-reactive protein (lbxhscrp) on the Y axis, and gammagliutamyltranspherase (GGT) (lbxsgtsi) on the X axis. Panel B: diagrams displaying the density and cumulative distribution functions for the first principal components of the combined datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Results of the second validation dataset. The second external validation dataset comprises 29 individuals with severe liver fibrosis assessed by Fibroscan®values ≥ 10,2 KPa and 26 controls, with Fibroscan®values ≤ 8,2 KPa.

	Accuracy	Precision	Sensitivity	Specificity	F1 score
NAIF	69 %	72 %	68 %	71 %	70 %
Fib4	60 %	65 %	50 %	71 %	56 %
APRI	56 %	100 %	15 %	100 %	26 %

individuals with severe LF (69 %) and 19/26 individuals without LF (74 %). Fib4 could correctly classify 8/26 individuals with severe LF (30 %), assigning 5 individuals to indeterminate classification (which could be considered true positive, leading overall to 13/26 (50 %) of correct identification of LF condition; in addition, Fib4 correctly identified 17/24 individuals without LF (71 %), but indicated 13/26 (50 %) false positives. Finally, APRI correctly ranked 4/26 (15 %) of individuals with LF and 24/24 (100 %) without LF. Also in this case, NAIF outperformed in sensitivity (68 %), as compared to Fib4 (50 %) and APRI (15 %) (Table 5).

The diagram in Fig. 4 shows the results of a *meta*-validation procedure that was performed on the internal kept-out validation dataset, as well as the first and second external datasets [23]. As expected, the internal validation dataset (which corresponds to the 5 % kept-out dataset from the training population, and which is labeled in Fig. 4 as InternaValidation-EPD, In) had a high level of similarity with the training set, and is located in the upper part of the diagram. On the other hand, both the first and second external datasets (which are indicated in Fig. 4 as Pa_external-EPD (Pa) and Fl_external-EPD (Fl)) are located in the very bottom portion of the diagram, which represents the area of low similarity. When the performance of a validation dataset falls into this region, the validation process can be considered conservative [23].

The third external validation dataset included a population of n = 13 biopsy-assessed fibrosis-stage patients (F3/F4 fibrosis stages). The serum markers presented in Supplementary Table S6 were correlated with biopsies using the closest available measurement in a ± 180-day interval. NAIF was challenged with biopsy-assessed fibrosis patients aimed to validate the algorithm with the gold-standard diagnostic of LF. NAIF confidently classified 12/13 patients correctly (92 %), Fib4 only 2/13 (15 %), even if assigned 10 individuals to the indetermined classification (overall 10 + 2/13, therefore 92 %), while APRI classified 6/13 patients correctly (46 %). Our newly developed NAIF algorithm has been confirmed to have an excellent diagnostic performance for identifying fibrosis in patients with F3/F4 biopsy-confirmed LF.

4. Discussion

This study aimed to identify severe liver fibrosis in adults using common laboratory parameters. We evaluated the accuracy of an AI model called NAIF, which was trained on the NHANES 2017–2020 pre-pandemic public database. NAIF uses 19 biochemical blood parameters, age, and sex and was compared against conventional scoring systems (Fib4 and APRI) on the same subsets. Our study intentionally excluded biometric and clinical parameters (such as body mass index, abdominal circumference, or blood pressure values) to create an easy-to-use tool for screening the general population.

Although the NHANES and the two external datasets showed significant differences as determined by the calculated degree of correspondence ($\Psi = 0.04$ and $\Psi = 0$, respectively) the NAIF algorithm could accurately classify most patients, proving its reliability. Indeed, when a validation set is different from the training&testing set, a predictive model with good results can still be considered dependable and robust.

In conclusion, compared with Fib4 and APRI scores, we demonstrated the higher sensitivity of NAIF in diagnosing severe fibrosis over Fib4 and APRI scores with using two internal validation subsets and confirmed the results with three external validation datasets.

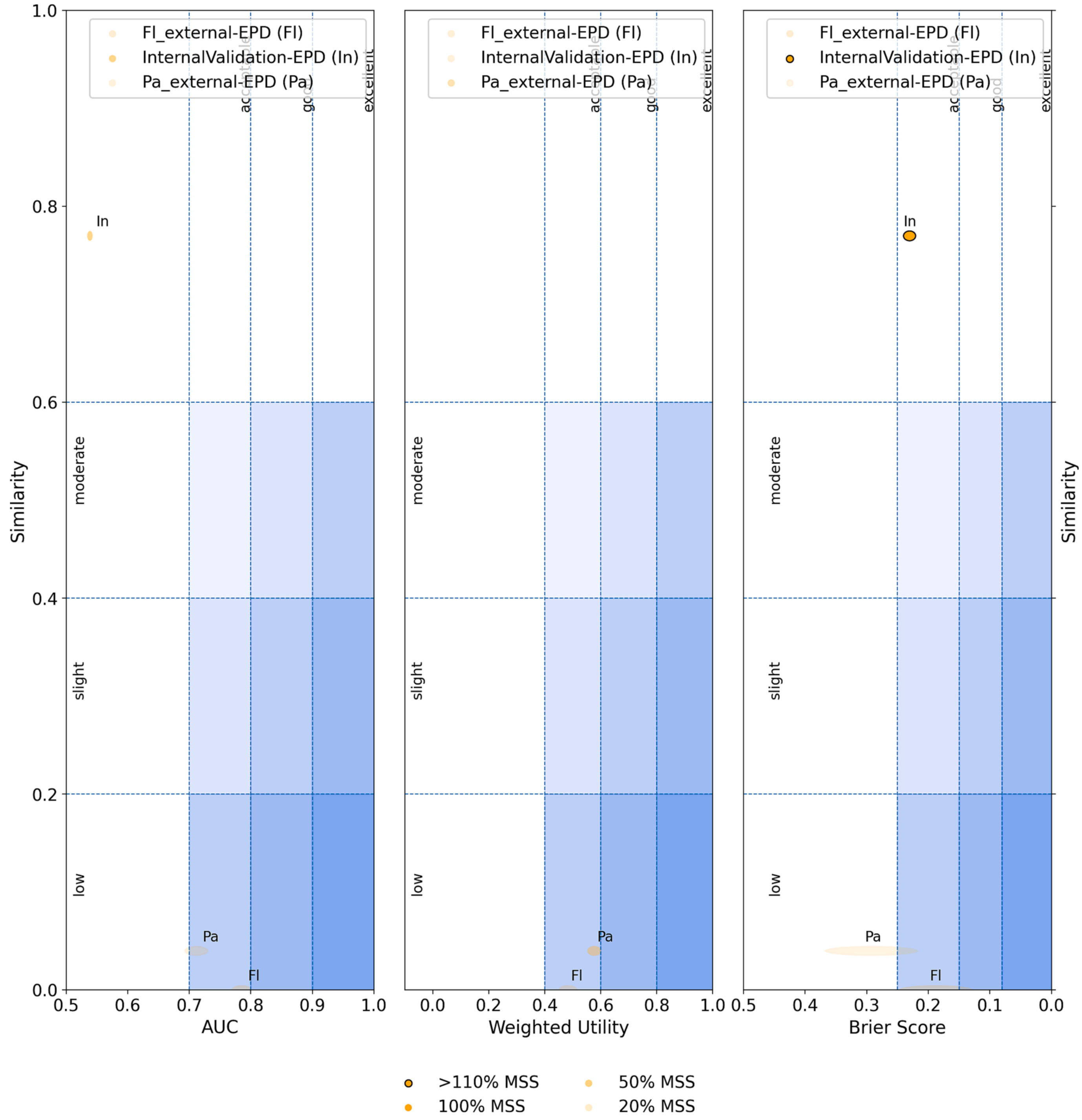


Fig. 4. The figure shows the diagram of the *meta*-validation process for diagnosing severe liver fibrosis. It involves three datasets: the internal kept-out validation dataset which is a randomly selected 5% from the NHANES study population (InternalValidation-EPD (In)), and two external validation datasets; the first from Policlinico Hospital of Palermo, Italy (Pa_external-EPD (Pa)) and the second from Azienda Ospedaliero-Universitaria Careggi (Florence, Italy) (Fl_external-EPD (Fl)). The external performance diagram displays the results of the three external validation studies. The color code indicates the minimum sample size (MSS) - the darker the blue, the higher the robustness. The width of the ellipses represents the width of the 95% confidence interval around the performance metrics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Ethical approval

The study conforms to the ethical guidelines of the 1975 Declaration of Helsinki (6th revision, 2008) as reflected in a priori approval by the institution's human research committee. The ethics committees of Palermo Hospital and Firenze hospital have previously approved the use of these data for non-invasive assessment of chronic liver disease (J Hepatol 2021 May;74(5):1109–1116).

CRediT authorship contribution statement

Samir Hassoun: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Chiara Bruckmann:** Writing – original draft, Visualization, Conceptualization. **Stefano Ciardullo:** Methodology, Data curation. **Gianluca Perseghin:** Validation, Investigation. **Fabio Marra:** Formal analysis, Data curation. **Armando Curto:** Validation, Formal analysis, Data curation. **Umberto Arena:** Methodology, Data curation. **Francesco Broccolo:** Writing – review & editing, Writing – original draft, Supervision, Methodology. **Francesca Di Gaudio:** Writing – review & editing, Writing – original draft, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2024.105373.

References

[1] A. Caligiuri, A. Gentilini, F. Marra, Molecular pathogenesis of NASH, Int. J. Mol. Sci. 17 (2016), https://doi.org/10.3390/ijms17091575.

[2] D. Berger, V. Desai, S. Janardhan, Con: liver biopsy remains the gold standard to evaluate fibrosis in patients with nonalcoholic fatty liver disease, Clin. Liver Dis. (Hoboken) 13 (2019) 114–116, https://doi.org/10.1002/cld.740.

[3] J. Lambrecht, S. Verhulst, I. Mannaerts, H. Reynaert, L.A. van Grunsven, Prospects in non-invasive assessment of liver fibrosis: Liquid biopsy as the future gold standard? Biochimica Et Biophysica Acta (BBA) – Mol. Basis Dis. 1864 (2018) 1024–1036, https://doi.org/10.1016/j.bbadis.2018.01.009.

[4] T. Pasha, S. Gabriel, T. Therneau, E.R. Dickson, K.D. Lindor, Cost-effectiveness of ultrasound-guided liver biopsy, Hepatology 27 (1998) 1220–1226, https://doi.org/10.1002/hep.510270506.

[5] I. Graupera, M. Thiele, A.T. Ma, M. Serra-Burriel, J. Pich, N. Fabrellas, L. Caballeria, R.J. de Knegt, I. Grgurevic, M. Reichert, D. Roulot, J.M. Schattenberg, J. M. Pericas, P. Angeli, E.A. Tsochatzis, I.N. Guha, M. Garcia-Retortillo, R.M. Morillas, R. Hernández, J. Hoyo, M. Fuentes, A. Madir, A. Juanola, A. Soria, M. Juan, M. Carol, A. Diaz, S. Detlefsen, P. Toran, C. Fournier, A. Llorca, P.N. Newsome, M. Manns, H.J. de Koning, F. Serra-Burriel, F. Cucchietti, A. Arslanow, M. Korenjak, L. van Kleef, J.L. Falcó, P.S. Kamath, T.H. Karlsen, L. Castera, F. Lammert, A. Krag, P. Ginès, M. Alvarez, P. Andersen, P. Angeli, A. Ardèvol, A. Arslanow, L. Beggiato, Z.B. Abdesselam, L. Bennett, B. Boutouria, A. Brocca, M.T. Broquetas, L. Caballeria, V. Calvino, J. Camacho, A. Capdevila, M. Carol, L. Castera, M. Cervera, F. Cucchietti, A. de Fuentes, R. de Knegt, S. Detlefsen, A. Diaz, J.D. Bande, V. Esnault, N. Fabrellas, J. lluis Falco, R. Fernández, C. Fournier, M. Fuentes, P. Galle, E. García, M. García-Retortillo, E. Garrido, P. Ginès, R.G. Medina, J. Gratacós-Gines, I. Graupera, I. Grgurevic, I.N. Guha, E. Guix, R. Harris, E.H. Boluda, R. Hernández-Ibañez, J. Hoyo, A. Ikram, S. Incicco, M. Israelsen, M. Juan, A. Juanola, R. Kaiser, P.S. Kamath, T.H. Karlsen, M. Kjærgaard, H.J. de Koning, M. Korenjak, A. Krag, J.K. Hansen, M. Krawczyk, I. Lambert, F. Lammert, P. Laboulaye, S.L. Sørensen, C. Laserna-Jiménez, S.L. Pi, E. Ledain, V. Levy, V. Londoño, G. Loyer, A. Llorca, A.T. Ma, A. Madir, M. Manns, D. Marshall, M.L. Martí, S. Martínez, R.M. Sala, R.M. Font, J.M. Jensen, R.M. Morillas, L. Muñoz, R. Nadal, L. Napoleone, J.M. Navarrete, P.N. Newsome, V. Nielsen, M. Pérez, J.M.P. Pulido, S. Piano, J. Pich, J.P. Escobet, E. Pose, K.P. Lindvig, M. Reichert, C. Riba, D. Roulot, A.B. Rubio, M. Sánchez-Morata, J. Schattenberg, F. Serra-Burriel, M. Serra-Burriel, L.S. Just, M. Sonneveld, A. Soria, C. Stern, P. Such, M. Thiele, P. Toran, A. Torrejón, M. Tonon, E.A. Tsochatzis, L. van Kleef, P. van Wijngaarden, V. Velázquez, A. Viu, S.N. Weber, T. Wildsmith, for the LiverScreen Consortium investigators, LiverScreen project: study protocol for screening for liver fibrosis in the general population in European countries, BMC Public Health 22 (2022) 1385. https://doi.org/10.1186/s12889-022-13724-6.

[6] L. Sandrin, B. Fourquet, J.-M. Hasquenoph, S. Yon, C. Fournier, F. Mal, C. Christidis, M. Ziol, B. Poulet, F. Kazemi, M. Beaugrand, R. Palau, Transient elastography: a new noninvasive method for assessment of hepatic fibrosis, Ultrasound Med. Biol. 29 (2003) 1705–1713, https://doi.org/10.1016/j.ultrasmedbio.2003.07.001.

[7] J. Joseph, Serum marker panels for predicting liver fibrosis - An update, Clin. Biochem. Rev. 41 (2020) 67–73, https://doi.org/10.33176/AACB-20-00002.

[8] R.K. Sterling, E. Lissen, N. Clumeck, R. Sola, M.C. Correa, J. Montaner, M. S. Sulkowski, F.J. Torriani, D.T. Dieterich, D.L. Thomas, D. Messinger, M. Nelson, APRICOT Clinical Investigators, development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection, Hepatology 43 (2006) 1317–1325, https://doi.org/10.1002/hep.21178.

[9] A. Vallet-Pichard, V. Mallet, B. Nalpas, V. Verkarre, A. Nalpas, V. Dhalluin-Venier, H. Fontaine, S. Pol, FIB-4: an inexpensive and accurate marker of fibrosis in HCV infection. Comparison with liver biopsy and fibrotest, Hepatology 46 (2007) 32–36, https://doi.org/10.1002/hep.21669.

[10] Z.-H. Lin, Y.-N. Xin, Q.-J. Dong, Q. Wang, X.-J. Jiang, S.-H. Zhan, Y. Sun, S.-Y. Xuan, Performance of the aspartate aminotransferase-to-platelet ratio index for the staging of hepatitis C-related fibrosis: an updated meta-analysis, Hepatology 53 (2011) 726–736, https://doi.org/10.1002/hep.24105.

[11] R. Catanzaro, A. Aleo, M. Sciuto, L. Zanoli, B. Balakrishnan, F. Marotta, FIB-4 and APRI scores for predicting severe liver fibrosis in chronic hepatitis HCV patients: a monocentric retrospective study, Clin. Exp. Hepatol. 7 (2021) 111–116, https://doi.org/10.5114/ceh.2021.104543.

[12] A.D. Schreiner, J. Zhang, W.P. Moran, D.G. Koch, J. Marsden, S. Livingston, P. D. Mauldin, M. Gebregziabher, FIB-4 and incident severe liver outcomes in patients with undiagnosed chronic liver disease: a fine-gray competing risk analysis, Liver Int. 43 (2023) 170–179, https://doi.org/10.1111/liv.15295.

[13] S.S. Sarvestany, J.C. Kwong, A. Azhie, V. Dong, O. Cerocchi, A.F. Ali, R.S. Karnam, H. Kuriry, M. Shengir, E. Candido, R. Duchen, G. Sebastiani, K. Patel, A. Goldenberg, M. Bhat, Development and validation of an ensemble machine learning framework for detection of all-cause advanced hepatic fibrosis: a retrospective cohort study, Lancet Digital Health 4 (2022) e188–e199, https://doi.org/10.1016/S2589-7500(21)00270-3.

[14] J. Lee, M. Westphal, Y. Vali, J. Boursier, R. Ostroff, L. Alexander, Y. Chen, C. Fournier, A. Geier, S. Francque, K. Wonders, D. Tiniakos, P. Bedossa, M. Allison, G. Papatheodoridis, H. Cortez-Pinto, R. Pais, J.-F. Dufour, D.J. Leeming, S. Harrison, J. Cobbold, A.G. Holleboom, H. Yki-Järvinen, J. Crespo, M. Ekstedt, G.P. Aithal, E. Bugianesi, M. Romero-Gomez, M. Karsdal, C. Yunis, J.M. Schattenberg, D. Schuppan, V. Ratziu, C. Brass, K. Duffin, K. Zwinderman, M. Pavlides, Q.M. Anstee, P.M. Bossuyt, on behalf of the LITMUS investigators, Machine learning algorithm improves detection of NASH (NAS-based) and at-risk NASH, a development and validation study, Hepatology (9900). https://journals.lww.com/hep/Fulltext/9900/Machine_learning_algorithm_improves_detection_of.356.aspx.

[15] V. Blanes-Vidal, K.P. Lindvig, M. Thiele, E.S. Nadimi, A. Krag, Artificial intelligence outperforms standard blood-based scores in identifying liver fibrosis patients in primary care, Sci. Rep. 12 (2022) 2914, https://doi.org/10.1038/s41598-022-06998-8.

[16] J.M. Schattenberg, M.-M. Balp, B. Reinhart, A. Tietz, S.A. Regnier, G. Capkun, Q. Ye, J. Loeffler, M.C. Pedrosa, M. Docherty, NASHmap: clinical utility of a machine learning model to identify patients at risk of NASH in real-world settings, Sci. Rep. 13 (2023) 5573, https://doi.org/10.1038/s41598-023-32551-2.

[17] P. Decharatanachart, R. Chaiteerakij, T. Tiyarattanachai, S. Treeprasertsuk, Application of artificial intelligence in chronic liver diseases: a systematic review and meta-analysis, BMC Gastroenterol. 21 (2021) 10, https://doi.org/10.1186/s12876-020-01585-5.

[18] R. Anteby, E. Klang, N. Horesh, I. Nachmany, O. Shimon, Y. Barash, U. Kopylov, S. Soffer, Deep learning for noninvasive liver fibrosis classification: a systematic review, Liver Int. 41 (2021) 2269–2278, https://doi.org/10.1111/liv.14966.

[19] R.A. Khan, Y. Luo, F.-X. Wu, Machine learning based liver disease diagnosis: a systematic review, Neurocomputing 468 (2022) 492–509, https://doi.org/10.1016/j.neucom.2021.08.138.

[20] Y. Wu, X. Yang, H.L. Morris, M.J. Gurka, E.A. Shenkman, K. Cusi, F. Bril, W. T. Donahoo, Noninvasive diagnosis of nonalcoholic steatohepatitis and advanced liver fibrosis using machine learning methods: comparative study with existing quantitative risk scores, JMIR Med. Inform. 10 (2022) e36997.

[21] J. Boursier, V. de Ledinghen, J.-P. Zarski, I. Fouchard-Hubert, Y. Gallois, F. Oberti, P. Calès, Comparison of eight diagnostic algorithms for liver fibrosis in hepatitis C: new algorithms are more precise and entirely noninvasive, Hepatology 55 (2012) 58–67, https://doi.org/10.1002/hep.24654.

[22] S. Hassoun, C. Bruckmann, S. Ciardullo, G. Perseghin, F. Di Gaudio, F. Broccolo, Setting up of a machine learning algorithm for the identification of severe liver fibrosis profile in the general US population cohort, Int. J. Med. Inf. 170 (2023) 104932, https://doi.org/10.1016/j.ijmedinf.2022.104932.

[23] F. Cabitza, A. Campagner, F. Soares, L. García de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, A. Carobene, The importance of being external. methodological insights for the external validation of machine learning models in medicine, Comput. Methods Programs Biomed. 208 (2021) 106288, https://doi.org/10.1016/j.cmpb.2021.106288.

[24] J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik, B. Zupan, Orange: data mining toolbox in Python, J. Mach. Learn. Res. 14 (2013) 2349–2353.

[25] F. Cabitza, A. Campagner, L.M. Sconfienza, As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI, BMC Medical Informatics Decision Making 20 (2020) 219, https://doi.org/10.1186/s12911-020-01224-9.

[26] https://psicorrespondence.pythonanywhere.com, (n.d.).