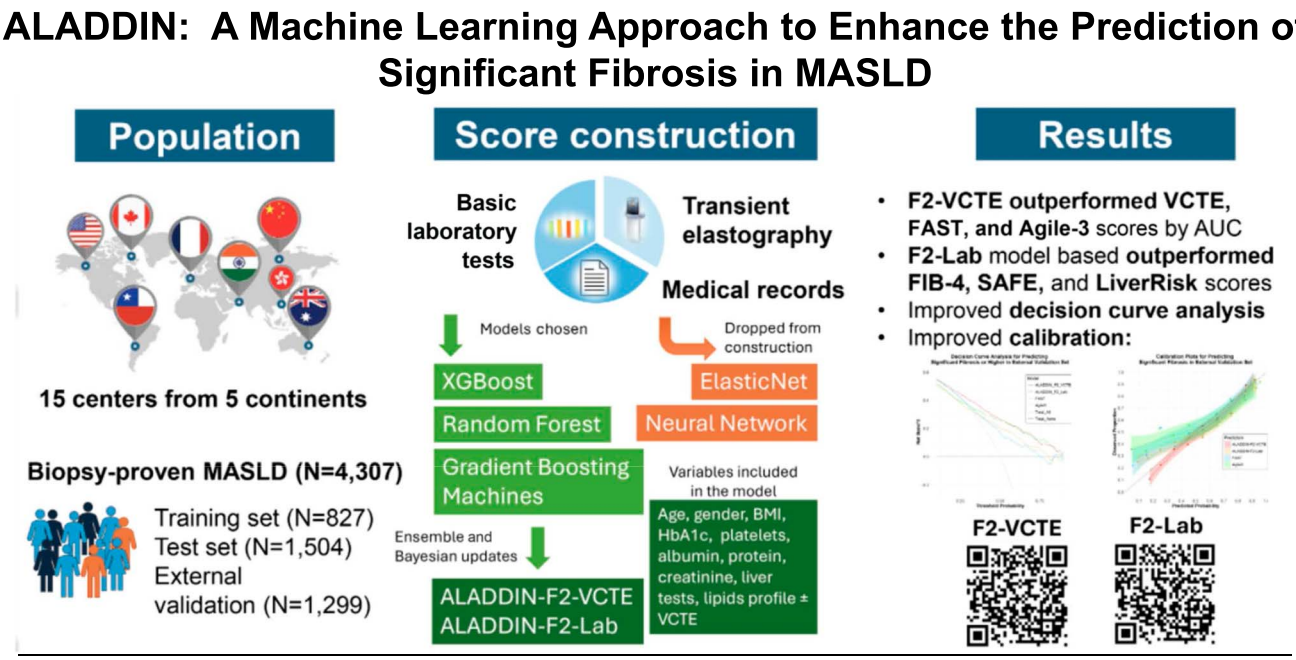
Open

ALADDIN: A Machine Learning Approach to Enhance the Prediction of Significant Fibrosis or Higher in Metabolic Dysfunction-Associated Steatotic Liver Disease

Naim Alkhouri, MD[1], Terry Cheuk-Fung Yip, MD[2], Laurent Castera, MD[3], Marina Takawy, MD[4], Leon A. Adams, MD[5], Nipun Verma, MD[6], Juan Pablo Arab, MD[7,8], Syed-Mohammed Jafri, MD[9], Bihui Zhong, MD[10], Julie Dubourg, MD[11], Vincent L. Chen, MD[12], Ashwani K. Singal, MD, MS, FACG[13], Luis Antonio Díaz, MD[8,14], Nicholas Dunn[15], Rida Nadeem, MD[1], Vincent Wai-Sun Wong, MD[2], Manal F. Abdelmalek, MD, MPH, FACG[4], Zhengyi Wang, MD[5], Ajay Duseja, MD[6], Yousef Almahanna, MD[6], Haya A. Omeish, MD[9], Junzhao Ye, MD[10], Stephen A. Harrison, MD[16], Jessica Cristiu, MD[12], Marco Arrese, MD[8], Sage Robert[17], Grace Lai-Hung Wong, MD[2], Amani Bajunayd, MD[6], Congxiang Shao, MD[10], Matthew Kubina, MD[12] and Winston Dunn, MD[17]

INTRODUCTION: **The recent US Food and Drug Administration approval of resmetirom for treating metabolic dysfunction-associated steatohepatitis in patients necessitates patient selection for significant fibrosis or higher (≥F2). No existing vibration-controlled transient elastography (VCTE) algorithm targets ≥F2.**



Alkhouri et al. *Am J Gastroenterol.* 2025. doi:10.14309/ajg.0000000000003432

AJG The American Journal of GASTROENTEROLOGY

[1]Department of Hepatology, Arizona Liver Health, Chandler, Arizona, USA; [2]Department of Medicine & Therapeutics, The Chinese University of Hong Kong, Hong Kong, China; [3]Université Paris-Cité, Department of Hepatology, Hospital Beaujon, AP-HP, Inserm UMR 1149, Clichy, France; [4]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota, USA; [5]Medical School, University of Western Australia, Perth, Washington, Australia; [6]Department of Hepatology, Postgraduate Institute of Medical Education and Research, Chandigarh, India; [7]Division of Gastroenterology, Hepatology, and Nutrition, Department of Internal Medicine, Virginia Commonwealth University School of Medicine, Richmond, Virginia, USA; [8]Departamento de Gastroenterología, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile; [9]Division of Gastroenterology and Hepatology, Henry Ford Hospital, Detroit, Michigan, USA; [10]Department of Gastroenterology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong, China; [11]Summit Clinical Research, San Antonio, Texas, USA; [12]Division of Gastroenterology and Hepatology, University of Michigan, Ann Arbor, Michigan, USA; [13]Department of Gastroenterology and Hepatology, University of Louisville, Louisville, Kentucky, USA; [14]MASLD Research Center, Division of Gastroenterology and Hepatology, University of California San Diego, San Diego, California, USA; [15]Pembroke Hill School, Kansas City, Missouri, USA; [16]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [17]Department of Gastroenterology, University of Kansas Medical Center, Kansas, USA. **Correspondence:** Winston Dunn, MD. E-mail: wdunn2@kumc.edu.
Received October 25, 2024; accepted February 27, 2025; published online March 27, 2025

METHODS: **The mAchine Learning ADvanceD fibrosis and rIsk metabolic dysfunction-associated steatohepatitis Novel predictor (ALADDIN) study addressed this gap by introducing a machine-learning-based web calculator that estimates the likelihood of significant fibrosis using routine laboratory parameters with and without VCTE. Our study included a training set of 827 patients, a testing set of 504 patients with biopsy-confirmed metabolic dysfunction-associated steatotic liver disease from 6 centers, and an external validation set of 1,299 patients from 9 centers. Five algorithms were compared using area under the curve (AUC) in the test set: ElasticNet, random forest, gradient boosting machines, XGBoost, and neural networks. The top 3 (random forest, gradient boosting machines, and XGBoost) formed an ensemble model.**

RESULTS: **In the external validation set, the ALADDIN-F2-VCTE model, using routine laboratory parameters with VCTE (AUC 0.791, 95% confidence interval [CI]: 0.764–0.819), outperformed VCTE alone (0.745, 95% CI 0.717–0.772, $P < 0.0001$), FibroScan-aspartate aminotransferase (0.710, 0.679–0.748, $P < 0.0001$), and Agile-3 model (0.740, 0.710–0.770, $P < 0.0001$) regarding the AUC, decision curve analysis, and calibration. The ALADDIN-F2-Lab model, using routine laboratory parameters without VCTE, achieved an AUC of 0.706 (95% CI: 0.668–0.749) and outperformed Fibrosis-4, steatosis-associated fibrosis estimator, and LiverRisk scores.**

DISCUSSION: **Along with the steatosis-associated fibrosis estimator model developed to target significant fibrosis or higher, ALADDIN-F2-VCTE (https://aihepatology.shinyapps.io/ALADDIN1) uniquely supports a refined noninvasive approach to patient selection for resmetirom without the need for liver biopsy. In addition, ALADDIN-F2-Lab (https://aihepatology.shinyapps.io/ALADDIN2) offers an effective alternative when VCTE is unavailable.**

KEYWORDS: artificial intelligence; VCTE; FibroScan; fibrosis; model; steatohepatitis; machine learning; outcome prediction

SUPPLEMENTARY MATERIAL accompanies this paper at http://links.lww.com/AJG/D621

Am J Gastroenterol 2026;121:362–374. https://doi.org/10.14309/ajg.0000000000003432

INTRODUCTION

Metabolic dysfunction-associated steatotic liver disease (MASLD) (1), which affects an estimated 30% of the global adult population (2), is now the predominant cause of chronic liver disease in Western countries, leading to an increase in liver transplantation in the United States (3). MASLD encompasses a spectrum that ranges from isolated steatosis to metabolic dysfunction-associated steatohepatitis (MASH), potentially progressing to fibrosis and cirrhosis (4). Patients with significant fibrosis (stage F2) or higher have a significantly increased risk of liver-related morbidity and mortality (5). The recent US Food and Drug Administration approval of resmetirom for the treatment of MASLD in patients with significant to advanced fibrosis specifically targets this subset of patients. Treatment generally requires precise patient selection and liver biopsy (6).

Liver biopsy is the gold standard for diagnosing fibrosis stage in MASLD (6); its invasiveness and variability in interpretation highlight the need for less invasive diagnostic methods (7). If the indication for resmetirom is solely based on liver biopsy, access to treatment for patients at the highest risk of disease progression would be severely restricted. Major gastroenterological societies, such as the American Gastroenterological Association (AGA) (8), American Association for the Study of Liver Diseases (AASLD) (9), and the European Association for the Study of Liver Diseases (10) advocate a sequential screening approach, initially using the Fibrosis-4 index (FIB-4), followed by vibration-controlled transient elastography (VCTE) for risk assessment of advanced (F3-4) fibrosis. FibroScan-aspartate aminotransferase(FAST) is a VCTE-based algorithm commonly used to diagnose at-risk MASH (at least significant fibrosis and Nonalcoholic fatty liver disease Activity Score ≥4) (11). This group is of particular interest for trials (12) because they are more likely to benefit from emerging treatments (13) aimed at inflammation and fibrosis. On the other hand, Agile-3 (14) uses VCTE and common laboratory parameters for the diagnosis of advanced fibrosis. Unfortunately, no existing VCTE-based algorithm effectively targets significant fibrosis or higher (≥F2), which would be optimal for targeting patients for current resmetirom treatment, as well as future therapeutics with similar indications. This drives the need for a VCTE-based algorithm, specifically for significant fibrosis or higher (≥F2). In addition, there is a need for a more accessible algorithm that uses routine laboratory parameters without VCTE in various clinical environments.

This study introduces the mAchine Learning ADvanceD fibrosis and rIsk MASH Novel predictor (ALADDIN), a cross-sectional study designed to bridge these diagnostic gaps. ALADDIN leverages a novel machine-learning-based web calculator to deliver comprehensive probability assessments for significant fibrosis, advanced fibrosis, and at-risk MASH. The findings on advanced fibrosis and at-risk MASH will be reported separately, whereas this study focused on significant fibrosis. The aims of this study were to (i) diagnose significant fibrosis or higher (≥F2) with 90% specificity adequate for resmetirom treatment consideration and (ii) identify patients at indeterminate risk of significant fibrosis or higher (≥F2) who can undergo further testing or follow-up. Notably, this model has various forms that accommodate scenarios with and without the VCTE data inputs. This feature significantly enhances the

applicability of the model across a wide range of healthcare environments, from local community clinics to advanced tertiary referral centers.

METHODS

Study design and participants

The ALADDIN study aggregated data from 15 global centers, with initial participants from 6 centers across various continents randomized 1:1 into training and test sets. This was supplemented by data from 9 additional centers for the external validation set. See Supplementary Table 1, http://links.lww.com/AJG/D621 for the center listings and characteristics. Inclusion criteria included patients with steatotic liver disease (1) and ≥1 cardiometabolic risk factor (body mass index [BMI] ≥25 kg/m^2, type 2 diabetes or impaired glucose tolerance, hypertension, hypertriglyceridemia, and low high-density lipoprotein cholesterol) with a liver biopsy within 6 months. Key exclusions were significant alcohol consumption and other etiologies of chronic liver diseases such as chronic viral hepatitis and hepatocellular carcinoma. In addition, we excluded patients with missing age, aspartate aminotransferase (AST), platelet, and gamma-glutamyl transpeptidase (GGTP) data from the train set to ensure model robustness. Data were transmitted to a central database managed by the main researchers of the study. With Institutional Review Board approval from each center and the retrospective nature of the study, the requirement for patient consent was waived. This study adhered to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis—Artificial Intelligence reporting guidelines and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Training set, test set, and external validation set

Our study used the training, test, and external validation sets to develop a robust methodology. Patients from the 6 global centers were split 1:1 into training and test sets. To ensure model accuracy and reliability, patients with missing key variables (age, AST, platelet, and GGTP) were excluded from the training set rather than imputing these variables. This approach helped create a cleaner and more reliable data set for training the predictive model. However, because both the training and test sets originated from the same centers, there was a risk of overfitting because of similar patient characteristics. To mitigate this and assess the model's wider generalizability on top of transportability, we included an external validation set from 8 additional centers in North America and Asia, offering diversity and independence from the derivation data. This smaller, yet diverse cohort is essential to evaluate the model's generalizability to new, unseen data, underpinning its real-world relevance and utility.

Data collection and definitions

Each participating center performed a medical chart review to obtain data on the patients who met the inclusion and exclusion criteria. Liver biopsies were graded and staged by a local pathologist(s) in accordance with the Nonalcoholic Steatohepatitis Clinical Research Network criteria (15). We included clinical data obtained within 6 months of liver biopsy. Significant fibrosis was defined as a stage ≥F2 fibrosis. The features considered in the modeling included patient demographics (age and sex) and common laboratory data (complete blood counts, comprehensive metabolic panel, lipid panel, and GGTP (16)). Protein (17) and albumin levels are part of a comprehensive metabolic panel. We used separate models that used (ALADDIN-F2-VCTE) and did not use VCTE (ALADDIN-F2-Lab) parameters, including VCTE liver stiffness measurement.

Missing data and imputation

We excluded patients with missing data on age, AST, platelet count, and GGTP from the training set because these were the most important predictors (other than VCTE) for the dependent variable (i.e., significant fibrosis or higher). Patients with missing VCTE were included in the ALADDIN-F2-Lab analysis only. Missing data were otherwise handled using the missForest algorithm in R for imputation based on random forest. This algorithm does not use data from the dependent variable or any biopsy-related data (e.g., Nonalcoholic fatty liver disease Activity Score) and relies solely on the independent variables included in the model. The rate of missing data is represented in Supplementary Digital Content (see Supplementary Table 2, http://links.lww.com/AJG/D621).

Machine learning, hyperparameters optimization, model selection, and ensemble model

From the derivation cohort, we developed models to predict significant fibrosis or higher, both with and without VCTE data, using 5 machine-learning algorithms: ElasticNet (EN), random forest (RF), gradient boosting machines (GBM), XGBoost (XGB), and neural networks (NN). The RF optimized the hyperparameters using a grid search based on the out-of-bag area under the curve (AUC), whereas EN, GBM, and XGB relied on the cross-validated AUC for hyperparameter tuning. By contrast, the NN was optimized based on the cross-validated accuracy of the training set. The top 3 algorithms regarding AUC in the test set were used to construct an ensemble model, which was calculated as the geometric mean of the predictions of these models.

Specifically, EN adjusted hyperparameters based on the balance between L1 and L2 regularization (alpha) and regularization strength (lambda). The RF tuned the number of variables to be considered at each split (mtry), sample size for each tree (sampsize), minimum size of terminal nodes (nodesize), number of trees (ntree), and maximum number of terminal nodes (maxnodes). GBM optimized the interaction depth, learning rate (shrinkage), minimum number of observations in a node (n.minobsinnode), and fraction of data used per tree (bag.fraction). XGB adjusted tree depth (max_depth), learning rate (eta), minimum child weight (min_child_weight), number of estimators (n_estimators), subsample rate (subsample), column sample rate per tree (colsample_bytree), and minimum split loss (gamma). The NN was fine-tuned by optimizing the learning rate (learn_rate), number of neurons per layer (neurons), dropout rate (dropoutrate), batch size (batchsize), number of epochs (epochs), and the optimizer type (optimizer).

Target imbalance and Bayesian updates

To address the target imbalance within our models, distinct approaches were used for RF, GBM, and XGB to optimize the performance. For the RF model, we balanced the data set by specifying the same sample size for both classes in the target variable, ensuring an equitable representation during model training. The aim of this approach was to mitigate bias toward the more prevalent class. Conversely, for the GBM and XGB, we adopted a weighting strategy to address this imbalance. The final

models, ALADDIN-F2-VCTE and ALADDIN-F2-Lab, incorporated Bayesian updates to adjust for the center-specific rate of significant fibrosis or higher (≥F2). We computed the baseline risk based on the prevalence in the training set.

For example, in a referral cohort with a 30% prevalence, a patient with a predicted probability of 45%. Calculation as follows.

$$\text{Predicted Odds} : 0.45/(1-0.45) = 0.8182$$

$$\text{Prevalence Odds} : 0.30/(1-0.30) = 0.4286$$

$$\text{Likelihood Ratio} : 0.8182/1.3529 = 0.605$$

$$\text{Posttest Odds} : 0.4286 \times 0.605 = 0.2592$$

$$\text{Updated Probability} : 0.2592/(1+0.2592) = 20.6\%$$

Statistical analysis

Data analysis was performed using R v4.4.0 (R Core Team 2024). Continuous variables were expressed as mean (SD), whereas categorical variables were presented as numbers (percentages). The discriminatory performance of the models was assessed using the area under the receiver operating characteristic curve with 95% confidence intervals (CIs).

Decision curve analysis

Decision curve analysis (DCA) was used to assess the clinical utility of predictive models for significant (≥F2) fibrosis or higher by calculating and comparing their net benefits across a range of decision thresholds. This method enabled us to evaluate the models against 2 baseline strategies: treating all patients and treating them based on their risk levels. We computed the net benefits of the models using a predefined function that factors the true and false positives for each threshold probability. The analysis was visualized using ggplot2, which demonstrated the net benefit of each model relative to baseline strategies. This process identified the most clinically useful models for predicting conditions in both the test and external validation sets, thereby guiding optimal decision making in clinical practice.

Calibration

Calibration analysis was conducted to evaluate the accuracy of the ALADDIN models for significant (≥F2) fibrosis or higher. Using ggplot2 in R, calibration plots were created by dividing the patient data into deciles based on the posterior probabilities from the ALADDIN models. These plots compare the observed condition rate with the median-predicted probability in each decile, identifying any over-predictions or under-predictions. The Brier Score, which measures the mean squared deviation between predictions and actual outcomes, further validated the accuracy of the models. Lower Brier Scores indicated higher accuracy, confirming the effectiveness and reliability of the ALADDIN models for practical use.

Dual cutoff approach

A sensitivity and specificity of 95% in the training set were targeted, with the goal of achieving 90% sensitivity and specificity in the test and external validation sets. Within the test set and external validation set, the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), percentage correctly classified, number of screening positive, negative, and indeterminate were determined.

RESULTS

Our study included 3,008 patients with biopsy-confirmed MASLD from 6 centers across 5 continents. These patients were divided into training and test sets in a 1:1 ratio. After excluding patients with missing key variables, 827 patients were retained in the training set, whereas 1,504 patients were included in the test set. In addition, an external validation set comprised 1,299 patients from 9 centers. Table 1 presents the patient characteristics of the training, test, and external validation sets. Supplementary Digital Content (see Supplementary Table 1, http://links.lww.com/AJG/D621) presents the characteristics of the participating centers across the 3 cohorts. Supplementary Digital Content (see Supplementary Table 3, http://links.lww.com/AJG/D621) presents the univariate analysis of significant fibrosis or higher in the training set. Log-transformed AST, alanine aminotransferase, platelet count (16), and GGTP had substantially lower *P* values than their counterparts, and therefore, log transformation was used.

Prediction of significant fibrosis or higher (≥F2) with VCTE: ALADDIN-F2-VCTE

This analysis included 752 patients in the training set, 1,281 patients in the test set, and 1,019 patients in the external validation set with VCTE data. In the test set, the RF model achieved an AUC of 0.789 (95% CI, 0.765–0.814), the GBM was 0.790 (95% CI, 0.766–0.814), and the XGB was 0.782 (95% CI, 0.758–0.807). These models outperformed VCTE alone with an AUC of 0.745 (95% CI, 0.717–0.772), EN with an AUC of 0.781 (95% CI 0.756–0.806), and NN with an AUC of 0.7286 (95% CI 0.701–0.756) and were thus included in the final ensemble model building. Table 2 presents the individual performances of the RF, GBM, and XGB models as well as the ensemble model, which integrates these algorithms with Bayesian updates across the training, test, and external validation sets.

The ensemble model using VCTE data, ALADDIN-F2-VCTE, achieved an AUC of 0.792 (95% CI, 0.768–0.817) in the test set and 0.791 (95% CI 0.764–0.819) in the external validation set, as shown in Figure 1a,b. Notably, this model significantly outperformed the VCTE alone (AUC 0.745, 95% CI 0.717–0.772, Delong test, $P < 0.0001$), FAST model (AUC 0.693, 95% CI 0.664–0.722, $P < 0.0001$), and the Agile-3 model (AUC 0.761, 95% CI 0.735–0.787, $P = 0.0016$) in the test set. A similar superiority was observed in the external validation set, with AUCs of 0.761 (95% CI 0.731–0.791, $P = 0.010$), 0.710 (95% CI 0.679–0.748, $P < 0.0001$), and 0.740 (95% CI 0.710–0.770, $P < 0.0001$), respectively. Figure 2 illustrates the ranking order of the variables used to derive the RF, GBM, and XGB models, highlighting that log VCTE is the most important, followed by FIB-4, log GGTP, and log AST.

Prediction of significant fibrosis or higher (≥F2) without VCTE: ALADDIN-F2-Lab

This analysis included all 827 patients in the training set, 1504 in the test set, and 1,299 in the external validation set. In the test set, the RF model achieved an AUC of 0.766 (95% CI, 0.743–0.790), GBM was 0.767 (95% CI, 0.744–0.791), and XGB was 0.764 (95% CI, 0.740–0.788). These models outperformed EN (AUC 0.747, 95% CI 0.722–0.772) and NN (AUC 0.723, 95% CI 0.700–0.746)

Table 1. Characteristics of the participants with metabolic dysfunction-associated steatotic liver disease included in the train set, test set, and external validation set

	Training set N = 827	Test set N = 1,504	External validation set N = 1,299	*P* value test set vs external validation set
Demographics				
Age (yr)	56 (46 to 62)	56 (48 to 63)	54 (41 to 65)	<0.0001
Sex–male (n, %)	366 (44.3%)	653 (43.4%)	642 (49.4%)	0.0017
Race (n, %)				0.0003
White	425 (51.4%)	693 (46.1%)	694 (46.1%)	
Hispanic	103 (12.6%)	236 (15.7%)	94 (6.4%)	
Black	4 (0.5%)	17 (1.1%)	30 (2.0%)	
Asian	224 (29.5%)	372 (24.7%)	364 (24.2%)	
Native American	4 (0.5%)	3 (0.2%)	1 (0.1%)	
Australian Native	2 (0.2%)	8 (0.5%)	0 (0%)	
Other	65 (7.9%)	175 (11.6%)	116 (8.9%)	
History and physical exam				
Type 2 diabetes (n, %)	501 (60.6%)	855 (56.8%)	565 (43.5%)	<0.0001
Hypertension (n, %)	449 (54.3%)	732 (48.7%)	644 (49.6%)	0.66
BMI (kg/m^2)	32 (28 to 37)	33 (28 to 38)	32 (26 to 37)	0.19
Complete blood count				
WBC (K/μL)	7.2 (5.9 to 8.8)	7.2 (6.0 to 8.9)	6.8 (5.6 to 8.2)	<0.0001
Platelet (k/μL)	236 (189 to 285)	273 (196 to 278)	226 (178 to 275)	<0.0001
Comprehensive metabolic panel				
AST (IU/L)	37 (26 to 54)	38 (27 to 54)	62 (36 to 99)	<0.0001
ALT (IU/L)	51 (33 to 81)	53 (33 to 76)	48 (32 to 71)	<0.0001
GGT (IU/L)	53 (33 to 89)	65 (40 to 97)	76 (47 to 121)	
Alkaline phosphatase (IU/L)	79 (65 to 100)	83 (66 to 99)	91 (72 to 119)	<0.0001
Creatinine (mg/dL)	0.8 (0.7 to 0.9)	0.8 (0.7 to 1.0)	0.9 (0.7 to 1.0)	0.0067
HbA1c (%)	6.3 (5.6 to 7.4)	6.3 (5.6 to 7.0)	6.0 (5.5 to 6.8)	
Albumin (mg/dL)	4.3 (4.1 to 4.6)	4.4 (4.2 to 4.5)	4.3 (3.9 to 4.5)	<0.0001
Protein (mg/dL)	7.3 (7.1 to 7.6)	7.3 (7.1 to 7.6)		<0.0001
Total bilirubin (mg/dL)	0.6 (0.5 to 0.8)	0.6 (0.4 to 0.8)	0.6 (0.4 to 0.8)	0.041
Globulin (mg/dL)	3.0 (2.7 to 3.3)	2.9 (2.7 to 3.3)	2.8 (2.4 to 3.2)	<0.0001
Lipid panel				
Total cholesterol (mg/dL)	182 (151 to 207)	189 (161 to 204)	189 (156 to 208)	0.90
LDL (mg/dL)	102 (76 to 124)	111 (83 to 124)	110 (83 to 127)	0.034
HDL (mg/dL)	45 (38 to 53)	45 (39 to 50)	44 (38 to 51)	0.15
Triglyceride (mg/dL)	151 (112 to 210)	159 (124 to 197)	155 (113 to 196)	0.36
VCTE (transient elastography)				
Liver stiffness on VCTE (kPa)	10.2 (7.0 to 14.1)	10.5 (7.6 to 14.3)	11.6 (7.8 to 17.1)	0.0002
CAP (dB/m)	321 (296 to 349)	323 (298 to 353)	319 (286 to 347)	0.0010
Composite scores				
FAST	0.51 (0.28 to 0.70)	0.55 (0.31 to 0.72)	0.61 (0.41 to 0.78)	<0.0001
Agile-3				0.25
FIB-4	1.21 (0.83 to 1.79)	1.27 (0.88 to 1.80)	1.47 (0.93 to 2.34)	0.0006

Table 1. (*continued*)

	Training set N = 827	Test set N = 1,504	External validation set N = 1,299	*P* value test set vs external validation set
SAFE	96 (18 to 166)	98 (23 to 163)	97 (−16 to 195)	<0.0001
LiverRisk score	8.7 (7.6 to 9.9)	8.5 (7.6 to 9.8)	8.9 (7.9 to 10.3)	0.88
Outcomes				
Significant fibrosis or higher	465 (56.2%)	851 (56.6%)	734 (56.5%)	0.99

Continuous variables are described using median and interquartile ranges, whereas categorical variables are presented through actual counts and percentages.
ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; CAP, controlled attenuation parameter; FAST, FibroScan-aspartate aminotransferase; FIB-4, Fibrosis-4; GGTP, Gamma-glutamyl transpeptidase; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SAFE, steatosis-associated fibrosis estimator; VCTE, vibration-controlled transient elastography; WBC, white blood cell.

and were thus included in the final ensemble model building. Table 1 presents the individual performances of the RF, GBM, and XGB models as well as the ensemble model, which integrates these algorithms with Bayesian updates across the training, test, and external validation sets.

The ensemble model using only common laboratory parameters, ALADDIN-F2-Lab, achieved an AUC of 0.779 (95% CI 0.776–0.802) in the test set and 0.717 (95% CI 0.689–0.744) in the external validation set. The performance of this model was significantly better than that of the FIB-4 Score (AUC 0.727, 95% CI 0.702–0.753, Delong test $P < 0.0001$), steatosis-associated fibrosis estimator (SAFE) score (AUC 0.749, 95% CI 0.724–0.773, Delong test $P = 0.0022$), and LiverRisk score (AUC 0.682, 95% CI 0.655–0.709, Delong test $P < 0.0001$) in the test set. The model also outperformed these scores in the external validation set, with AUCs of 0.655 (95% CI 0.625–0.684, Delong test, $P < 0.0001$) for FIB-4, 0.680 (95% CI 0.651–0.709, Delong test, $P = 0.0005$) for SAFE, and 0.632 (95% CI 0.601–0.662, Delong test, $P < 0.0001$) for the LiverRisk score, as depicted in Figure 1a,b. Figure 2 shows the ranking order of the variables used to derive the RF, GBM, and XGB models. Without VCTE elasticity, the most important variables in the RF, GBM, and XGB models are FIB-4, log GGTP, followed by in various orders of BMI, total cholesterol, and log AST.

Table 2. **Area under the curve of several machine learning models in derivation, internal validation, and external validation sets**

	Significant fibrosis or higher including VCTE	Significant fibrosis or higher without VCTE available
Training set		
RF	0.819 (0.790–0.849)	0.768 (0.737–0.800)
GBM	0.824 (0.795–0.853)	0.764 (0.733–0.796)
XGB	0.790 (0.795–0.853)	0.761 (0.729–0.793)
Ensemble	0.824 (0.801–0.859)	0.781 (0.753–0.809)
Test set		
RF	0.789 (0.765–0.789)	0.766 (0.743–0.790)
GBM	0.790 (0.766–0.814)	0.767 (0.744–0.791)
XGB	0.782 (0.758–0.807)	0.764 (0.740–0.788)
Ensemble	0.792 (0.768–0.817)	0.779 (0.756–0.802)
External validation set		
RF	0.773 (0.745–0.802)	0.688 (0.659–0.716)
GBM	0.779 (0.751–0.807)	0.678 (0.649–0.707)
XGB	0.777 (0.748–0.805)	0.683 (0.654–0.712)
Ensemble	0.791 (0.764–0.819)	0.717 (0.690–0.744)

We included the random forest (RF), gradient boosting machines (GBM), and XGBoost (XGB) models before their integration into the ALADDIN models. The 95% confidence intervals are calculated using the DeLong method, providing a statistical measure of the precision of the AUC values for each model.
VCTE, vibration-controlled transient elastography.

A subset of patients with available VCTE data (n = 1,281 in the test set and n = 1,019 in the external validation set) was used to compare the performance of the ALADDIN-F2-Lab model against the FAST and Agile-3 scores. In the test set, the ALADDIN-F2-Lab model achieved an AUC of 0.764 (95% CI: 0.738–0.790), which was significantly better than the FAST score (AUC: 0.693, 95% CI: 0.664–0.722, $P < 0.0001$) but comparable with the Agile-3 score (AUC: 0.761, 95% CI: 0.735–0.787, $P = 0.82$). In the external validation set, the ALADDIN-F2-Lab model achieved an AUC of 0.720 (95% CI: 0.689–0.751), which was comparable with both the FAST score (AUC: 0.710, 95% CI: 0.679–0.742, $P = 0.58$) and the Agile-3 score (AUC: 0.740, 95% CI: 0.710–0.770, $P = 0.18$).

Decision curve

DCA revealed that ALADDIN-F2-VCTE consistently shows a higher net benefit compared with FAST in the test set and external validation set across a wide range of threshold probabilities, especially between 0.1 and 0.7 (Figure 3). This suggests that the ALADDIN-F2-VCTE model is better at balancing the benefits of true-positives while minimizing the harm of false-positives than the FAST models in this range. Although ALADDIN-F2-VCTE shows superiority over Agile-3, this advantage is modest. The ALADDIN-F2-Lab performed similar to ALADDIN-F2-VCTE in the test set and similarly to FAST and Agile-3 in the external validation set.

Calibration

The calibration of observed vs expected probabilities demonstrated improvement with the ALADDIN-F2-VCTE model, as reflected by the Brier Score. In the test set, ALADDIN-F2-VCTE achieved a lower Brier Score (0.184) than to the FAST score (0.232) and Agile-3 (0.209), indicating better calibration (lower

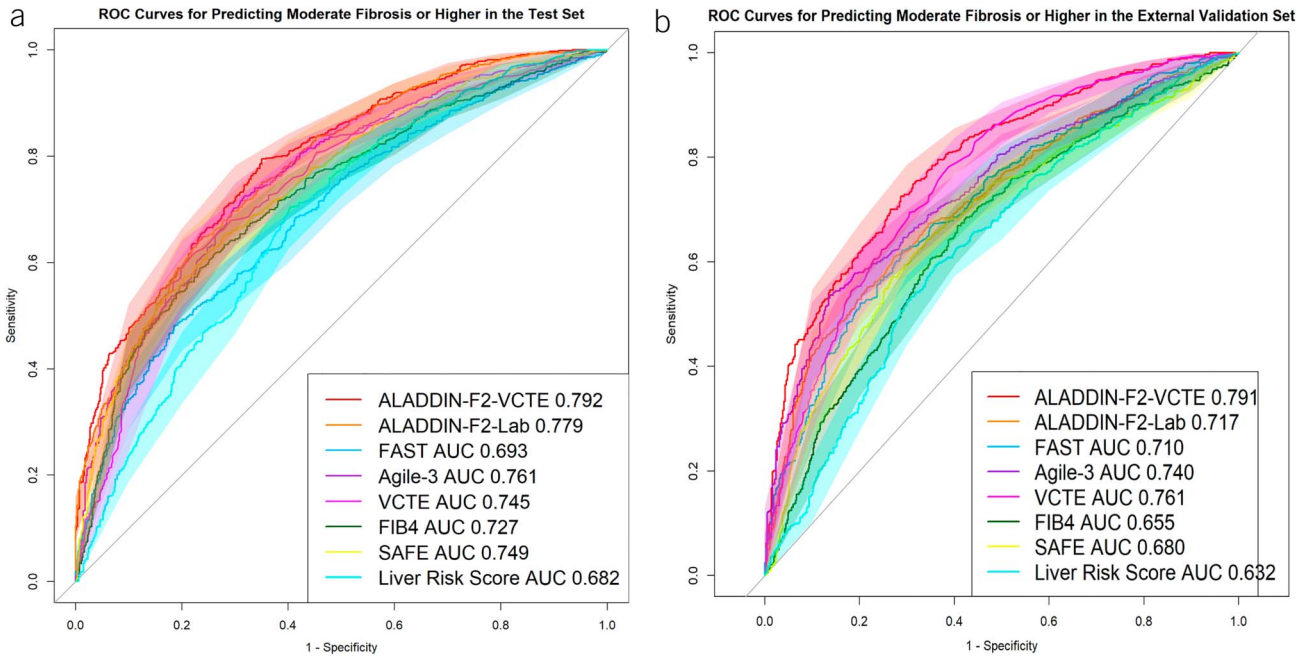


Figure 1. ROC Curves. In test (**a**) and external validation (**b**) sets, the ALADDIN-F2-VCTE model shows superior predictive accuracy for significant fibrosis or higher, significantly outperforming FAST. ALADDIN-F2-VCTE and FAST were calculated from test set of 1,203 and external validation set of 876 with VCTE data. The ALADDIN-F2-Lab, FIB4, SAFE, and LiverRisk score were calculated from the entire test set of 1,307 and external validation set of 1,135. ALADDIN, mAchine Learning ADvanceD fibrosis and at-risk mash Novel predictor; FIB-4, Fibrosis-4; ROC, receiver operating characteristic curve; SAFE, steatosis-associated fibrosis estimator; VCTE, vibration-controlled transient elastography.

values indicate better performance). Similarly, in the external validation set, ALADDIN-F2-VCTE outperformed the other models with a Brier Score of 0.183, compared with 0.217 for the FAST score and 0.229 for Agile-3. The ALADDIN-F2-Lab model achieved a comparable Brier Score with ALADDIN-F2-VCTE in the test set (0.189) and performed better than the FAST score in the external validation set (0.215). The calibration curve of the observed vs expected probabilities is shown in Figure 4a,b. The ALADDIN-F2-VCTE is very closely aligned with the ideal calibration, as shown by the dotted diagonal line, indicating exact matches between predictions and outcomes.

Dual cutoff approach

We used a dual cutoff strategy to enhance the diagnostic performance in predicting significant fibrosis or higher. For the ALADDIN-F2-VCTE model, we used a rule-out cutoff designed to achieve 95% sensitivity and a rule-in cutoff aimed at 95% specificity in the training set. By contrast, the ALADDIN-F2-Lab model aimed for 90% sensitivity and specificity. Supplementary Digital Content (see Supplementary Figures 1A and 1B, http://links.lww.com/AJG/D621) illustrate these cutoffs in the training set, whereas Table 3 outlines the corresponding rule-in and rule-out thresholds for the ALADDIN models, alongside the VCTE, FIB-4, FAST, Agile-3, SAFE, and LiverRisk scores, based on commonly used thresholds.

The ALADDIN-F2-VCTE model, with cutoffs of 0.36 and 0.77, demonstrated notable improvements in sensitivity, specificity, NPV, and PPV compared with the FAST model. However, this led to approximately 20% more patients in the testing cohort and 2% more patients in the external validation cohort falling into the indeterminate zone (as presented in Table 3 and Figure 5a,b). Importantly, the ALADDIN-F2-VCTE model achieved ≥90% sensitivity and specificity in both the testing and external validation cohorts. Similarly, the ALADDIN-F2-Lab model, with its cutoffs of 0.37 and 0.72, also achieved ≥80% sensitivity and specificity in both cohorts. In the external validation set, the performance metrics (sensitivity, specificity, NPV, and PPV) of the ALADDIN-F2-Lab model were comparable with those of the FAST and Agile-3 models.

DISCUSSION

In this study, we developed and introduced ALADDIN, a groundbreaking machine-learning-based web calculator to deliver comprehensive and nuanced probability assessments for significant fibrosis or higher (≥F2). The training and test sets were sourced from 6 global centers, with subsequent external validation conducted at 8 additional centers, encompassing 2,677 patients. This model, adaptable to include or exclude VCTE data, relies solely on commonly available laboratory parameters and uses Bayesian updates to cater to a wide range of healthcare settings from community clinics to specialized tertiary referral centers. In external validation set, the ALADDIN model with VCTE outperformed the FAST score, the closest existing VCTE model, regarding AUC, DCA, and calibration. The ALADDIN model with common laboratory parameters without VCTE was noninferior to the FAST score and superior to the FIB-4, SAFE, and LiverRisk scores. Using a dual cutoff approach, ALADDIN provided high diagnostic accuracy; the rule-in cutoff achieved over 90% specificity in referral and tertiary referral settings, whereas the rule-out cutoff demonstrated over 90% sensitivity. When this algorithm is applied to referral and tertiary referral settings, over half of patients fall into the indeterminate zone. However, when applied to a population-

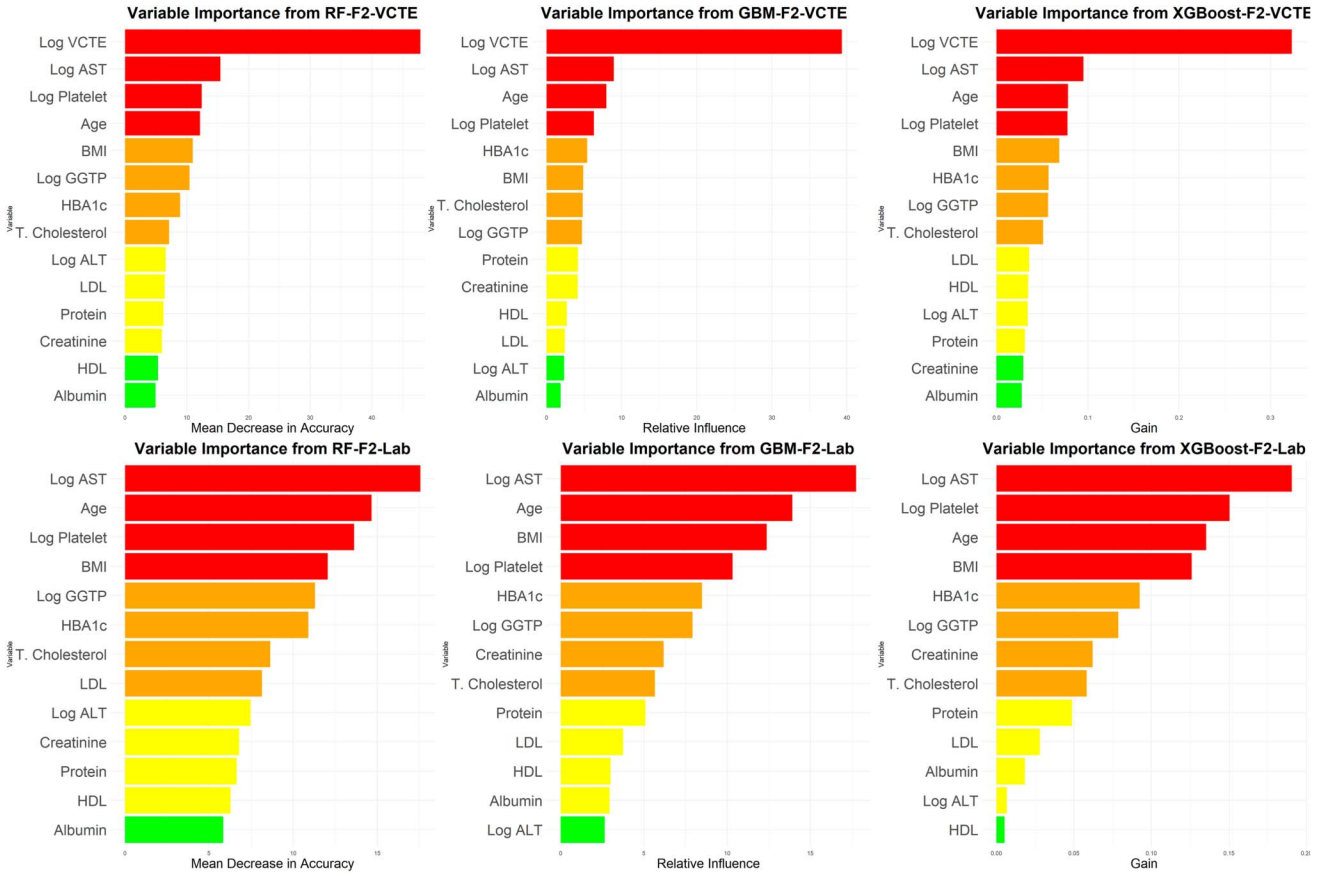


Figure 2. Ranking importance of variables used in models. Figure 2 displays the variable importance rankings across machine-learning algorithms with VCTE (first row) and without VCTE (second row). The columns represent 3 distinct machine-learning algorithms: random forest (left), gradient boosting machines (middle), and XGBoost (right). In the machine-learning algorithms with VCTE, log VCTE is the most important, followed by log Aspartate Aminotransferase-based Fibrosis Score-4 and various orders of log platelet, age log GGTP, and BMI. In the machine-learning algorithms without VCTE, FIB-4 is the most important variables include log AST, age, log platelet, log GGTP, and BMI. BMI, body mass index; FIB-4, Fibrosis-4; GGTP, gamma-glutamyl transpeptidase; VCTE, vibration-controlled transient elastography.

based cohort, this approach allowed for the exclusion of significant fibrosis or higher in >90% of the patients.

Of the initial contenders, RF, GBM, and XGB outperformed EN and NN in the test set and proceeded with ensemble model building. RF, GBM, and XGB are the 3 decision-tree-based methods. RF improves accuracy using bagging, where multiple deep decision trees are trained on different random subsets of the data. GBM and XGB are based on the boosting approach, where a series of shallow trees are trained, and each minimizes the error made from the previous tree. Compared with conventional decision trees, these 3 tree-based methods can avoid overfitting and improve the accuracy. They can also better handle nonlinear relationships and complex interactions between covariates and clinical outcomes than traditional linear methods, such as ridge regression, least absolute shrinkage and selection operator regression, and EN. In addition, NNs, which are powerful for complex data, often require extensive tuning and large amounts of data to generalize well and avoid overfitting, which can lead to underperformance when our training set is approximate one thousand in size. This could explain why tree-based methods, owning to their robustness to diverse data structures and less demanding hyperparameter tuning, outperformed NNs in the test set.

In the ALADDIN-F2 models, with and without VCTE, age, AST, platelet, GGTP, and protein and total cholesterol emerge as key features. GGTP and total cholesterol are also being used in the LiverRisk score (16). Globulin, derived from the difference between protein and albumin, is included in the SAFE score (17). Additional moderately important variables include BMI, creatinine, and HbA1c. Although removing less important features might streamline data entry into the website, doing so incrementally reduces the AUC in both the training and test sets. Furthermore, eliminating these features does not reduce costs, as laboratory tests are typically ordered in bundles, such as the comprehensive metabolic panel and lipid panel.

Patients with significant fibrosis or higher (≥F2) because of MASH are at an increased risk of progression and morbidity (4) and may benefit from newly approved resmetirom treatments (13). Given the limitations of VCTE availability, only a few thousand VCTE are available in the United States, therefore a score that is not dependent on access to VCTE for providers who do not have access to point-of-care VCTE. We introduced 2 novel algorithms, ALADDIN-F2-VCTE and ALADDIN-F2-Lab. The ALADDIN-F2-VCTE algorithm offers a highly specific diagnosis, achieving greater than 90% specificity for significant fibrosis or higher in both the test and external validation sets. This high

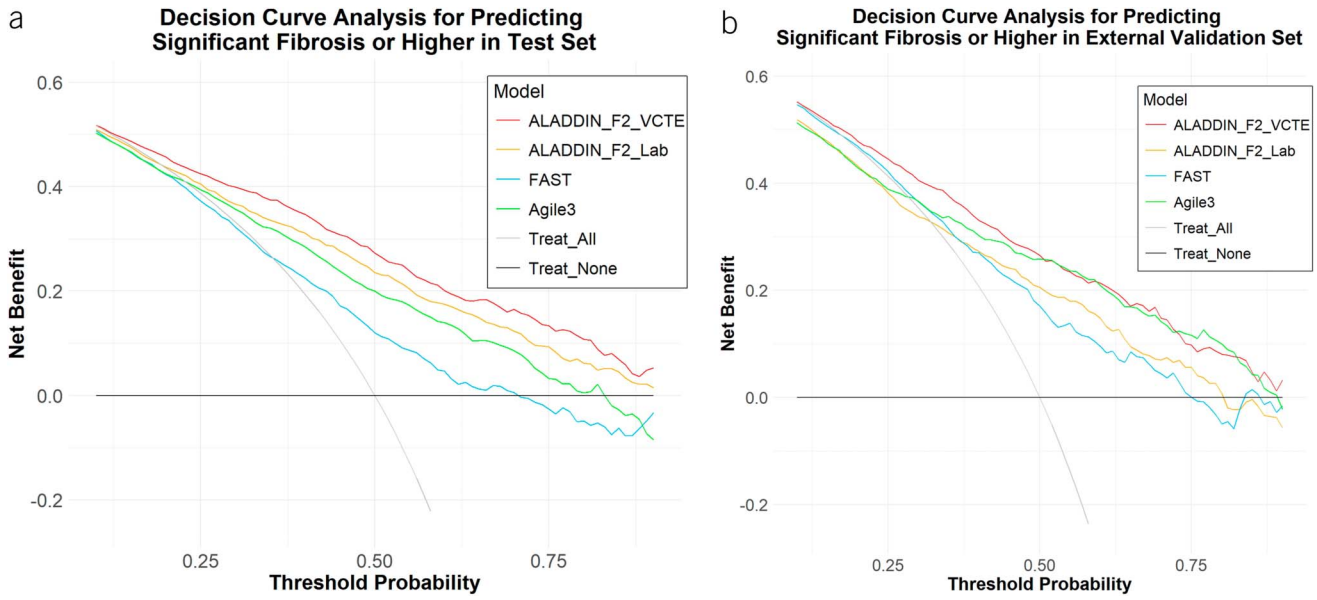


Figure 3. Decision curve analysis. X-axis (threshold probability) represents the probability threshold for classifying a patient as positive for significant fibrosis or higher. Y-axis (net benefit) is a measure that balances true-positives and false-positives. ALADDIN-F2-VCTE (red line) consistently shows a higher net benefit compared with FAST (blue line) in both test set and external validation set across a wide range of threshold probabilities, especially between 0.1 and 0.5. This suggests that the ALADDIN_VCTE model is better at balancing the benefits of true-positives while minimizing the harm of false-positives compared with the FAST model in this range. While ALADDIN-F2-VCTE shows superiority over Agile-3 (green line), this advantage is modest. Between threshold probabilities of approximately 0.1 and 0.5, the ALADDIN-F2-VCTE outperforms both FAST and the "treat all" strategy, meaning that it is the most beneficial in this clinically relevant range. As the threshold probability increases (above ~0.5), the net benefit of all models decreases sharply and eventually converge toward zero, similar to "treat none" (the black line) strategy. ALADDIN, mAchine Learning ADvanceD fibrosis and at-risk mash Novel predictor; FAST, FibroScan-aspartate aminotransferase; VCTE, vibration-controlled transient elastography.

specificity could reduce the necessity for liver biopsy in up to a third of patients in referral and tertiary referral centers, thereby broadening accessibility for resmetirom and other future treatments. The ALADDIN-F2-Lab algorithm, which does not require VCTE data, provides a sensitivity and specificity >85% for significant or advanced fibrosis.

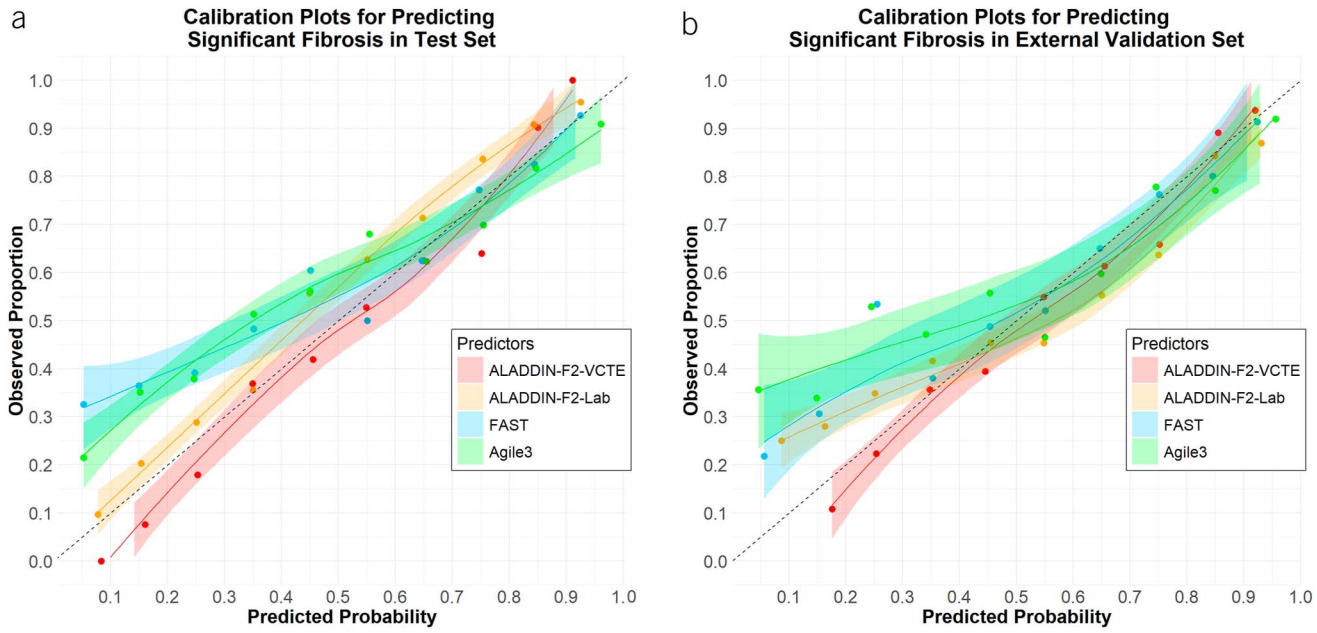


Figure 4. Calibration plots. The calibration plot visualizes the model's calibration, comparing predicted probabilities with observed frequencies using a locally weighted regression (Loess) for smoothing. This method highlights areas of suboptimal calibration within the predicted range. The shaded region represents the 95% confidence interval, indicating the certainty of the calibration curve. The dots represent the observed probabilities for each decile of predicted probabilities. Ideal calibration is shown by the dotted diagonal line, indicating exact matches between predictions and outcomes. Deviations from the diagonal reflect discrepancies, with the line above suggesting risk underestimation and below for overestimation.

Table 3. Comparative analysis of classification accuracy based on dual cut points approach

	Prediction of significant fibrosis							
	Internal validation cohort							
	ALADDIN-F2-VCTE	ALADDIN-F2-lab	FIB-4	VCTE	FAST	Agile 3	SAFE	Liver risk
Rule-out cut point	0.36	0.37	1.3	8.2	0.35	0.451	0	10
Rule-out zone % patients	19.0%	20.5%	59.6%	45.6%	41.8%	34.8%	19.0%	77.9%
Sensitivity	93.1%	92.5%	63.6%	80.2%	69.2%	81.6%	90.6%	29.0%
NPV	79.1%	79.2%	60.0%	67.3%	57.6%	64.2%	72.0%	72.0%
Indeterminate zone % patients	50.0%	47.5%	28.1%	29.7%	29.7%	18.3%	31.7%	32.4%
Rule-in cut point	0.77	0.72	2.67	12.1	0.67	0.679	100	15
Rule-in zone % patients	31.0%	31.0%	12.3%	35.5%	28.2%	36.1	49.3%	2.5%
Specificity	90.1%	87.3%	96.3%	84.2%	86.8%	83.9%	71.7%	98.5%
PPV	86.4%	82.7%	85.0%	81.1%	80.0%	81.0%	75.0%	73.7%
	Prediction of significant fibrosis							
	External validation cohort							
	ALADDIN-F2-VCTE	ALADDIN-F2-lab	FIB-4	VCTE	FAST	Agile 3	SAFE	Liver risk
Rule-out cut point	0.36	0.37	1.3	8.2	0.35	0.451	0	10
Rule-out zone % patients	19.8%	22.1%	42.3%	35.8%	32.6%	35.8%	28.5%	70.1%
Sensitivity	91.8%	87.1%	68.7%	79.9%	78.6%	66.7&	80.1%	36.1%
NPV	75.7%	66.9%	58.1%	67.1%	61.4%	67.1%	60.5%	48.5%
Indeterminate zone % patients	48.2%	43.0%	39.0%	27.9%	32.4%	16.0%	22.0%	32.4%
Rule-in cut point	0.77	0.72	2.67	12.1	0.67	0.679	100	15
Rule-in zone % patients	32.0%	34.9%	18.7%	36.3%	35.0%	36.5%	49.5%	2.3%
Specificity	90.3%	83.3%	89.6%	83.4%	82.7%	86.5%	67.4%	98.8%
PPV	87.4%	79.2%	75.7%	81.1%	79.6%	84.7%	71.4%	76.7%

ALADDIN, mAchine Learning ADvanceD fibrosis and at-risk mash Novel predictor; FAST, FibroScan-aspartate aminotransferase; FIB-4, Fibrosis-4; NPV, negative predictive value; PPV, positive predictive value; SAFE, steatosis-associated fibrosis estimator.

Similar to the FAST and Agile-3 score, both the ALADDIN-F2-VCTE and ALADDIN-F2-Lab models provide probability estimates for significant fibrosis or higher. While a specific cut point for sensitivity and specificity is offered, it is less clinically meaningful for individual patients. Instead, the probability score offers a personalized estimate of significant fibrosis for any given patient. As shown in Figure 4, the observed probabilities closely match the predicted probabilities of the ALADDIN-F2-VCTE model, largely because of the accurate estimation of the pretest probabilities. When users input data into the web-based calculator, they are prompted to provide the prevalence of significant fibrosis in their local center, calculated to the nearest 10th percentile. This actionable insight enhances clinical decision making by providing more refined diagnostic precision.

Although major gastroenterological societies, including AGA (8), AASLD (9), and European Association for the Study of Liver Diseases (10,18) advocate using the FIB-4 as the initial screening test, our study shows that a rule-out cutoff of 1.3 has suboptimal sensitivity of 63%–68% for the diagnosis of significant fibrosis or higher in our patient setting originating from tertiary referral centers. On the other hand, we found an FIB-4 of >2.66 is highly specific for the diagnosis of significant fibrosis or higher, although relatively few patients fall within the rule-in zone. In addition, a recent study evidenced that AGA and AASLD clinical care algorithms yielded high false-negative rates, especially in Hispanic participants (19). By contrast, ALADDIN algorithms were performed including Hispanic population. Given the specificity, the use of ALADDIN-F2-VCTE or ALADDIN-F2-Lab could significantly decrease the need for further risk assessment with more expansive techniques.

Our study primarily explored the ALADDIN model, combining serum tests with and without VCTE, to diagnose significant fibrosis and higher. In addition, the literature describes various techniques for similar diagnoses. An Indian multicenter study with comprehensive training, testing, and external validation sets demonstrated that a random forest algorithm using common laboratory parameters without VCTE is superior to traditional parameters including FIB-4, nonalcoholic fatty liver disease fibrosis score, and SAFE Score for diagnosing significant fibrosis or higher (20). To diagnose advanced fibrosis and at-risk MASH using VCTE-based algorithms, substantial literature supports Agile-3, Agile-4 (14,21,22), and FAST (11,23). Chang's introduction of a random forest model for at-risk MASH (24) was notable, although limited by internal validation. In addition to VCTE, serum-based algorithms such as enhanced liver fibrosis combined with FIB-4 (25) and nonalcoholic fatty liver disease fibrosis score (26) and imaging-based methods such as the magnetic resonance elastography AST score using magnetic

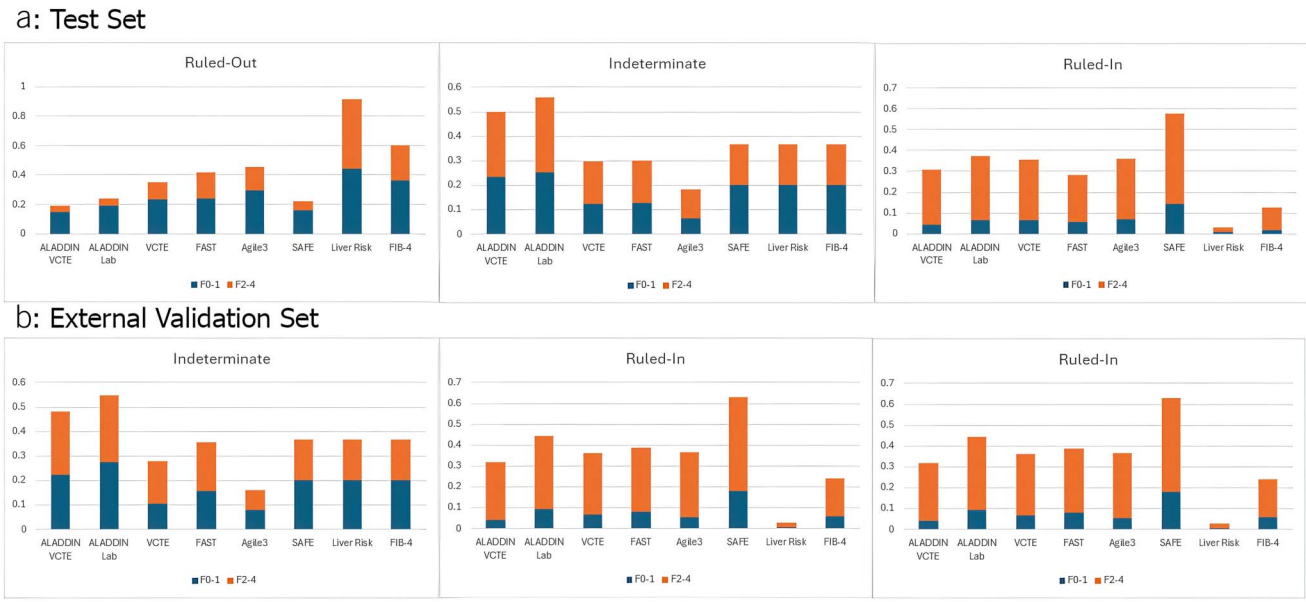


Figure 5. Patient distribution across diagnostic zones for significant fibrosis or higher. The bar charts present the distribution of patients among the rule-out, indeterminate, and rule-in categories for significant fibrosis or more, according to the diagnostic characteristics in Table 3. The bars are color-coded to represent negative (blue) and positive (orange) cases. In the left panels, the blue fraction corresponds to the negative predictive value (NPV), whereas in the right panels, the orange fraction represents the positive predictive value (PPV). The total height of each bar indicates the percentage of patients classified in the respective zones: Rule-out (left), indeterminate (middle), and rule-in (right). The top panel (**a**) displays results from the test set, whereas the bottom panel (**b**) shows results from the external validation set. The bar charts highlight that ALADDIN-F2-VCTE improved NPV and PPV compared with other models. However, it rules out fewer patients and places more in the indeterminate zone. In addition, ALADDIN-F2-VCTE demonstrates superior PPV and a comparable number of patients in the rule-in zone. The PPV of ALADDIN-F2-Lab matches closely with the VCTE-based FAST and Agile-3 scores. ALADDIN, mAchine Learning ADvanceD fibrosis and at-risk mash Novel predictor; FAST, FibroScan-aspartate aminotransferase; VCTE, vibration-controlled transient elastography.

resonance elastography (27) have been explored. The Magnetic Resonance Elastography plus Fibrosis-4 Index approach, which integrates magnetic resonance elastography (MRE) and FIB-4, is superior to magnetic resonance elastography AST (28) and FAST (27,29,30) for predicting significant fibrosis. Long-term perspectives include integrating the enhanced liver fibrosis test, Chronic Liver Disease Risk Score, and Polygenic Risk Score-5 to predict severe liver-related outcomes (31). For at-risk MASH, the Noninvasive Steatohepatitis 4 score (32) and Metabolomics-Advanced Steatohepatitis Fibrosis Score outperform FAST (33). The Agile-3 model is notable for predicting liver-related events (34), and the Liver Investigation: Testing Marker Utility in Steatohepatitis project introduces high-performance tests such as SomaSignal and Age, Diabetes, Pro-C3, and Platelets for diagnosing non-alcoholic steatohepatitis and significant fibrosis (35).

Our study has limitations similar to others in this field (11,14). Local pathologists interpreted liver biopsies, not a consensus of experts, affecting reliability, but typically affecting the identification of lobular inflammation and ballooning (15) more than fibrosis and, therefore, should have less impact in this study. The inclusion of patients with liver biopsy in referral centers and the retrospective nature of the study introduced potential biases and limited the extrapolation of this model to all patients with MASLD. Despite the inclusion of Bayesian updates to adjust for pretest probability, the accuracy of this model in screening and community centers requires further studies. Although ALADDIN models achieve better predictive performance by combining a complex combination of decision trees, they trade off interpretability relative to linear models such as logistic regression. To address this problem, our study includes a web-based calculator that provides user-friendly, personalized predictions and variable importance estimates to enhance interpretability. Although high performance was shown in the test set and external validation set, the potential for overfitting is a concern, especially when considering complicated models trained with heterogeneous medical data. To minimize this concern, we used an external validation set that was substantially different from the test set, as presented in Table 1. In addition, we used an ensemble method that aggregates predictions of several algorithms, thereby reducing variance and overfitting by averaging the unique patterns observed in individual models. We also acknowledge that our target for comparison: FAST score was designed for diagnosing at-risk MASH, Agile-3 and FIB-4 was designed for diagnosing advanced fibrosis, LiverRisk score was designed for primary care setting, and only SAFE score was specifically designed for significant fibrosis or higher.

In conclusion, the ALADDIN model, trained and validated through a large global consortium of 14 centers and 2,677 patients across 5 continents, stands out as a groundbreaking tool specifically designed for diagnosing significant fibrosis or higher (≥F2) using commonly available laboratory parameters with and without VCTE. By using a dual cutoff approach, ALADDIN achieved high diagnostic accuracy, with the rule-in cutoff offering over 90% specificity in one-third of patients in referral and tertiary settings and the rule-out cutoff providing over 90% sensitivity, effectively excluding significant fibrosis or higher in over 90% of a population-based cohort. The ALADDIN model

without VCTE, ALADDIN-F2-Lab, can serve as a reasonable alternative to ALADDIN-F2-VCTE in cases where VCTE is unavailable given the high degree of agreement and small mean difference between the models, thereby serving as an essential tool in clinical practices that do not otherwise have access to VCTE to guide prescription decisions for newly approved drugs for MASH for at-risk (≥F2) disease. ALADDIN-F2-Lab is superior to the currently available common laboratory parameter-based algorithms, including FIB-4, SAFE, and LiverRisk score, and noninferior to the FAST Score. These results support the use of ALADDIN-F2-Lab in clinical practice to assess liver fibrosis when the VCTE access is limited. Its versatility, functioning both with and without VCTE data, combined with an accessible web-based calculator, underscores its potential to significantly enhance the noninvasive diagnosis and management of MASLD across various healthcare environments.

ACKNOWLEDGEMENTS

We acknowledge that Stephen Harrison has passed away and honor their significant contribution to this work.

CONFLICTS OF INTEREST

Guarantor of the article: Winston Dunn, MD.

Specific author contributions: W.D., N.A., and T.C.F.Y. conceived and designed the study; all authors collected the data and contributed to data analysis and interpretation; W.D., N.A., T.C.F.Y., L.C., L.A.D., N.V., J.P.A., S.-M.J., J.D., A.S., V.W.-S.W., V.L.C., and M.A.: performed the final analysis and drafted the manuscript; N.D. designed website calculators; all authors participated in drafting the article and revising it critically for important intellectual content; all authors gave final approval of the submitted version.

Financial support: W.D. is a recipient of the National Institute of Diabetes and Digestive and Kidney award (K23DK10929401A1). A.K.S. is funded by grants U01 AA026980-06 from NIAAA, P20 GM103436 supplement from NIGMS, and PON27462400004956 from DHHS Department of health.

Potential competing interests: None to report.

Data availability statement: The datasets generated and analyzed during the current study are not publicly available, but are available from the corresponding author upon reasonable request.

Study Highlights

WHAT IS KNOWN

- ✓ FAST score is currently available for liver fibrosis assessment using VCTE.
- ✓ FIB-4, SAFE, and LiverRisk score assess fibrosis using common laboratory parameters.

WHAT IS NEW HERE

- ✓ Global study with robust external validation.
- ✓ ALADDIN-F2-VCTE outperforms FAST for predicting ≥F2 fibrosis.
- ✓ ALADDIN-F2-Lab outperforms FIB-4, SAFE, and LiverRisk score
- ✓ ALADDIN-F2-Lab offers a reliable alternative without VCTE.
- ✓ High sensitivity and specificity with a dual cutoff approach.

REFERENCES

1. Rinella ME, Lazarus JV, Ratziu V, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. Hepatology 2023;78(6):1966–86.
2. Younossi ZM, Golabi P, Paik JM, et al. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): A systematic review. Hepatology 2023;77(4):1335–47.
3. Younossi ZM, Stepanova M, Ong J, et al. Nonalcoholic steatohepatitis is the most rapidly increasing indication for liver transplantation in the United States. Clin Gastroenterol Hepatol 2021;19(3):580–9.e5.
4. Sanyal AJ, Van Natta ML, Clark J, et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. N Engl J Med 2021;385(17):1559–69.
5. Ng CH, Lim WH, Hui Lim GE, et al. Mortality outcomes by fibrosis stage in nonalcoholic fatty liver disease: A systematic review and meta-analysis. Clin Gastroenterol Hepatol 2023;21(4):931–9.e5.
6. Rockey DC, Caldwell SH, Goodman ZD, et al. Liver biopsy. Hepatology 2009;49(3):1017–44.
7. Davison BA, Harrison SA, Cotter G, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. J Hepatol 2020;73(6):1322–32.
8. Kanwal F, Shubrook JH, Adams LA, et al. Clinical care pathway for the risk stratification and management of patients with nonalcoholic fatty liver disease. Gastroenterology 2021;161(5):1657–69.
9. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023;77(5):1797–835.
10. European Association for the Study of the Liver. EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis - 2021 update. J Hepatol 2021;75(3):659–89.
11. Newsome PN, Sasso M, Deeks JJ, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: A prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020;5(4):362–73.
12. Noureddin M, Muthiah MD, Sanyal AJ. Drug discovery and treatment paradigms in nonalcoholic steatohepatitis. Endocrinol Diabetes Metab 2020;3(4):e00105.
13. Harrison SA, Bedossa P, Guy CD, et al. A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. N Engl J Med 2024;390(6):497–509.
14. Sanyal AJ, Foucquier J, Younossi ZM, et al. Enhanced diagnosis of advanced fibrosis and cirrhosis in individuals with NAFLD using FibroScan-based Agile scores. J Hepatol 2023;78(2):247–59.
15. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005;41(6):1313–21.
16. Serra-Burriel M, Juanola A, Serra-Burriel F, et al. Development, validation, and prognostic evaluation of a risk score for long-term liver-related outcomes in the general population: A multicohort study. Lancet 2023;402(10406):988–96.
17. Sripongpun P, Kim WR, Mannalithara A, et al. The steatosis-associated fibrosis estimator (SAFE) score: A tool to detect low-risk NAFLD in primary care. Hepatology 2023;77(1):256–67.
18. European Association for the Study of the Liver EASLEuropean Association for the Study of Diabetes EASDEuropean Association for the Study of Obesity EASO. EASL-EASD-EASO Clinical Practice Guidelines on the management of metabolic dysfunction-associated steatotic liver disease (MASLD). Obes facts 2024;17(4):374–444.
19. Lazarus JV, Han H, Mark HE, et al. The global fatty liver disease Sustainable Development Goal country score for 195 countries and territories. Hepatology 2023;78(3):911–28.
20. Verma N, Duseja A, Mehta M, et al. Machine learning improves the prediction of significant fibrosis in Asian patients with metabolic dysfunction-associated steatotic liver disease–the Gut and Obesity in Asia (GO-ASIA) Study. Aliment Pharmacol Ther 2024;59(6):774–88.
21. Pennisi G, Enea M, Pandolfo A, et al. AGILE 3+ score for the diagnosis of advanced fibrosis and for predicting liver-related events in NAFLD. Clin Gastroenterol Hepatol 2023;21(5):1293–302.e5.

22. Noureddin M, Mena E, Vuppalanchi R, et al. Increased accuracy in identifying NAFLD with advanced fibrosis and cirrhosis: Independent validation of the agile 3+ and 4 scores. Hepatol Commun 2023;7(5):e0055.
23. Woreta TA, Van Natta ML, Lazo M, et al. Validation of the accuracy of the FAST score for detecting patients with at-risk nonalcoholic steatohepatitis (NASH) in a North American cohort and comparison to other non-invasive algorithms. PLoS One 2022;17(4):e0266859.
24. Chang D, Truong E, Mena EA, et al. Machine learning models are superior to noninvasive tests in identifying clinically significant stages of NAFLD and NAFLD-related cirrhosis. Hepatology 2023;77(2):546–57.
25. Kjaergaard M, Lindvig KP, Thorhauge KH, et al. Using the ELF test, FIB-4 and NAFLD fibrosis score to screen the population for liver disease. J Hepatol 2023;79(2):277–86.
26. Younossi ZM, Stepanova M, Felix S, et al. The combination of the enhanced liver fibrosis and FIB-4 scores to determine significant fibrosis in patients with nonalcoholic fatty liver disease. Aliment Pharmacol Ther 2023;57(12):1417–22.
27. Tamaki N, Imajo K, Sharpton S, et al. Magnetic resonance elastography plus Fibrosis-4 versus FibroScan-aspartate aminotransferase in detection of candidates for pharmacological treatment of NASH-related fibrosis. Hepatology 2022;75(3):661–72.
28. Noureddin M, Truong E, Gornbein JA, et al. MRI-based (MAST) score accurately identifies patients with NASH and significant fibrosis. J Hepatol 2022;76(4):781–7.
29. Kim BK, Tamaki N, Imajo K, et al. Head-to-head comparison between MEFIB, MAST, and FAST for detecting stage 2 fibrosis or higher among patients with NAFLD. J Hepatol 2022;77(6):1482–90.
30. Jung J, Loomba RR, Imajo K, et al. MRE combined with FIB-4 (MEFIB) index in detection of candidates for pharmacological treatment of NASH-related fibrosis. Gut 2021;70(10):1946–53.
31. Åberg F, Saarinen K, Jula A, et al. Combined use of the ELF test and CLivD score improves prediction of liver-related outcomes in the general population. Liver Int 2023;43(10):2107–15.
32. Harrison SA, Ratziu V, Boursier J, et al. A blood-based biomarker panel (NIS4) for non-invasive diagnosis of non-alcoholic steatohepatitis and liver fibrosis: A prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020;5(11):970–85.
33. Noureddin M, Truong E, Mayo R, et al. Serum identification of at-risk MASH: The metabolomics-advanced steatohepatitis fibrosis score (MASEF). Hepatology 2024;79(1):135–48.
34. Pennisi G, Enea M, Romero-Gomez M, et al. Risk of liver-related events in metabolic dysfunction-associated steatohepatitis (MASH) patients with fibrosis: A comparative analysis of various risk stratification criteria. Hepatology 2024;79(4):912–25.
35. Vali Y, Lee J, Boursier J, et al. Biomarkers for staging fibrosis and non-alcoholic steatohepatitis in non-alcoholic fatty liver disease (the LITMUS project): A comparative diagnostic accuracy study. Lancet Gastroenterol Hepatol 2023;8:714–25.