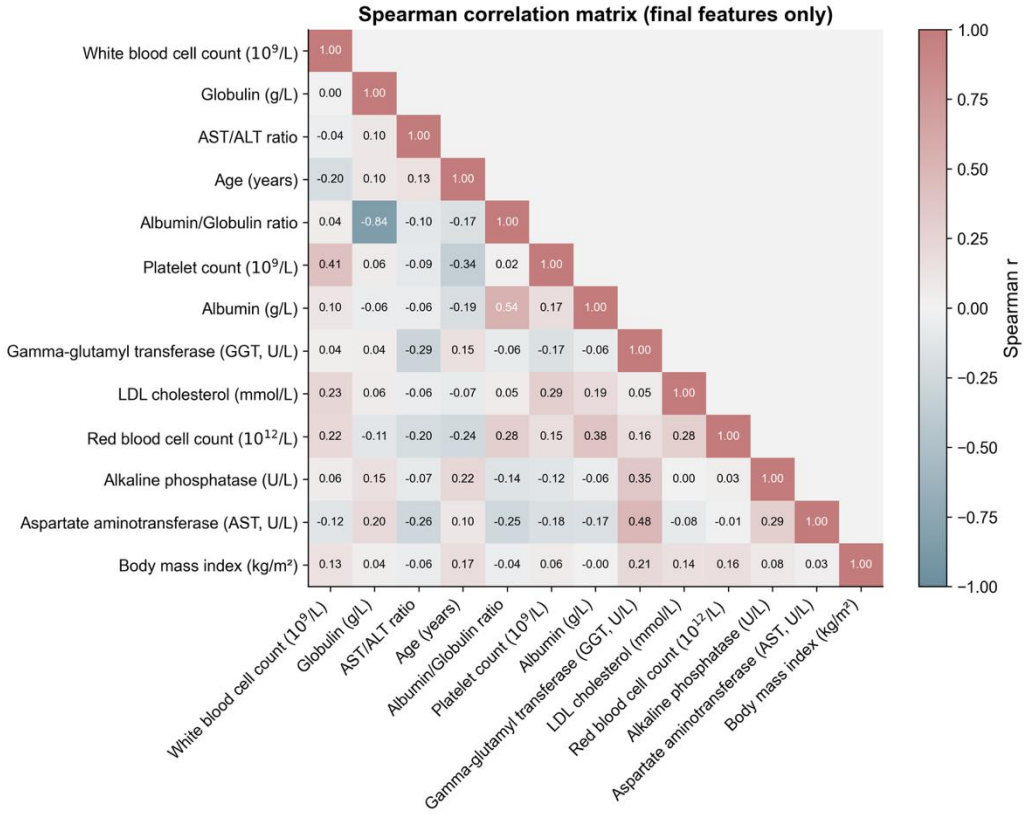
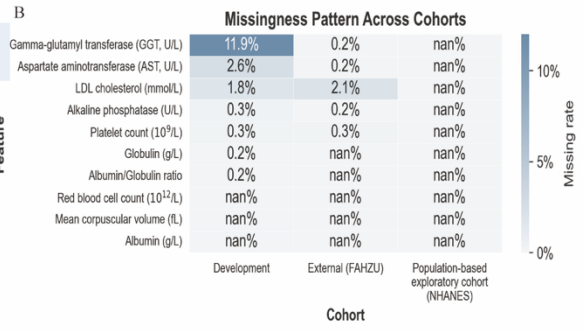


Supplementary Figure 1 Spearman correlation matrix of final predictors. Spearman correlation coefficients among the 13 final predictors included in the stacking ensemble model. Color intensity reflects the magnitude and direction of pairwise correlations (red = positive correlation; blue = negative correlation). As expected, a strong inverse correlation was observed between globulin and the albumin-to-globulin (A/G) ratio due to their mathematical relationship. No other pairwise correlations exceeded conventional thresholds for severe collinearity. This analysis supports the absence of problematic multicollinearity among the final feature set.

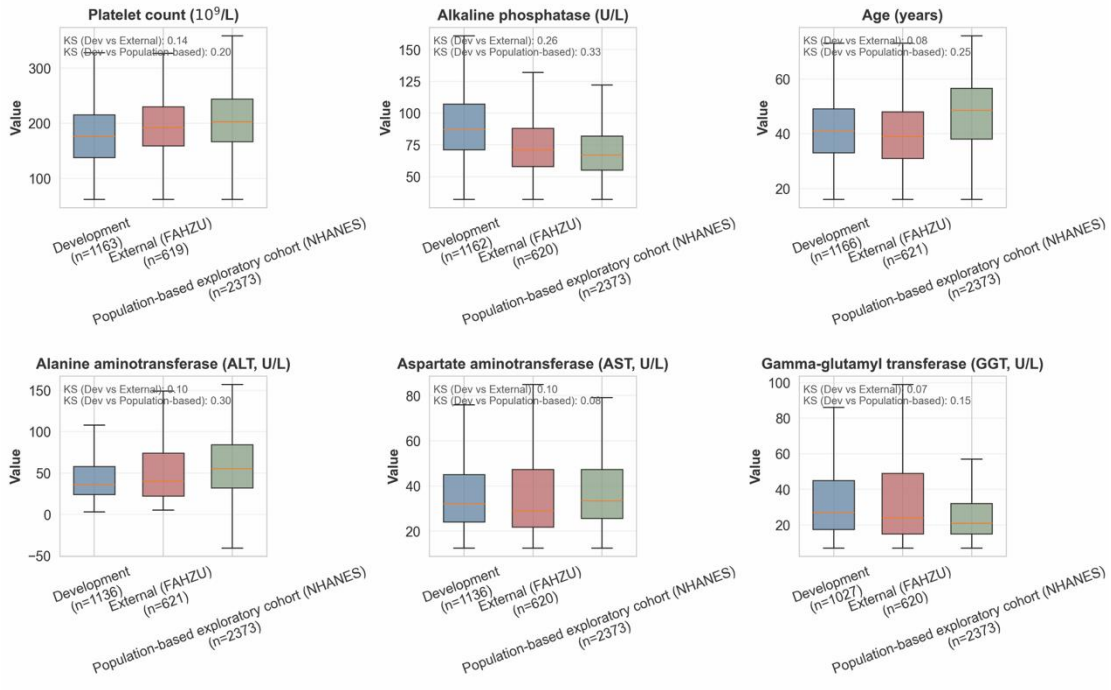
A

feature	KS (Dev vs FAHZU)	KS P-value (Dev vs FAHZU)	KS (Dev vs NHANES)	KS P-value (Dev vs NHANES)
AST/ALT ratio	0.562	<0.001	0.985	<0.001
Globulin (g/L)	0.170	<0.001	0.991	<0.001
Albumin (g/L)	0.185	<0.001	0.995	<0.001
Platelet count (10^9/L)	0.138	<0.001	0.204	<0.001
Age (years)	0.084	0.006	0.262	<0.001
Gamma-glutamyl transferase (U/L)	0.118	<0.001	0.223	<0.001
White blood cell count (10^9/L)	0.045	0.383	0.285	<0.001
Albumin/Globulin ratio	0.248	<0.001	0.994	<0.001
Red blood cell count (10^{12}/L)	0.046	0.348	0.095	<0.001
Aspartate aminotransferase (U/L)	0.110	<0.001	0.062	0.005
LDL-cholesterol (mmol/L)	0.079	0.012	0.992	<0.001
Alkaline phosphatase (U/L)	0.263	0.000	0.328	<0.001
Mean corpuscular volume (fL)	0.036	0.638	0.129	<0.001

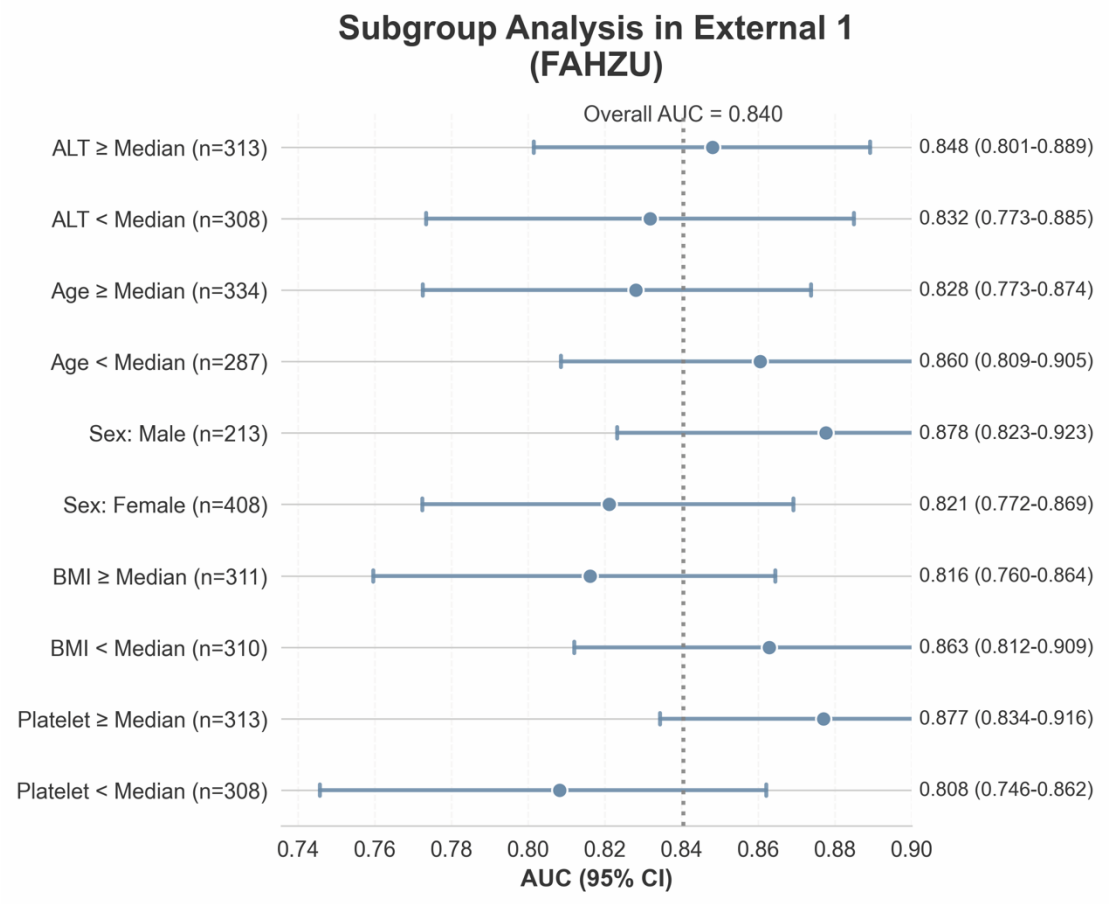
B



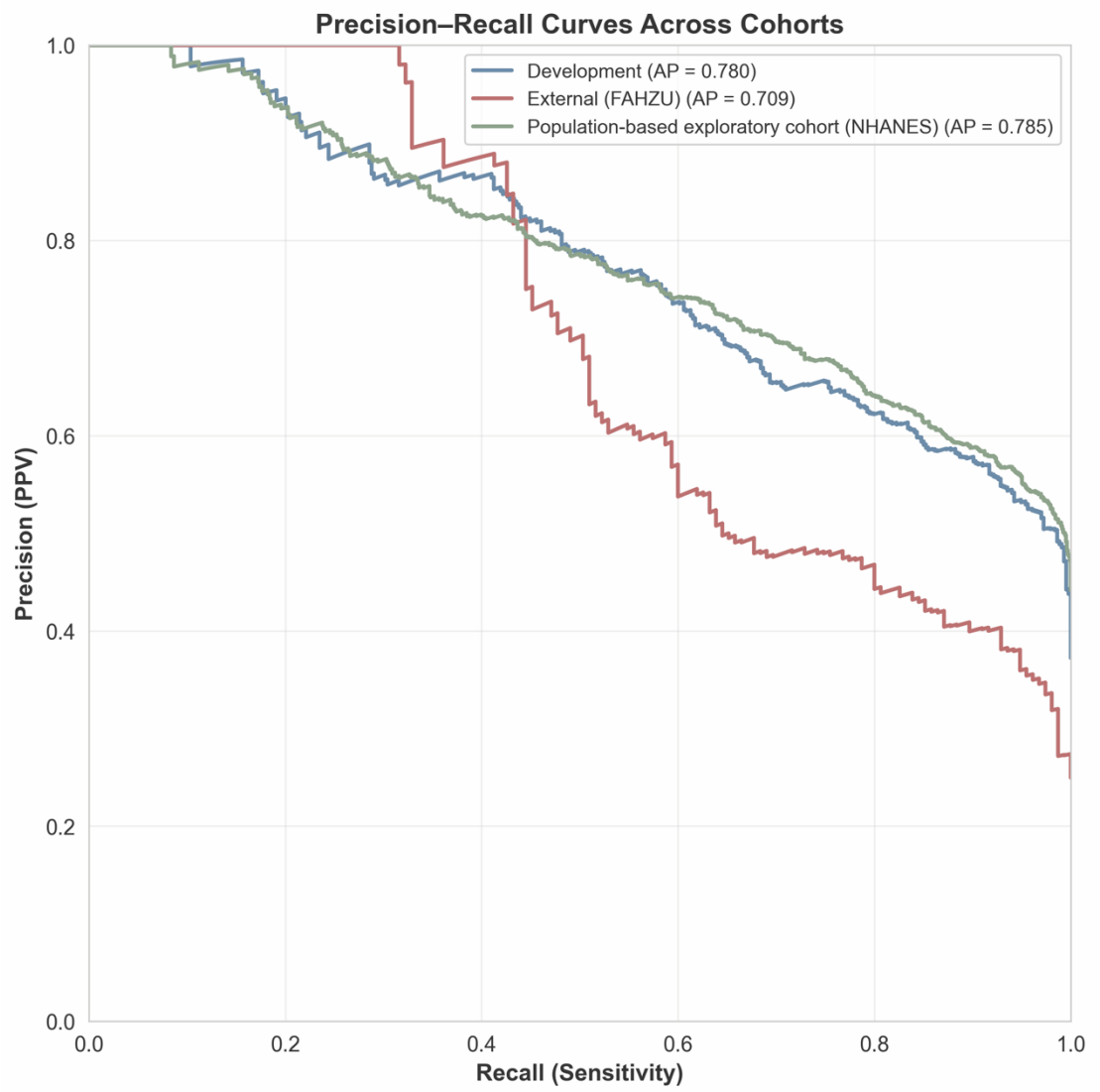
Supplementary Figure 2 Distributional shift and missingness of candidate predictors across cohorts. (A) Kolmogorov–Smirnov (KS) statistics comparing the empirical distributions of candidate predictors between the development cohort and the biopsy-confirmed external cohort (FAHZU), and between the development cohort and the population-based NHANES cohort. Larger KS values indicate greater distributional divergence. (B) Variable-level missingness across cohorts before imputation. Missingness proportions are shown for the development cohort, FAHZU cohort, and NHANES cohort. Variables derived through predefined formulas (e.g., globulin in NHANES) are shown as available after derivation. This figure illustrates substantial cross-cohort heterogeneity in feature distributions, particularly between hospital-based cohorts and NHANES, while also demonstrating that all final model predictors were available or derivable across cohorts after harmonization.



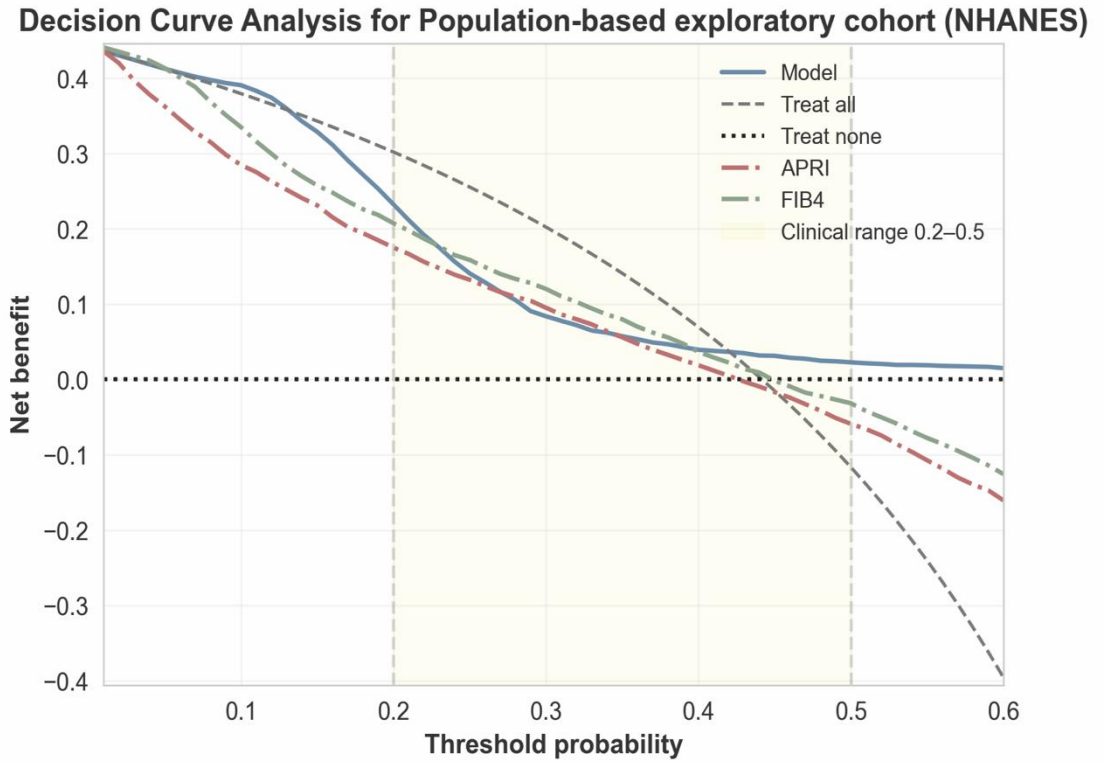
Supplementary Figure 3 Distribution of key predictors across cohorts. Boxplots compare six predictors used in the final model—platelet count (PLT), alkaline phosphatase (ALP), age, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma-glutamyl transferase (GGT)—across the development cohort, the biopsy-confirmed external cohort (FAHZU), and the population-based exploratory cohort (NHANES). To improve visual comparability and reduce the influence of extreme values, each variable was clipped at unified percentiles (0.5th–99.5th) prior to plotting. For each panel, Kolmogorov–Smirnov (KS) statistics are reported for development vs FAHZU and development vs NHANES, quantifying distributional shift across cohorts.



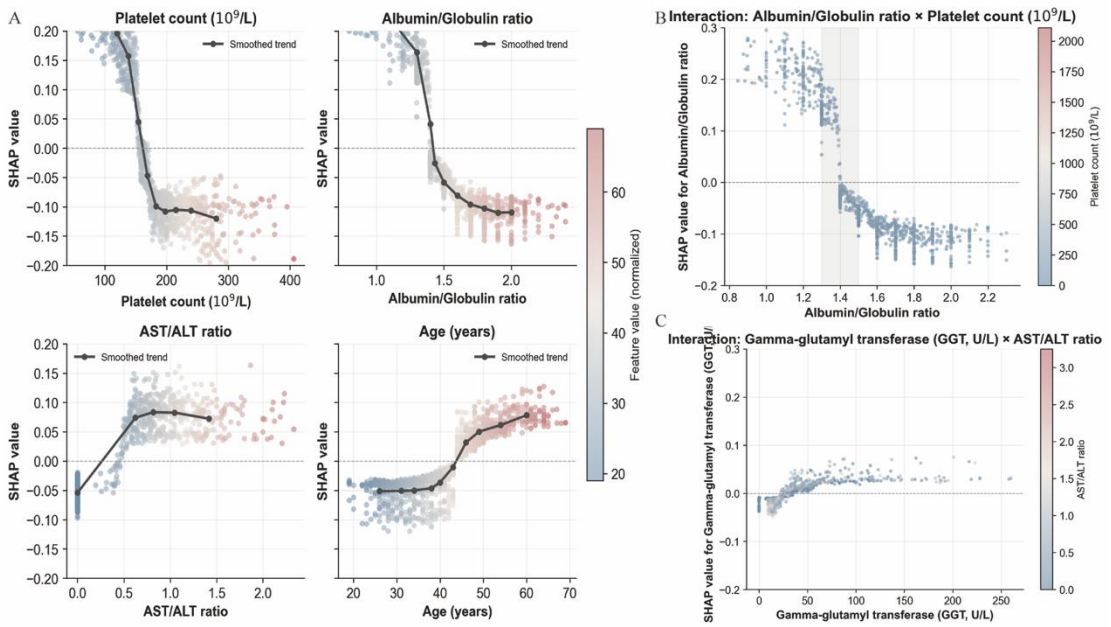
Supplementary Figure 4 Subgroup discrimination performance in the external validation cohort (FAHZU). Forest plot showing the area under the receiver operating characteristic curve (AUC) of the final ensemble model across predefined clinical subgroups in the biopsy-confirmed external cohort (FAHZU). Subgroups were defined by age, sex, body mass index, alanine aminotransferase level, and platelet count. Points indicate subgroup-specific AUC estimates, and horizontal lines represent 95% confidence intervals. The vertical dashed line denotes the overall cohort AUC. Consistent AUCs with overlapping confidence intervals indicate stable model discrimination across clinically relevant subgroups.



Supplementary Figure 5 Precision–recall performance of the ensemble model across cohorts. Precision–recall (PR) curves for the ensemble model in the development cohort, the biopsy-confirmed external validation cohort (FAHZU), and the population-based exploratory cohort (NHANES). Curves depict the trade-off between precision and recall across probability thresholds. Average precision (AP) is reported for each cohort. In NHANES, PR performance reflects discrimination against APRI/FIB-4–derived surrogate fibrosis labels rather than histologically confirmed outcomes.



Supplementary Figure 6 Decision curve analysis in the population-based NHANES cohort (surrogate-labeled). Decision curve analysis showing the net benefit of the ensemble model compared with "treat-all" and "treat-none" strategies across a range of threshold probabilities in the NHANES cohort. Fibrosis status in NHANES was defined using APRI/FIB-4–based surrogate criteria rather than histology; therefore, results are presented for exploratory interpretation only. Because the prevalence of surrogate-defined significant fibrosis in NHANES is low, the net benefit of the "treat-none" strategy appears higher than that of "treat-all" across much of the threshold range. This reflects the underlying class distribution and does not indicate clinical superiority. The ensemble model does not introduce net harm relative to either reference strategy and shows comparable performance under surrogate labeling, supporting its exploratory transportability in a population-based setting.



Supplementary Figure 7 SHAP-based feature dependence plots for key predictors. SHAP dependence plots illustrating the relationship between individual predictor values and their contributions to the model output, based on the surrogate model. Each point represents a participant, with the SHAP value indicating the direction and magnitude of the feature's effect on predicted fibrosis risk. Lower platelet counts and lower albumin-to-globulin (A/G) ratios are associated with higher predicted risk of significant fibrosis, while higher AST/ALT ratios and older age show monotonic increases in risk. Color gradients indicate the value of an interacting feature, highlighting potential feature–feature interactions. These plots demonstrate non-linear yet clinically coherent effects of core laboratory variables on model predictions and support the biological plausibility and interpretability of the model.

Supplementary Table 1 Availability and missingness of candidate variables across cohorts

Variable	Development (Taizhou Hospital)	Development (Zhejiang Provincial People's Hospital)	External Validation (FAHZU)	NHANES	Final model
Age (years)	0	0	0.06	Available	Yes
Sex (male)	0	0	0	Available	No
Body mass index (kg/m^2)	7.7	15.18	9.53	Available	No
Height (cm)	7.61	14.85	9.53	Available	No
Weight (kg)	2.92	14.61	8.63	Available	No
Alanine aminotransferase (ALT, U/L)	4.78	0.11	0	Available	Yes
Aspartate aminotransferase (AST, U/L)	4.78	0.11	0.39	Available	Yes
Gamma-glutamyl transferase (GGT, U/L)	15.58	0.22	0.39	Available	Yes
Alkaline phosphatase (ALP, U/L)	0	0.66	0.45	Available	Yes
Albumin (g/L)	0	0.11	0	Available	Yes
Globulin (g/L)	0	0.33	0	Derived using standard clinical formula	Yes

				(TP – Albumin)	
Platelet count ($\times 10^9$/L)	0.18	0.11	0.39	Available	Yes
Hemoglobin (g/L)	0.09	0.11	0.19	Available	No
Red blood cell count ($\times 10^{12}$/L)	0.18	0.11	0.19	Available	Yes
White blood cell count ($\times 10^9$/L)	0.09	0	0.19	Available	Yes
Mean corpuscular volume (fL)	0.09	0.22	0.19	Available	Yes
Mean platelet volume (MPV, fL)	0.09	—	3.89	Available	No
Mean corpuscular hemoglobin (MCH, pg)	0.09	—	0.19	Available	No
Mean corpuscular hemoglobin concentration (MCHC, g/L)	0.09	—	0.26	Available	No
Lactate dehydrogenase (LDH, U/L)	1.59	8.88	50.39	Not measured	No
Creatinine (μmol/L)	0.18	4.5	1.75	Available	No
Urea (mmol/L)	6.99	4.82	1.75	Available	No
Total bilirubin (μmol/L)	0	0.11	0.06	Available	No
Indirect bilirubin (μmol/L)	0	—	0	Available	No
Total protein (g/L)	0	—	0	Available	No
Total calcium (mmol/L)	0.35	—	3.31	Available	No

Inorganic phosphate (mmol/L)	0.35	—	3.24	Available	No
Sodium (mmol/L)	0.09	—	2.08	Available	No
Potassium (mmol/L)	0.09	—	2.14	Available	No
Chloride (mmol/L)	0.09	—	2.14	Available	No
Magnesium (mmol/L)	0.35	—	89.3	Available	No
Uric acid (μ mol/L)	0.18	—	1.75	Available	No
Fasting plasma glucose (mmol/L)	—	65.46	—	Available	No
Total cholesterol (mmol/L)	0.97	1.54	2.66	Available	No
HDL cholesterol (mmol/L)	0.97	1.43	2.66	Available	No
LDL cholesterol (mmol/L)	0.97	1.86	3.11	Available	Yes
Triglycerides (mmol/L)	0.97	1.54	2.72	Available	No
Creatine kinase (CK, U/L)	1.33	—	36.58	Available	No
Platelet distribution width (PDW, %)	0.09	—	3.89	Available	No
Plateletcrit (PCT, %)	0.09	—	3.96	Available	No
Alpha-L-fucosidase (AFU, U/L)	21.59	4.28	4.6	Not measured	No
Alpha-fetoprotein (AFP, ng/mL)	6.02	16.45	6.87	Not measured	No
HBV DNA (IU/mL)	64.51	11.29	14.14	Not measured	No
Hepatitis B surface antigen (HBsAg)	12.3	0	6.29	Available	No

Hepatitis B e antigen (HBeAg)	2.65	0	0.19	Not measured	No
Antibody to hepatitis B core antigen (anti-HBc)	2.3	0	0.13	Not measured	No
Antibody to hepatitis B e antigen (anti-HBe)	3.1	6.14	0.32	Not measured	No
Antibody to hepatitis B surface antigen (anti-HBs)	2.48	6.03	0	Not measured	No

Values shown for the hospital-based cohorts represent the percentage of missing observations for each variable before imputation.

The development cohort consisted of two independent hospital datasets: Taizhou Hospital and Zhejiang Provincial People's Hospital. Missingness is therefore reported separately for each development site to illustrate cross-center data availability and heterogeneity prior to harmonization.

The External Validation cohort was derived from the First Affiliated Hospital of Zhejiang University (FAHZU). For the NHANES cohort, variable availability is indicated qualitatively ("Available" or "Not measured"), as laboratory testing followed a standardized survey protocol rather than routine clinical practice.

Variables marked as "Derived" in NHANES were constructed using standard clinical formulas when all required components were available (e.g., globulin = total protein − albumin).

The "Final model" column indicates whether a variable was retained in the final stacking ensemble after feasibility screening, harmonization, and feature selection. Variables were excluded primarily due to limited cross-cohort availability, non-harmonizable definitions, or redundancy rather than predictive irrelevance.

Supplementary Table 2 Variable harmonization and pre-specified plausibility rules

Variable (standardized name)	Harmonized unit	Pre-specified plausibility rule (applied before imputation)	Handling of implausible values
Age	years	18–85	Set to missing
Platelet count (PLT)	$\times 10^9$/L	20–600	Set to missing
Albumin (ALB)	g/L	20–55	Set to missing
Globulin (GLB)	g/L	10–60	Set to missing
Albumin-to-globulin ratio (A/G)	unitless	0.2–5.0	Set to missing
Aspartate aminotransferase (AST)	U/L	5–1000	Set to missing
AST/ALT ratio	unitless	0.1–10.0	Set to missing
Alkaline phosphatase (ALP)	U/L	20–1000	Set to missing
Gamma-glutamyl transferase (GGT)	U/L	5–1000	Set to missing
White blood cell count (WBC)	$\times 10^9$/L	1.0–30.0	Set to missing
Red blood cell count (RBC)	$\times 10^{12}$/L	2.0–7.5	Set to missing

Mean corpuscular volume (MCV)	fL	60–120	Set to missing
LDL-cholesterol (LDL-C)	mmol/L	0.2–10.0	Set to missing

All variables were semantically harmonized to standardized names and converted to harmonized units prior to analysis.

Pre-specified plausibility ranges were defined a priori based on established physiological limits and routine clinical laboratory practice, and were intentionally set wider than the observed distributions in all cohorts (Table 1). Values outside these ranges were considered clinically implausible and were set to missing before imputation.

Outlier handling was performed uniformly across cohorts, without reference to outcome status, and prior to any model fitting.

Derived variables (A/G ratio and AST/ALT ratio) were calculated only when all required components were available.

Supplementary Table 3 Sensitivity analysis of model performance across imputation methods

Imputation method	Dev AUC (95% CI)	Ext AUC (95% CI)	Dev F1	Ext F1	Dev Brier	Ext Brier
Median (primary)	0.853 (0.830–0.872)	0.838 (0.800–0.872)	0.703	0.578	0.164	0.178
KNN	0.858 (0.835–0.877)	0.841 (0.804–0.875)	0.709	0.585	0.16	0.176
MICE	0.861 (0.838–0.880)	0.842 (0.806–0.876)	0.712	0.588	0.158	0.175

Model performance was re-estimated after replacing the primary median imputation with alternative strategies, including k-nearest neighbors (KNN) imputation and multiple imputation by chained equations (MICE), while keeping the feature set, model architecture, and evaluation procedures unchanged. Imputation was performed in a leakage-free manner within each training fold of the stratified five-fold cross-validation framework, and external cohorts were processed using the fitted preprocessing parameters without using any external outcome information. AUC values are reported with 95% confidence intervals; F1-score and Brier score are reported at the pre-specified operating point used in the main analysis. Dev, development cohort; Ext, biopsy-confirmed external validation cohort.

Supplementary Table 4 Variance inflation factors (VIFs) for the final predictors

Predictor	VIF
White blood cell count	1.42
Globulin	3.18
AST/ALT ratio	1.76
Age	1.53
Albumin/Globulin ratio	3.64
Platelet count	1.89
Albumin	2.47
Gamma-glutamyl transferase	1.58
LDL cholesterol	1.36
Red blood cell count	1.72
Alkaline phosphatase	1.44
Aspartate aminotransferase	1.91
Body mass index	1.29

Variance inflation factors (VIFs) were calculated for the 13 predictors included in the final model to assess multicollinearity. VIF values close to 1 indicate minimal collinearity, and no predictor exceeded commonly used thresholds suggestive of severe multicollinearity. The relatively higher VIFs for globulin and the albumin-to-globulin (A/G) ratio reflect their expected mathematical dependency rather than model instability.

Supplementary Table 5 Hyperparameters of base learners and meta-learner in the stacking ensemble

Learner	Implementation	Fixed hyperparameters
Logistic regression (L2)	LogisticRegression(solver='liblinear')	penalty="l2"; C=1.0; max_iter=2000
Random forest	RandomForestClassifier()	n_estimators=500; max_depth=None; min_samples_split=4; random_state=42
Gradient boosting	GradientBoostingClassifier()	n_estimators=400; learning_rate=0.03; max_depth=3
SVM (RBF kernel)	SVC(kernel='rbf', probability=True)	C=1.0; kernel="rbf"; probability=True
Extra Trees	ExtraTreesClassifier()	n_estimators=600; max_depth=None; random_state=42
CatBoost	CatBoostClassifier()	iterations=400; learning_rate=0.05; depth=4; loss_function='Logloss'
Meta-learner	LogisticRegression(solver='lbfgs')	penalty="l2"; C=1.0; max_iter=1000

All hyperparameters were pre-specified and kept fixed across cross-validation folds and cohorts. No grid search, random search, or Bayesian optimization was performed. Base learners were trained using stratified five-fold cross-validation (shuffle=True, random_state=42). Out-of-fold predicted probabilities were used to construct the level-1 meta-feature matrix. The meta-

learner was fitted on concatenated OOF predictions, and final models were refit on the full development cohort before external evaluation.

Supplementary Table 6 Exploratory operating characteristics of the model in the NHANES cohort

Cohort	AUC	Sensitivity	Specificity	F1
NHANES (surrogate-labeled, original model output)	0.817	0.173	0.995	0.293
NHANES (surrogate-labeled, recalibrated)	0.817	0.724	0.741	0.706

Performance metrics are shown for the population-based NHANES cohort, in which fibrosis status was defined using APRI/FIB-4–based surrogate criteria rather than histological confirmation.

"Original model output" refers to performance obtained using the unchanged model probabilities. "Recalibrated" results reflect a simple post hoc probability adjustment applied to improve operating characteristics under the surrogate-defined phenotype, without altering the relative ranking of predictions; consequently, the area under the receiver operating characteristic curve (AUC) remains unchanged

These operating characteristics are presented for exploratory purposes only and should be interpreted cautiously, as they are specific to the surrogate labeling strategy used in NHANES and do not represent biopsy-validated clinical performance.

This recalibration does not change the ranking of predictions and therefore does not affect discrimination (AUC).

Supplementary Table 7 Discriminative performance of individual base learners and the final ensemble across cohorts

Model	Development AUC	External AUC (FAHZU)	NHANES	
Logistic	0.806	0.792	0.869	baseline
Random Forest	0.849	0.838	0.791	best single model
Gradient Boosting	0.839	0.812	0.575	
SVM (RBF)	0.718	0.736	0.754	
Extra Trees	0.848	0.832	0.772	
CatBoost	0.841	0.815	0.746	
Stacking Ensemble	0.853	0.838	0.817	Final model

Area under the receiver operating characteristic curve (AUC) is reported for each individual base learner and for the final stacking ensemble in the development cohort, the biopsy-confirmed external validation cohort (FAHZU), and the population-based NHANES cohort.

All models were trained using the same leakage-free preprocessing and cross-validation pipeline within the development cohort and were evaluated unchanged in external cohorts.

In NHANES, fibrosis status was defined using APRI/FIB-4–based surrogate criteria; therefore, AUC reflects discrimination against surrogate outcomes rather than biopsy-confirmed fibrosis.

The stacking ensemble was selected as the final model based on its balanced and stable performance across cohorts rather than maximal performance in any single dataset.