Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Second-Step Risk Stratification Within the FIB-4 Gray Zone in MASLD: Derivation and Temporal Validation of a Four-Variable Clinical Model

Zhiyi Qiu
University of Chinese Medicine The fourth Clinical Medical College of Guangzhou
Xinyu Zhang
University of Chinese Medicine The fourth Clinical Medical College of Guangzhou
Yufeng Xing
Yufeng000729@163.com

Shenzhen Traditional Chinese Medicine Hospital

Research Article

Keywords: MASLD, FIB-4, risk stratification, liver stiffness, VCTE

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11481076/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background & Aims:

FIB-4 is recommended for first-line fibrosis risk assessment in MASLD, but many patients fall into the indeterminate gray zone (1.30–2.67), requiring further stratification. We tested whether a simple model using routine clinical variables improves second-step risk stratification after indeterminate FIB-4, especially when VCTE is unavailable.

Methods

Using NHANES data, we developed a weighted ridge logistic model in 2017–March 2020 (n = 2,709) and validated it in 2021–2023 (n = 1,543). Elevated liver stiffness was LSM ≥ 8.0 kPa. Predictors were age, BMI, ALT, and GGT. Performance was assessed by discrimination, calibration, prediction error, reclassification, decision-curve analysis, and prespecified gray-zone analyses.

Results

In temporal validation, the model had AUC 0.722 (95% CI 0.690–0.754) versus 0.581 (0.544–0.617) for FIB-4 (P < 0.001); survey-weighted AUCs were 0.735 and 0.562. Calibration was preserved (intercept 0.213; slope 1.014). Within the FIB-4 gray zone (n = 420; weighted LSM ≥ 8.0 kPa prevalence 23.7%), AUC was 0.697 versus 0.542 for FIB-4 (P < 0.001). Discrimination was higher than SAFE but similar to MAF-5. Thresholds 0.0923 and 0.2623 defined lower-, intermediate-, and higher-priority groups with event prevalences 8.2%, 23.8%, and 45.8%. At validation, sensitivity at the lower threshold was 90.8%; specificity at the upper threshold was 86.9%. At ~ 90% sensitivity, model-based prioritization reduced VCTE referral from 86.6% to 69.0%.

Conclusion

A simple four-variable model improved second-step risk stratification beyond FIB-4, particularly in the gray zone, with acceptable temporal calibration and clinically useful risk separation. It may help prioritize VCTE when access is constrained, but should complement, not replace, definitive fibrosis assessment, and requires external validation.

1. Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) affects more than one-third of adults worldwide, and fibrosis severity is the strongest predictor of liver-related outcomes [1–3]. Current guidelines therefore recommend sequential noninvasive risk stratification rather than universal histologic characterization [4, 5]. FIB-4 is the recommended first-line blood test; however, values in the

intermediate gray zone (1.30–2.67) do not reliably separate risk and mandate additional evaluation [4–7]. Second-line tools such as VCTE, enhanced liver fibrosis tests, and newer scores (SAFE, MAF-5) improve risk classification but require equipment, assays, or workflows that are often unavailable, and implementation of confirmatory testing remains incomplete in primary care [8–12]. Existing scores have also been designed primarily for general case-finding rather than for prioritizing patients with an indeterminate FIB-4 result when further testing is constrained. We hypothesized that a small set of routinely available variables could improve risk discrimination within this FIB-4 gray zone. Using temporally separated NHANES cohorts, we evaluated a prespecified four-variable model (age, BMI, ALT, GGT) as a second-step prioritization strategy. We assessed discrimination, calibration, prediction error, decision curves, reclassification, and exploratory probability-based strata; compared the model with FIB-4, SAFE, MAF-5, and NFS; and examined model-based prioritization under constrained VCTE capacity.

2. Materials and methods

2.1. Data Source and Study Design

Using publicly available NHANES data, we developed the model in the 2017–March 2020 prepandemic cycle and temporally validated it in the August 2021–August 2023 cycle, which was not used for predictor selection, model fitting, penalty selection, or recalibration [13, 14]. Analyses incorporated cycle-appropriate survey weights (WTMECPRP; WTPH2YR for blood-based analyses; fasting weights for NFS) and design variables. The NHANES data collection protocols were approved by the NCHS Research Ethics Review Board (ERB) of the National Center for Health Statistics, Centers for Disease Control and Prevention, and written informed consent was obtained from all participants at enrollment [15]. This study is a secondary analysis of de-identified, publicly available NHANES data and was therefore exempt from additional informed consent and local ethics review. Reporting followed TRIPOD + AI [16]. The primary analysis used complete cases; missingness is summarized in Supplementary Table S14. As a sensitivity analysis, multiple imputation by chained equations [17] (predictive mean matching, 20 imputations) was performed in development only; the outcome informed imputation of predictors but was not itself imputed, survey variables were included but not imputed, and survey-weighted logistic models were combined using Rubin's rules.

2.2. Study Population

Eligible participants were adults (≥ 18 years) with an NHANES operationalization of MASLD [18]: CAP ≥ 274 dB/m plus at least one of four harmonized cardiometabolic domains (adiposity, dysglycemia, blood pressure, lipids). Weekly alcohol was required to be < 210 g in men and < 140 g in women; participants with unclassifiable alcohol, hepatitis B surface antigen, or detectable HCV RNA were excluded. VCTE was included if the examination was complete, median LSM was positive, and IQR/median LSM < 0.30 [19, 20]. The primary outcome was VCTE-defined LSM ≥ 8.0 kPa, interpreted as elevated liver stiffness/fibrosis risk rather than histologic F2. Sensitivity analyses used LSM thresholds ≥ 6.0, ≥ 10.0,

and ≥ 13.0 kPa. NFS was analyzed in a separate fasting common subset and did not affect cohort membership.

2.3. Predictor Variables and Model Development

A broader candidate set comprised 12 clinical/biochemical variables; age, BMI, ALT, and GGT were specified a priori as the frozen four-variable model because they are routinely available and capture complementary biological dimensions. Tenfold cross-validated LASSO using glmnet [21] (normalized development weights) was run as a supplementary stability assessment and did not alter the prespecified predictors. The model was logit(p) = β0 + β1·Age + β2·BMI + β3·ALT + β4·GGT. Survey-weighted logistic regression provided population-representative associations; weighted ridge logistic regression (penalty λ chosen by outcome-stratified tenfold cross-validation on binomial deviance; lambda.min primary, lambda.1se sensitivity) generated predicted probabilities (coefficients in Supplementary Table S13). A survey-weighted Platt recalibration model was fitted in development and applied unchanged in validation as a secondary calibration analysis.

2.4. Comparator Scores

FIB-4 was calculated using the original formula [22] and categorized as low (< 1.30), gray zone (1.30–2.67), or high (> 2.67); an age-adjusted lower threshold of 2.00 was used among participants ≥ 65 years in sensitivity. NFS [23] was evaluated only in its fasting-weight subset (WTSAFPRP/WTSAF2YR). FAST [24] served as a VCTE-dependent reference benchmark. SAFE and MAF-5 were implemented as fixed comparators using published equations [10, 11], with diabetes operationalized separately for each. For probability-scale analyses, FIB-4 and FAST were mapped to outcome probability by survey-weighted logistic regression in development only; NFS was mapped in its fasting subset. Direct comparisons were restricted to the all-comparator common subset with complete data for every strategy.

2.5. Model Validation and Performance Assessment

Discrimination was quantified by AUCs with 95% CIs and compared with paired DeLong tests [25]; survey-weighted AUC point estimates were also reported. Calibration in temporal validation was assessed using survey-weighted calibration intercept [26] (predicted log-odds as offset), calibration slope, calibration plots, and Brier scores. Reclassification was explored with non-weighted IDI and continuous/categorical NRI [27]; the model threshold used Youden's index in development and the FIB-4 threshold was 1.30 after probability mapping. Survey-weighted decision-curve analysis [28] compared net benefit against treat-all and treat-none across threshold probabilities 0.01–0.40 in the gray zone; FAST was excluded because it already requires VCTE.

2.6. FIB-4 Gray Zone Analyses

The prespecified clinical focus was the FIB-4 gray zone (1.30–2.67); primary discrimination, calibration, prediction-error, and decision-curve analyses were repeated in this subgroup to assess second-step risk separation after an indeterminate FIB-4 result. Additional gray-zone sensitivity analyses examined alternative FIB-4 definitions, the age-adjusted gray zone, and restriction to adults aged ≥ 35 years.

2.7. Exploratory dual-cutoff risk stratification

Within the development-cohort gray zone, lower and upper thresholds for predicted probability were derived to achieve survey-weighted sensitivity ≥ 90% and specificity ≥ 90%, respectively, selecting the threshold pair with the smallest intermediate group; thresholds were locked before validation. Validation participants were classified into lower-priority, intermediate, and higher-priority groups and summarized by sensitivity, specificity, predictive values, population shares, and outcome prevalence. A parallel 85% sensitivity/85% specificity strategy was a sensitivity analysis. For VCTE resource benchmarking in validation, matched-sensitivity (80%, 90%) and fixed-capacity (20–50% referred) analyses compared the four-variable model with FIB-4, SAFE, and MAF-5.

2.8. Subgroup and Sensitivity Analyses

Prespecified subgroups (age < 60 vs ≥ 60, sex, BMI < 30 vs ≥ 30, diabetes status) were assessed descriptively using AUCs, AUC differences, and paired tests where estimable. Additional sensitivity analyses excluded FIB-4 > 2.67, used alternative LSM thresholds, the lambda.1se ridge solution, NFS in its fasting subset, and a three-variable (age/BMI/ALT) ridge model. Fivefold cross-validation grouped by primary sampling unit assessed internal stability.

2.9. Statistical software and reproducibility

Analyses used R 4.5.3 with the survey, glmnet, pROC, caret, and mice packages. The TRIPOD + AI checklist and reproducibility outputs are provided as supplements; an interactive Shiny application implements the frozen model and development-derived thresholds. Two-sided $P < 0.05$ was used; because SAFE/MAF-5 comparisons, subgroups, and resource-capacity analyses were secondary, nominal P values were interpreted with effect sizes, confidence intervals, and consistency.

3. Results

3.1. Study Population and Baseline Characteristics

After applying eligibility, VCTE-quality, alcohol, viral-hepatitis, survey-weight, and complete-case criteria, 4,252 participants were analyzed (Fig. 1). The development cohort included 2,709 adults from 2017–March 2020 (456 with LSM ≥ 8.0 kPa) and the temporal-validation cohort included 1,543 adults from 2021–2023 (313 events), representing 76.5 and 65.3 million US adults. Weighted prevalence of elevated liver stiffness was 16.7% and 19.4%. Baseline characteristics were similar across periods (Table 1); survey-weighted standardized mean differences for age, BMI, ALT, and GGT were ≤ 0.05 (Supplementary Table S12).

Table 1

Survey-weighted baseline characteristics of the development and temporal validation cohorts

Characteristic	Development cohort (2017–March 2020), n = 2,709	Temporal validation cohort (2021–2023), n = 1,543
Age, years	50.83 (SE 0.74)	50.56 (SE 0.66)
BMI, kg/m²	33.65 (SE 0.27)	33.74 (SE 0.28)
Waist circumference, cm	110.83 (SE 0.65)	111.19 (SE 0.71)
Systolic blood pressure, mmHg	124.24 (SE 0.50)	122.39 (SE 0.49)
Diastolic blood pressure, mmHg	76.42 (SE 0.33)	77.09 (SE 0.42)
HbA1c, %	5.97 (SE 0.04)	5.98 (SE 0.04)
HDL cholesterol, mg/dL	47.73 (SE 0.45)	47.86 (SE 0.45)
AST, U/L	22.38 (SE 0.29)	23.21 (SE 0.46)
ALT, U/L	26.83 (SE 0.55)	27.00 (SE 0.76)
GGT, U/L	34.05 (SE 0.90)	32.28 (SE 0.69)
Platelet count, ×10⁹/L	249.94 (SE 2.39)	262.40 (SE 2.49)
FIB-4	1.01 (SE 0.02)	1.01 (SE 0.02)
CAP, dB/m	322.67 (SE 1.21)	323.65 (SE 1.82)
LSM, kPa	6.75 (SE 0.22)	7.26 (SE 0.33)
Female, %	44.3 (SE 1.5)	44.6 (SE 1.0)
Non-Hispanic White, %	62.5 (SE 2.7)	61.6 (SE 2.6)
Non-Hispanic Black, %	8.3 (SE 1.2)	9.3 (SE 2.1)
Non-Hispanic Asian, %	5.1 (SE 0.9)	5.5 (SE 1.6)
Mexican American, %	12.1 (SE 1.7)	8.7 (SE 3.2)
Other Hispanic, %	7.5 (SE 0.9)	9.8 (SE 1.7)
Other/multiracial, %	4.6 (SE 0.6)	5.2 (SE 0.8)
Diagnosed diabetes, %	18.8 (SE 1.3)	17.6 (SE 1.2)
Hypertension criterion, %	55.9 (SE 1.8)	53.3 (SE 1.1)

Note. Continuous variables are survey-weighted means (SE); categorical variables are survey-weighted percentages (SE). No between-cohort significance testing is presented because the cohorts represent temporally separated development and validation samples. CAP, controlled attenuation parameter; LSM, liver stiffness measurement.

Characteristic	Development cohort (2017–March 2020), n = 2,709	Temporal validation cohort (2021–2023), n = 1,543
LSM ≥ 8.0 kPa, %	16.7 (SE 1.5)	19.4 (SE 1.1)

Note. Continuous variables are survey-weighted means (SE); categorical variables are survey-weighted percentages (SE). No between-cohort significance testing is presented because the cohorts represent temporally separated development and validation samples. CAP, controlled attenuation parameter; LSM, liver stiffness measurement.

3.2. FIB-4 Clinical Categories and Residual Risk in the Gray Zone

In validation, the weighted proportions with low (< 1.30), gray-zone (1.30–2.67), and high (> 2.67) FIB-4 were 76.7%, 20.8%, and 2.5% (Table 2). Weighted prevalence of LSM ≥ 8.0 kPa increased progressively from 17.3% (95% CI, 14.5–20.6) to 23.7% (95% CI, 21.0–26.6) to 47.0% (95% CI, 32.5–62.0). Among 420 gray-zone participants, 100 had elevated liver stiffness, a weighted population of approximately 13.6 million US adults, confirming substantial residual risk after indeterminate FIB-4.

Table 2
FIB-4 risk categories and VCTE-defined elevated liver stiffness in the temporal validation cohort

FIB-4 category	Unweighted n	LSM ≥ 8.0 kPa events, n	Survey-weighted population, %	Survey-weighted population, millions	Survey-weighted LSM ≥ 8.0 kPa prevalence, %
Low (< 1.30)	1,076	192	76.7	50.08	17.3
Gray zone (1.30–2.67)	420	100	20.8	13.56	23.7
High (> 2.67)	47	21	2.5	1.64	47.0

Note. The standard FIB-4 gray zone was defined as 1.30–2.67. Survey-weighted population shares were calculated using the primary temporal-validation survey weights. The outcome represents VCTE-defined elevated liver stiffness and should not be interpreted as histologic fibrosis stage.

3.3. Temporal Validation and Overall Model Performance

The frozen four-variable model achieved temporal-validation AUC 0.722 (95% CI, 0.690–0.754) versus 0.581 (95% CI, 0.544–0.617) for FIB-4 (P < 0.001; Table 3 and Fig. 2A); survey-weighted AUCs were 0.735 and 0.562. Visual comparisons of development and validation ROC curves in the overall and gray-zone populations are shown in Supplementary Figure S1. In the all-comparator common subset (n = 1,513; 300 events), AUCs were 0.721 for the model, 0.578 for FIB-4, 0.665 for SAFE (P = 0.002 vs model), and 0.733 for MAF-5 (P = 0.302) (Fig. 2C). FAST, as a VCTE-dependent benchmark, had AUC 0.808 (survey-weighted 0.788; Supplementary Figure S2). In the fasting subset, the model was not significantly better than NFS

(0.718 vs 0.675; P = 0.109; Supplementary Figure S3). Calibration was maintained, with survey-weighted intercept 0.213 (95% CI, 0.073–0.354) and slope 1.014 (95% CI, 0.790–1.238; Fig. 3A). Brier score was lower than development-mapped FIB-4 (0.135 vs 0.154); Platt scaling did not materially change prediction error. Decision-curve analysis in the full cohort is shown in Supplementary Figure S4.

Table 3. **Temporal validation performance of the four-variable model and comparator scores**

Table 3

A. Primary temporal-validation cohort: n = 1,543; 313 events

Strategy	n / events	Conventional AUC (95% CI)	Survey-weighted AUC	Survey-weighted Brier score	P vs four-variable model
Four-variable model	1,543 / 313	0.722 (0.690–0.754)	0.735	0.135	Reference
FIB-4	1,543 / 313	0.581 (0.544–0.617)	0.562	0.154	< 0.001
FAST (VCTE-dependent reference)	1,543 / 313	0.808 (0.780–0.836)	0.788	0.121	< 0.001

Table 3

B. All-comparator common subset: n = 1,513; 300 events

Strategy	n / events	Conventional AUC (95% CI)	Survey-weighted AUC	P vs four-variable model
Four-variable model	1,513 / 300	0.721 (0.689–0.754)	0.738	Reference
FIB-4	1,513 / 300	0.578 (0.542–0.615)	0.561	< 0.001
SAFE	1,513 / 300	0.665 (0.630–0.699)	0.660	0.002
MAF-5	1,513 / 300	0.733 (0.700–0.765)	0.734	0.302

Table 3
C. NFS fasting common subset: n = 790; 146 events

Strategy	n / events	Conventional AUC (95% CI)	Survey-weighted AUC	Survey-weighted Brier score	P vs four-variable model
Four-variable model	790 / 146	0.718 (0.669–0.768)	0.719	0.120	Reference
NFS	790 / 146	0.675 (0.626–0.724)	0.696	0.128	0.109

Note. In the full temporal-validation cohort, the four-variable model had a survey-weighted calibration intercept of 0.213 (95% CI 0.073–0.354) and calibration slope of 1.014 (95% CI 0.790–1.238). Reclassification relative to FIB-4 yielded an IDI of 0.107, continuous NRI of 0.527, and categorical NRI of 0.198 (event component, 0.268; nonevent component, − 0.070). SAFE and MAF-5 comparisons were restricted to participants with complete data for all comparator strategies. NFS was evaluated separately in its fasting-weighted common subset. FAST was treated as a VCTE-dependent reference benchmark.

3.4. Performance Within the FIB-4 Gray Zone

Within the prespecified gray zone (n = 420; 100 events), the model AUC was 0.697 (95% CI, 0.639–0.756) versus 0.542 (95% CI, 0.479–0.606) for FIB-4 (P < 0.001; Table 4 and Fig. 2B). In the gray-zone comparator subset (n = 403; 94 events), AUCs were 0.686 for the model, 0.544 for FIB-4 (P < 0.001), 0.630 for SAFE (P = 0.046), and 0.676 for MAF-5 (P = 0.662) (Fig. 2D); the model also outperformed NFS (0.689 vs 0.578; P = 0.010). FAST had higher discrimination (AUC, 0.831) but was not an independent serum-based strategy. Brier score was lower than FIB-4 (0.165 vs 0.180), and calibration and decision-curve analyses supported clinical utility (Fig. 3B–C).

Table 4. **Model performance within the standard FIB-4 gray zone (1.30–2.67)**

Table 4
A. Primary gray-zone cohort: n = 420; 100 events

Strategy	n / events	Conventional AUC (95% CI)	Survey-weighted AUC	Survey-weighted Brier score	P vs four-variable model
Four-variable model	420 / 100	0.697 (0.639–0.756)	0.707	0.165	Reference
FIB-4	420 / 100	0.542 (0.479–0.606)	0.569	0.180	< 0.001
FAST (VCTE-dependent reference)	420 / 100	0.831 (0.785–0.877)	0.834	—	< 0.001

Table 4
B. SAFE/MAF-5 common subset: n = 403; 94 events

Strategy	n / events	Conventional AUC (95% CI)	Survey-weighted AUC	P vs four-variable model
Four-variable model	403 / 94	0.686 (0.625–0.746)	0.690	Reference
FIB-4	403 / 94	0.544 (0.479–0.610)	0.573	< 0.001
SAFE	403 / 94	0.630 (0.565–0.695)	0.644	0.046
MAF-5	403 / 94	0.676 (0.614–0.739)	0.673	0.662

Table 4
C. NFS gray-zone fasting subset: n = 224; 54 events

Strategy	n / events	Conventional AUC (95% CI)	Survey-weighted AUC	P vs four-variable model
Four-variable model	224 / 54	0.689 (0.607–0.772)	0.706	Reference
NFS	224 / 54	0.578 (0.492–0.665)	0.613	0.010

Note. The primary gray-zone analysis included all 420 eligible participants. SAFE/MAF-5 comparisons used the all-comparator common subset (n = 403). NFS was evaluated separately in the fasting common subset (n = 224). FAST was treated as a VCTE-dependent reference benchmark rather than an independent second-step serum strategy.

3.5. Incremental Reclassification

Participant-level reclassification analyses were supportive: compared with development-mapped FIB-4, the model yielded an IDI of 0.107, continuous NRI of 0.527, and categorical NRI of 0.198, driven by improved event classification (event NRI, 0.268) despite a modest unfavourable non-event shift (–0.070) (Table 3). These metrics were interpreted as complementary to discrimination and calibration.

3.6. Exploratory Dual-cutoff Stratification

Development-derived dual thresholds (0.0923 and 0.2623) met prespecified ≥ 90% sensitivity/specificity targets, leaving 53.8% of the weighted gray-zone population intermediate. Transported unchanged to validation, 109/420 participants were lower priority, 233 intermediate, and 78 higher priority (26.6%, 55.0%, 18.4% of weighted population; Table 5A–B). Weighted LSM ≥ 8.0 kPa prevalence increased from 8.2% (95% CI, 5.2–12.6) to 23.8% (95% CI, 19.4–28.8) to 45.8% (95% CI, 33.7–58.3); the corresponding FIB-4 gradient and model-derived strata are shown in Fig. 4A–C. At the lower threshold, sensitivity was 90.8% (95% CI, 86.0–94.1) and NPV 91.8% (95% CI, 87.4–94.8); at the upper threshold, specificity was 86.9% (95% CI, 80.8–91.3) and PPV 45.8% (95% CI, 33.7–58.3). The 85%/85% variant reduced the intermediate group to 36.3% with sensitivity 80.7% (95% CI, 69.8–88.4) and specificity 80.0% (95% CI, 72.7–85.8) (Supplementary Table S9). A Shiny application implements the frozen model and transported thresholds (Supplementary Figure S11).

Table 5

A. Validation-cohort distribution across development-derived probability priority strata

Priority group	Predicted-probability range	n	Events, n	Survey-weighted population share, % (95% CI)	Survey-weighted LSM ≥ 8.0 kPa prevalence, % (95% CI)
Lower-priority	< 0.0923	109	12	26.6 (21.4–32.5)	8.2 (5.2–12.6)
Intermediate	0.0923–0.2623	233	52	55.0 (48.7–61.1)	23.8 (19.4–28.8)
Higher-priority	> 0.2623	78	36	18.4 (14.1–23.7)	45.8 (33.7–58.3)

Note. Cutoffs were selected exclusively in the development-cohort FIB-4 gray zone and transported unchanged to temporal validation. Priority groups represent relative prioritization for subsequent fibrosis assessment and should not be interpreted as validated rule-out or rule-in categories.

Table 5

B. Transported operating characteristics of the 90%/90% development-derived dual-cutoff strategy

Metric	Development cohort	Temporal validation cohort, estimate (95% CI)
Lower predicted-probability cutoff	0.0923	0.0923
Upper predicted-probability cutoff	0.2623	0.2623
Sensitivity at lower cutoff, %	91.8	90.8 (86.0–94.1)
Specificity at upper cutoff, %	90.3	86.9 (80.8–91.3)
NPV of lower-priority group, %	95.4	91.8 (87.4–94.8)
PPV of higher-priority group, %	46.5	45.8 (33.7–58.3)
Intermediate population share, %	53.8	55.0 (48.7–61.1)

Note. The lower cutoff was selected to achieve survey-weighted sensitivity ≥ 90%, and the upper cutoff was selected to achieve survey-weighted specificity ≥ 90% in the development cohort while minimizing the intermediate group. Neither cutoff was re-optimized in temporal validation.

3.7. Prioritization Under Limited VCTE Capacity

In matched-sensitivity benchmarking within the gray-zone comparator subset, the model referred 60.3% for VCTE at 80% sensitivity versus 69.0% for FIB-4 (8.7-point reduction); at 90% sensitivity, referral was 69.0% versus 86.6% (17.5-point reduction) (Supplementary Table S7; Fig. 5A). Under fixed VCTE capacity of 20–50%, model-based prioritization captured 36.3–70.5% of elevated-liver-stiffness events versus 22.1–61.2% for FIB-4 (absolute gains 9.3–14.2 points), similar to MAF-5 and, at several operating points,

SAFE (Supplementary Table S8; Fig. 5B). These analyses were exploratory and not used to validate clinical cutoffs.

3.8. Subgroup Analyses

The model outperformed FIB-4 in most prespecified subgroups (all P ≤ 0.012): age < 60 (0.737 vs 0.556), age ≥ 60 (0.706 vs 0.607), women (0.719 vs 0.583), men (0.726 vs 0.578), with diabetes (0.686 vs 0.570), without diabetes (0.720 vs 0.571), and BMI ≥ 30 (0.693 vs 0.589) (Fig. 6 and Supplementary Table S4). In BMI < 30, discrimination was similar (0.647 vs 0.628; P = 0.632). Survey-weighted calibration plots by BMI category are shown in Supplementary Figure S5. These were descriptive assessments; no formal interaction tests were performed.

3.9. Sensitivity and Stability Analyses

Main findings were robust. Excluding FIB-4 > 2.67 (n = 1,496; 292 events) yielded AUC 0.718 versus 0.563 (Supplementary Table S1). Across LSM thresholds of 6.0, 10.0, and 13.0 kPa, model AUCs were 0.707–0.790 versus 0.545–0.648 for FIB-4 (all P < 0.001; Supplementary Table S3A and Supplementary Figure S6). Alternative gray-zone definitions (0.90–2.67; 1.30–2.00; 1.00–2.00) gave model AUCs 0.668–0.727 versus 0.549–0.552 (all P ≤ 0.013; Supplementary Table S2); the age-adjusted gray zone and restriction to ≥ 35 years also favored the model (0.736 vs 0.543; 0.715 vs 0.570) (Supplementary Table S3B–C and Supplementary Figure S7). Supplementary LASSO did not reproduce the prespecified predictor set (ALT not retained at lambda.min; GGT alone at lambda.1se) and was not used to alter the frozen model (Supplementary Table S10 and Supplementary Figure S8A–B). The lambda.1se solution (AUC 0.718) and a three-variable model without GGT (AUC 0.713) performed similarly to the primary model (0.722); the comparison with the three-variable model reached nominal significance (P = 0.049; Supplementary Table S13 and Supplementary Table S5). A conditional GGT prediction profile is shown in Supplementary Figure S9. Multiple-imputation coefficients were similar to complete-case estimates (Supplementary Table S11), and PSU-grouped fivefold cross-validation showed stable discrimination (fold AUCs 0.732–0.818; Supplementary Table S6 and Supplementary Figure S10).

4. Discussion

In this nationally representative study with temporal validation, we evaluated a simple four-variable second-step risk-stratification model designed to address the uncertainty that persists when initial FIB-4 assessment in adults with MASLD remains indeterminate. The model, based on age, BMI, ALT, and GGT, improved discrimination relative to FIB-4 alone in both the full temporal-validation cohort (AUC, 0.722 vs 0.581) and, more importantly, within the FIB-4 gray zone (AUC, 0.697 vs 0.542), where clinical decision-making is most challenging. Its calibration slope was close to 1, and comparisons with contemporary scores showed better discrimination than SAFE and comparable discrimination to MAF-5, while NFS was lower in its smaller fasting subset and FAST remained a VCTE-dependent benchmark. The central

contribution of this study, however, is not to replace established fibrosis assessment paths but to offer an accessible second-step prioritization when first-line FIB-4 has not provided a definitive answer.

4.1. Principal Findings and Clinical Implications

The clinical relevance of this study derives from the persistent uncertainty of intermediate FIB-4 values. In temporal validation, approximately one-fifth of the weighted MASLD population fell within the conventional gray zone, and the weighted prevalence of LSM ≥ 8.0 kPa in that group was 23.7%. Although current guidance recommends further evaluation with elastography or other second-line tests after an indeterminate result [4–7], access to VCTE is highly variable, and confirmatory testing is often not performed in routine practice [8, 9]. A strategy based on variables already available in the same consultation may therefore provide useful risk separation when immediate specialized testing is unavailable.

Most importantly, the model restored meaningful discrimination specifically within the FIB-4 gray zone, where FIB-4 alone was almost non-informative. This supports the premise that the model adds information at precisely the point where first-line screening has become inconclusive rather than reproducing the existing FIB-4 classification.

The dual-cutoff analysis further illustrates the potential application of this approach. Development-derived thresholds (0.0923 and 0.2623), transported unchanged to validation, divided the gray-zone population into strata with weighted LSM ≥ 8.0 kPa prevalence of 8.2%, 23.8%, and 45.8%. The lower threshold retained 90.8% sensitivity and an NPV of 91.8%, whereas the upper threshold achieved 86.9% specificity with a PPV of 45.8%. Nevertheless, 55% of participants remained in the intermediate group, and specificity at the upper threshold did not fully retain the prespecified development target. These thresholds should therefore be used for relative prioritization, not as definitive rule-out or rule-in criteria.

The resource-allocation analyses provide a further perspective. At a matched sensitivity of approximately 90%, the model required referral of 69.0% of the gray-zone comparator subset compared with 86.6% for FIB-4, an absolute reduction of 17.5 percentage points. When VCTE capacity was limited to 20%, model-based ranking captured 36.3% of events versus 22.1% for FIB-4. These analyses are exploratory but illustrate how a simple ranking could improve allocation of constrained elastography resources.

4.2. Comparison With Existing Fibrosis Assessment Strategies

The main practical advantage of the present model is its simplicity. All four variables are routinely recorded and require no additional equipment or patient visits. Unlike elastography-based strategies, it can be applied in primary care or resource-limited settings where VCTE is not immediately available. Compared with SAFE and MAF-5, the model performed better than SAFE and similarly to MAF-5 in

common subsets. The potential distinction of this approach lies in its positioning specifically after an indeterminate FIB-4 result and in its deliberately minimal data requirements.

NFS was assessed only in a smaller fasting-weighted subset and showed lower discrimination than the model; this comparison is constrained by the reduced sample. FAST had higher discrimination because it incorporates LSM directly, but it answers a different clinical question and was treated as a reference benchmark rather than a competing serum-based strategy.

4.3. Calibration, Predictor Stability, and Subgroup Considerations

Adequate calibration is necessary for risk-based prioritization. The calibration slope of 1.014 indicates preserved spread of predicted risk, whereas the positive intercept of 0.213 indicates modest average underestimation of absolute risk. The lower Brier score compared with FIB-4 supports better overall prediction. As with any model derived in NHANES, recalibration may be needed before the probability estimates are used in other populations with different prevalence or clinical composition.

The incremental contribution of GGT was modest: excluding it decreased the temporal-validation AUC from 0.722 to 0.713. The supplementary LASSO did not reproduce the four-variable set, emphasizing that the model was clinically prespecified rather than data-driven. Sensitivity analyses were consistent across alternative LSM thresholds, gray-zone definitions, age adjustment, and age restriction. Descriptively, the model outperformed FIB-4 in most subgroups, but the gain was attenuated among participants with BMI < 30 kg/m²; this leaner phenotype should be examined further.

4.4. Implementation, Strengths, and Limitations

For implementation, predicted probabilities can be generated automatically in electronic health records or laboratory systems. We provide a Shiny application that implements the frozen model and thresholds to facilitate reproducible application, but it is a research tool, not a diagnostic system. Patients with higher probabilities could be prioritized for earlier VCTE or specialist assessment.

The main strengths are the nationally representative data, temporal separation between development and validation, focus on the clinically relevant gray zone, and evaluation that went beyond discrimination to include calibration, decision curves, reclassification, subgroups, and multiple sensitivity analyses. Important limitations include the use of VCTE-defined stiffness rather than histology, the NHANES origin of both cohorts requiring external validation, complete-case primary analysis, the smaller NFS subset, and the cross-sectional design that precludes conclusions about progression or long-term outcomes.

4.5. Conclusions and Future Directions

In summary, this four-variable model offers an accessible second-step prioritization strategy among patients with MASLD and indeterminate FIB-4 values. It does not replace VCTE or specialist evaluation, but it may help direct limited resources toward those at highest risk. Future research should externally validate the model and its thresholds in diverse primary-care cohorts, compare it prospectively with

SAFE and MAF-5 in identical clinical workflows, and assess whether implementation reduces unnecessary testing and improves clinically meaningful outcomes.

5. Conclusions

In conclusion, a simple four-variable second-step risk-stratification model based on age, BMI, ALT, and GGT provided additional risk discrimination beyond FIB-4, particularly among adults with MASLD in the FIB-4 gray zone. The model showed reasonable temporal calibration and maintained informative risk separation across multiple sensitivity analyses. Its performance was similar to MAF-5 in direct head-to-head comparisons, suggesting that its potential value lies primarily in simplicity, availability, and its specific use after an indeterminate FIB-4 result rather than in universal superiority over existing fibrosis scores.

Development-derived probability strata further identified clinically distinct levels of elevated liver stiffness risk, while resource-allocation analyses suggested that model-based ranking may reduce VCTE referral burden or improve event capture when elastography capacity is constrained. However, the proposed thresholds remain exploratory, and independent external and prospective validation is required before clinical implementation.

Declarations

Funding

This work was financially supported by Shenzhen Medical Health "Three Famous" Project （SZZYSM202411013）; Shenzhen Science and Technology Program (ZDSYS20210623092000002)

Data Availability Statement

The NHANES data used in this study are publicly available through the Centers for Disease Control and Prevention (CDC) website. Analysis code is available from the corresponding author upon reasonable request.

Ethics approval

The NHANES data collection protocols were approved by the NCHS Research Ethics Review Board (ERB) of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Consent to participate

Written informed consent was obtained from all NHANES participants at enrollment. This study is a secondary analysis of de-identified, publicly available data; therefore, additional informed consent for this specific analysis was not required, and the study was deemed exempt from further ethics review.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors thank the National Center for Health Statistics and all NHANES participants. This study used publicly available data from the National Health and Nutrition Examination Survey.

References

1. Miao L, Targher G, Byrne CD, Cao YY, Zheng MH (2024) Current status and future trends of the global burden of MASLD. Trends Endocrinol Metab 35(8):697–707. 10.1016/j.tem.2024.02.007. Epub 2024 Feb 29. PMID: 38429161
2. Sanyal AJ, Van Natta ML, Clark J, Neuschwander-Tetri BA, Diehl A, Dasarathy S, Loomba R, Chalasani N, Kowdley K, Hameed B, Wilson LA, Yates KP, Belt P, Lazo M, Kleiner DE, Behling C, Tonascia J, NASH Clinical Research Network (CRN) (2021) Prospective Study of Outcomes in Adults with Nonalcoholic Fatty Liver Disease. N Engl J Med 385(17):1559–1569. 10.1056/NEJMoa2029349. PMID: 34670043; PMCID: PMC8881985
3. Dulai PS, Singh S, Patel J, Soni M, Prokop LJ, Younossi Z, Sebastiani G, Ekstedt M, Hagstrom H, Nasr P, Stal P, Wong VW, Kechagias S, Hultcrantz R, Loomba R (2017) Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: Systematic review and meta-analysis. Hepatology 65(5):1557–1565 Epub 2017 Mar 31. PMID: 28130788; PMCID: PMC5397356
4. European Association for the Study of the Liver (EASL); European Association for the Study of Diabetes (EASD); European Association for the Study of Obesity (EASO) (2024) EASL-EASD-EASO Clinical Practice Guidelines on the management of metabolic dysfunction-associated steatotic liver disease (MASLD). J Hepatol 81(3):492–542 Epub 2024 Jun 7. PMID: 38851997
5. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, Abdelmalek MF, Caldwell S, Barb D, Kleiner DE, Loomba R (2023) AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 77(5):1797–1835. 10.1097/HEP.0000000000000323. Epub 2023 Mar 17. PMID: 36727674; PMCID: PMC10735173
6. Kanwal F, Shubrook JH, Adams LA, Pfotenhauer K, Wai-Sun Wong V, Wright E, Abdelmalek MF, Harrison SA, Loomba R, Mantzoros CS, Bugianesi E, Eckel RH, Kaplan LM, El-Serag HB, Cusi K (2021) Clinical Care Pathway for the Risk Stratification and Management of Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology 161(5):1657–1669 Epub 2021 Sep 20. PMID: 34602251; PMCID: PMC8819923
7. European Association for the Study of the Liver (2021) EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis – 2021 update. J Hepatol 75(3):659–689. 10.1016/j.jhep.2021.05.025. Epub 2021 Jun 21. PMID: 34166721
8. Moore JA, Wheless WH, Zhang J, Marsden J, Mauldin PD, Moran WP, Schreiner AD (2023) Gaps in Confirmatory Fibrosis Risk Assessment in Primary Care Patients with Nonalcoholic Fatty Liver

Disease. Dig Dis Sci 68(7):2946–2953. 10.1007/s10620-023-07959-5. Epub 2023 May 16. PMID: 37193930; PMCID: PMC10659111

9. Davyduke T, Tandon P, Al-Karaghouli M, Abraldes JG, Ma MM (2019) Impact of Implementing a FIB-4 First Strategy on a Pathway for Patients With NAFLD Referred From Primary Care. Hepatol Commun 3(10):1322–1333. 10.1002/hep4.1411. PMID: 31592044; PMCID: PMC6771169
10. Sripongpun P, Kim WR, Mannalithara A, Charu V, Vidovszky A, Asch S, Desai M, Kim SH, Kwong AJ (2023) The steatosis-associated fibrosis estimator (SAFE) score: A tool to detect low-risk NAFLD in primary care. Hepatology 77(1):256–267 Epub 2022 May 24. PMID: 35477908; PMCID: PMC9613815
11. van Kleef LA, Francque SM, Prieto-Ortiz JE, Sonneveld MJ, Sanchez-Luque CB, Prieto-Ortiz RG, Kwanten WJ, Vonghia L, Verrijken A, De Block C, Gadi Z, Janssen HLA, de Knegt RJ, Brouwer WP (2024) Metabolic Dysfunction-Associated Fibrosis 5 (MAF-5) Score Predicts Liver Fibrosis Risk and Outcome in the General Population With Metabolic Dysfunction. Gastroenterology 167(2):357–367e9. 10.1053/j.gastro.2024.03.017. Epub 2024 Mar 19. PMID: 38513745
12. Li M, Lin Y, Yu H, Lin W, Chen J, Yang Y, Wu B (2024 Sep-Oct) The steatosis-associated fibrosis estimator (SAFE) outperformed the FIB-4 score in screening the population for liver disease. Ann Hepatol 29(5):101516. 10.1016/j.aohep.2024.101516. Epub 2024 Jun 6. PMID: 38851395
13. Akinbami LJ, Chen TC, Davy O, Ogden CL, Fink S, Clark J, Riddles MK, Mohadjer LK (2022) National Health and Nutrition Examination Survey, 2017-March 2020 Prepandemic File: Sample Design, Estimation, and Analytic Guidelines. Vital Health Stat 1(190):1–36 PMID: 35593699
14. National Center for Health Statistics Brief Overview of Sample Design, Nonresponse Bias Assessment, and Analytic Guidelines for NHANES August 2021–August 2023. Centers for Disease Control and Prevention
15. National Center for Health Statistics Ethics Review Board Approval. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention
16. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. ;385:e078378. 10.1136/bmj-2023-078378. Erratum in: BMJ. 2024;385:q902. doi: 10.1136/bmj.q902. PMID: 38626948; PMCID: PMC11019967
17. van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 45(3):1–67. https://doi.org/10.18637/jss.v045.i03
18. Rinella ME, Lazarus JV, Ratziu V, Francque SM, Sanyal AJ, Kanwal F, Romero D, Abdelmalek MF, Anstee QM, Arab JP, Arrese M, Bataller R, Beuers U, Boursier J, Bugianesi E, Byrne CD, Narro GEC, Chowdhury A, Cortez-Pinto H, Cryer DR, Cusi K, El-Kassas M, Klein S, Eskridge W, Fan J, Gawrieh S, Guy CD, Harrison SA, Kim SU, Koot BG, Korenjak M, Kowdley KV, Lacaille F, Loomba R, Mitchell-Thain

R, Morgan TR, Powell EE, Roden M, Romero-Gómez M, Silva M, Singh SP, Sookoian SC, Spearman CW, Tiniakos D, Valenti L, Vos MB, Wong VW, Xanthakos S, Yilmaz Y, Younossi Z, Hobbs A, Villota-Rivas M, Newsome PN (2024 Jan-Feb) NAFLD Nomenclature consensus group. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. Ann Hepatol 29(1):101133. 10.1016/j.aohep.2023.101133. Epub 2023 Jun 24. PMID: 37364816

19. National Center for Health Statistics NHANES 2017–March 2020 Pre-Pandemic: Liver Ultrasound Transient Elastography (P_LUX). Centers for Disease Control and Prevention
20. National Center for Health Statistics NHANES August 2021–August 2023: Liver Ultrasound Transient Elastography (LUX_L). Centers for Disease Control and Prevention
21. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33(1):1–22 PMID: 20808728; PMCID: PMC2929880
22. Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, Sulkowski S, Torriani M, Dieterich FJ, Thomas DT, Messinger DL, Nelson D (2006) M; APRICOT Clinical Investigators. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology. ;43(6):1317-25. 10.1002/hep.21178. PMID: 16729309
23. Angulo P, Hui JM, Marchesini G, Bugianesi E, George J, Farrell GC, Enders F, Saksena S, Burt AD, Bida JP, Lindor K, Sanderson SO, Lenzi M, Adams LA, Kench J, Therneau TM, Day CP (2007) The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology. ;45(4):846 – 54. 10.1002/hep.21496. PMID: 17393509
24. Newsome PN, Sasso M, Deeks JJ, Paredes A, Boursier J, Chan WK, Yilmaz Y, Czernichow S, Zheng MH, Wong VW, Allison M, Tsochatzis E, Anstee QM, Sheridan DA, Eddowes PJ, Guha IN, Cobbold JF, Paradis V, Bedossa P, Miette V, Fournier-Poizat C, Sandrin L, Harrison SA (2020) 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. ;5(4):362–373. 10.1016/S2468-1253(19)30383-8. Epub 2020 Feb 3. Erratum in: Lancet Gastroenterol Hepatol. 2020;5(4):e3. doi: 10.1016/S2468-1253(20)30055-8. PMID: 32027858; PMCID: PMC7066580
25. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837–845 PMID: 3203132
26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 74:167–176 Epub 2016 Jan 6. PMID: 26772608
27. Pencina MJ, D'Agostino RB, Sr D, Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. ;27(2):157 – 72; discussion 207 – 12. 10.1002/sim.2929. PMID: 17569110
28. Vickers AJ, Elkin EB (2006 Nov-Dec) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak 26(6):565–574. 10.1177/0272989X06295361. PMID: 17099194;

Figures

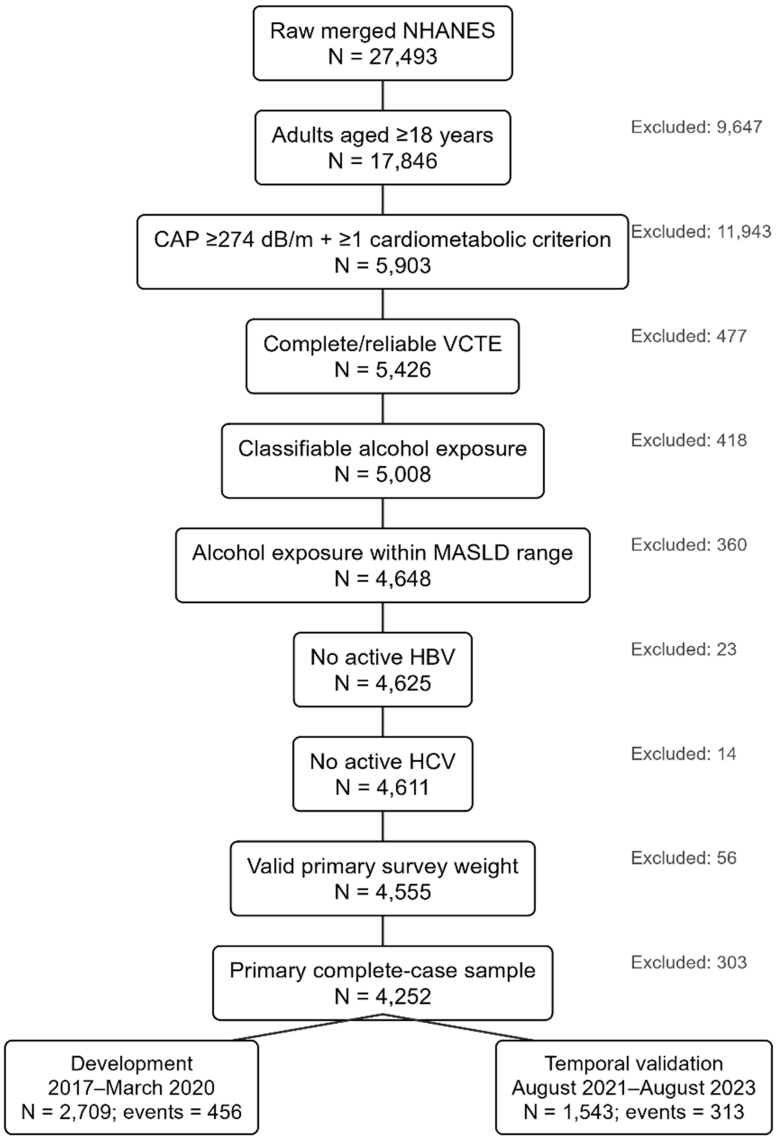


Figure 1

Participant flow and construction of the development and temporal-validation cohorts. Participants were sequentially restricted to adults aged ≥18 years, those meeting the operational MASLD definition based on CAP ≥274 dB/m plus at least one cardiometabolic criterion, those with complete and reliable VCTE,

classifiable alcohol exposure within the MASLD range, no active hepatitis B or C, a valid primary survey weight, and complete data for the primary analysis. The final primary complete-case sample included 4,252 participants, comprising 2,709 participants (456 events) in the 2017–March 2020 development cohort and 1,543 participants (313 events) in the August 2021–August 2023 temporal-validation cohort. CAP, controlled attenuation parameter; HBV, hepatitis B virus; HCV, hepatitis C virus; MASLD, metabolic dysfunction-associated steatotic liver disease; VCTE, vibration-controlled transient elastography.

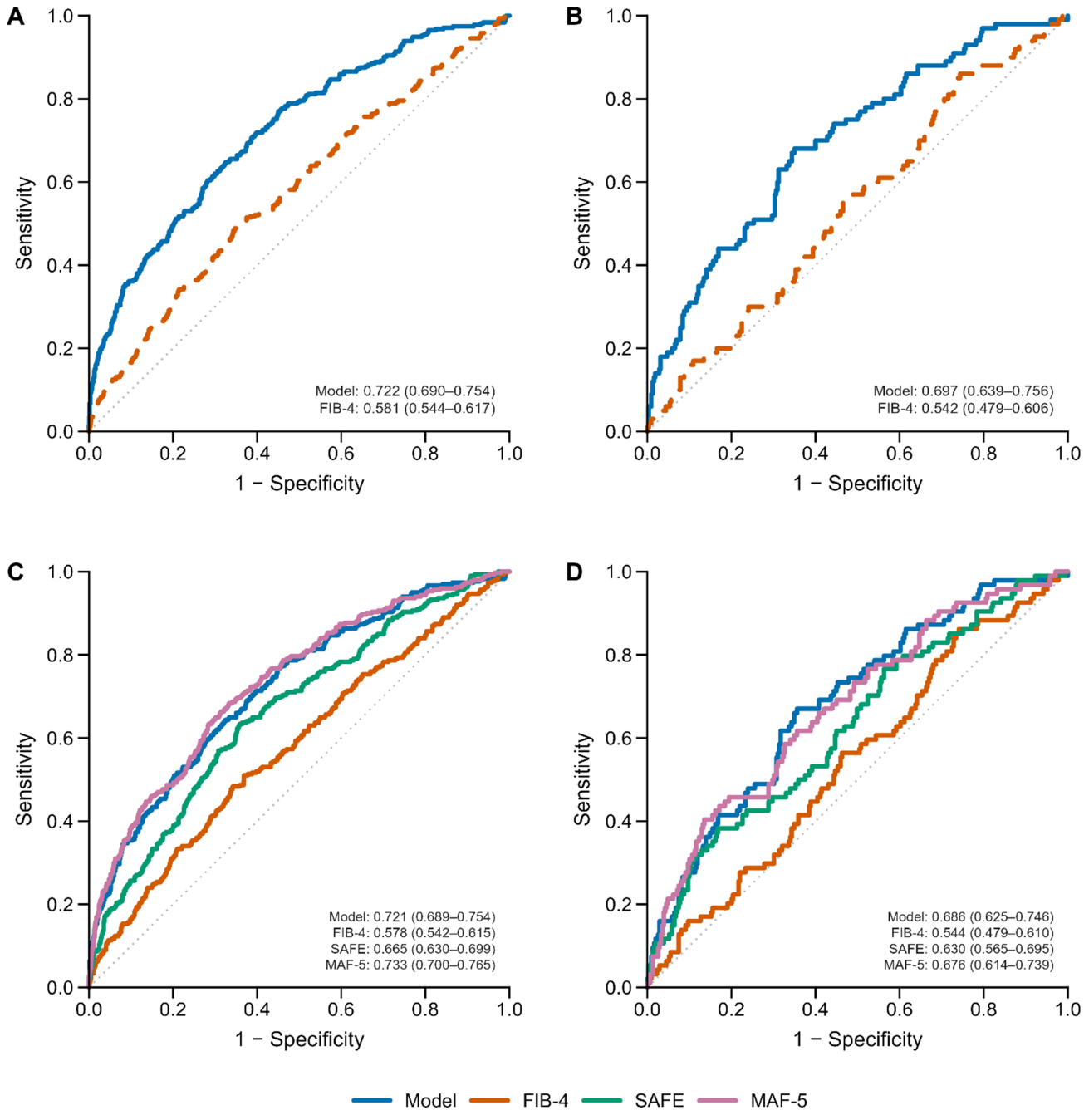


Figure 2

Discrimination of the four-variable model and comparator scores in temporal validation. (A) Receiver operating characteristic (ROC) curves for the four-variable model and FIB-4 in the full temporal-validation cohort (n=1,543; 313 LSM ≥8.0-kPa events). (B) ROC curves for the four-variable model and FIB-4 within the standard FIB-4 gray zone (1.30–2.67; n=420; 100 events). (C) Head-to-head ROC comparison of the four-variable model, FIB-4, SAFE, and MAF-5 in the all-comparator common subset of the temporal-validation cohort (n=1,513; 300 events). (D) Corresponding head-to-head comparison within the standard FIB-4 gray-zone common subset (n=403; 94 events). Conventional AUCs with 95% confidence intervals are displayed in each panel; survey-weighted AUC estimates are reported in Tables 3 and 4. AUC, area under the receiver operating characteristic curve; LSM, liver stiffness measurement; MAF-5, metabolic dysfunction-associated fibrosis-5; SAFE, steatosis-associated fibrosis estimator.

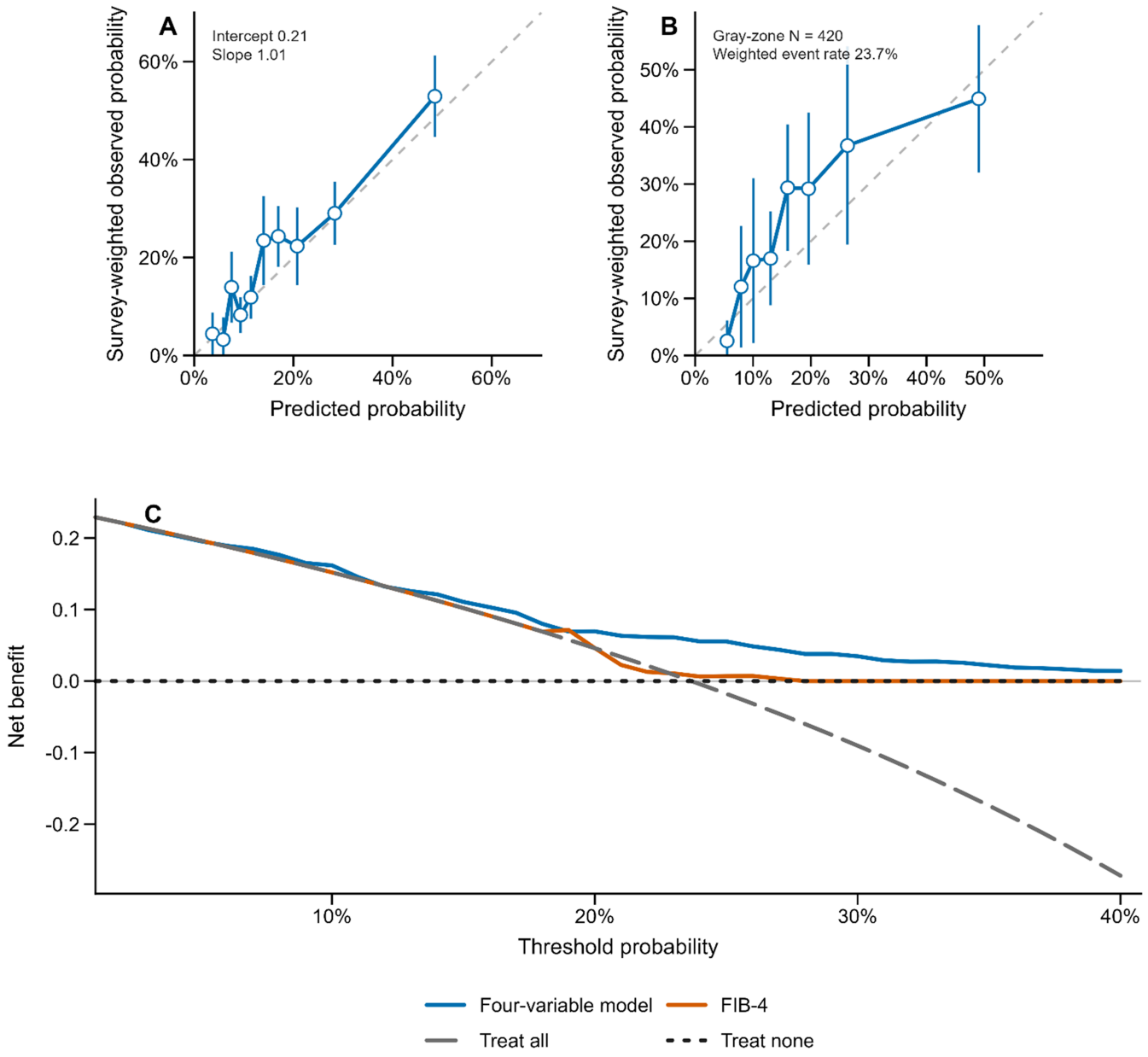


Figure 3

Calibration and clinical utility of the four-variable model in temporal validation. (A) Survey-weighted calibration in the full temporal-validation cohort (n=1,543), with observed outcome probabilities plotted against mean predicted probabilities across predicted-risk groups; vertical bars indicate 95% confidence intervals and the diagonal line indicates perfect calibration. The survey-weighted calibration intercept and slope were 0.21 and 1.01, respectively. (B) Corresponding survey-weighted calibration within the standard FIB-4 gray zone (n=420; weighted event prevalence 23.7%). (C) Survey-weighted decision-curve analysis within the standard FIB-4 gray zone, comparing the four-variable model, FIB-4, treat-all, and treat-none strategies across threshold probabilities from 0.01 to 0.40. Net benefit is shown on the vertical axis.

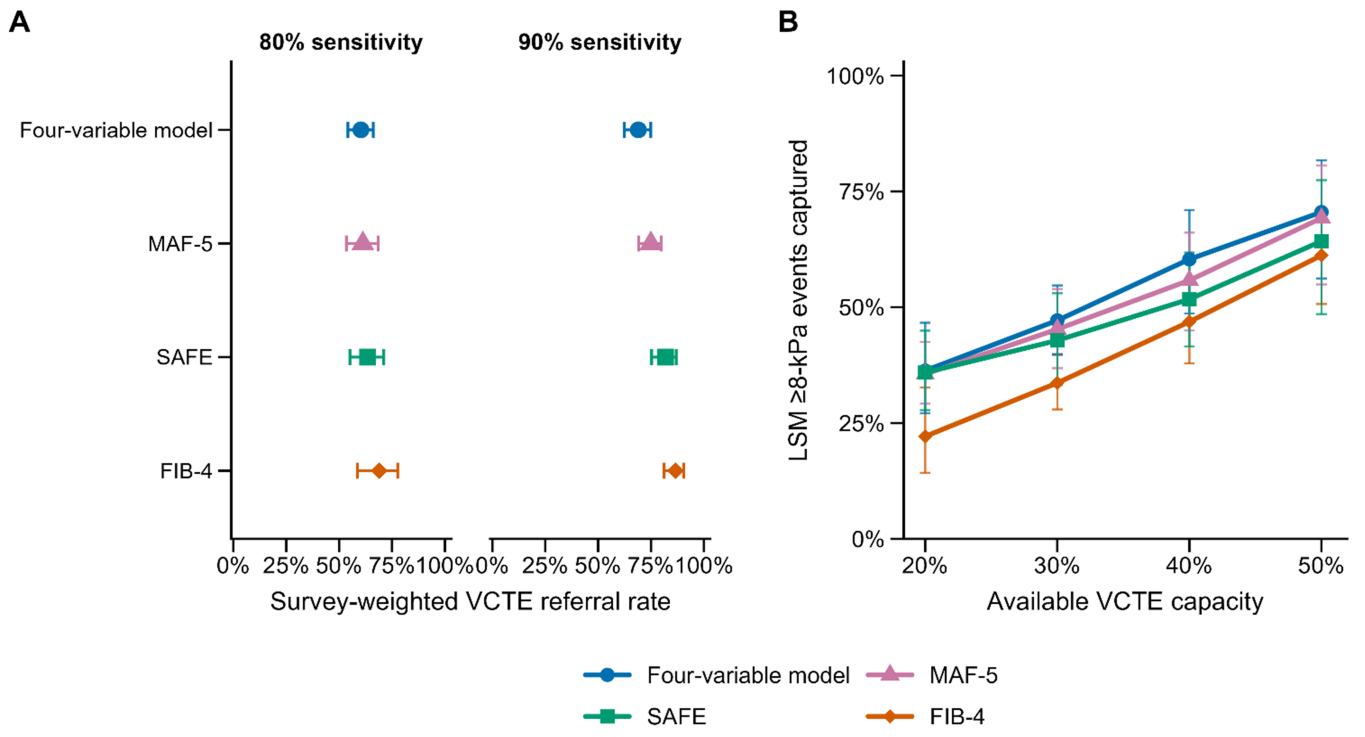


Figure 4

FIB-4 risk gradient and development-derived probability priority strata in temporal validation. (A) Survey-weighted prevalence of LSM ≥8.0 kPa across low (<1.30), gray-zone (1.30–2.67), and high (>2.67) FIB-4 categories in the full temporal-validation cohort; error bars indicate 95% confidence intervals. (B) Survey-weighted LSM ≥8.0-kPa prevalence across lower-priority, intermediate, and higher-priority groups within the standard FIB-4 gray zone; labels indicate the survey-weighted share of the gray-zone population assigned to each group. (C) Survey-weighted distribution of predicted probabilities within the standard FIB-4 gray zone. The development-derived lower and upper cutoffs (0.0923 and 0.2623,

respectively) were transported unchanged to temporal validation and define the three priority strata. These strata represent relative prioritization for subsequent fibrosis assessment and should not be interpreted as validated rule-out or rule-in categories. LSM, liver stiffness measurement.

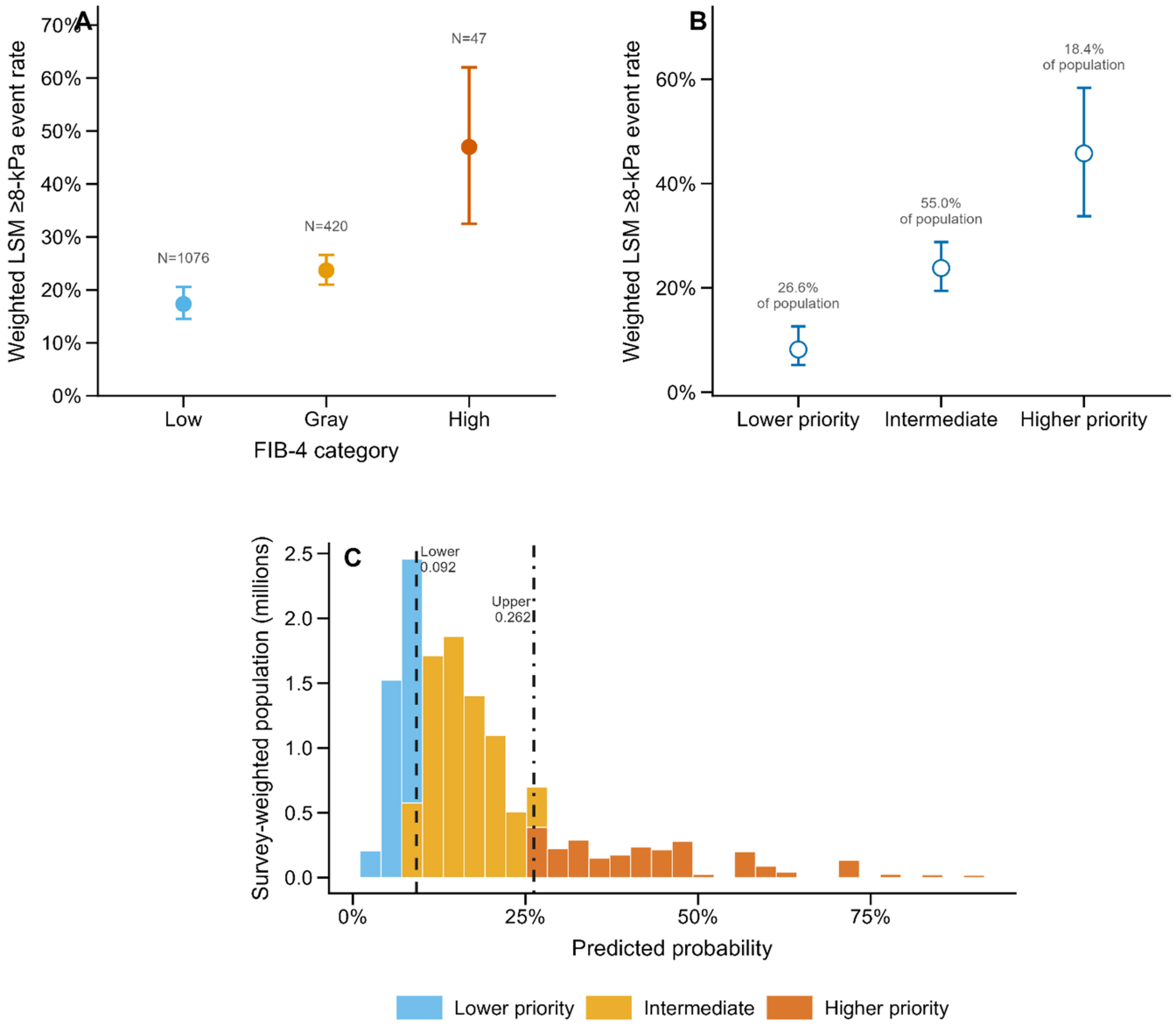


Figure 5

Prioritization of VCTE use within the standard FIB-4 gray zone. Analyses were restricted to the all-comparator common subset (n=403; 94 LSM ≥8.0-kPa events). (A) Survey-weighted VCTE referral rates, with 95% confidence intervals, required by the four-variable model, MAF-5, SAFE, and FIB-4 at matched sensitivity targets of approximately 80% and 90%. Thresholds were selected within the temporal-validation cohort only to create common benchmarking operating points and are not proposed as validated clinical cutoffs. (B) Survey-weighted proportion of LSM ≥8.0-kPa events captured when

available VCTE capacity was restricted to 20%, 30%, 40%, or 50% of the gray-zone population; error bars indicate 95% confidence intervals. Participants were ranked by each strategy for the fixed-capacity analysis. MAF-5, metabolic dysfunction-associated fibrosis-5; SAFE, steatosis-associated fibrosis estimator; VCTE, vibration-controlled transient elastography.

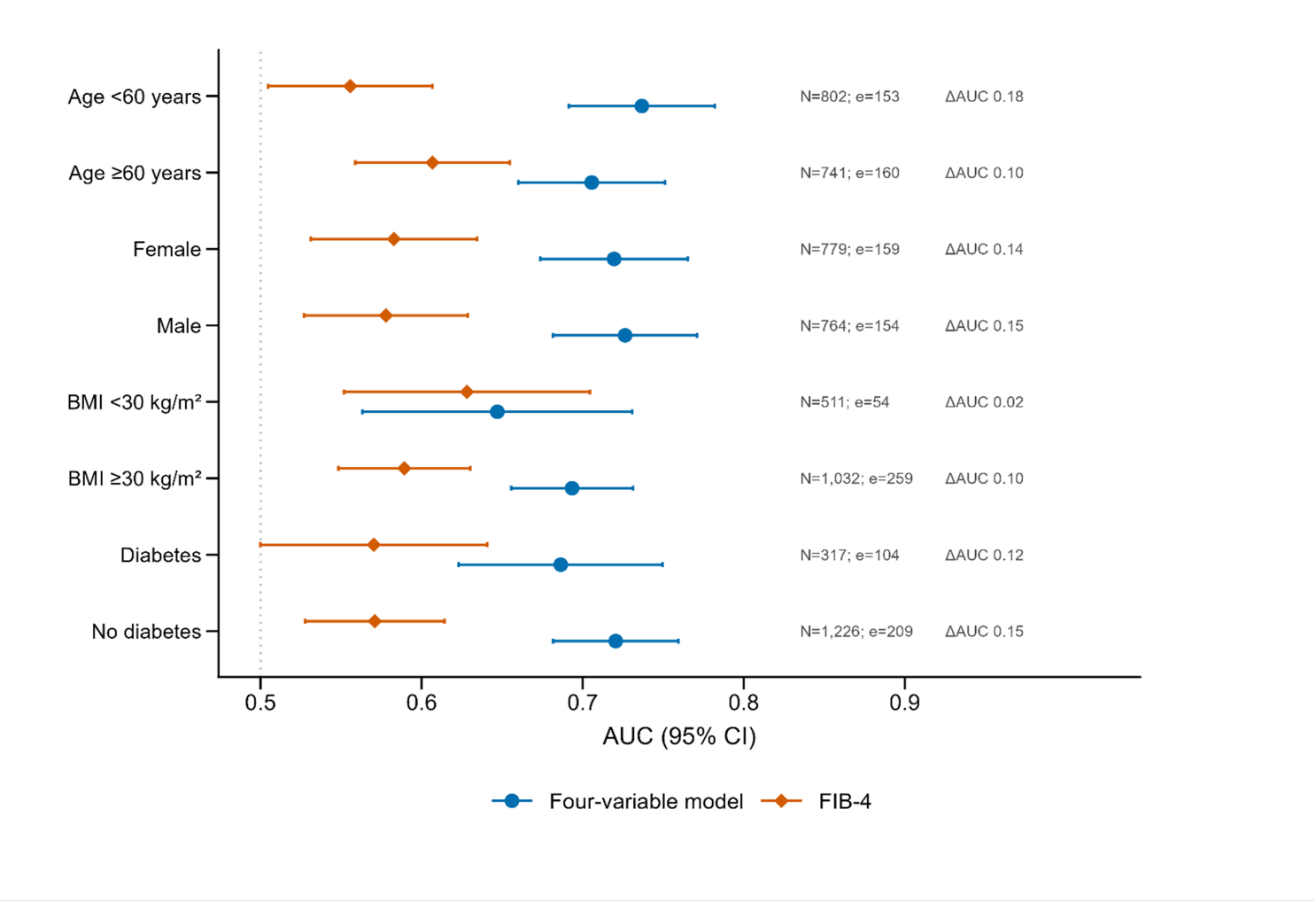


Figure 6

Subgroup discrimination of the four-variable model and FIB-4 in temporal validation. Conventional AUCs with 95% confidence intervals are shown for the four-variable model and FIB-4 according to age (<60 vs ≥60 years), sex, BMI (<30 vs ≥30 kg/m²), and physician-diagnosed diabetes status. The right-hand annotations show subgroup sample size (N), number of LSM ≥8.0-kPa events (e), and the absolute AUC difference (ΔAUC) between the four-variable model and FIB-4. Subgroup analyses were prespecified descriptive assessments of consistency and were not formal interaction tests. AUC, area under the receiver operating characteristic curve; BMI, body mass index; LSM, liver stiffness measurement.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx