

Body mass index modifies the diagnostic performance of liver stiffness measurement for histologically significant fibrosis: a multicentre BMI-aware prediction-model study

Le Wang

the Affiliated Huai'an Hospital of Xuzhou Medical University

Lei Ding

Huai'an Fifth People's Hospital

Guangnian Ji

Huaian First People's Hospital

Yujie Gao

the Affiliated Huai'an Hospital of Xuzhou Medical University

Jie Zhang

the Affiliated Huai'an Hospital of Xuzhou Medical University

Chuanrong Zhu

Huaian First People's Hospital

Wan'e Wang

Huai'an Fourth People's Hospital

Min Xu

304501202@qq.com

the Affiliated Huai'an Hospital of Xuzhou Medical University

Research Article

Keywords: MASLD, liver stiffness measurement, liver fibrosis, obesity, prediction model

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11193667/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Body mass index modifies the diagnostic performance of liver stiffness measurement for histologically significant fibrosis: a multicentre BMI-aware prediction-model study

Le Wang^{1*}, Lei Ding^{2*}, Guangnian Ji^{3*}, Yujie Gao⁴, Jie Zhang⁴, Chuanrong Zhu^{3#}, Wan'e Wang^{5#}, Min Xu^{6#}

¹ Department of Ultrasound Medicine, the Affiliated Huai'an Hospital of Xuzhou Medical University, Huai'an, Jiangsu, China

² Department of Endocrinology and Metabolism, Huai'an Fifth People's Hospital, Huai'an, Jiangsu, China

³ Department of Hepatobiliary and Pancreatic Surgery, The Affiliated Huai'an No.1 People's Hospital of Nanjing Medical University, Northern Jiangsu Institute of Clinical Medicine, Huai'an, Jiangsu, China

⁴ Department of Endocrinology and Metabolism, the Affiliated Huai'an Hospital of Xuzhou Medical University, Huai'an, Jiangsu, China

⁵ Department of Hepatology, Huai'an Fourth People's Hospital, Huai'an, Jiangsu, China

⁶ Department of Medical Imaging, the Affiliated Huai'an Hospital of Xuzhou Medical University, Huai'an, Jiangsu, China

*These authors contributed equally.

#Corresponding authors: Min Xu, E-mail: 304501202@qq.com; Wan'e Wang, E-mail: wangwane72@126.com; Chuanrong Zhu, E-mail: zcr33333@126.com

Abstract

Background: Liver stiffness measurement (LSM) is widely used to identify fibrosis in metabolic dysfunction-associated steatotic liver disease (MASLD), but its performance may vary with adiposity. We examined how BMI affects the diagnostic performance of LSM for $F \geq 2$ and evaluated whether BMI-aware prediction models restore performance at high BMI. **Methods:** In this retrospective, multicentre diagnostic-accuracy and prediction-model study, we analysed 520 adults with paired histological fibrosis staging and LSM from three prespecified hospital cohorts. Histological $F \geq 2$ was the primary outcome. We assessed pooled and cohort-stratified discrimination, six prespecified LSM thresholds, BMI-stratified performance and misclassification. For model development, participants were randomly assigned 7:3 to a training set ($n=364$) and an internal-validation set ($n=156$), with outcome stratification and a fixed random seed. Penalized logistic, linear interaction, spline varying-coefficient, BMI-balanced spline and monotonic gradient-boosting models were compared using AUC, calibration, Brier score and decision-curve analysis. **Results:** $F \geq 2$ was present in 255 participants (49.0%). Pooled LSM discrimination was 0.861 (95% CI 0.829-0.891). At 8 kPa, sensitivity was 0.929 and specificity 0.668. BMI-stratified LSM AUC declined from 0.880 (0.808-0.939) below 28 kg/m² to 0.777 (0.659-0.880) at ≥ 42 kg/m². Each 5-kg/m² BMI increase was associated with higher odds of false-negative classification (OR 1.30, 95% CI 0.99-1.72; $P=0.062$), although the estimate was imprecise. In internal validation, AUCs were 0.760 for the clinical model, 0.839 for LSM alone, 0.857 for the reference LSM-clinical model, 0.860 for the linear LSM $\times$ BMI interaction model, 0.855 for the standard spline varying-coefficient model, 0.859 for the BMI-balanced spline model and 0.876 for monotonic gradient boosting. The gradient-boosting model had a large training-validation gap (0.991 versus 0.876). No BMI-aware model consistently restored discrimination in the highest BMI stratum. **Conclusions:** LSM provided

good overall discrimination for histological $F \geq 2$, but performance weakened with increasing BMI. BMI-aware models represented effect heterogeneity and yielded local improvements, yet did not eliminate high-BMI uncertainty. External validation with locked models is required before clinical use.

Keywords: MASLD; liver stiffness measurement; liver fibrosis; obesity; prediction model

Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) is closely linked to obesity, insulin resistance and type 2 diabetes and has become a major cause of chronic liver disease worldwide^[1,2]. Advanced fibrosis is prevalent in metabolically high-risk populations^[3], and because clinical outcomes are strongly related to fibrosis stage, reliable identification of significant fibrosis is central to risk stratification, referral and prognostic assessment^[4,5,6].

Liver biopsy remains the reference standard for fibrosis staging but is invasive, samples only a small tissue volume and is unsuitable for repeated population-level assessment^[2,7]. Liver stiffness measurement (LSM), acquired by vibration-controlled transient elastography, provides rapid and repeatable assessment and is incorporated into current non-invasive testing pathways^[7-12]. A threshold near 8 kPa is often used to identify patients who may require further assessment, but performance depends on disease spectrum, inflammation, technical quality and body habitus^[2,7-12].

Obesity is particularly relevant: increased skin-to-liver distance, subcutaneous adipose tissue, altered mechanical coupling and probe selection may affect the elastography signal^[10,11], while metabolic comorbidity changes the prior probability of fibrosis. A global AUC can therefore conceal clinically important variation across BMI strata. Adding BMI as an ordinary covariate may adjust baseline risk without allowing the diagnostic contribution of LSM itself to change with BMI. Interaction and varying-coefficient models provide a direct way to represent this heterogeneity, but added flexibility increases variance and overfitting risk.

We therefore addressed three questions in a pooled multicentre pathology cohort with histological reference staging. First, how well does LSM discriminate histological $F \geq 2$ overall and across prespecified diagnostic thresholds, cohorts and BMI strata? Second, is higher BMI associated with misclassification, particularly false-negative classifications that carry prognostic risk? Third, do BMI-aware prediction models, including linear interaction, spline varying-coefficient, BMI-balanced and monotonic gradient-boosting approaches, restore diagnostic performance in severe obesity? All models were developed in a stratified training set and evaluated in an internal-validation set with locked fitted parameters, separating local gains from the evidence required to claim transportability.

Methods

Study design and participants

This retrospective, multicentre diagnostic-accuracy and prediction-model study used 520 participants with paired liver histology, LSM (Echosens) and clinical data^[13-15]. The prespecified cohorts were Cohort 1 (the Affiliated Huai'an No.1 People's Hospital of Nanjing Medical University, $n=321$), Cohort 2 (Huai'an Fourth People's Hospital, $n=88$) and Cohort 3 (the Affiliated Huai'an Hospital of Xuzhou Medical University and Huai'an Fifth People's Hospital, $n=111$). Recruitment spanned January 2020 to December 2025 across all three centres. Baseline characteristics are summarized in Table 1.

The following participants were excluded before analysis: those aged <18 years; those with inadequate liver biopsy specimens; those with invalid liver stiffness measurement (LSM) results; those with known coexisting chronic liver disease of another cause (hepatitis B or C, autoimmune hepatitis, Wilson disease, haemochromatosis, alcohol-associated liver disease exceeding 30 g/day, or drug-induced liver injury); and those lacking complete covariate data.. Of 726 potentially eligible records, 206 were excluded for these reasons, leaving 520 participants for the present analysis. A patient-flow diagram is provided in Supplementary Figure S1.

Liver biopsies were scored by 2 experienced pathologists blinded to LSM and clinical data using the Kleiner NAFLD Activity Score^[16] and the METAVIR fibrosis staging system. Discrepant readings were adjudicated by a third senior pathologist. LSM was performed by trained operators using a FibroScan (Echosens) after an overnight fast of at least 4 h. The M or XL probe was selected by experienced ultrasound specialists according to each participant's body habitus, following the manufacturer's recommendations. The median of ten valid acquisitions was used; the controlled attenuation parameter (CAP) was recorded simultaneously with LSM. The study protocol was approved by the ethics committees of these hospitals. All analyses used de-identified data, and each participant contributed one independent observation.

Clinical variables and classical scores

Candidate variables comprised age, sex, BMI, fasting glucose, AST, ALT, platelet count, LSM and routinely derived non-invasive scores. Conventional scores were calculated as follows:

$$\text{FIB-4} = \text{age (years)} \times \text{AST (U/L)} / [\text{platelets (10}^9\text{/L)} \times \sqrt{\text{ALT (U/L)}}];$$

$$\text{APRI} = [(\text{AST/AST upper limit of normal})/\text{platelets (10}^9\text{/L)}] \times 100;$$

$$\text{NFS} = -1.675 + 0.037 \times \text{age} + 0.094 \times \text{BMI} + 1.13 \times (\text{diabetes or impaired fasting glucose}) + 0.99 \times (\text{AST/ALT ratio}) - 0.013 \times \text{platelets (10}^9\text{/L)} - 0.66 \times \text{albumin (g/dL)};$$

$$\text{ZJU} = \text{BMI} + 2 \times \text{fasting glucose (mmol/L)} + 0.66 \times \text{triglycerides (mmol/L)} + 0.02 \times \text{age};$$

$$\text{TyG} = \ln[\text{fasting triglycerides (mg/dL)} \times \text{fasting glucose (mg/dL)} / 2];$$

$$\text{TyG-BMI} = \text{TyG} \times \text{BMI};$$

$$\text{BARD} = \text{BMI} \geq 28 \text{ (1 point)} + \text{AST/ALT ratio} \geq 0.8 \text{ (2 points)} + \text{diabetes (1 point)}.$$

Training and internal validation

The 520 participants were randomly split 7:3 with stratification by $F \geq 2$ status. The training set included 364 participants and 178 events; the internal-validation set included 156 participants and 77 events. The training set provided 178 events for 8 candidate predictors (events per variable ≈ 22), exceeding the minimum sample size recommended for prediction-model development^[17]. All preprocessing, transformations, hyperparameter selection and model fitting used the training data, and fitted parameters were applied unchanged to internal validation.

Prediction models

The clinical model used age/10, sex, BMI/5, diabetes-range hyperglycaemia, $\log_2(\text{AST})$, $\log_2(\text{ALT})$ and $\log_2(\text{platelet count})$. The LSM-only model used $\log_2(\text{LSM})$. The reference LSM-clinical model combined $\log_2(\text{LSM})$ with the clinical variables; these three models used L1-penalized (lasso) logistic regression^[17-20]. The linear interaction model additionally included $\log_2(\text{LSM}) \times \text{BMI}/5$. The standard varying-coefficient model used restricted cubic-spline terms for BMI and their products with $\log_2(\text{LSM})$. The BMI-balanced model used the same spline structure with inverse-frequency weights for the five prespecified BMI strata. The machine-learning sensitivity model used histogram gradient boosting with a monotonic increasing constraint for LSM. The flexible models were treated as sensitivity analyses.

Model equations

For all logistic models, the predicted probability was $p=1/[1+\exp(-\eta)]$. Indicator variables were coded as male=1 versus female=0 and diabetes-range hyperglycaemia=1 versus 0.

The fitted training coefficients were:

clinical model: $-11.572 + 0.050(\text{age}/10) - 0.211(\text{male}) + 0.137(\text{BMI}/5) - 0.217(\text{diabetes}) + 0.669\log_2(\text{AST}) + 0.813\log_2(\text{ALT}) + 0.27\log_2(\text{platelets})$;

LSM-only model: $-8.650 + 2.717\log_2(\text{LSM})$;

reference LSM-clinical model: $-15.548 + 2.349\log_2(\text{LSM}) + 0.037(\text{age}/10) - 0.059(\text{male}) + 0.023(\text{BMI}/5) - 0.319(\text{diabetes}) + 0.648\log_2(\text{AST}) + 0.585\log_2(\text{ALT}) + 0.112\log_2(\text{platelets})$;

linear LSM×BMI interaction model: $-20.563 + 3.541\log_2(\text{LSM}) + 0.517(\text{BMI}/5) - 0.150[\log_2(\text{LSM})\times\text{BMI}/5] + 0.075(\text{age}/10) - 0.110(\text{male}) - 0.431(\text{diabetes}) + 0.721\log_2(\text{AST}) + 0.617\log_2(\text{ALT}) + 0.168\log_2(\text{platelets})$.

We established five models to represent increasing flexibility in how BMI modifies the LSM effect using machine learning. Spline complexity was determined via repeated five-fold cross-validation on the training set. The BMI-balanced model incorporated inverse-frequency weights to ensure the five pre-specified BMI strata contributed more equally to the loss function. The structural characteristics, strengths, and limitations of each model are summarized in Table S1.

Statistical analysis

Continuous variables are summarized as median (IQR) and categorical variables as n (%). Across-cohort comparisons used Kruskal-Wallis or Pearson chi-square tests as appropriate. AUC 95% CIs were estimated by patient-level bootstrap, and BMI-stratified AUCs were estimated within each prespecified stratum. Threshold sensitivity, specificity, predictive values, accuracy, Youden index and likelihood ratios were calculated from 2×2 tables. Calibration was assessed with observed-versus-predicted plots, calibration intercept, calibration slope and Brier score^[21-23]. Decision-curve analysis quantified net benefit across risk thresholds from 0.05 to 0.80^[24].

Results

Cohort characteristics and histological distribution

Among 520 participants, 255 (49.0%) had $F\geq 2$. The median age was 44.0 years (IQR 33.0-55.0), median BMI 32.2 kg/m² (28.9-37.7), median LSM 8.4 kPa (6.7-11.9) and median CAP 316 dB/m (287.8-334.2). The cohort included 252 men (48.5%). Stage frequencies were F0 69 (13.3%), F1 196 (37.7%), F2 163 (31.3%), F3 66 (12.7%) and F4 26 (5.0%) (Table 1; Figure 1). Cohort 1, Cohort 2 and Cohort 3 differed in age, BMI, LSM, CAP, liver enzymes and fibrosis prevalence, indicating marked case-mix heterogeneity. LSM increased progressively across F0-F4, with a Spearman correlation of 0.650 ($P<0.001$), although adjacent stages overlapped (Figure 2). CAP increased across histological steatosis grades with a weaker monotonic association ($r=0.396$, $P<0.001$). These patterns support construct validity for both measurements while showing that neither marker provides perfect histological separation.

Primary LSM discrimination and threshold trade-offs

Continuous LSM discriminated $F\geq 2$ with a pooled AUC of 0.861 (95% CI 0.829-0.891). Cohort-stratified AUCs were 0.852 for Cohort 1, 0.892 for Cohort 2 and 0.848 for Cohort 3 (Figure 4). At 8 kPa, pooled sensitivity was 0.929 (0.891-0.958), specificity 0.668 (0.608-0.724), PPV 0.729 and NPV 0.908. Lower thresholds prioritized sensitivity and NPV, whereas higher thresholds

increased specificity and PPV. Across the six prespecified values, 8 kPa had the highest Youden index (0.597), while 12 kPa yielded the highest specificity (0.936) but sensitivity fell to 0.443. A threshold selected for ruling out disease therefore cannot be treated as a universal rule-in threshold (Figure 3).

Subgroup variation and BMI-related misclassification

Prespecified subgroup AUCs ranged from approximately 0.81 to 0.92 (Figure 5). BMI showed the clearest clinically ordered attenuation: AUC decreased from 0.880 (0.808-0.939) below 28 kg/m² to 0.777 (0.659-0.880) at ≥ 42 kg/m² (Figure 6). At 8 kPa, sensitivity remained high across strata, but specificity was lowest in the highest BMI stratum (0.567), indicating a shift in the balance between missed and overcalled disease. Among $F \geq 2$ cases, each 5-kg/m² BMI increase was associated with 30% higher odds of a false-negative classification after centre adjustment (OR 1.30, 95% CI 0.99-1.72; $P=0.062$). The corresponding false-positive estimate among $F < 2$ participants was OR 1.17 (0.96-1.42; $P=0.128$). These imprecise estimates do not establish causality, but the directionally consistent false-negative signal and declining AUC identify severe obesity as an important source of diagnostic instability.

Primary four-model development and internal validation

In the training set, AUCs were 0.822 (95% CI 0.778-0.863) for the clinical model, 0.869 (0.831-0.904) for LSM alone, 0.908 (0.876-0.936) for the LSM-clinical model and 0.909 (0.877-0.937) for the linear LSM $\times$ BMI interaction model. Internal-validation AUCs were 0.760 (0.683-0.835), 0.839 (0.772-0.898), 0.857 (0.795-0.912) and 0.860 (0.799-0.915), respectively (Figure 7). Thus, adding clinical context to LSM improved discrimination relative to either clinical variables or LSM alone, whereas the linear interaction produced only a small incremental gain over the LSM-clinical model (delta AUC +0.003, 95% CI -0.003 to +0.010 in internal validation). In internal validation, the clinical model had a calibration intercept of 0.128, calibration slope of 0.795 and Brier score of 0.202; the LSM-only model had corresponding values of 0.117, 0.938 and 0.168. The reference LSM-clinical and linear interaction models had Brier scores of 0.156, with calibration slopes of 0.855 and 0.806, respectively. Calibration plots showed greater irregularity than in training, consistent with the smaller internal-validation sample and the centre- and BMI-related case-mix differences. Decision curves indicated positive net benefit over treat-none and, across most of the risk-threshold range, over treat-all; the curves of the four models were close, so the clinical utility gain from the interaction term was modest (Figure 7).

BMI-stratified performance of the four primary models

The LSM-clinical and linear interaction models performed well below BMI 32 kg/m² in internal validation, with AUCs of 0.912-0.942, but discrimination fell at higher BMI (Figure 8). At BMI ≥ 42 kg/m², both models had an AUC of 0.778. The conditional coefficient of log₂(LSM) decreased as BMI increased, indicating that a two-fold increase in LSM carried less modelled diagnostic information at higher BMI. The linear interaction therefore represented the observed effect modification but did not restore the lost rank separation.

Flexible BMI-aware and machine-learning models

As the linear interaction model failed to restore the lost rank separation despite capturing the observed effect modification, we further developed the three more flexible models, including a machine-learning approach, to mitigate the decline in LSM diagnostic accuracy at higher BMI and deliver potential diagnostic benefits. The standard spline varying-coefficient, BMI-balanced spline and monotonic gradient-boosting models were evaluated as sensitivity analyses (Figure 9). Internal-validation AUCs were 0.855, 0.859 and 0.876, compared with 0.857 for the reference model and

0.860 for the linear interaction model. Gradient boosting achieved an AUC of 0.991 in training but 0.876 in validation, a large optimism gap. Its validation Brier score was lowest (0.142), although the calibration slope remained below 1 (0.776), indicating residual overfitting or over-dispersion of fitted predictions. The model was therefore not treated as evidence that BMI-related measurement limitations had been eliminated.

Whether BMI-aware modelling overcame BMI-related degradation

Flexible modelling did not produce uniform high-BMI gains (Figure 10; Table S2). The BMI-balanced spline model improved AUC locally in the 37-41.999 kg/m² stratum (0.795 versus 0.768 for the reference), but its AUC was lower at ≥ 42 kg/m² (0.742 versus 0.778). Gradient boosting improved the 32.0-36.999 and ≥ 42 kg/m² strata but performed poorly at 37-41.999 kg/m² (AUC 0.696). Confidence intervals were wide because high-BMI validation strata contained only 23-29 participants, and all subgroup comparisons were exploratory. The models therefore captured BMI-dependent effects and sometimes improved local ranking, but did not eliminate the influence of BMI on discrimination.

Discussion

This multicentre, three-cohort pathology-linked study shows that LSM is a useful marker for significant fibrosis, with good pooled discrimination (AUC 0.861) consistent with pooled estimates from systematic reviews of elastography in fatty liver disease^[8,12] and high sensitivity at the widely used 8-kPa threshold. Its limitations are equally clear. Specificity was moderate, neighbouring histological stages overlapped, and discrimination weakened as BMI increased. The practical implication is that a low LSM can contribute to ruling out significant fibrosis, whereas an elevated value should be interpreted with clinical context and, particularly in severe obesity, may require confirmation.

The BMI analyses add information beyond a standard overall ROC curve. The decrease in AUC from 0.880 below 28 kg/m² to 0.777 at ≥ 42 kg/m² indicates loss of rank ordering, not merely a shift in predicted probability. An ordinary BMI covariate can recalibrate baseline risk but cannot by itself correct a change in the LSM-outcome relationship. The linear and spline varying-coefficient models allowed the LSM coefficient to depend on BMI, directly representing effect heterogeneity. Nevertheless, a conditional coefficient changes the prediction equation; it cannot guarantee higher AUC within every subgroup, particularly when the measured LSM distribution is compressed or noisier in severe obesity.

The model comparison illustrates the cost of flexibility. Penalized LSM-clinical regression was comparatively stable, and the linear interaction added little overall discrimination. Spline and BMI-balanced models produced local gains but no consistent improvement in the most obese group. Monotonic gradient boosting achieved the highest internal-validation AUC and lowest Brier score, but the near-perfect training AUC signals substantial overfitting. It should therefore be viewed as a sensitivity model rather than the preferred clinical model until tested in an independent cohort with locked preprocessing and sufficient severe-obesity representation.

Classical scores performed less well than LSM in both the training and internal-validation sets (Supplementary Figure S2). In internal validation, AUCs were 0.469 for FIB-4, 0.588 for APRI, 0.448 for TyG, 0.607 for TyG-BMI, 0.433 for NFS, 0.621 for ZJU, 0.509 for BARD and 0.839 for LSM; the corresponding training AUCs were 0.479, 0.655, 0.568, 0.560, 0.466, 0.571, 0.436 and 0.869. These comparisons should be interpreted within the pathology-enriched case mix and the available variable definitions. NFS and BARD depend on clinical and laboratory assumptions that may not align with this cohort, while glucose-based diabetes proxies may differ from adjudicated

diagnoses. The inverted direction of several scores is consistent with the ALT-driven referral pathway in a biopsy-enriched cohort: median ALT was 62 U/L, FIB-4 is inversely weighted by the square root of ALT, and the NFS and BARD incorporate the AST/ALT ratio, so marked inflammatory activity among participants without significant fibrosis can depress these scores. The inverted TyG estimate parallels the cohort composition, in which younger and more obese participants carried a high metabolic burden but a lower prevalence of significant fibrosis. The result supports the added value of direct stiffness measurement in this dataset but does not invalidate established scores in their intended sequential-testing pathways^[25-29].

This study has strengths, including histological reference staging across three prespecified hospital-defined cohorts, prespecified LSM thresholds, explicit analysis of false positives and false negatives, and reporting of discrimination, calibration and decision utility. The fixed stratified split prevents direct information leakage from validation into model fitting, and monotonic constraints preserve the clinically plausible direction of the LSM effect. Reporting model coefficients and transformations makes the four primary logistic models reproducible within the analysed dataset.

Several limitations constrain interpretation. The 7:3 validation set is internal and drawn from the same pooled cohort; it does not test geographical or temporal transportability. High-BMI strata contain few participants and events, yielding wide confidence intervals. Multiple models and subgroup analyses were exploratory, and no adjustment can recover information that is absent or noisy in the original measurement. The three centres also differed markedly in case mix (Table 1): Cohort 3 participants were younger (median age 33.0 years), more obese (median BMI 36.3 kg/m²) and had substantially lower transaminases (median ALT 18.1 U/L) and a lower prevalence of significant fibrosis (33.3%) than Cohorts 1 and 2, consistent with different referral and recruitment pathways across centres. This heterogeneity broadens the BMI range represented but means that pooled performance reflects a specific pathology-enriched case mix, and between-centre differences in assay platforms may contribute to the observed differences in liver enzymes. Prospective external validation with locked coefficients, standardized elastography quality criteria and sufficient representation of severe obesity is required.

In conclusion, LSM provides good overall discrimination for histologically significant fibrosis, but BMI materially affects its diagnostic stability. LSM-clinical modelling improves performance, and BMI-aware approaches can describe or locally reduce heterogeneity, yet the current data do not show that BMI effects have been eliminated. The reference penalized LSM-clinical model offers the most defensible balance of accuracy, calibration and complexity pending external validation; the interaction and flexible models are useful for hypothesis generation and future validation rather than proof of complete BMI correction.

Declarations

Ethical Approval

This study was performed in compliance with the Declaration of Helsinki. The protocol was approved by the Medical Ethics Committee of Huai'an First People's Hospital (No. JS-2022-002-01) and the local ethics committees of all participating centers. Clinical trial number: not applicable.

Funding

This study was supported by the Natural Science Research Project of Huai'an City (HAB202318).

Data availability

The data supporting the findings of this study are available within the article and its supplementary materials.

Author contributions

Le Wang and Lei Ding contributed to LSM data acquisition and formal statistical analysis. Le Wang was additionally responsible for drafting the original manuscript. Guangnian Ji, Chuanrong Zhu, and Wan'e Wang participated in patient recruitment, sample collection, and curation of associated clinical datasets. Yujie Gao and Jie Zhang carried out data quality validation and partial data analysis. Min Xu supervised the overall data coordination workflow, provided strategic guidance for analytical approaches, and critically reviewed and edited the manuscript for important intellectual content.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank all study participants and the staff of all participating centers for their contributions.

References

1. Rinella ME, Lazarus JV, Ratziu V, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. *Hepatology*. 2023;78:1966-1986. doi:10.1097/HEP.0000000000000520.
2. 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;77:1797-1835. doi:10.1097/HEP.0000000000000323.
3. Ciardullo S, Monti T, Perseghin G. High prevalence of advanced liver fibrosis assessed by transient elastography among US adults with type 2 diabetes. *Diabetes Care*. 2021;44:519-525. doi:10.2337/dc20-1778.
4. Dulai PS, Singh S, Patel J, et al. Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: systematic review and meta-analysis. *Hepatology*. 2017;65:1557-1565. doi:10.1002/hep.29085.
5. Angulo P, Kleiner DE, Dam-Larsen S, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2015;149:389-397.e10. doi:10.1053/j.gastro.2015.04.043.
6. Mozes FE, Lee JA, Vali Y, et al. Performance of non-invasive tests and histology for the prediction of clinical outcomes in patients with non-alcoholic fatty liver disease: an individual participant data meta-analysis. *Lancet Gastroenterol Hepatol*. 2023;8:704-713. doi:10.1016/S2468-1253(23)00141-3.
7. 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;75:659-689. doi:10.1016/j.jhep.2021.05.025.
8. Selvaraj EA, Mozes FE, Jayaswal ANA, et al. Diagnostic accuracy of elastography and magnetic resonance imaging in patients with NAFLD: a systematic review and meta-analysis. *J Hepatol*. 2021;75:770-785. doi:10.1016/j.jhep.2021.04.044.

9. Wong VWS, Vergniol J, Wong GLH, et al. Diagnosis of fibrosis and cirrhosis using liver stiffness measurement in nonalcoholic fatty liver disease. *Hepatology*. 2010;51:454-462. doi:10.1002/hep.23312.
10. Myers RP, Pomier-Layrargues G, Kirsch R, et al. Feasibility and diagnostic performance of the FibroScan XL probe for liver stiffness measurement in overweight and obese patients. *Hepatology*. 2012;55:199-208. doi:10.1002/hep.24624.
11. Eddowes PJ, Sasso M, Allison M, et al. Accuracy of FibroScan controlled attenuation parameter and liver stiffness measurement in assessing steatosis and fibrosis in patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2019;156:1717-1730. doi:10.1053/j.gastro.2019.01.042.
12. Castera L, Friedrich-Rust M, Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2019;156:1264-1281.e4. doi:10.1053/j.gastro.2018.12.036.
13. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi:10.1136/bmj.h5527.
14. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD). *Ann Intern Med*. 2015;162:55-63. doi:10.7326/M14-0697.
15. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370:1453-1457. doi:10.1016/S0140-6736(07)61602-X.
16. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*. 2005;41:1313-1321. doi:10.1002/hep.20701.
17. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. *Stat Med*. 2019;38:1276-1296. doi:10.1002/sim.7992.
18. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B*. 1996;58:267-288. doi:10.1111/j.2517-6161.1996.tb02080.x.
19. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51-58. doi:10.7326/M18-1376.
20. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54:774-781. doi:10.1016/S0895-4356(01)00341-9.
21. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
22. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
23. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78:1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
24. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565-574. doi:10.1177/0272989X06295361.

25. Wai CT, Greenson JK, Fontana RJ, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology*. 2003;38:518-526. doi:10.1053/jhep.2003.50346.
26. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology*. 2006;43:1317-1325. doi:10.1002/hep.21178.
27. Angulo P, Hui JM, Marchesini G, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. *Hepatology*. 2007;45:846-854. doi:10.1002/hep.21496.
28. Harrison SA, Oliver D, Arnold HL, Gogia S, Neuschwander-Tetri BA. Development and validation of a simple NAFLD clinical scoring system for identifying patients without advanced disease. *Gut*. 2008;57:1441-1447. doi:10.1136/gut.2007.146019.
29. Guerrero-Romero F, Simental-Mendia LE, Gonzalez-Ortiz M, et al. The product of triglycerides and glucose, a simple measure of insulin sensitivity: comparison with the euglycemic-hyperinsulinemic clamp. *J Clin Endocrinol Metab*. 2010;95:3347-3351. doi:10.1210/jc.2010-0288.

Figure legends

Figure 1 | Histological fibrosis-stage composition of the multicentre pathology cohort. A, pooled counts and percentages across F0-F4. B, within-cohort stage composition; numbers inside segments are patient counts. Cohort 1 comprised the Affiliated Huai'an No.1 People's Hospital of Nanjing Medical University; Cohort 2 comprised Huai'an Fourth People's Hospital; Cohort 3 combined the Affiliated Huai'an Hospital of Xuzhou Medical University and Huai'an Fifth People's Hospital. n=520 independent participants.

Figure 2 | Fibrosis-stage distribution, LSM and CAP-steatosis features. A, pooled F0-F4 counts. B, individual LSM values and boxplots across fibrosis stages, with medians connected to show the ordinal trend. C, CAP values across histological steatosis grades S0-S3; five values originally entered as grade 4 were confirmed as data-entry errors and corrected to S3 before analysis. Boxes show the IQR, centre lines the median and whiskers 1.5xIQR; points are individual participants.

Figure 3 | Diagnostic trade-offs across prespecified LSM thresholds. A, sensitivity and specificity. B, positive and negative predictive values. C, Youden index and accuracy. D, combined display of sensitivity, specificity, positive predictive value and negative predictive value. The vertical dashed line marks the prespecified 8-kPa reference threshold. n=520 independent participants.

Figure 4 | ROC performance of LSM for histologically significant fibrosis ($F \geq 2$). A, pooled multicentre cohort. B, cohort strata. Cohort 1 comprised Huai'an First People's Hospital; Cohort 2 comprised Huai'an Fourth People's Hospital; Cohort 3 combined Huai'an Second People's Hospital and Huai'an Fifth People's Hospital. Circles mark the 8-kPa operating point. AUC intervals were obtained by stratified bootstrap.

Figure 5 | LSM discrimination and CAP-steatosis associations in prespecified subgroups. A, LSM AUC for $F \geq 2$ with 95% CI. B, Spearman correlations between CAP and histological steatosis grade with 95% CI. Subgroup analyses are exploratory.

Figure 6 | Influence of BMI on LSM diagnosis of $F \geq 2$. A, LSM distributions by fibrosis status within BMI strata. B, BMI-stratified AUC with 95% CI. C, sensitivity, specificity, PPV and NPV at

8 kPa. D, exploratory LSM×BMI interaction expressed as the conditional OR per two-fold increase in LSM.

Figure 7 | Development and internal validation of four primary prediction models. A and B, ROC curves in training and internal validation. C and D, calibration plots. E and F, decision curves. Models comprise clinical variables alone, LSM alone, LSM plus clinical variables and the linear LSM×BMI interaction model.

Figure 8 | BMI-stratified performance and conditional LSM effect of the four primary models. A and B, AUC heatmaps with 95% CI in training and internal validation. C, BMI-dependent conditional LSM effect. D, interaction-model AUC across BMI strata in training and internal validation.

Figure 9 | Comparison of reference, BMI-aware and monotonic machine-learning models. A and B, ROC curves in training and internal validation. C and D, calibration plots. E and F, decision curves. Model selection and fitting used the training set; validation predictions were generated with locked fitted models.

Figure 10 | BMI robustness of five prediction models. A and B, BMI-stratified AUCs with 95% CI in training and internal validation. C, BMI-varying coefficient for log₂(LSM) with bootstrap 95% CI. D, internal-validation AUC across BMI strata for the reference and BMI-aware varying-coefficient models. All subgroup analyses are exploratory and are not evidence of complete BMI correction.

Supplementary Figure S1 | Patient Flow Diagram.

Supplementary Figure S2 | Classical-score and LSM ROC comparisons in the training and internal-validation sets. Curves show FIB-4, APRI, TyG, TyG-BMI, NFS, ZJU, BARD and LSM; legends report AUC with 95% CI.

Integrated Tables and Figures

Table 1 | Baseline characteristics of the pooled multicentre pathology cohort.

Characteristic	Overall (n=520)	Cohort 1 (n=321)	Cohort 2 (n=88)	Cohort 3 (n=111)
Age, years	44.0 (33.0-55.0)	49.0 (38.0-57.0)	42.0 (32.8-53.0)	33.0 (28.5-41.0)
BMI, kg/m ²	32.2 (28.9-37.7)	31.1 (28.2-34.7)	32.9 (30.1-38.0)	36.3 (31.6-42.1)
NAS score	4.0 (3.0-5.0)	4.0 (3.0-6.0)	5.0 (4.0-6.0)	4.0 (3.0-5.0)
LSM, kPa	8.4 (6.7-11.9)	8.4 (6.6-12.3)	9.8 (8.0-14.1)	7.8 (6.5-9.9)
ALT, U/L	62.0 (28.8-104.2)	73.0 (39.0-110.0)	89.5 (53.2-128.8)	18.1 (10.6-35.1)
AST, U/L	41.0 (26.0-72.0)	46.0 (29.0-76.0)	40.0 (28.0-70.0)	30.1 (20.0-50.4)
Platelets, x10 ⁹ /L	253.5 (206.9-311.1)	249.0 (204.0-306.0)	253.3 (196.2-324.7)	258.0 (225.5-301.0)
Male, n (%)	252 (48.5%)	167 (52.0%)	47 (53.4%)	38 (34.2%)
Lobular inflammation grade ≥2, n (%)	195 (37.5%)	124 (38.6%)	54 (61.4%)	17 (15.3%)
Significant fibrosis (F≥2), n (%)	255 (49.0%)	156 (48.6%)	62 (70.5%)	37 (33.3%)
Histological F0, n (%)	69 (13.3%)	31 (9.7%)	5 (5.7%)	33 (29.7%)
Histological F1, n (%)	196 (37.7%)	134 (41.7%)	21 (23.9%)	41 (36.9%)
Histological F2, n (%)	163 (31.3%)	85 (26.5%)	52 (59.1%)	26 (23.4%)
Histological F3, n (%)	66 (12.7%)	49 (15.3%)	9 (10.2%)	8 (7.2%)
Histological F4, n (%)	26 (5.0%)	22 (6.9%)	1 (1.1%)	3 (2.7%)

Figure 1 | Histological fibrosis-stage composition of the multicentre pathology cohort.

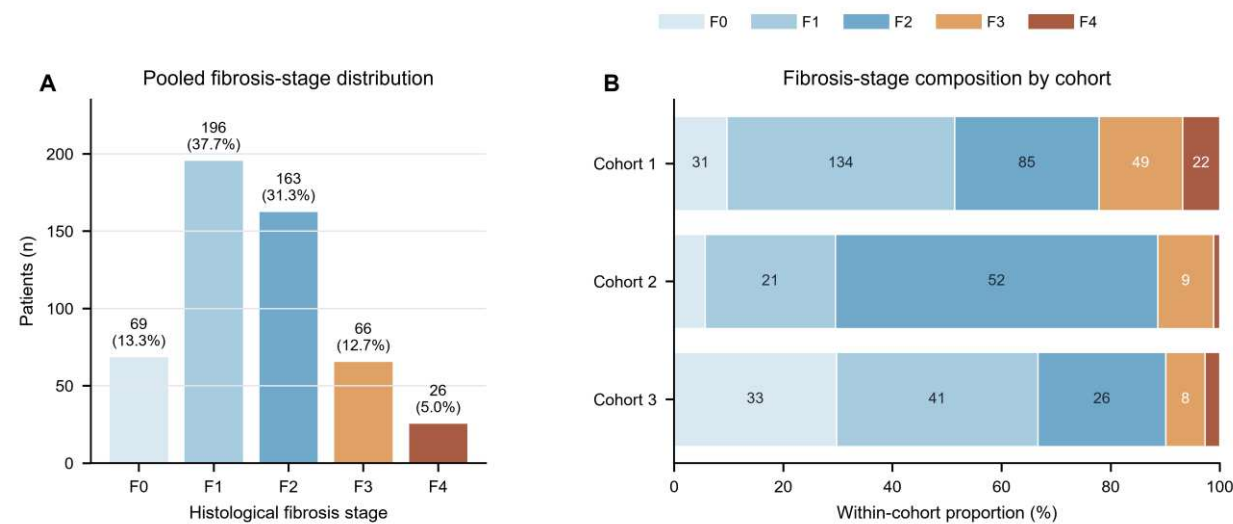


Figure 2 | Fibrosis-stage distribution, LSM and CAP-steatosis features.

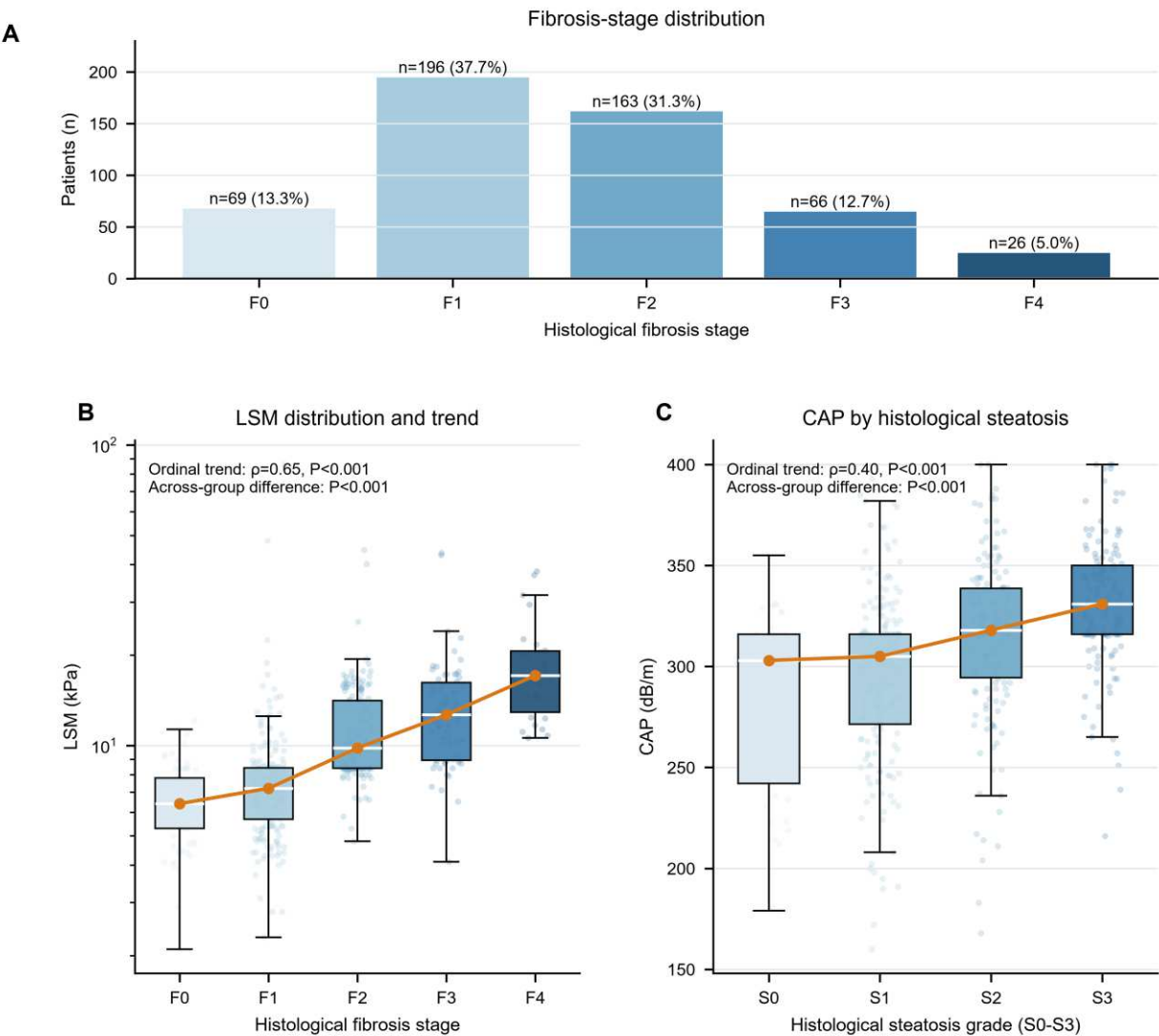


Figure 3 | Diagnostic trade-offs across prespecified LSM thresholds.

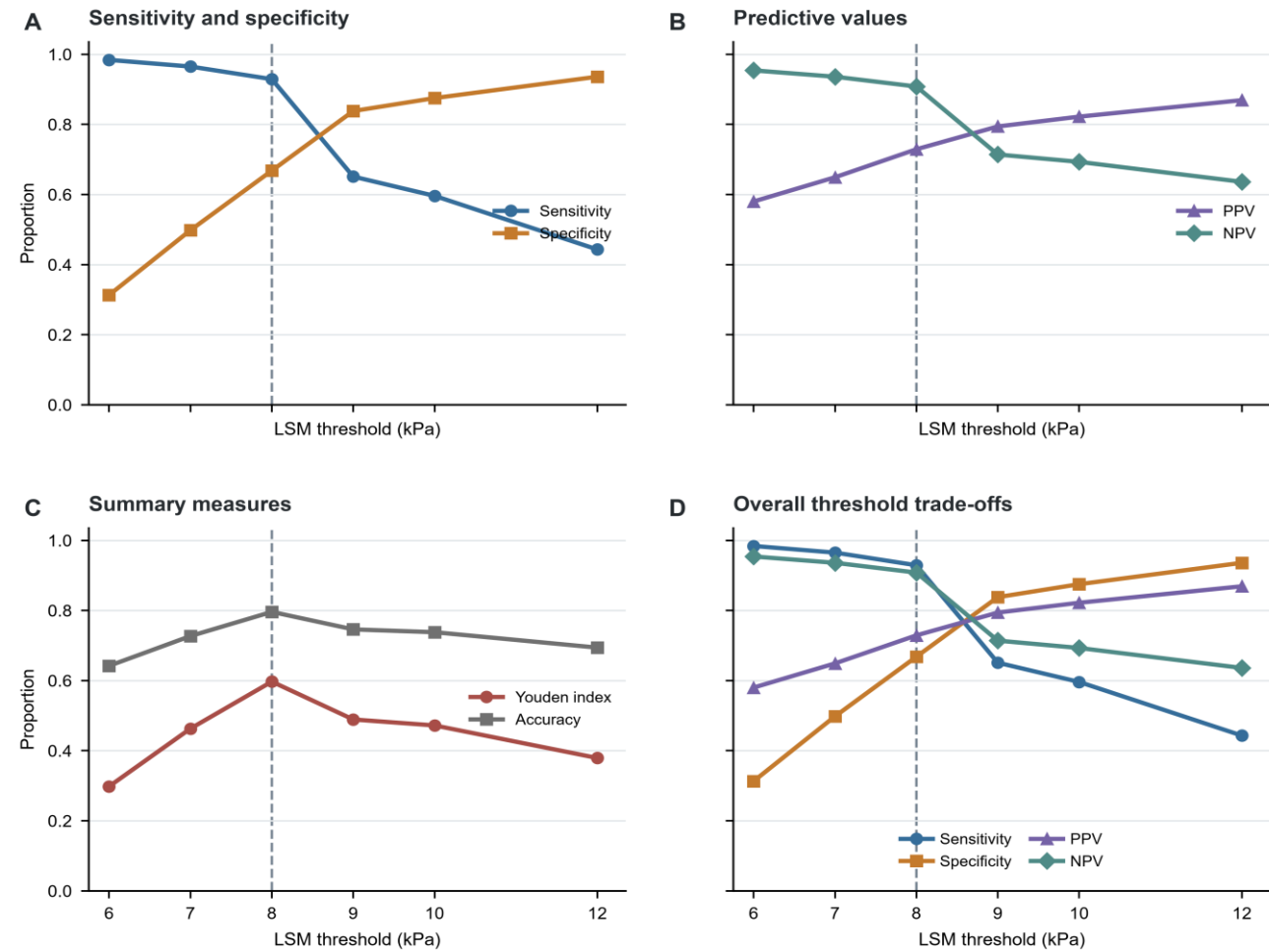


Figure 4 | ROC performance of LSM for histologically significant fibrosis ($F \geq 2$).

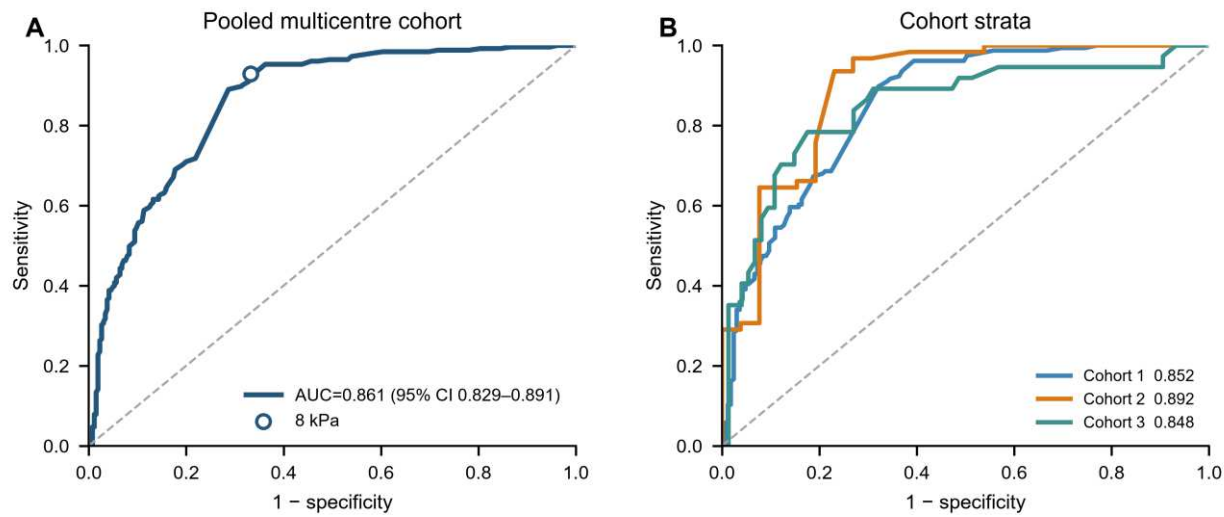


Figure 5 | LSM discrimination and CAP-steatosis associations in prespecified subgroups.

LSM discrimination and CAP-steatosis associations in prespecified subgroups

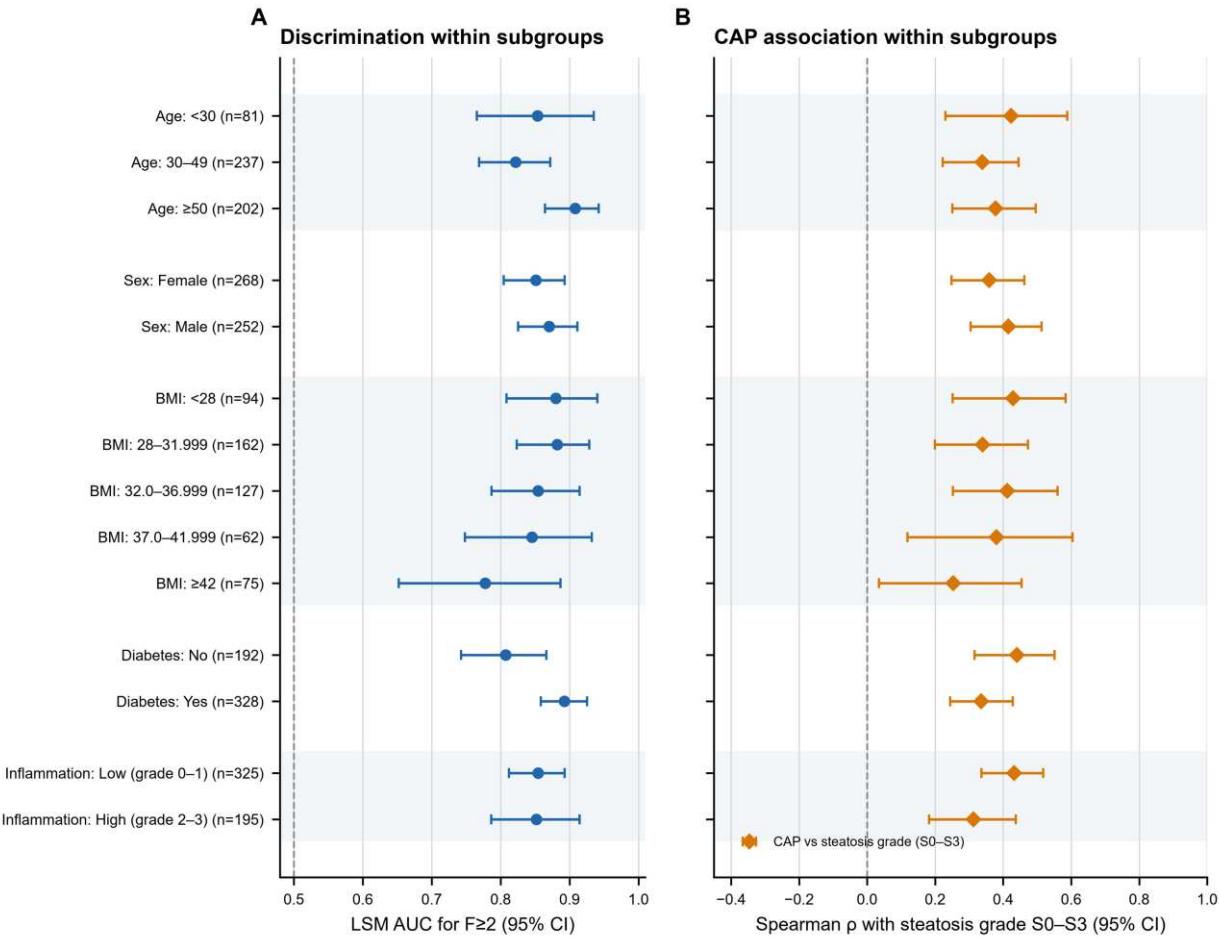


Figure 6 | Influence of BMI on LSM diagnosis of $F \geq 2$.

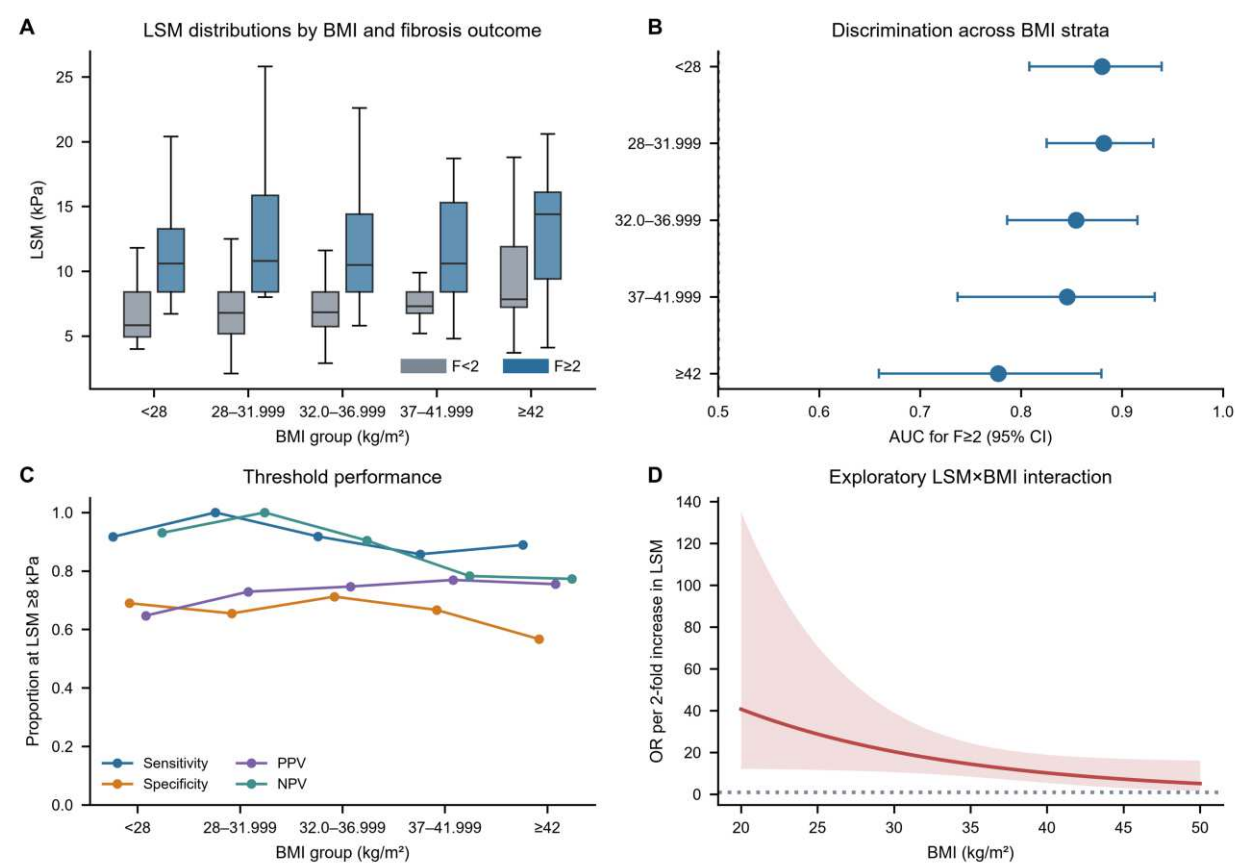


Figure 7 | Development and internal validation of four primary prediction models.

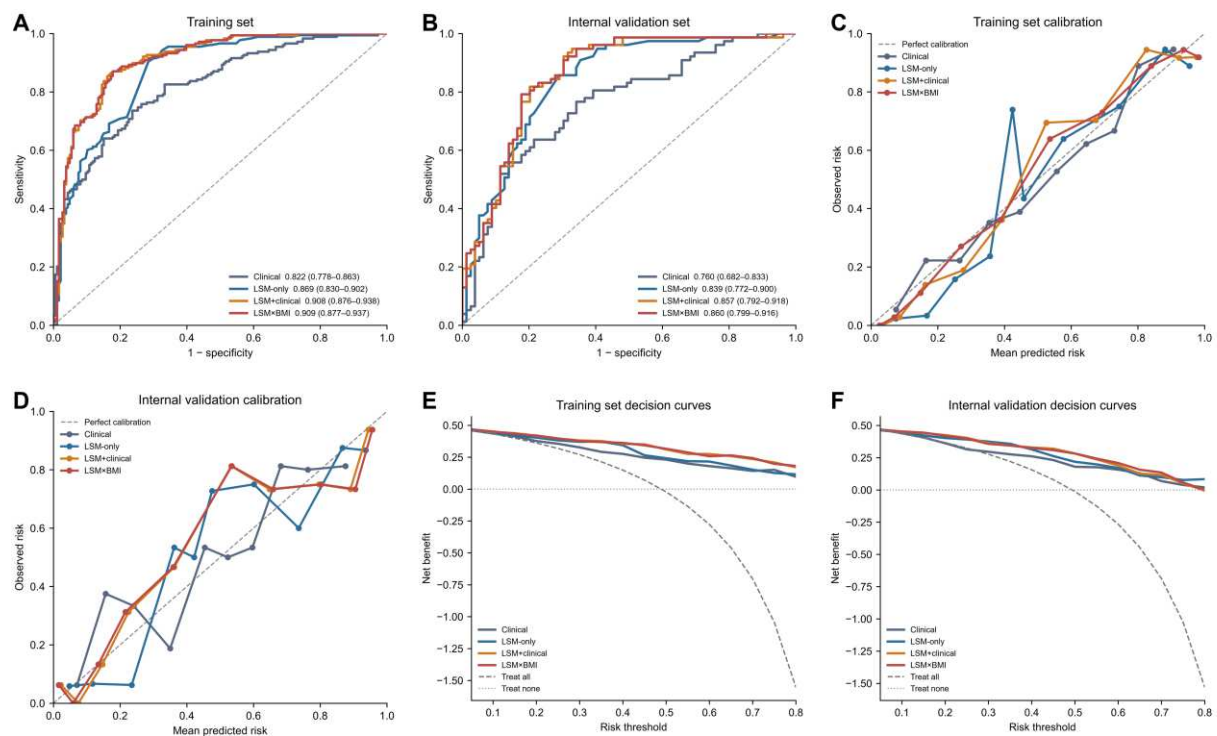


Figure 8 | BMI-stratified performance and conditional LSM effect of the four primary models.

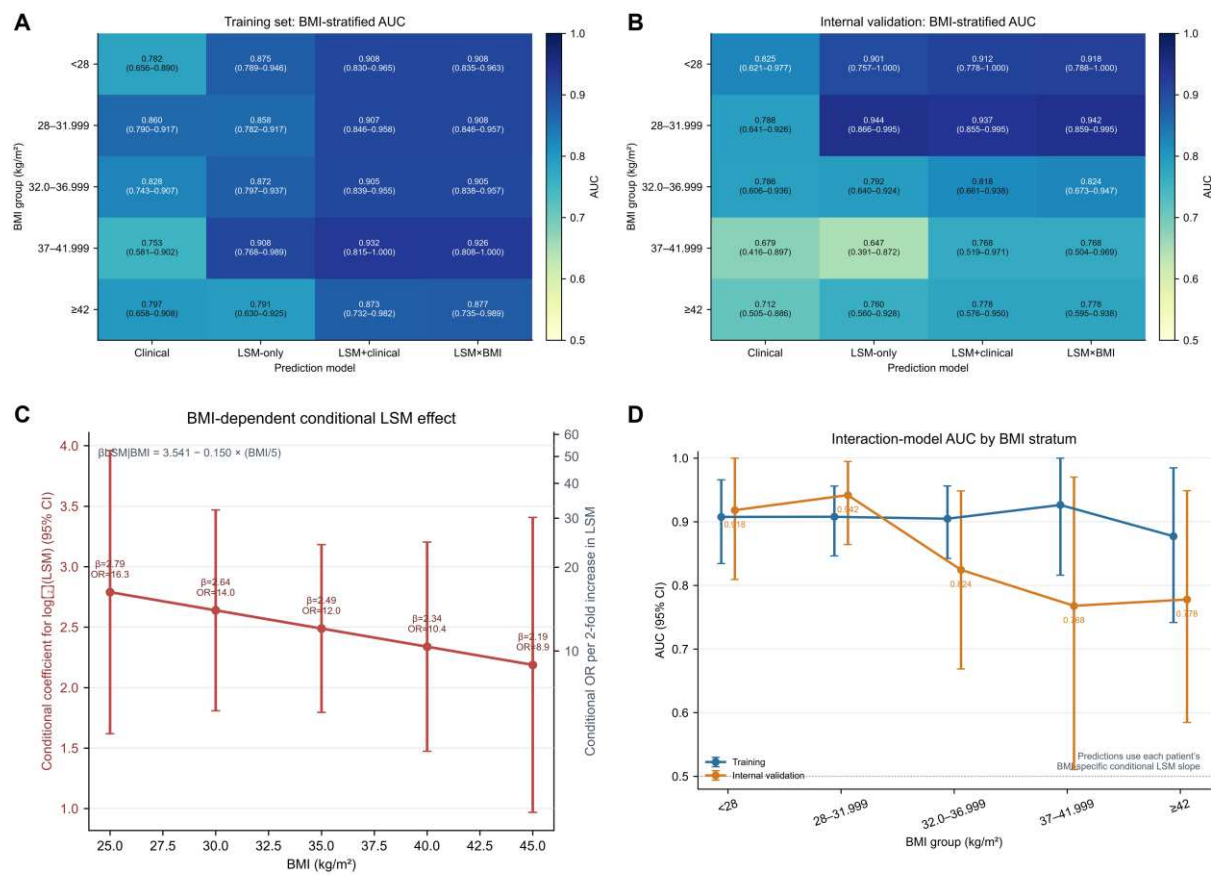


Figure 9 | Comparison of reference, BMI-aware and monotonic machine-learning models.

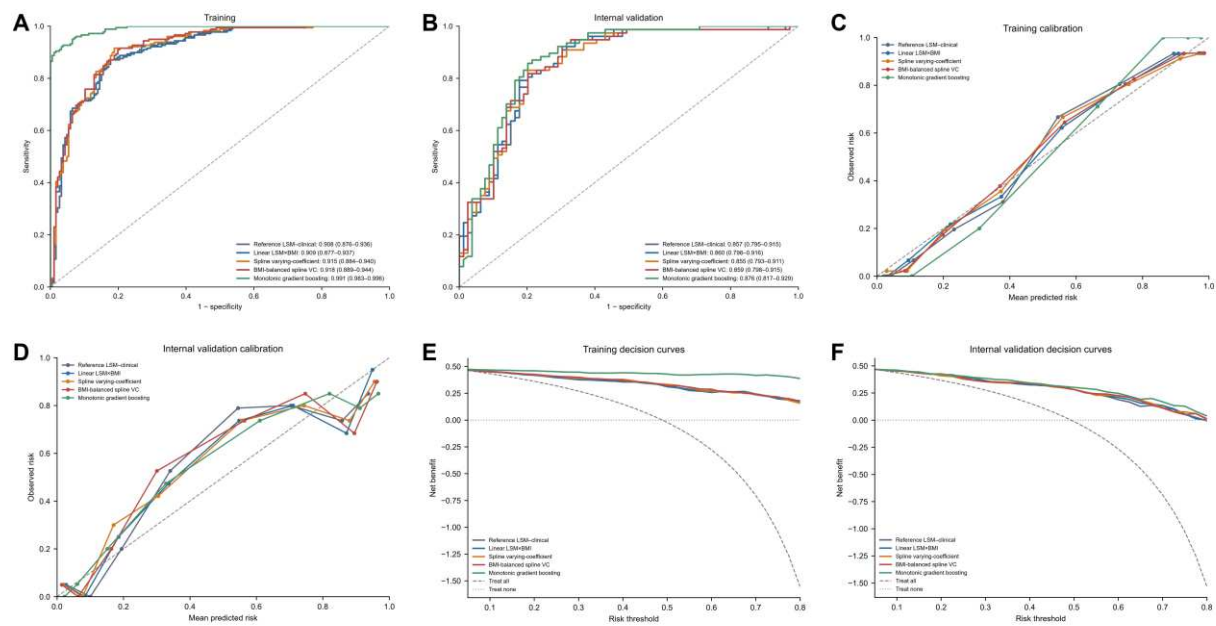
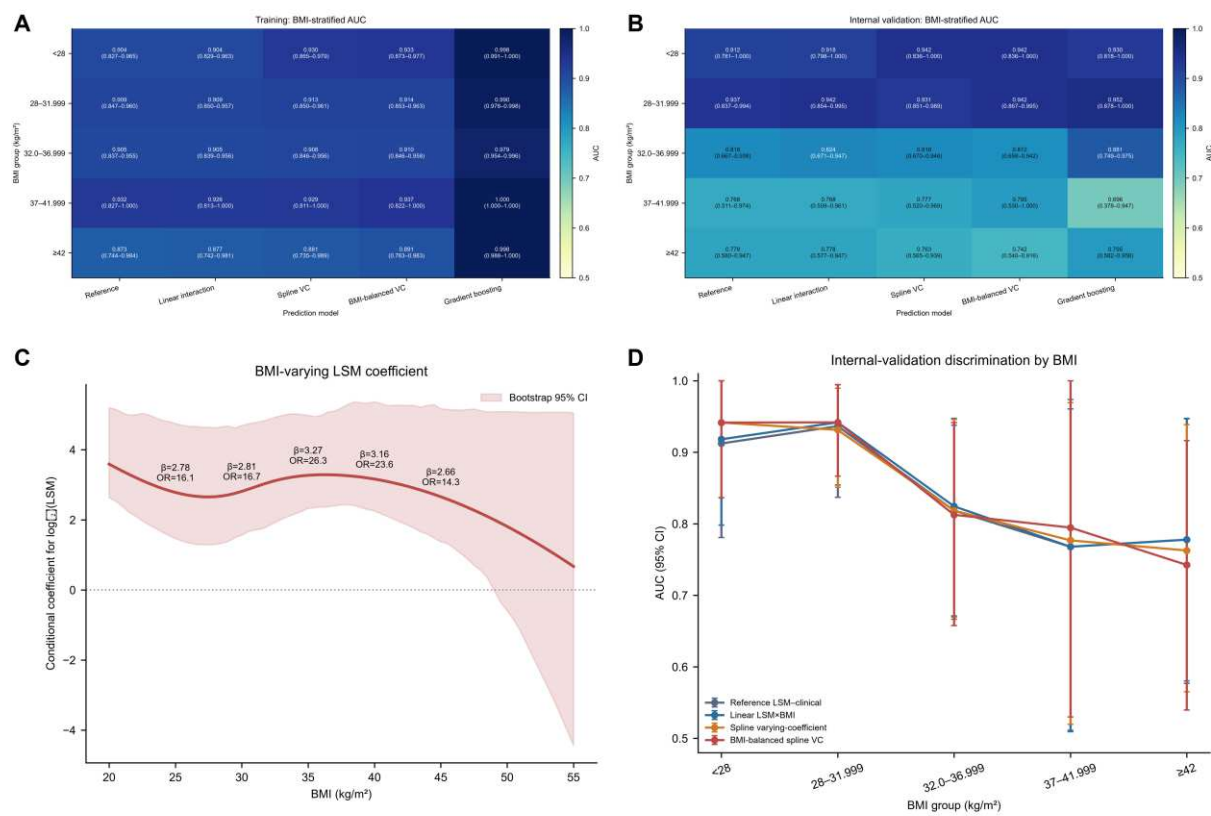


Figure 10 | BMI robustness of five prediction models.



Supplementary Files

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

- [SupplementaryTablesS1S2.docx](#)
- [FigureS2.tiff](#)
- [FigureS1.tif](#)