MASH IN FOCUS

Current Developments in the Management of Metabolic Dysfunction-Associated Steatohepatitis

Section Editor: Naim Alkhouri, MD

Novel Risk Scores in Metabolic Dysfunction-Associated Steatohepatitis



Winston Dunn, MD
Professor Emeritus of Medicine
University of Kansas Medical Center
Kansas City, Kansas
Director of Liver Research
Summit Clinical Research Solutions
San Antonio, Texas

G&H Why has there been a push to develop new risk scores in metabolic dysfunction-associated steatohepatitis when there are already commonly used scores like the Fibrosis-4 score?

WD The clinical objective has broadened. The initial Fibrosis-4 (FIB-4) score was created for use in HIV/hepatitis C coinfection. It used easily obtained parameters such as age, aspartate aminotransferase (AST), alanine aminotransferase (ALT), and platelet count. Since then, its utility has expanded mainly to be used as a first-line rule-out of advanced fibrosis, usually defined as stages F3 to F4, in metabolic dysfunction-associated steatotic liver disease (MASLD). Simplicity is not a limitation but a strength.[1-3]

The change is that there are now earlier treatment opportunities. Resmetirom (Rezdiffra, Madrigal) and semaglutide (Wegovy, Novo Nordisk) have shown histologic improvements and have received accelerated approval in the United States for noncirrhotic metabolic dysfunction-associated steatohepatitis (MASH) with F2 to F3 fibrosis. With available treatment, the objective is to identify MASH patients before they develop cirrhosis.[4-7] At this time, tools are needed that are appropriate for different decisions, namely identifying potential significant fibrosis for assessing treatment candidacy.

G&H What are the components of the Steatosis-Associated Fibrosis Estimator score?

WD The Steatosis-Associated Fibrosis Estimator (SAFE) score merges age, body mass index (BMI), diabetes status, AST, ALT, platelet count, and globulin (total serum protein minus albumin). It was developed to distinguish F0 and F1 from significant fibrosis, defined as F2 or higher.[8]

Diabetes and BMI provide context for interpreting the laboratory results. For example, otherwise similar results may have different implications in a patient with diabetes and obesity than in someone without those characteristics. These variables contribute information about metabolic risk; their inclusion does not make the SAFE score a cardiovascular risk calculator.[2,8]

Globulin provides a routinely available laboratory parameter (not specific for fibrosis) as a component. Patients are then classified into 3 risk classes: low risk (<0), intermediate risk (0-<100), and high risk (≥100). These risk levels do not represent identical equivalents of individual histologic stage. An increased score is, therefore, a trigger for further evaluation and does not substitute for independent diagnostic findings.[8]

G&H How does the intended use of the LiverRisk score compare with FIB-4's traditional role?

WD The LiverRisk score was developed for risk assessment in the general population, including people not already selected for hepatology evaluation. The main factors considered are age, sex, AST, ALT, gamma-glutamyl transferase (GGT), fasting glucose, total cholesterol, and platelet count.[9]

The key distinction is its development target and setting. The LiverRisk score was derived to predict liver stiffness and was evaluated for future liver-related outcomes. It should not be described as a biopsy-based F2 diagnostic

score. Its potential role is to identify people who warrant additional evaluation before advanced disease becomes clinically apparent.[9]

Referral triage is not new to the LiverRisk score. FIB-4 already does this in a supported guideline pathway. They can be applied together to identify who needs further testing; it is simply whether the LiverRisk score will improve detection, calibration, and referral efficiency in the target population. A general population model still needs to be tested in a higher-risk population, such as people with diabetes, rather than assuming equivalence.[2,3,9]

G&H How do the two versions of the ALADDIN score differ from each other and from other risk scores?

WD ALADDIN has two alternative configurations. ALADDIN-F2-Lab employs clinical and laboratory data without elastography. ALADDIN-F2-VCTE includes liver stiffness by vibration-controlled transient elastography (VCTE). Both aim for meaningful fibrosis or higher in MASLD, and the version with elastography has outperformed the published scores in external validation.[10] The purpose of the version without elastography is to ease access. A laboratory-based model can be considered where VCTE is unavailable, whereas the VCTE model combines information already obtained from an elastography examination with the patient's other characteristics.[10]

From a methodologic perspective, ALADDIN uses tree-based machine-learning algorithms, including random forests and gradient-boosting methods. These can represent nonlinear and interaction effects, instead of assuming a univariate additive relationship between each predictor and the risk of fibrosis. That flexibility is potentially useful, but the designation of machine learning does not itself establish clinical superiority. The relevant evidence is performance in independent patients, especially at thresholds used to decide whether further testing is needed.[10]

G&H What is LiverPRO designed to predict, and how does it differ from other scores?

WD LiverPRO was developed as a blood-based decision-support tool for detecting significant fibrosis in primary care. Histologic F2 or higher– and F3 or higher–guided model development validates against increased liver stiffness and liver-related events. These outputs should not be used as a prescriptive assignment of a single biopsy stage.[11]

LiverPRO's distinctive feature is flexibility when laboratory panels are incomplete. The published platform comprises 466 models using age plus 3 to 9 tests selected from AST, alkaline phosphatase, GGT, international normalized ratio (INR), albumin, sodium, bilirubin, platelet count, and cholesterol.[11] Thus, even if an INR is unavailable, that does not mean calculation is impossible; the software may select a suitable model that can be derived from the tests obtained. This is distinct from presuming negative tests are normal or asserting the risk can be computed without laboratory results. The key practical issue is which inputs are accessible, which model is chosen, and how well that setup works in the relevant population.[11]

G&H How do these new scores perform across different races and ethnic groups, as well as different BMIs?

WD No score can be described as universally superior across racial and ethnic groups. The SAFE score has been evaluated in US populations and in Hong Kong. LiverPRO's initial cohorts were European, and ALADDIN used multinational development and validation cohorts. Geographic diversity is valuable, but it does not establish equivalent sensitivity or calibration within every racial or ethnic subgroup.[10-13]

BMI also takes careful consideration. Patients who are not obese may still develop MASLD and significant fibrosis. However, obesity by itself does not confirm the presence of fibrosis. Excluding BMI from a model does not ensure unbiased results across all weight ranges; other predictors and the characteristics of the development cohort are also important.[2,3]

In the SAFE score, BMI enters the model as an additive term. Values above 40 kg/m² are capped at 40. The effect stops growing after that point. ALADDIN's tree-based approach can fit nonlinear effects. That does not mean BMI only matters a little or that lean patients cannot be missed.[10,13]

Another important variation is age. A US validation of the SAFE score identified limited sensitivity in younger adults and a high proportion requiring further assessment among older adults. Age-related limitations are also well documented for FIB-4. Clinicians should examine subgroup sensitivity, missed cases, and referral rates, not just an overall area under the receiver operating characteristic curve (AUROC).[13,14]

G&H Could you discuss any direct comparisons of the new scores vs older scores like FIB-4?

WD Several comparisons are informative, provided that their reference standards are kept separate. In a Hong Kong biopsy cohort, the SAFE score had an AUROC of

0.773 for significant fibrosis, compared with 0.746 for FIB-4 and 0.844 for VCTE. The SAFE score was numerically better than FIB-4, but that particular AUROC comparison did not reach statistical significance.[12]

In ALADDIN's external validation cohort, ALADDIN-F2-VCTE achieved an AUROC of 0.791 for biopsy-defined significant fibrosis or higher, compared with 0.745 for VCTE alone. ALADDIN-F2-Lab achieved 0.706 and outperformed the blood-based comparators, including FIB-4, in that analysis. Adding elastography clearly provided additional discrimination.[10]

LiverPRO was tested against a different endpoint in the 6468-participant DECIDE validation cohort: liver stiffness of at least 8 kPa. Its AUROC was 0.80, whereas for FIB-4 it was 0.69, 0.78 for the Enhanced Liver Fibrosis (ELF) test, and 0.81 for the LiverRisk score. That means LiverPRO performed better than FIB-4 but not much better than other alternatives.[11]

A 2026 cross-sectional evaluation of 12,050 National Health and Nutrition Examination Survey participants also confirmed higher AUROCs for LiverRisk in comparison with FIB-4 (0.718 vs 0.621 for stiffness of 8 kPa and 0.748 vs 0.654 for 10 kPa). However, discrimination was still moderate, and thresholds prioritized for sensitivity resulted in significant false-positive testing.[15]

Prognostic comparisons address another question. In a 2026 UK Biobank study of 385,738 participants, LiverRisk had the best-performing routine score for severe cirrhosis-related morbidity, with a 10-year C-index of 0.812. However, rankings were different when precision and the benefit of referral decisions were considered.[16]

The practical point is that biopsy-defined fibrosis, increased stiffness, and future complications are related, but with separate endpoints. Scores within the same population and the same endpoint should be compared, in addition to missed cases and unnecessary referrals. A higher AUROC by itself does not prove that a pathway leads to better clinical decisions.[17]

G&H Where would you place the new scores in current clinical practice?

WD Rather than being treated as interchangeable standalone diagnostic tests, these new scores can be put into a structured case-finding pathway. Existing guidance recommends evaluating fibrosis among people at higher risk and those people with type 2 diabetes or obesity with additional metabolic risk factors. FIB-4 is already an established first-line screening/diagnostic pathway, and abnormal or borderline scores often warrant referral for VCTE, ELF testing, or specialist assessment.[2,3]

The new blood-based scores are potentially suitable to refine primary selection, whereas ALADDIN-F2-VCTE is more naturally suited to follow available elastography. Their use should be dependent on local validation and whether they actually aid detection without saturating testing capability. A low-risk result designed to rule out advanced fibrosis should not be assumed to rule out the entire F2 disease.[2,3,10]

Bayesian adjustment is proposed for ALADDIN to address the influence of pretest probability. The same test information usually implies a lower posttest probability in a lower-prevalence setting. However, changing the assumed prevalence does not correct every difference between a specialist cohort and primary care; calibration still needs to be demonstrated.[10]

Treatment decisions should aim for reliable detection of noncirrhotic F2 to F3 disease using the right tool (clinical assessment combined with proper use of validated noninvasive tests), reserving biopsy for some areas of uncertainty. A new score should not be a reason for starting therapy and not being evaluated for cirrhosis.[6,7]

G&H How feasible is it for the average clinician to calculate and use these new scores in current clinical practice?

WD Calculation can be performed using electronic calculators or decision-support software, but it alone does not constitute implementation. The greater challenge lies in obtaining the necessary data, using the correct units and dates, and providing an actionable result within the electronic health record. The system should recognize missing values rather than silently treating them as normal.

An ideal workflow would flag eligible patients, perform risk estimation, advise on the next assessment, and monitor if it was carried out. During a hectic consultation, this might be hard to achieve with manual input alone. It is worth noting that structured primary-care pathways have already achieved better detection and fewer ill-advised referrals through known tests; any newer algorithm should be tested in that context.[18]

Computational complexity should not be mistaken for major per-patient computing expense. High costs include extra blood tests, software licensing, integration, elastography, and downstream referrals. These costs will have to be assessed when considering where the score will be applied.

Finally, a normal aminotransferase level must not discourage risk assessment. Liver biopsy research shows advanced fibrosis and cirrhosis despite normal AST and ALT. Case-finding criteria should be based on clinical risk factors in combination with an appropriate composite score.[2,19]

G&H What further research is needed?

WD Independent, prospective comparisons in intended-use populations, with prespecified thresholds and sufficient numbers of people across age, diabetes status, BMI, and race/ethnicity, are required. These should include calibration, sensitivity, predictive values, and numbers referred or missed, apart from discrimination. Decision-curve analysis can reveal whether better prediction represents a clinically useful trade-off between missed disease and unjustified testing.[17]

Detection should also be separated from prognosis. The fibrosis stage correlates with clinical outcomes. However, the detection of fibrosis at one given time point does not define how well a given model predicts decompensation, hepatocellular carcinoma, transplantation, or death. Hence, there should be focus on further longitudinal validation.[20]

Treatment monitoring is the next concern. A decreasing score might be due to changes in weight, aminotransferases, or other inputs rather than fibrosis regression. Modifications to these newer scores should be validated based on histology and clinical outcomes before they are used as treatment-response surrogates.[6,7]

Finally, I would consider implementation to be successful if it detects treatable disease at an earlier stage, ensures that the patient receives appropriate treatment, and improves outcomes without imposing an unsustainable burden of unnecessary testing.

Disclosures

Dr Dunn has no relevant conflicts of interest to disclose.

References

1. Sterling RK, Lissen E, Clumeck N, et al; APRICOT Clinical Investigators. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology*. 2006;43(6):1317-1325.

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(5):1797-1835.

3. 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). EASL-EASD-EASO Clinical Practice Guidelines on the management of metabolic dysfunction-associated steatotic liver disease (MASLD). *J Hepatol*. 2024;81(3):492-542.

4. Harrison SA, Bedossa P, Guy CD, et al; MAESTRO-NASH Investigators. A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. *N Engl J Med*. 2024;390(6):497-509.

5. Sanyal AJ, Newsome PN, Kliers I, et al; ESSENCE Study Group. Phase 3 trial of semaglutide in metabolic dysfunction-associated steatohepatitis. *N Engl J Med*. 2025;392(21):2089-2099.

6. Bansal MB, Patton H, Morgan TR, Carr RM, Dranoff JA, Allen AM. Semaglutide therapy for metabolic dysfunction-associated steatohepatitis: November 2025 updates to AASLD Practice Guidance. *Hepatology*. 2026;83(5):1326-1340.

7. European Association for the Study of the Liver; European Association for the Study of Diabetes; European Association for the Study of Obesity. EASL-EASD-EASO guidance for the use of resmetirom and semaglutide as MASH-targeted therapy [published online September 28, 2026]. *J Hepatol*. doi:10.1016/j.jhep.2026.08.028.

8. Sripongpun P, Kim WR, Mannalithara A, et al. The steatosis-associated fibrosis estimator (SAFE) score: a tool to detect low-risk NAFLD in primary care. *Hepatology*. 2023;77(1):256-267.

9. Serra-Burriel M, Juanola A, Serra-Burriel F, et al; LiverScreen Consortium Investigators. Development, validation, and prognostic evaluation of a risk score for long-term liver-related outcomes in the general population: a multicohort study. *Lancet*. 2023;402(10406):988-996.

10. Alkhouri N, Cheuk-Fung Yip T, Castera L, et al. ALADDIN: a machine learning approach to enhance the prediction of significant fibrosis or higher in metabolic dysfunction-associated steatotic liver disease. *Am J Gastroenterol*. 2026;121(2):362-374.

11. Lindvig KP, Thorhauge KH, Hansen JK, et al. Development, validation, and prognostic evaluation of LiverPRO for the prediction of significant liver fibrosis in primary care: a prospective cohort study. *Lancet Gastroenterol Hepatol*. 2025;10(1):55-67.

12. Li G, Lin H, Sripongpun P, et al. Diagnostic and prognostic performance of the SAFE score in non-alcoholic fatty liver disease. *Liver Int*. 2024;44(1):15-26.

13. van Kleef LA, de Knegt RJ, Ayada I, Pan Q, Brouwer WP. The steatosis-associated fibrosis estimator (SAFE) score: validation in the general US population. *Hepatol Commun*. 2023;7(4):e0075.

14. McPherson S, Hardy T, Dufour JF, et al. Age as a confounding factor for the accurate non-invasive diagnosis of advanced NAFLD fibrosis. *Am J Gastroenterol*. 2017;112(5):740-751.

15. Rinella ME, Charlton MR, Kim Y, Hosmane S, Woolley JJ, O'Connell T. Accuracy of the FIB-4 and LiverRisk score for use in primary risk assessment for liver fibrosis, and variation by age. *Dig Liver Dis*. 2026;58(7):935-942.

16. Liang Z, Jin H, Gao W, et al. Prospective head-to-head comparison of routine non-invasive scores for predicting severe cirrhosis-related morbidity in the general population [published online March 23, 2026]. *Hepatology*. doi:10.1097/HEP.0000000000001749.

17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.

18. Srivastava A, Gailer R, Tanwar S, et al. Prospective evaluation of a primary care referral pathway for patients with non-alcoholic fatty liver disease. *J Hepatol*. 2019;71(2):371-378.

19. Gawrieh S, Wilson LA, Cummings OW, et al; NASH Clinical Research Network. Histologic findings of advanced fibrosis and cirrhosis in patients with nonalcoholic fatty liver disease who have normal aminotransferase levels. *Am J Gastroenterol*. 2019;114(10):1626-1635.

20. Sanyal AJ, Van Natta ML, Clark J, et al; NASH Clinical Research Network (CRN). Prospective study of outcomes in adults with nonalcoholic fatty liver disease. *N Engl J Med*. 2021;385(17):1559-1569.