scientific reports

Check for updates

OPEN

Machine learning models using non-invasive tests & B-mode ultrasound to predict liver-related outcomes in metabolic dysfunction-associated steatotic liver disease

Heather Mary-Kathleen Kosick[1,2]✉, Chris McIntosh[2,3], Chinmay Bera[1,2], Mina Fakhriyehasl[3], Mohamed Shengir[4], Oyedele Adeyi[6], Leila Amiri[1,2], Giada Sebastiani[4,5], Kartik Jhaveri[2,3,7] & Keyur Patel[1,2,7]

Advanced metabolic-dysfunction-associated steatotic liver disease (MASLD) fibrosis (F3-4) predicts liver-related outcomes. Serum and elastography-based non-invasive tests (NIT) cannot yet reliably predict MASLD outcomes. The role of B-mode ultrasound (US) for outcome prediction is not yet known. We aimed to evaluate machine learning (ML) algorithms based on simple NIT and US for prediction of adverse liver-related outcomes in MASLD. Retrospective cohort study of adult MASLD patients biopsied between 2010–2021 at one of two Canadian tertiary care centers. Random forest was used to create predictive models for outcomes—hepatic decompensation, liver-related outcomes (decompensation, hepatocellular carcinoma (HCC), liver transplant, and liver-related mortality), HCC, liver-related mortality, F3-4, and fibrotic metabolic dysfunction-associated steatohepatitis (MASH). Diagnostic performance was assessed using area under the curve (AUC). 457 MASLD patients were included with 44.9% F3-4, diabetes prevalence 31.6%, 53.8% male, mean age 49.2 and BMI 32.8 kg/m^2. 6.3% had an adverse liver-related outcome over mean 43 months follow-up. AUC for ML predictive models were—hepatic decompensation 0.90(0.79–0.98), liver-related outcomes 0.87(0.76–0.96), HCC 0.72(0.29–0.96), liver-related mortality 0.79(0.31–0.98), F3-4 0.83(0.76–0.87), and fibrotic MASH 0.74(0.65–0.85). Biochemical and clinical variables had greatest feature importance overall, compared to US parameters. FIB-4 and AST:ALT ratio were highest ranked biochemical variables, while age was the highest ranked clinical variable. ML models based on clinical, biochemical, and US-based variables accurately predict adverse MASLD outcomes in this multi-centre cohort. Overall, biochemical variables had greatest feature importance. US-based features were not substantial predictors of outcomes in this study.

Keywords Steatosis, Artificial intelligence, Fibrosis, Liver biopsy, Hepatic decompensation

Metabolic dysfunction-associated steatotic liver disease (MASLD), previously termed non-alcoholic fatty liver disease (NAFLD) is now a global health epidemic and leading cause of liver-related mortality[1–3]. The presence of advanced MASLD fibrosis (F3-4) predicts worse liver-related outcomes[4–8]. Liver biopsy remains the reference

[1]Division of Gastroenterology, University Health Network Toronto, Toronto General Hospital, 200 Elizabeth Street, Toronto, ON M5G 2C4, Canada. [2]University of Toronto, 27 King's College Circle, Toronto, ON M5S 1A1, Canada. [3]Joint Department of Medical Imaging, University Health Network, Mount Sinai Hospital, Women's College Hospital, 200 Elizabeth Street, Toronto, ON M5G 2C4, Canada. [4]Division of Gastroenterology and Hepatology, McGill University Health Centre, 1001 Boul Decarie, Montreal, QC H4A 3J1, Canada. [5]McGill University, 805 Rue Sherbrooke O, Montreal, QC H3A 0B9, Canada. [6]Department of Laboratory Medicine and Pathology, University of Minnesota, Minnesota, MN 55455, USA. [7]Kartik Jhaveri and Keyur Patel authors jointly supervised this work. ✉email: heather.evans@medportal.ca; heather.kosick@gmail.com

standard to diagnose metabolic dysfunction-associated steatohepatitis (MASH) and stage MASLD fibrosis, however, is limited by its invasive nature, sampling heterogeneity, and poor suitability as a screening tool[2,9]. Markov models for MASLD burden of disease from several countries project a marked increase in liver-related outcomes over the next decade[10,11]. Identifying patients at risk of disease progression prior to developing symptoms is crucial. Research efforts have focussed on the development of non-invasive methods to predict F3-4. Diagnostic algorithms based on simple serum-based tests such as the NAFLD-fibrosis score (NFS), FIB-4, aspartate aminotransferase (AST) to platelet ratio index (APRI), patented blood tests, and point-of-care imaging-based tests, such as vibration-controlled transient elastography (VCTE™, EchoSens, Paris, France)) have been developed to further improve diagnostic accuracy[12]. However, a recent meta-analysis of 9 studies with simple blood tests and histologic scores for predicting clinical outcomes in MASLD indicated only NFS > 0.676 was associated with all-cause mortality, but not liver-related outcomes[13].

With increasing prevalence of F3-4, it is important to identify simple, cost-effective testing strategies to allow for risk stratification in an integrated healthcare model with non-specialist providers. B-mode ultrasound (US) is routinely performed on patients investigated for liver disease. In patients with metabolic risk factors for MASLD, US is recommended for initial assessment of steatosis[12,14,15], although limitations exist for detecting mild steatosis compared to more advanced techniques including VCTE controlled attenuation parameter and magnetic resonance imaging proton density fat fraction[14,16,17]. US can also detect signs of advanced liver disease (nodularity, coarse echotexture, ascites) and portal hypertension[18–20]. Surface nodularity was shown to predict F3-4 and F4 in a MASLD-predominant population of patients with liver disease of mixed etiologies[21]. Its low cost and wide availability make it an ideal screening test[18]. Despite this, B-mode US has yet to be studied for prediction of liver-related outcomes in MASLD.

Artificial intelligence encompasses multiple techniques, including machine learning (ML), neural networks, deep learning, and natural language processing[22]. ML has been applied to develop diagnostic and risk-prediction models for chronic liver disease severity [22–24]. ML methods using non-invasive serum markers have been developed to identify MASH and fibrosis stage in chronic liver disease[22,25–28]. Use of ML to predict steatosis using US is an area of interest[29]. Han et al. describe use of a neural network technique to analyze raw radiofrequency data available from US analysis in MASLD patients to quantify hepatic steatosis; previous methods using B-mode data were limited to qualitative analysis[30]. ML has also been applied to MASLD histopathology allowing for quantitative assessment to monitor disease progression[31,32].

In this study, we aimed to investigate the role of ML algorithms, using simple serum-based non-invasive tests (NIT), including FIB-4, NFS and APRI, combined with B-mode US, for non-invasive prediction of (1) all liver-related outcomes (hepatic decompensation, hepatocellular carcinoma (HCC), liver transplant, and/or liver-related death), (2) hepatic decompensation, (3) HCC, (4) liver-related mortality, and (5) advanced MASLD, including F3/4 fibrosis and "fibrotic" MASH.

Methods

Study design and population

This was a retrospective, observational cohort study of biopsy-proven MASLD patients from two Canadian tertiary-care centres (University Health Network, Toronto; McGill University Health Centre, Montreal) between January 1, 2010–July 1, 2021. Inclusion criteria were: (1) age $\geq$ 18 at time of biopsy, (2) histologic diagnosis of MASLD, and (3) ≥ 2 months of follow-up. Exclusion criteria were: (1) alternate causes of chronic liver disease or steatosis (viral hepatitis, significant alcohol use (women – > 14 units/week, men – > 21 units/week) (1 unit = 12 oz (oz) 5% beer, 1.5 oz 40% liquor, 5 oz glass of 12% wine), steatogenic medications), (2) non-HCC malignancy within the past 5 years, (3) immunosuppression within the past 3 years, (4) Human Immunodeficiency Virus, (5) inadequate liver biopsy (< 10 mm or based on pathologist assessment), and (6) hepatic decompensation (ascites, jaundice, hepatic encephalopathy, variceal bleed) or HCC at time zero. Anthropometric data, bloodwork, and B-mode US data were included if available within ± 6 months of liver biopsy. A six month interval was selected as minimal histologic changes in MASLD are expected during this timeframe, and to allow for inclusion of a greater proportion of imaging data for our study cohort.

Clinical and imaging data acquisition

Patient level data were collected from our electronic medical record. Baseline clinical parameters included age, gender, comorbidities, and anthropometrics (height, weight, body mass index (BMI)). Laboratory data included complete blood count, electrolytes, creatinine, liver enzymes AST, alanine aminotransferase (ALT), alkaline phosphatase, and liver function tests (bilirubin, International Normalized Ratio, albumin).

B-mode US reports were collected for each subject. Two investigators (MF, MS) reviewed and recorded variables of interest including degree of hepatic steatosis, liver/spleen size, liver nodularity, and features of portal hypertension.

The following outcomes were determined for all patients: (1) hepatic decompensation, (2) HCC, (3) liver transplant, (4) liver-related mortality (5) F3-4 fibrosis and (6) fibrotic MASH (NAFLD Activity Score (NAS) $\geq$ 4 and F2-4). Hepatic decompensation was defined as the presence of any of the following: ascites, jaundice, hepatic encephalopathy, or variceal bleeding. All outcomes were determined by review of the electronic medical record (HK, CB, MS), based on physician documentation and/or endoscopy records. Date of final follow-up, and time elapsed from biopsy to each outcome/decompensating event was recorded for all patients. Liver transplant and death were considered terminal outcomes for analysis.

Non-invasive prediction of advanced (F3-4) fibrosis

Anthropometric and biochemical data were used to calculate scores for serum-based NIT for MASLD-fibrosis, including NFS, FIB-4, BARD, APRI and AST/ALT ratio. All NIT were calculated using published formulae[33–36].

Histologic analysis

Liver biopsies were assessed by experienced tertiary Hepatology referral center histopathologists at each institution. As such, agreement on histologic scoring systems or consensus on discordant results was not feasible for this retrospective study. Biopsy report summaries were then verified (HK, MS) to ensure alternate causes of chronic liver disease were excluded. MASH and fibrosis were scored using the NASH Clinical Research Network (NASH-CRN) Scoring System[37]. 'Advanced fibrosis' was defined as a CRN score of 3–4 (bridging fibrosis or cirrhosis).

Statistical analysis

Statistical analysis was performed using MedCalc (*MedCalc Software Version 19.0.7, Ostend, Belgium*). Continuous variables were expressed as mean and standard deviation (SD). Ordinal variables were expressed as median and interquartile range (IQR). Quantitative data was assessed using Student's T test. Chi-squared test was used to compare frequency data. Ordinal data was compared using the Mann–Whitney U test. Area under the receiver operating curve (AUROC), as described by DeLong et al.[38], was used to determine diagnostic performance of individual NIT. A p value < 0.05 was considered significant.

ML analysis

Analyses were performed using 'Random Forest' (RF) technique[39]. Repeated randomized stratified k-fold cross-validation was used to perform 10 runs of fivefold cross-validation (i.e. stratified into 80% training, 20% testing per fold). Categorical features were hot encoded. Missing variables were recorded as additional categories for categorical variables, and -1.0 for numerical features. Each outcome variable was separately analyzed. Patients with target variable = NA were excluded. Allowable predictors for each outcome were predetermined. For each experiment a Random Forest model was fit to the training dataset and evaluated on the independent testing set. Model performance was assessed by AUROC, calculated on each testing set. Average area under the curve (AUC) and 95% confidence intervals (CI) are calculated across the 500 runs using the percentile method. Feature importance, representing the contribution of a given variable ('feature') to the model, was estimated using Shapley feature importance with concatenation across the 5 folds and averaging across the 10 repeats (i.e. the average feature importance per patient across all 10 repeats)[40]. Sensitivity and specificity were determined according to the point closest to (0,1) on the AUC curve.

Results

Baseline demographics

Following assessment of study eligibility, 457 patients with biopsy-proven MASLD were included in this study (Supplementary Fig. 1). Overall, 53.8% of patients were male, with mean age (± SD) at time of biopsy 49.2 ± 12.9 years, mean BMI 32.8 ± 7.0 kg/m^2, 31.6% had diabetes, and biopsy prevalence of F3-4 fibrosis was 44.9%. Rate of fibrotic MASH was 51.0%. Patients with F3-4 fibrosis (45%) were generally older, female, with higher BMI, and higher rates of metabolic comorbidities. Rates of smoking did not differ significantly between groups. Patients were well-compensated at time of biopsy, and all NIT differed significantly between F0-2 vs. F3-4 (Table 1). Median duration of follow-up for this study was 71 months (2–170 months).

Baseline demographics for individual cohorts is available in Supplementary Table 1.

Ultrasound features

Overall, the mean liver span was 16.2 ± 2.7 cm, with 86.0% of patients having features of fatty infiltration on US. Compared to F0-2, patients with F3-4 also had higher rates of hepatic nodular contour, lobar redistribution, and greater average spleen size (Table 1).

Liver-related outcomes

Overall, 6.3% (29/457) patients experienced 'liver-related outcomes', defined as development of hepatic decompensation, HCC, transplant, or liver-related death. Patient timeline for occurrence of outcomes was within 10 years. The first-occurring liver-related event was HCC (n = 9), ascites (n = 15), encephalopathy (n = 4), jaundice (n = 1), variceal bleeding (n = 3), and other/unspecified (n = 1). A total of one transplant occurred during the study period. The median time to development of a 'liver-related outcome' was 38 months (IQR 17.9–60.4 months). Hepatic decompensation occurred in 4.8% (22/457) within 114 months (median (IQR) 40 (22.9–55.8) months). HCC occurred in 2.0% (9/457) of patients during the study period within 91 months (median 22 (11.4–84.7) months). Liver-related mortality occurred in 1.1% (5/457) within 117 months (median 74 (35.0–79.0) months. Liver-related outcomes occurred more frequently among patients with F3-4 on baseline liver biopsy vs F0-2 (11.2% vs 2.4%; p = 0.0001). There were no differences in incidence of HCC between F0-2 (n = 3) and F3-4 (n = 6) (1.3% vs 2.9%; p = 0.19) (Supplementary Table 2).

Machine learning models

ML models were generated for prediction of liver-related outcomes, hepatic decompensation, HCC, liver-related mortality, F3-4 fibrosis, and fibrotic MASH. Each model was created using a set of pre-selected allowable predictive variables. These included variables listed in Table 1. A complete list of included predictive variables for each model is included in Supplementary Table 3. Variables were further subdivided based on description as 'clinical', 'biochemical' or 'radiographic' predictors, to determine the impact of each variable sub-class on the model accuracy, with importance reported as the summed importance across all features in each sub-class.

1. Liver-Related Outcomes

Patient characteristic	Combined (n = 457)	F0-2 (n = 252)	F3-4 (n = 205)	*p*
Fibrosis stage	F0-2—55.1% (252) F3-4—44.9% (205)	F0—31.0% (n = 78) F1—36.1% (n = 91) F2—32.9% (n = 83)	F3—49.8% (n = 102) F4—50.2% (n = 103)	–
% Males (n)	53.8% (246)	63.1% (159)	42.4% (87)	* < 0.0001
Age (mean ± SD) [years]	49.2 ± 12.9	45.9 ± 12.7	53.2 ± 12.1	* < 0.0001
% Hypertension (n)	32.7% (147/450)	24.4% (60/246)	42.6% (87/204)	* < 0.0001
% Diabetes (n)	31.6% (142/449)	19.6% (48/245)	46.1% (94/204)	* < 0.0001
% Smoking (reported) (n = 314)	25.4% (105/414)	27.8% (61/220)	22.7% (44/194)	0.2348
BMI (mean ± SD) [kg/m^2]	32.8 ± 7.0 (271)	31.4 ± 6.2 (137)	34.2 ± 7.4 (134)	*0.0008
AST (mean ± SD) [U/L]	60.2 ± 55.0 (395)	50.6 ± 48.6 (222)	72.6 ± 60.2 (173)	*0.0001
ALT (mean ± SD) [U/L]	86.2 ± 74.1 (404)	81.5 ± 66.7 (227)	92.2 ± 82.4 (177)	0.1500
NAS (median, IQR)	4 (3–5) (453)	4 (2–5) (249)	4 (4–5) (204)	* < 0.0001
MELD (median, IQR)	7 (6–8) (371)	6 (6–7) (210)	7 (7–8) (161)	* < 0.0001
NaMELD (median, IQR)	8 (7–10) (275)	8(6–9) (164)	9 (7–10) (111)	*0.0230
NFS (n = 163)	-1.41 ± 2.13 (231)	-2.26 ± 1.63 (124)	-0.42 ± 2.22 (107)	* < 0.0001
FIB-4 (n = 303)	1.81 ± 1.75 (390)	1.26 ± 1.41 (220)	2.52 ± 1.89 (170)	* < 0.0001
BARD (median, IQR) (n = 182)	2 (1–3) (260)	1 (1–2) (136)	2.5 (1–4) (124)	* < 0.0001
APRI (n = 303)	0.77 ± 0.76 (392)	0.59 ± 0.67 (222)	1.01 ± 0.80 (170)	* < 0.0001
AST/ALT ratio (n = 305)	0.80 ± 0.42 (393)	0.71 ± 0.40 (220)	0.92 ± 0.41 (173)	* < 0.0001
Liver Span (mean ± SD) [cm] (n = 148)	16.2 ± 2.7 (149)	15.9 ± 2.5 (74)	16.6 ± 2.9 (75)	0.1169
% Fatty Liver (n = 161)	86.0% (221/257)	88.4% (122/138)	83.2% (99/119)	0.2320
% Hepatic Nodularity Contour (n = 161)	21.2% (46/217)	7.1% (8/112)	36.2% (38/105)	* < 0.0001
% Hepatic Vein Nodularity (n = 161)	8.2% (13/158)	4.9% (4/82)	11.8% (9/76)	0.1158
% Lobar Redistribution (n = 161)	8.2% (13/158)	2.4% (2/82)	14.5% (11/76)	*0.0058
% Patent Para-Umbilical Vein (n = 161)	1.3% (2/158)	0% (0/82)	2.5% (2/76)	0.1510
Spleen Length (mean ± SD) [cm] (n = 153)	12.0 ± 2.7 (220)	11.2 ± 2.1 (111)	12.7 ± 3.1 (109)	* < 0.0001

Table 1. Baseline clinical characteristics, combined cohort. n, number of patients; *– $p < 0.05$; p calculated using Mann–Whitney test. BMI, body mass index; AST, aspartate aminotransferase; ALT, alanine aminotransferase; NAS, NAFLD Activity Score; MELD, Model for End-Stage Liver Disease; NaMELD, MELD sodium score; NFS, NAFLD Fibrosis Score; APRI, AST to Platelet Ratio Index; US, ultrasound; SD, standard deviation; IQR, interquartile range; US, ultrasound.

The AUC for prediction of liver-related outcomes (hepatic decompensation, HCC, liver transplant, and/or liver-related death) using the ML model was 0.87 (95% CI 0.76–0.96), sensitivity 0.77, specificity 0.78 (Supplementary Table 4). Feature importance and directional association for each allowable predictive variable is shown in Fig. 1a and b. The top five features with respect to contribution to AUC were AST:ALT ratio, FIB-4, age, platelet count, and APRI. Overall, biochemical variables had the greatest feature importance, as compared to clinical and imaging-based variables (Supplementary Fig. 2a).

When histologic F3-4 was included as a variable, AUC for prediction of liver-related outcomes was unchanged (0.86 (95% CI 0.70–0.97)). Sensitivity was 0.82 and specificity 0.75. The top five predictive features were also unchanged. Biochemical variables remained most important contributors to overall AUC, followed by clinical and imaging-based variables. Biopsy F3-4 had lower feature importance than spleen length and simple NITs for liver-related outcomes (Supplementary Fig. 2b, c).

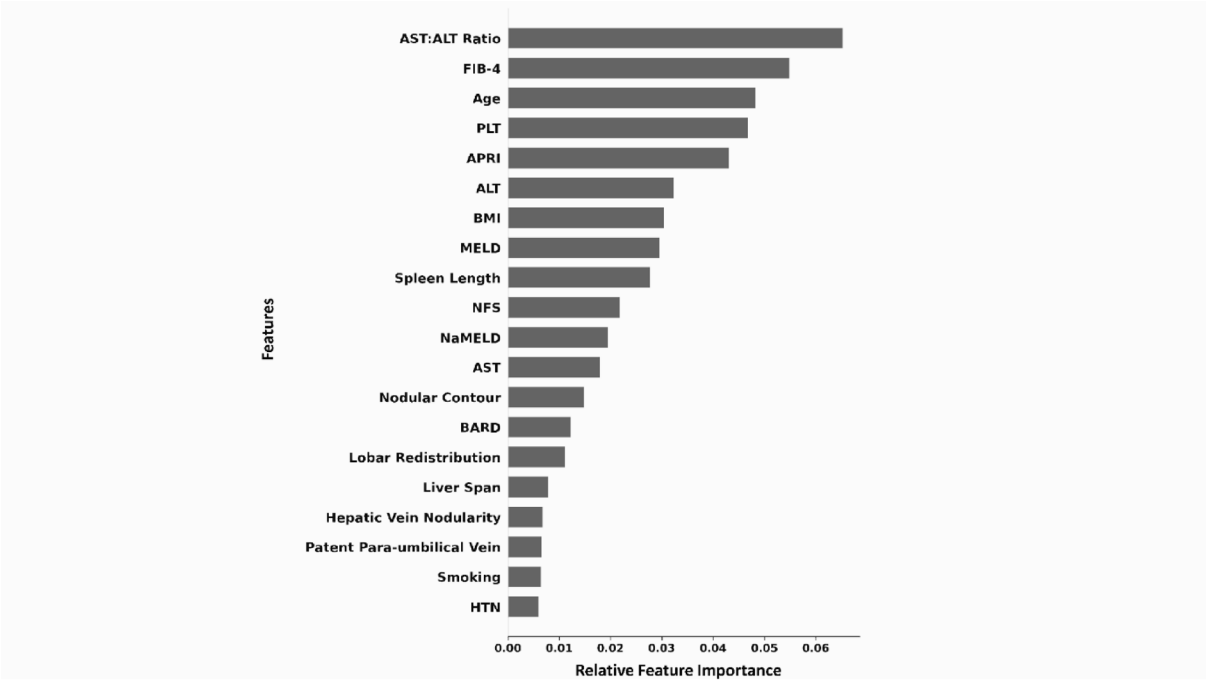
2. Hepatic Decompensation

The AUC for prediction of hepatic decompensation was 0.90 (95% CI 0.79–0.98), sensitivity 0.88, specificity 0.78 (Supplementary Table 5). Feature importance and directional association for each allowable predictive variable is shown in Fig. 2a and b. Top five performing features were FIB-4, AST:ALT ratio, APRI, age, and platelet count, respectively. Overall, biochemical features were most important to overall AUC, compared to clinical and imaging-based features (Supplementary Fig. 3a).

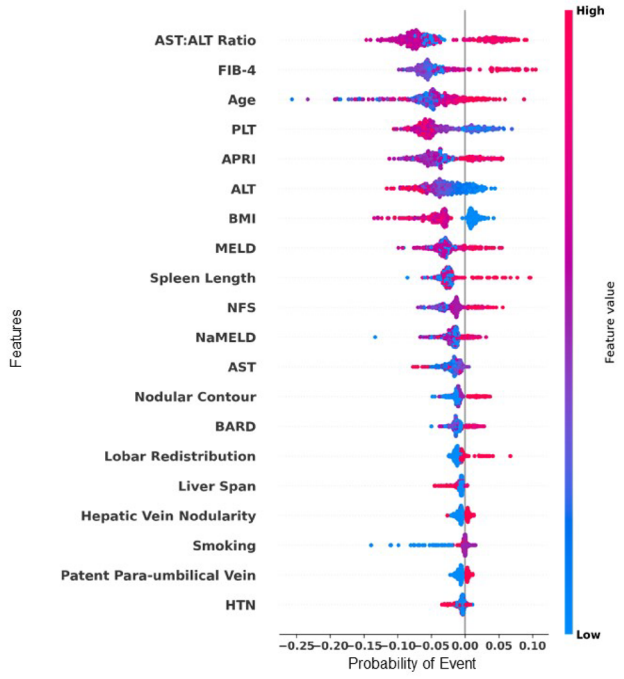
When histologic F3-4 was included as a variable, AUC for prediction of F3-4 was unchanged (0.90 (95% CI 0.81–0.98). Sensitivity and specificity were unchanged, 0.88 and 0.78 respectively. Top performing features were also unchanged. Biochemical variables again outperformed clinical and imaging-based variables. (Supplementary Fig. 3b, c).

3. HCC

The AUC for prediction of HCC using the ML model was 0.72 (95% CI 0.29–0.96), sensitivity 0.69, specificity 0.71 (Supplementary Table 6). Feature importance and directional association for each allowable predictive

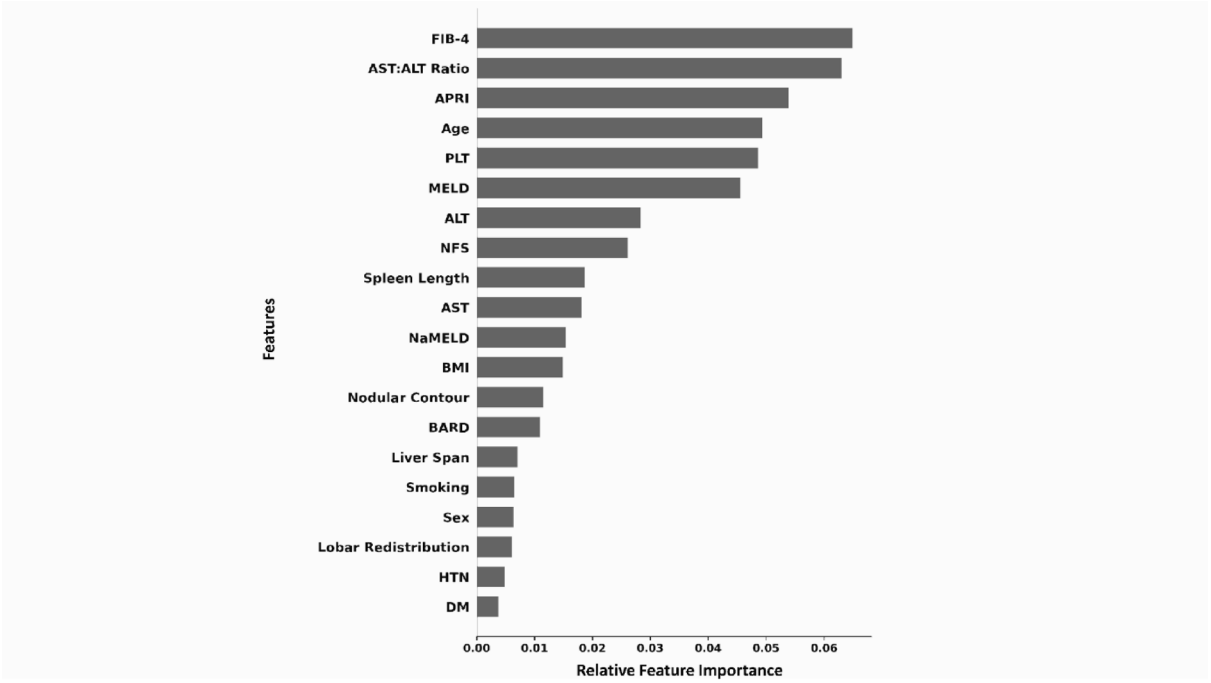


(a)

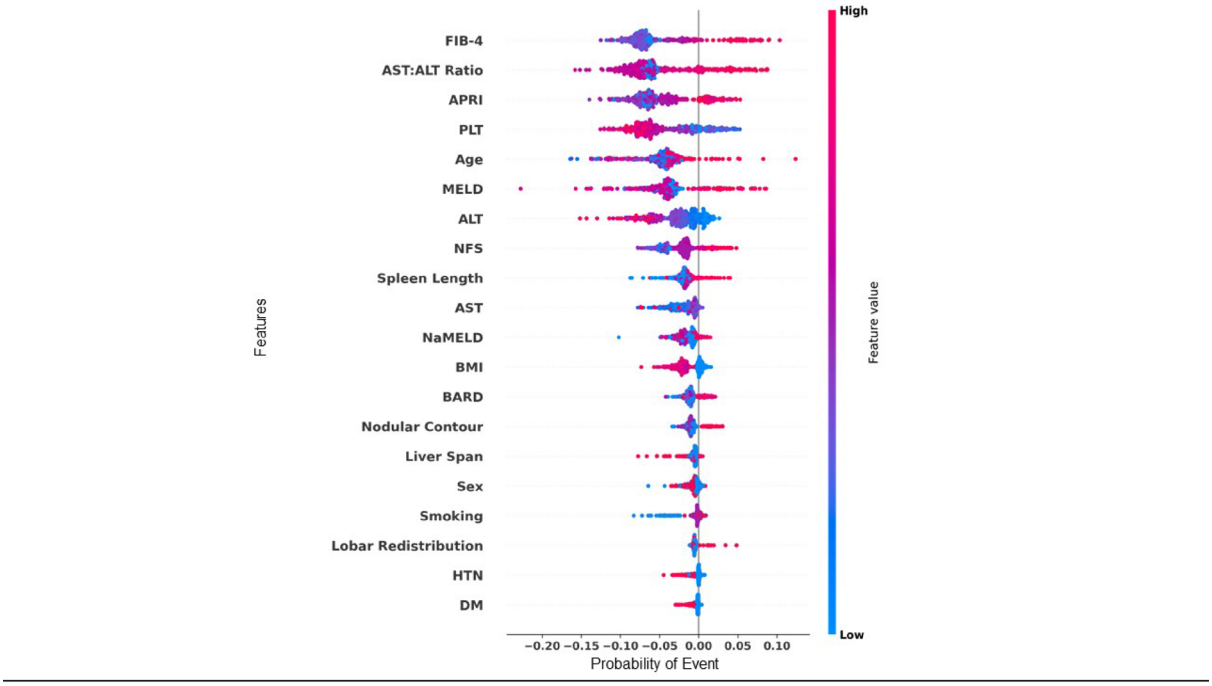


(b)

Fig. 1. ML feature importance for outcome 'liver-related outcomes'. (**a**) Shows the relative feature importance of allowable predictive variables included in the machine learning model for prediction of outcome 'Liver-Related Outcomes', a composite outcome including hepatic decompensation, hepatocellular carcinoma, liver transplant, and liver-related mortality; Categorical variables for directional change to high probability and high feature value include smoking (yes); (**b**) Illustrates directional change from left-to-right for a high probability of event, and blue-to-red representing transition from low to high feature value.



(a)



(b)

Fig. 2. ML feature importance for outcome 'hepatic decompensation'. (**a**) Shows the relative feature importance of allowable predictive variables included in the machine learning model for prediction of outcome 'Hepatic Decompensation, a composite outcome including ascites, jaundice, hepatic encephalopathy, and variceal bleeding. Categorical variables for directional change to high probability and high feature value include Smoking (yes), Sex (male), and presence of Lobar Redistribution (yes), Hypertension (HTN), and Diabetes mellitus (DM); (**b**) Illustrates directional change from left-to-right for a high probability of event, and blue-to-red representing transition from low to high feature value.

variable is shown in Fig. 3a and b. Age, BMI, APRI, ALT, and platelet count were the top five predictive features, respectively. Overall, biochemical and clinical variables had the greatest feature importance for prediction of HCC, as compared to imaging-based variable. (Supplementary Fig. 4a).

As expected, based on the incidence of HCC relative to F0-2, when histologic F3-4 was included as a variable, AUC for prediction of HCC was essentially unchanged (0.74 (95% CI 0.42–0.95)). Top four performing features were unchanged, and AST:ALT ratio overtook PLT count. Sensitivity was improved at 0.78, with specificity 0.65. Overall, biochemical and clinical variables had greatest feature importance for prediction of HCC as compared to imaging-based variables. (Supplementary Fig. 4b, c).

4. Liver-Related Mortality

For prediction of liver-related mortality, AUC using the ML model was 0.79 (95% CI 0.31–0.98), sensitivity 0.75, specificity 0.80 (Supplementary Table 7). Feature importance and directional association for each allowable predictive variable is shown in Fig. 4a and b. The top five features contributing to AUC were ALT, platelets, AST:ALT ratio, AST and NFS, respectively. For grouped variables, biochemical variables outperformed imaging-based and clinical variables. (Supplementary Fig. 5a).

When histologic F3-4 was included as a variable, AUC for prediction of liver-related mortality was marginally increased (0.82 (95% CI 0.32–0.99)). Sensitivity improved to 0.81, and specificity was essentially stable at 0.81. F3-4 as a variable became an important feature. With respect to individual feature importance, top five features included ALT, platelet count, F3-4, AST:ALT ratio, and AST. There was no difference in grouped feature importance. Biochemical variables outperformed clinical and imaging-based variables Supplementary Figs. 5b, c).

5. Advanced Fibrosis

For prediction of advanced F3-4 fibrosis, AUC from the ML model was 0.83 (95% CI 0.76–0.87), sensitivity 0.79, specificity 0.70 (Supplementary Table 8). Feature importance and directional association for each allowable predictive variable is shown in Fig. 5a and b. FIB-4 had the greatest contribution towards overall AUC, followed by age, AST:ALT ratio, APRI and AST. Overall, biochemical features had the greatest contribution to AUC, followed by clinical, then imaging-based features. (Supplementary Fig. 6).

6. Fibrotic MASH

AUC for prediction of Fibrotic MASH histology using the ML model was 0.74 (95% CI 0.65–0.85), sensitivity 0.71, specificity 0.66 (Supplementary Table 9). Feature importance and directional association is shown in Fig. 5c and d. AST had the greatest feature contribution, followed by age, BMI, sex, and diabetes. Overall, clinical and biochemical features contributed most to overall AUC, followed by imaging-based features (Supplementary Fig. 7).

Simple biochemical markers for prediction of liver-related outcomes

Performance of individual NIT as compared to ML algorithms for prediction of outcomes are summarized in Table 2 and Supplementary Tables 4–9.

For prediction of liver-related outcomes, AUROC ranged from 0.75 (APRI) to 0.88 (NFS). For prediction of hepatic decompensation, AUROC ranged from 0.77 (BARD) to 0.93 (NFS). For HCC, AUROC were lower, ranging from 0.61 (APRI) to 0.80 (BARD). For liver-related mortality, AUROC ranged from 0.66 (APRI) to 0.97 (NFS). For prediction of F3-4, FIB-4 > 1.32 had the highest AUROC of 0.78. For prediction of fibrotic MASH, APRI > 0.46 performed best (AUROC 0.71).

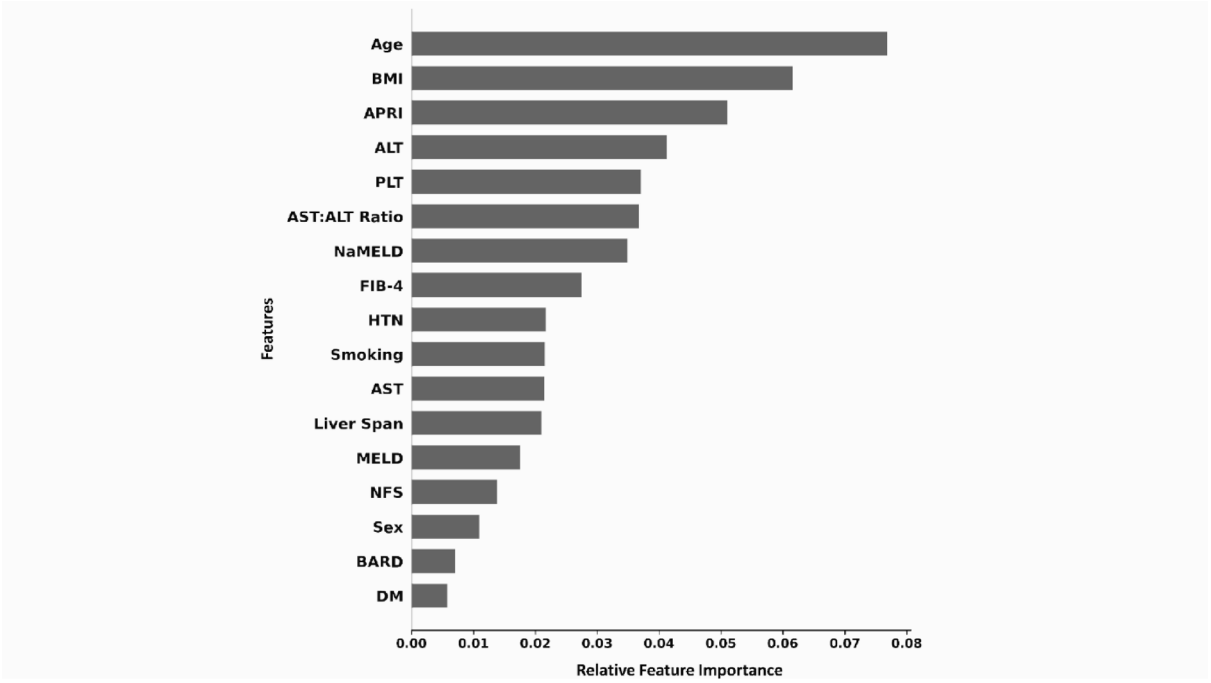
Biopsy F3-4 and clinical outcomes

AUROC was determined for each clinical outcome based on histologic F3-4 alone. AUROC for prediction of liver-related outcomes was 0.68 (95% CI 0.64–0.73), hepatic decompensation 0.69 (0.65–0.74), HCC 0.61 (0.57–0.66), and liver-related mortality 0.78 (0.74–0.82). ML algorithms had higher AUROC for all clinical outcomes except for liver-related mortality. Table 2.

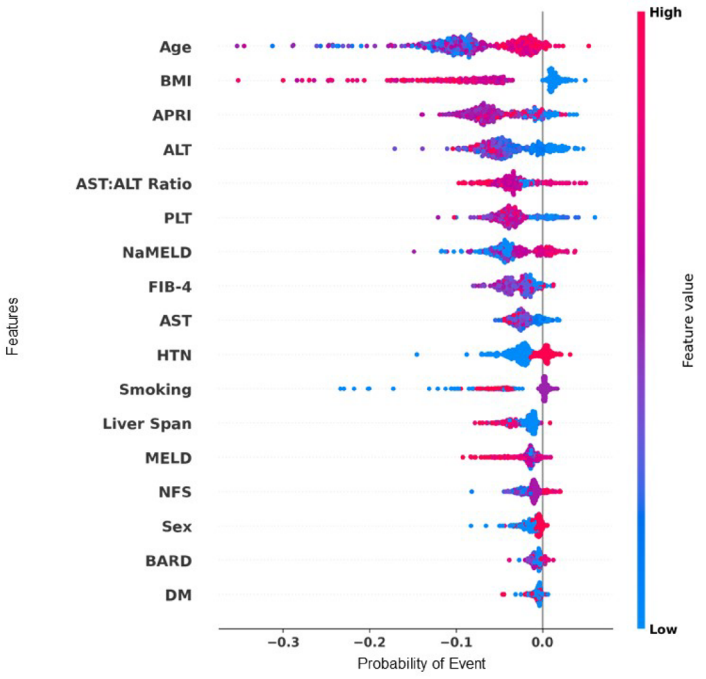
Discussion

Our study demonstrates the utility of ML for prediction of liver-related outcomes in a cohort of biopsy-proven MASLD patients, using B-mode US parameters and clinical data. ML models combining simple, readily available clinical, biochemical, and US-based variables predicted liver-related outcomes and hepatic decompensation with good accuracy, matching individual NIT, with AUC approaching 0.9. ML algorithms improved accuracy for prediction of liver-related outcomes such as hepatic decompensation and HCC as compared to histologic F3-4. Compared to simple NIT, ML algorithms had lower diagnostic performance for less frequently occurring outcomes such as HCC and liver-related mortality; however, accuracy was improved for prediction of F3-4 and fibrotic MASH.

Our ML models identified biochemical variables as having greatest feature importance for both liver-related outcomes and hepatic decompensation. FIB-4 had greatest feature importance for prediction of F3-4, in keeping with its validated use for prediction of advanced fibrosis. Of all clinical variables, age had greatest feature contribution to AUC for each ML outcome. Imaging based features had the lowest contribution to AUC for predicting outcomes. Of all imaging-based variables, spleen length performed best for prediction of liver-related outcomes, including for outcomes of hepatic decompensation, HCC and liver-related mortality, along with F3-4 on biopsy. Fibrotic MASH, uniquely, had significant feature importance contribution from clinical variables

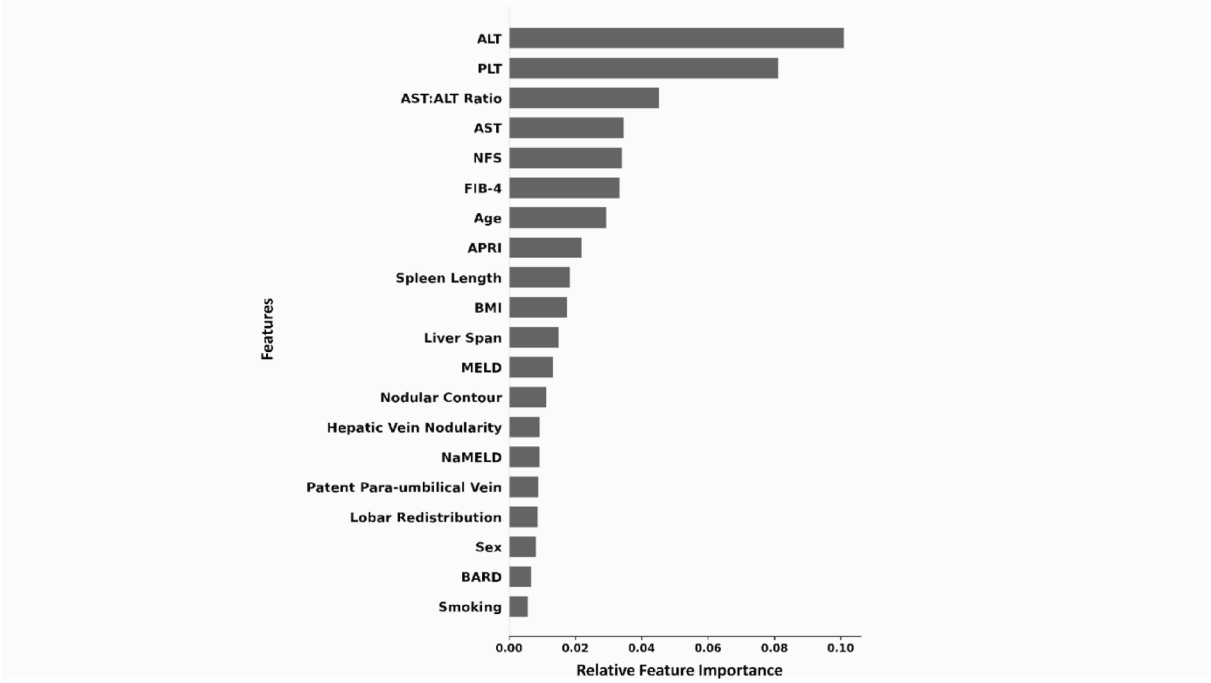


(a)

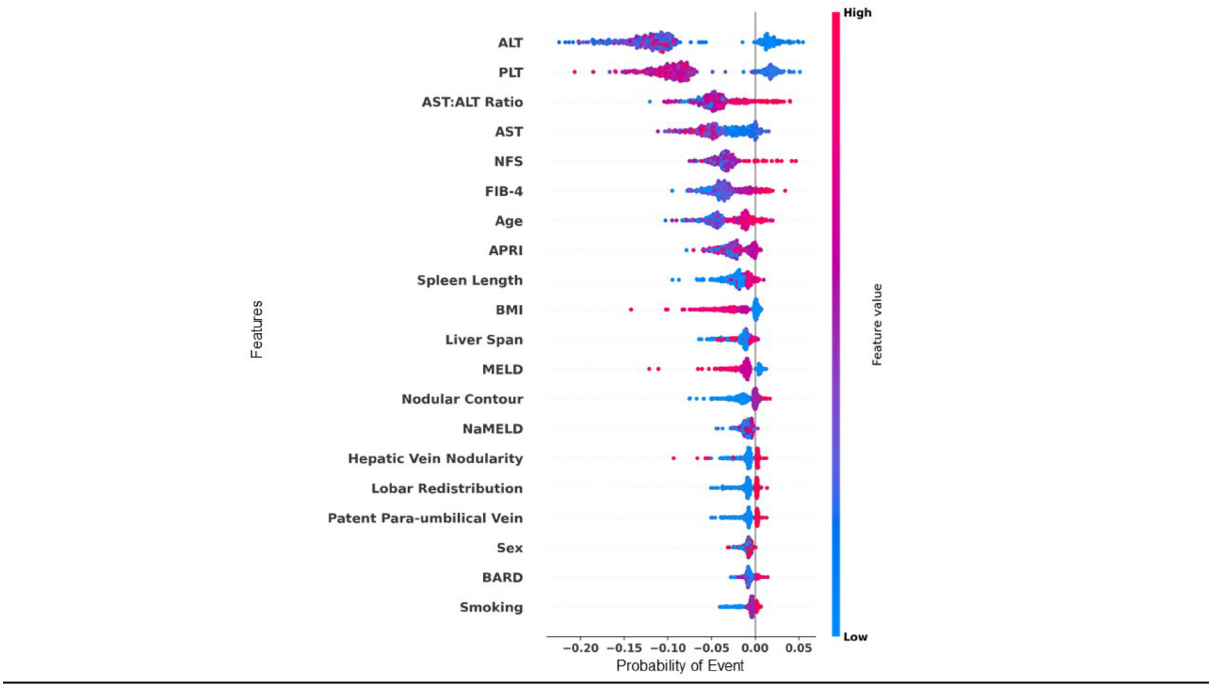


(b)

Fig. 3. ML Feature Importance for Outcome 'Hepatocellular Carcinoma'. (**a**) Shows the relative feature importance of allowable predictive variables included in the machine learning model for prediction of outcome 'Hepatocellular Carcinoma'; Categorical variables for directional change to high probability and high feature value include Smoking (yes), Sex (male), and presence of Hypertension (HTN), and Diabetes mellitus (DM); (**b**) Illustrates directional change from left-to-right for a high probability of event, and blue-to-red representing transition from low to high feature value.



(a)



(b)

Fig. 4. ML feature importance for Outcome 'Liver-Related Mortality' (**a**) shows the relative feature importance of allowable predictive variables included in the machine learning model for prediction of outcome 'Liver-Related Mortality'; Categorical variables for directional change to high probability and high feature value include Sex (male), smoking (yes) and presence of Nodular Contour, Patent Para-umbilical vein, Hepatic Vein Nodularity, Lobar Redistribution; (**b**) illustrates directional change from left-to-right for a high probability of event, and blue-to-red representing transition from low to high feature value.

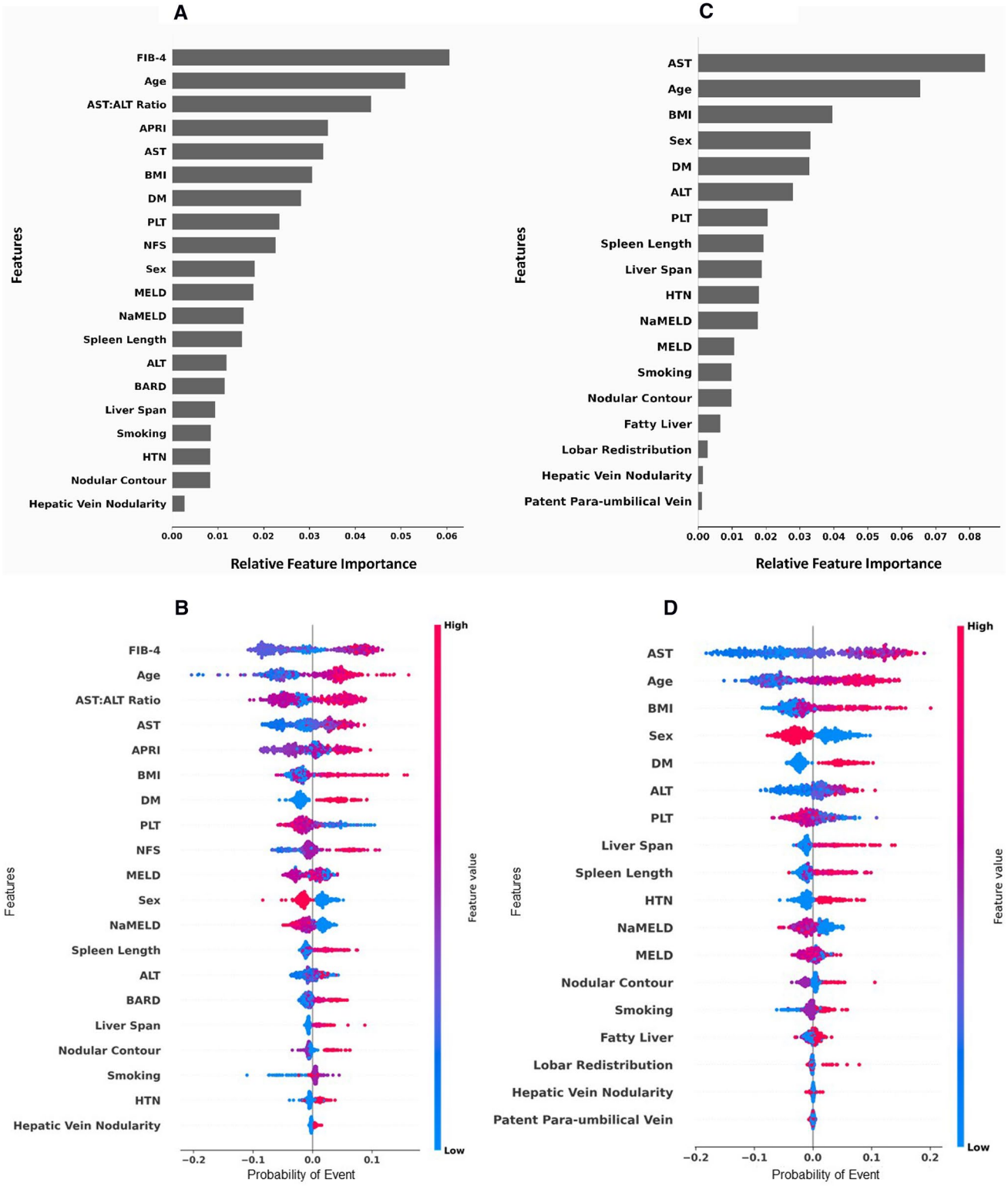


Fig. 5. ML Feature Importance for Outcomes 'F3-4' and 'Fibrotic MASH'. Relative feature importance of allowable predictive variables included in the machine learning model for prediction of outcome 'F3-4' (**A**) and 'Fibrotic MASH' (NAS ≥ 4 and F2-4) (**C**). Categorical variables for directional change to high probability and high feature value include Sex (male), smoking (yes) and presence of Hypertension (HTN), Diabetes mellitus (DM), Nodular Contour, Patent Para-umbilical vein, Hepatic Vein Nodularity, and Lobar Redistribution; Corresponding directional change from left-to-right for a high probability of event, and blue-to-red representing transition from low to high feature value are shown in (**B**) and (**D**) respectively. MASH, Metabolic dysfunction-associated steatohepatitis; NAS, NAFLD activity score.

NIT	Outcomes					
	Liver-related outcomes	Hepatic decompensation	HCC	Liver-related mortality	F3-4	Fibrotic MASH
FIB-4	0.84 (0.80–0.88)	0.88 (0.85–0.91)	0.69 (0.64–0.74)	0.88 (0.84–0.91)	0.78 (0.74–0.82)	0.68 (0.63–0.73)
NFS	0.88 (0.83–0.92)	0.93 (0.88–0.96)	0.70 (0.64–0.76)	0.97 (0.94–0.99)	0.76 (0.70–0.81)	0.62 (0.55–0.68)
APRI	0.75 (0.71–0.79)	0.80 (0.76–0.84)	0.61 (0.56–0.65)	0.66 (0.61–0.71)	0.74 (0.70–0.79)	0.71 (0.66–0.75)
AST:ALT	0.84 (0.79–0.87)	0.85 (0.81–0.88)	0.75 (0.70–0.79)	0.91 (0.88–0.94)	0.71 (0.66–0.75)	0.60 (0.55–0.65)
BARD	0.78 (0.73–0.83)	0.77 (0.71–0.82)	0.80 (0.75–0.85)	0.82 (0.77–0.86)	0.71 (0.65–0.76)	0.57 (50–0.63)
Biopsy F3-4	0.68 (0.64–0.73)	0.69 (0.65–0.74)	0.61 (0.57–0.66)	0.78 (0.74–0.82)	–	–
ML Algorithm Combined Cohort	0.87 (0.76–0.96)	0.90 (0.79–0.98)	0.72 (0.29–0.96)	0.79 (0.31–0.98)	0.83 (0.76–0.87)	0.74 (0.65–0.85)

Table 2. AUROC for Non-Invasive Tests for Prediction of Outcomes and F3-4 Fibrosis. NIT, non-invasive serum-based tests; HCC, hepatocellular carcinoma; NFS, NAFLD Fibrosis Score; APRI, AST to Platelet Ratio Index; AST:ALT, AST to ALT ratio.

including age, BMI, sex, and diabetes status, providing validation of known important clinical risk factors for MASH.

As a feature, histologic F3-4 did not change the AUC for any outcome, although it became a top predictive feature when included as a variable in the ML model to predict liver-related mortality, in keeping with its known association with this outcome[4,6–8].

Individual simple NIT accurately predicted adverse liver related outcomes in our cohort. FIB-4, NFS, and AST:ALT ratio predicted liver-related outcomes and hepatic decompensation with AUC 0.84–0.93. NIT, apart from BARD, did not perform as well for prediction of HCC. Overall rates of HCC were low and occurred with similar frequency between F0-2 and F3-4 groups, which may have accounted for reduced performance. Interestingly, all 9 HCC patients in our cohort had BARD score > 2. The BARD score contains four of the top five performing features identified using our ML algorithm, including age, BMI, ALT, and AST:ALT ratio, likely accounting for its performance. Except for APRI, other simple NIT could accurately predict liver-related mortality, with AUC > 0.9 and 0.95 for AST:ALT ratio and NFS, respectively. Interestingly, ML algorithms had higher AUC as compared to individual NIT for prediction of both F3-4 and fibrotic MASH, likely due to inclusion of unweighted variables such as age, BMI, and diabetes status, which are important clinical predictors of advanced fibrosis, and as also identified by ML feature importance analysis. Additional performance characteristics of our ML algorithms, including sensitivity, specificity, positive/negative likelihood ratios, were generally comparable to individual NIT.

Prior studies of serum-based NIT, including NFS, FIB-4, and APRI have shown mixed results in prediction of both liver-related outcomes and mortality[5,13,41–54]. In a post-hoc analysis of outcomes data from four multicenter clinical trials of Simtuzumab and Selonsertib[51], higher baseline NIT scores were associated with poorer outcomes, with Enhanced Liver Fibrosis Test (ELF) (Siemens Healthineers, Erlangen, Germany), NFS and FIB-4 performing best. This study, however, was limited to F3-4 disease, with a highly selected clinical trial cohort with F3-4 and only had a median follow-up of 16 months.

A retrospective multicenter cohort study with 594 patients with baseline liver biopsy and F3-4 prevalence 30.3%, noted a composite of end-stage liver disease related complications (ascites, spontaneous bacterial peritonitis, hepatorenal syndrome, varices, variceal bleeding, liver failure, encephalopathy) and HCC in 42 patients after a median of 2.2 years. Both FIB-4 and VCTE predicted the composite outcome of liver-related events (Harrell's C index 0.775–0.88). Despite a smaller patient cohort, our study had a longer median follow-up time with standard defined hepatic decompensation and comparable outcome rates. Although VCTE was included, additional NIT and simple clinical and imaging-based variables were not assessed.

A multicenter study of 1773 patients from the NASH Clinical Research Network followed for 4 years indicated MASLD F3-4 was associated with increased risk of liver-related complications and death but included 19 patients with history of hepatic decompensation[5]. Our cohort included patients of similar median age and BMI, higher F3-4 prevalence without prior history of liver-decompensation and followed for a longer period. As such, we noted higher rates liver-related mortality (1.1%), HCC (2%), and higher proportion (11.2% versus 6.6% for NASH CRN) of F3-4 patients with subsequent decompensation. Higher event rates allowed us to examine ML models for liver-related outcomes.

Liver stiffness measure determined by VCTE has been shown to predict mortality and liver-related outcomes in MASLD[4,54,55]. VCTE is not routinely available in primary care clinical practice and is often associated with additional out-of-pocket cost to the patient. Although US-based variables generally had lower feature-importance as compared to NIT and clinical variables, this study is one of the first to use readily available US findings in a predictive model for MASLD outcomes. Standard B-mode US is routinely obtained in most patients with elevated liver tests and suspected chronic liver disease, providing an initial diagnosis of fatty liver disease. Several US parameters such as surface nodularity, portal vein flow, or spleen size may indicate advanced chronic liver disease, but have not been previously evaluated in comparison to other NIT for the diagnosis of outcomes in advanced MASLD fibrosis. US-based parameters such as spleen diameter-to-platelet ratio, in combination with LSM by VCTE, has been successfully used in hepatitis B virus-related cirrhosis to predict high risk varices, variceal bleeding, and hepatic decompensation, suggesting they may still have a role for risk stratification in MASLD[56–59].

ML techniques are now being increasingly investigated for their predictive capabilities in hepatology and MASLD[22–32,60,61]. Chang et al. demonstrated the utility of RF using simple clinical and biochemical markers for prediction of MASLD fibrosis and fibrotic MASH in a multicentre population of 1370 biopsy-proven MASLD patients[60]. In this study, RF had greater AUC compared to logistic regression, artificial neural network techniques, and other standard non-invasive techniques. Our study, however, represents the first to demonstrate the role of ML for the prediction of adverse liver-related outcomes in MASLD, as well as the first to assess the role of B-mode US for this purpose.

Our study does have several strengths. We included a large population of over 450 patients with biopsy-proven MASLD from two tertiary Canadian centers. This is one of the first studies to assess readily available US-based variables for prediction of liver-related outcomes in MASLD. Our study also used novel ML based tools for creation of predictive models. ML algorithms allow for inclusion and assessment of multiple predictive variables simultaneously. RF in particular are known as non-parametric models in that they do not assume any particular distribution or prior relationship (e.g. linearity) on the variables enabling the discovery of more nuanced relationships than support vector machines or linear classifiers. ML models permit direct selection of allowable predictors, without subjective preselection, aiding to minimize bias[26].

Our study does have limitations. Data were collected retrospectively to maximize the number of identified outcomes. Outcome studies of MASLD patients often require long follow-up periods to obtain enough outcomes, due to the natural history of the disease, long asymptomatic period, and low rates of outcomes amongst patients without advanced fibrosis[2]. Biopsies were performed 'for cause' and were not per protocol. US were performed at a tertiary centre and read by radiologists with expertise in hepatobiliary imaging, potentially limiting extrapolation to US reported in the community. Mortality may be incompletely captured due to inability to link to a local death register. There were relatively few outcomes over a median of 38 months, and we were unable to perform time-to-event analysis to interrogate several covariates of interest. As such, we were not able to develop a time-dependent prognostic model for clinical use. Our study aimed to capture the full range of the disease progression in the context of real-world clinical data, where outcomes can vary considerably over time. Despite the absence of fixed time windows, the observed time intervals in our study does provide valuable insight into the progression of liver-related events in this cohort. The time from biopsy-to-event was used for analysis, and even though we did not define a specific pre-determined time window for the outcomes, we believe this approach is more reflective of the real-world progression of MASLD. The minimum follow-up duration in our study was 2 months, and some patients may have been right-censored as they did not experience an event within that period. However, the RF algorithm, through repeated randomized cross-validation (as used in our study), can naturally accommodate censored data without requiring exclusion of these patients based on their follow-up duration. The VCTE was not included in this study as many patients did not have available LSM performed within 6 months of liver biopsy. Our study does have a high rate of F3-4 which limits application of results to lower prevalence cohorts, and we did not further assess sequential NIT or other proprietary NIT for liver-related outcomes.

Although ML algorithms used established cross-validation methods, it will be important to validate these results in external cohorts, including those with lower prevalence of F3-4, with a goal of determining a simple clinical algorithm based on top feature importance variables to predict adverse liver-related outcomes in MASLD patients. Identification of predictive variables for adverse liver-related outcomes will help avoid need for liver biopsy and better risk stratify patients at time of referral to a specialist/hepatologist. Patients at high risk for decompensation and HCC can be identified early in their disease course, and appropriately referred to specialist care for close monitoring. Unfortunately the number of clinical outcomes in our retrospective study was relatively small (n = 29 liver-related outcomes). We selected binary outcomes, as time-to-event modeling typically requires an order of magnitude higher number of events, and thus much larger datasets followed prospectively for several years, in order to observe the number of outcomes required to effectively train ML models. Use of ML-techniques to predict outcomes in a time-dependent manner, as opposed to binary risk prediction, along with incorporating time-varying covariates, and nested k-fold cross-validation to further enhance the robustness of the model evaluation process, will be important for developing MASLD clinical risk predictors in the future[62]. Increased availability of VCTE and other US-based elastography as a point-of-care test will enable development of models using this variable for prediction of outcomes[14,17,63]. MRE, alone or in combination with NIT, has also been shown to accurately predict adverse liver-related outcomes, including decompensation and death in MASLD patients[63–65]. Perhaps future inclusion of MRE in predictive models may improve their accuracy for prediction of liver-related outcomes and further help to identify high-risk patients, but cost and access will remain a limitation.

In conclusion, ML algorithms based on a combination of simple NIT, clinical, and standard B-mode US-based variables accurately predicted adverse liver-related outcomes and hepatic decompensation in this tertiary center cohort of patients with biopsy-proven MASLD. FIB-4 and AST:ALT ratio were the highest ranked NIT based on feature-importance, while age was the most important clinical variable. US-based parameters were not substantial predictors of clinical outcomes in this study. External validation of these results will be important in MASLD cohorts with lower prevalence of F3-4.

Data availability

There are no sponsors that played a role in the study design, data collection or analysis, interpretation of data, in the writing of the report, or in the decision to submit the manuscript for publication. All data, analytic methods and study materials will be made available to other researchers upon request (contact corresponding author HMK; heather.kosick@gmail.com).

Received: 12 January 2024; Accepted: 26 June 2025

Published online: 08 July 2025

References

1. Younossi, Z. et al. Global burden of NAFLD and NASH: Trends, predictions, risk factors and prevention. *Nat. Rev. Gastroenterol. Hepatol.* **15**(1), 11–20. https://doi.org/10.1038/nrgastro.2017.109 (2018).
2. Chalasani, N. et al. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. *Hepatology* **67**(1), 328–357. https://doi.org/10.1002/hep.29367 (2018).
3. Paik, J. M., Golabi, P., Younossi, Y., Mishra, A. & Younossi, Z. M. Changes in the global burden of chronic liver diseases from 2012 to 2017: The growing impact of NAFLD. *Hepatology* **72**(5), 1605–1616. https://doi.org/10.1002/hep.31173 (2020).
4. Ekstedt, M. et al. Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up. *Hepatology* **61**(5), 1547–1554. https://doi.org/10.1002/hep.27368 (2015).
5. Sanyal, A. J. et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. *N. Engl. J. Med.* **385**(17), 1559–1569. https://doi.org/10.1056/nejmoa2029349 (2021).
6. Taylor, R. S. et al. Association between fibrosis stage and outcomes of patients with nonalcoholic fatty liver disease: A systematic review and meta-analysis. *Gastroenterology* **158**(6), 1611-1625.e12. https://doi.org/10.1053/j.gastro.2020.01.043 (2020).
7. Angulo, P. et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology* **149**(2), 389-397.e10. https://doi.org/10.1053/j.gastro.2015.04.043 (2015).
8. Hagström, H. et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. *J. Hepatol.* **67**(6), 1265–1273. https://doi.org/10.1016/j.jhep.2017.07.027 (2017).
9. Bedossa, P. & Patel, K. Biopsy and Noninvasive methods to assess progression of nonalcoholic fatty liver disease. *Gastroenterology* **150**(8), 1811-1822.e4. https://doi.org/10.1053/j.gastro.2016.03.008 (2016).
10. Estes, C., Razavi, H., Loomba, R., Younossi, Z. & Sanyal, A. J. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology* **67**(1), 123–133. https://doi.org/10.1002/hep.29466 (2018).
11. Swain, M. G. et al. Burden of nonalcoholic fatty liver disease in Canada, 2019–2030: A modelling study. *CMAJ Open* **8**(2), E429–E436. https://doi.org/10.9778/cmajo.20190212 (2020).
12. Berzigotti, A. et al. EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis—2021 update. *J. Hepatol.* **75**(3), 659–689. https://doi.org/10.1016/j.jhep.2021.05.025 (2021).
13. Liu, C. H. et al. Simple non-invasive scoring systems and histologic scores in predicting mortality in NAFLD patients. A systematic review and meta-analysis. *J. Gastroenterol. Hepatol.* https://doi.org/10.1111/jgh.15431 (2021).
14. Grat, K., Grat, M. & Rowiński, O. Usefulness of different imaging modalities in evaluation of patients with non-alcoholic fatty liver disease. *Biomedicines* **8**(9), 298. https://doi.org/10.3390/biomedicines8090298 (2020).
15. Petzold, G. et al. Diagnostic accuracy of B-Mode ultrasound and Hepatorenal Index for graduation of hepatic steatosis in patients with chronic liver disease. *PLoS ONE* **15**(5), 1–13. https://doi.org/10.1371/journal.pone.0231044 (2020).
16. Hernaez, R. et al. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: A meta-analysis. *Hepatology* **54**(3), 1082–1090. https://doi.org/10.1002/hep.24452 (2011).
17. Park, C. C. et al. Magnetic resonance elastography vs transient elastography in detection of fibrosis and noninvasive measurement of steatosis in patients with biopsy-proven nonalcoholic fatty liver disease. *Gastroenterology* **152**(3), 598-607.e2. https://doi.org/10.1053/j.gastro.2016.10.026 (2017).
18. Sharma, S., Khalili, K. & Nguyen, G. C. Non-invasive diagnosis of advanced fibrosis and cirrhosis. *World J. Gastroenterol.* **20**(45), 16820–16830. https://doi.org/10.3748/wjg.v20.i45.16820 (2014).
19. Maruyama, H. & Yokosuka, O. Ultrasonography for noninvasive assessment of portal hypertension. *Gut Liver* **11**(4), 464–473. https://doi.org/10.5009/gnl16078 (2017).
20. Kelly, E. M. M. et al. An assessment of the clinical accuracy of ultrasound in diagnosing cirrhosis in the absence of portal hypertension. *Gastroenterol. Hepatol. (N Y).* **14**(6), 367–373 (2018).
21. Moini, M. et al. Combination of FIB-4 with ultrasound surface nodularity or elastography as predictors of histologic advanced liver fibrosis in chronic liver disease. *Sci. Rep.* **11**(1), 19275. https://doi.org/10.1038/s41598-021-98776-1 (2021).
22. Ahn, J. C., Connell, A., Simonetto, D. A., Hughes, C. & Shah, V. H. The application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology* **73**, 2546. https://doi.org/10.1002/hep.31603 (2020).
23. Konerman, M. A., Beste, L. A., Van, T et al. ML_CHC_PLoS_2019.pdf. *PLoS One*. **14** (1), 1–14. https://doi.org/10.5281/zenodo.1490242.Funding (2019).
24. Patel, K. et al. Multiplex protein analysis to determine fibrosis stage and progression in patients with chronic hepatitis C. *Clin. Gastroenterol. Hepatol.* **12**(12), 2113-2120.e3. https://doi.org/10.1016/j.cgh.2014.04.037 (2014).
25. Canbay, A. et al. Non-invasive assessment of NAFLD as systemic disease—A machine learning perspective. *PLoS ONE* **14**(3), 1–15. https://doi.org/10.1371/journal.pone.0214436 (2019).
26. Yip, T. C. F. et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. *Aliment Pharmacol. Ther.* **46**(4), 447–456. https://doi.org/10.1111/apt.14172 (2017).
27. Wu, C. C. et al. Prediction of fatty liver disease using machine learning algorithms. *Comput. Methods Programs Biomed.* **170**, 23–29. https://doi.org/10.1016/j.cmpb.2018.12.032 (2019).
28. Perakakis, N. et al. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: A proof of concept study. *Metabolism* **101**, 154005. https://doi.org/10.1016/j.metabol.2019.154005 (2019).
29. Lockhart, M. E. & Smith, A. D. Fatty liver disease: Artificial intelligence takes on the challenge. *Radiology* **295**(2), 351–352. https://doi.org/10.1148/radiol.2020200058 (2020).
30. Han, A. et al. Noninvasive diagnosis of nonalcoholic fatty liver disease and quantification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks. *Radiology* **295**(2), 342–350. https://doi.org/10.1148/radiol.2020191160 (2020).
31. Taylor-Weiner, A. et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. *Hepatology* **74**(1), 133–147. https://doi.org/10.1002/hep.31750 (2021).
32. Forlano, R. et al. High-throughput, machine learning-based quantification of steatosis, inflammation, ballooning, and fibrosis in biopsies from patients with nonalcoholic fatty liver disease. *Clin. Gastroenterol. Hepatol.* **18**(9), 2081-2090.e9. https://doi.org/10.1016/j.cgh.2019.12.025 (2020).
33. Angulo, P. et al. The NAFLD fibrosis score: A noninvasive system that identifies liver fibrosis in patients with NAFLD. *Hepatology* **45**(4), 846–854. https://doi.org/10.1002/hep.21496 (2007).
34. Sterling, R. K. et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology* **43**(6), 1317–1325. https://doi.org/10.1002/hep.21178 (2006).
35. Harrison, S. A., Oliver, D., Arnold, H. L., Gogia, S. & Neuschwander-Tetri, B. A. Development and validation of a simple NAFLD clinical scoring system for identifying patients without advanced disease. *Gut* **57**(10), 1441–1447. https://doi.org/10.1136/gut.2007.146019 (2008).
36. Wai, C. T. et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology* **38**(2), 518–526. https://doi.org/10.1053/jhep.2003.50346 (2003).
37. Kleiner, D. E. et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**(6), 1313–1321. https://doi.org/10.1002/hep.20701 (2005).

38. DeLong, E. R. & DeLong, D. M. C. P. D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).
39. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
40. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0 (2018).
41. Angulo, P. et al. Simple noninvasive systems predict long-term outcomes in patients with nonalcoholic fatty liver disease. *Gastroenterology* **145**(4), 782–789. https://doi.org/10.1016/j.earlhumdev.2006.05.022 (2013).
42. Hagström, H. et al. Accuracy of noninvasive scoring systems in assessing risk of death and liver-related endpoints in patients with nonalcoholic fatty liver disease. *Clin. Gastroenterol. Hepatol.* **17**(6), 1148-1156.e4. https://doi.org/10.1016/j.cgh.2018.11.030 (2019).
43. Treeprasertsuk, S., Björnsson, E., Enders, F., Suwanwalaikorn, S. & Lindor, K. D. NAFLD fibrosis score: A prognostic predictor for mortality and liver complications among NAFLD patients. *World J. Gastroenterol.* **19**(8), 1219–1229. https://doi.org/10.3748/wjg.v19.i8.1219 (2013).
44. Sebastiani, G. et al. Prognostic value of non-invasive fibrosis and steatosis tools, hepatic venous pressure gradient (HVPG) and histology in nonalcoholic steatohepatitis. *PLoS ONE* **10**(6), 1–15. https://doi.org/10.1371/journal.pone.0128774 (2015).
45. Le, M. H. et al. Prevalence of non-Alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States. *PLoS ONE* **12**(3), 1–13. https://doi.org/10.1371/journal.pone.0173499 (2017).
46. Unalp-Arida, A. & Ruhl, C. E. Liver fibrosis scores predict liver disease mortality in the United States population. *Hepatology* **66**(1), 84–95. https://doi.org/10.1016/j.physbeh.2017.03.040 (2017).
47. Kim, D., Kim, W. R., Kim, H. J. & Therneau, T. M. Association between non-invasive fibrosis markers and mortality among adults with non-alcoholic fatty liver disease in the United States. *Hepatology* **57**(4), 1357–1365. https://doi.org/10.1002/hep.26156.Association (2013).
48. Xun, Y. H. et al. Non-alcoholic fatty liver disease (NAFLD) fibrosis score predicts 6.6-year overall mortality of Chinese patients with NAFLD. *Clin. Exp. Pharmacol. Physiol.* **41**(9), 643–649. https://doi.org/10.1111/1440-1681.12260 (2014).
49. Jaruvongvanich, V., Wijarnpreecha, K. & Ungprasert, P. The utility of NAFLD fibrosis score for prediction of mortality among patients with nonalcoholic fatty liver disease: A systematic review and meta-analysis of cohort study. *Clin. Res. Hepatol. Gastroenterol.* **41**(6), 629–634. https://doi.org/10.1016/j.clinre.2017.03.010 (2017).
50. Lee, J. et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score and APRI for NAFLD-related events: A systematic review. *Liver Int.* **41**(2), 261–270. https://doi.org/10.1111/liv.14669 (2021).
51. Younossi, Z. M. et al. The Association of histologic and noninvasive tests with adverse clinical and patient-reported outcomes in patients with advanced fibrosis due to nonalcoholic steatohepatitis. *Gastroenterology* **160**(5), 1608-1619.e13. https://doi.org/10.1053/j.gastro.2020.12.003 (2021).
52. Salomone, F., Micek, A. & Godos, J. Simple scores of fibrosis and mortality in patients with NAFLD: A systematic review with meta-analysis. *J. Clin. Med.* **7**(8), 219. https://doi.org/10.3390/jcm7080219 (2018).
53. Paik, J., Younossi, E., Keo, W., Allawi, H., Henry, L. Y. Z. High risk fibrosis 4-score is predictive of all-cause, cardiovascular and liver-related mortality among adults non-alcoholic fatty liver disease (NAFLD) in the United States (U.S.). Abstract. *AASLD The Liver Meeting*. (2021).
54. Boursier, J. et al. Non-invasive tests accurately stratify patients with NAFLD based on their risk of liver-related events. *J. Hepatol.* **76**(5), 1013–1020. https://doi.org/10.1016/j.jhep.2021.12.031 (2022).
55. Shili-Masmoudi, S. et al. Liver stiffness measurement predicts long-term survival and complications in non-alcoholic fatty liver disease. *Liver Int.* **40**(3), 581–589. https://doi.org/10.1111/liv.14301 (2020).
56. Kim, B. et al. A liver stiffness measurementbased, noninvasive prediction model for high-risk esophageal varices in B-viral liver cirrhosis. *Am. J. Gastroenterol.* **105**, 1382–1390 (2010).
57. Kim, B. et al. Risk assessment of esophageal variceal bleeding in B-viral liver cirrhosis by a liver stiffness measurement-based model. *Am. J. Gastroenterol.* **106**(9), 1654–1662 (2011).
58. Kim, B. K. et al. Risk assessment of development of hepatic decompensation in histologically proven hepatitis B viral cirrhosis using liver stiffness measurement. *Digestion* **85**(3), 219–227. https://doi.org/10.1159/000335430 (2012).
59. Berzigotti, A. et al. Elastography, spleen size, and platelet count identify portal hypertension in patients with compensated cirrhosis. *Gastroenterology* **144**(1), 102-111.e1. https://doi.org/10.1053/j.gastro.2012.10.001 (2013).
60. Chang, D. et al. Machine learning models are superior to noninvasive tests in identifying clinically significant stages of NAFLD and NAFLD-related cirrhosis. *Hepatology* https://doi.org/10.1002/hep.32655 (2022).
61. Lee, J. et al. Machine learning algorithm improves the detection of NASH (NAS-based) and at-risk NASH: A development and validation study. *Hepatology* **78**(1), 258–271. https://doi.org/10.1097/HEP.0000000000000364 (2023).
62. Cygu, S., Seow, H., Dushoff, J. & Bolker, B. M. Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. *Sci. Rep.* **13**(1), 1370 (2023).
63. Gidener, T. et al. Liver stiffness by magnetic resonance elastography predicts future cirrhosis, decompensation, and death in NAFLD. *Clin. Gastroenterol. Hepatol.* https://doi.org/10.1016/j.cgh.2020.09.044 (2020).
64. Han, M. A. T. et al. MR elastography-based liver fibrosis correlates with liver events in nonalcoholic fatty liver patients: A multicenter study. *Liver Int.* **40**(9), 2242–2251. https://doi.org/10.1111/liv.14593 (2020).
65. Tamaki, N. et al. MRE plus FIB-4 (MEFIB) versus FAST in detection of candidates for pharmacological treatment of NASH-related fibrosis. *Hepatology* https://doi.org/10.1002/hep.32145 (2021).

Author contributions

HMK contributed to study concept and design, performed data collection, statistical analysis, and contributed to writing and editing the manuscript; CM contributed to study concept and design, performed the machine learning analysis and contributed to writing of the manuscript; MS performed data collection for the McGill dataset; MF performed data collection and review of all ultrasound data for the Toronto cohort; CB performed data collection; OA served as the lead pathologist reviewing liver biopsies from the Toronto cohort, GS contributed to study concept and design, served as the lead investigator for the McGill cohort, and contributed to the writing and editing of the manuscript; KJ contributed to study concept and design, served as the radiology lead for the Toronto cohort, and contributed to the writing and editing of the manuscript; KP contributed to study concept and design, served as the lead investigator for the study and contributed to the statistical analysis, writing and editing of the manuscript.

Funding

This study was supported by a research grant for gastroenterology residents/fellows, provided by the University of Toronto. GS is supported by a Senior Salary Award from *Fonds de Recherche du Quebec – Sante (FRQS)* (#296306).

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval and consent statement

This study was performed according to the 'Good Clinical Practice' guidelines based upon principles outlined in the Declaration of Helsinki, as well as local and national guidelines regarding the conduct of clinical research studies. This study was approved by the Institutional Research Ethics Board at the University Health Network and McGill University Health Centre. All patient data used for this study was de-identified and anonymized; accordingly, patient consent was not required.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-09288-1.

Correspondence and requests for materials should be addressed to H.M.-K.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.