Title: Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD

Short Title: Liver fibrosis severity assessment

Authors:

Gong Feng,[1#] Kenneth I. Zheng,[2#] Yang-Yang Li,[3] Rafael S. Rios,[2] Pei-Wu Zhu,[4] Xiao-Yan Pan,[5] Gang Li,[2] Hong-Lei Ma,[2] Liang-Jie Tang,[2] Christopher D. Byrne,[6] Giovanni Targher,[7] Na He,[8] Man Mi,[1] Yong-Ping Chen[2,9,10] and Ming-Hua Zheng[2,9,10*]

Affiliations:

[1]Xi'an Medical University, Xi'an, China;

[2]NAFLD Research Center, Department of Hepatology, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China;

[3]Department of Pathology, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China;

[4]Department of Laboratory Medicine, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China;

[5]Department of Endocrinology, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China;

[6]Southampton National Institute for Health Research Biomedical Research Centre,

University Hospital Southampton, Southampton General Hospital, Southampton, UK;

[7]Section of Endocrinology, Diabetes and Metabolism, Department of Medicine, University and Azienda Ospedaliera Universitaria Integrata of Verona, Verona, Italy;

[8]Department of Gastroenterology, the First Affiliated Hospital of Xi'an Medical University, Xi'an, China;

[9]Institute of Hepatology, Wenzhou Medical University, Wenzhou, China;

[10]Key Laboratory of Diagnosis and Treatment for The Development of Chronic Liver Disease in Zhejiang Province, Wenzhou, China.

[#]**Co-first author:** Gong Feng and Kenneth I. Zheng

***Corresponding author:**

Ming-Hua Zheng, MD, PhD

NAFLD Research Center, Department of Hepatology, the First Affiliated Hospital of Wenzhou Medical University; No. 2 Fuxue Lane, Wenzhou 325000, China.

E-mail: zhengmh@wmu.edu.cn; fax: (86) 577-55578522; tel: (86) 577-55579622.

Number of figures and tables: 3 figures and 4 tables

Abbreviation list

A/G = albumin/globulin, ALT = alanine aminotransferase, AST = aspartate

aminotransferase, APRI = AST-to-platelet ratio index, AUROC = area under the receiver operator characteristic curve, BMI = body mass index, CK-18 M30 = cytokeratine-18 neoepitope M30, DCA = decision curve analysis, FIB-4 = fibrosis-4 index, GGT = γ-glutamyltransferase, HbA1c = glycated hemoglobin, HOMA-IR = homeostasis model assessment of insulin resistance, IV-C = type IV collagen, LASSO = least absolute shrinkage and selection operator, LRM =logistic regression model, MLA = machine learning algorithm, NAFL = nonalcoholic fatty liver, NAFLD = non-alcoholic fatty liver disease, NAS = NAFLD activity score, NASH = non-alcoholic steatohepatitis, NFS = NAFLD fibrosis score, NPV = negative predictive value, OR = odds ratio, PC-III = procollagen type III, PPV = positive predictive value, PLT = platelet count, T2DM = type 2 diabetes, WHR = waist-to-hip ratio.

Funding sources

This work was supported by grants from the National Natural Science Foundation of China (81500665), High Level Creative Talents from Department of Public Health in Zhejiang Province and Project of New Century 551 Talent Nurturing in Wenzhou. Na He is supported in part by Special Research Fund of Youan Medical Alliance for the Liver and Infectious Diseases (LM202003). GT is supported in part by grants from the School of Medicine, University of Verona, Verona, Italy. CDB is supported in part by the Southampton NIHR Biomedical Research Centre (IS-BRC-20004), UK.

Acknowledgments

We thank Prof. Ji-Min Liu, a liver pathologist from McMaster University, who conducted quality control of liver pathology data.

Conflicts of interest

All authors: nothing to declare.

ABSTRACT

Background: The presence of significant liver fibrosis is a key determinant of long-term prognosis in non-alcoholic fatty liver disease (NAFLD). We aimed to develop a novel machine learning algorithm (MLA) to predict fibrosis severity in NAFLD and compared it with the most widely used non-invasive fibrosis biomarkers.

Methods: We used a cohort of 553 adults with biopsy-proven NAFLD, who were randomly divided into a training cohort (n=278) for the development of both logistic regression model (LRM) and MLA, and a validation cohort (n=275). Significant fibrosis was defined as fibrosis stage F≥2. MLA and LRM were derived from variables that were selected using a least absolute shrinkage and selection operator (LASSO) logistic regression algorithm.

Results: In the training cohort, the variables selected by LASSO algorithm were body mass index, pro-collagen type III, collagen type IV, aspartate aminotransferase and albumin-to-globulin ratio. The diagnostic accuracy of MLA showed the highest values of area under the receiver operator characteristic curve (AUROC: 0.902, 95%CI 0.869–0.904) for identifying fibrosis F≥2. The LRM AUROC was 0.764, 95%CI 0.710-0.816) and significantly better than the AST-to-Platelet ratio (AUROC 0.684, 95%CI 0.605-0.762), FIB-4 score (AUROC 0.594, 95%CI 0.503-0.685) and NAFLD Fibrosis Score (AUROC 0.557, 95%CI 0.470-0.644). In the validation cohort, MLA also showed the highest AUROC (0.893, 95%CI 0.864–0.901). The diagnostic accuracy of MLA outperformed that of LRM in all subgroups considered.

Conclusions: Our newly developed MLA algorithm has excellent diagnostic performance for predicting fibrosis F≥2 in patients with biopsy-confirmed NAFLD.

Keywords: Fibrosis, NAFLD, Machine learning algorithm, Diagnosis, Liver biopsy

Key Summary

- Fibrosis stage is a key determinant of adverse liver-related outcomes in NAFLD.
- The diagnostic accuracy of the machine learning algorithm (MLA) had the highest AUROC in predicting significant fibrosis in NAFLD and outperformed non-invasive biomarker scores.
- The MLA showed excellent diagnostic accuracy for potential clinical use.

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is the leading cause of chronic liver diseases around the world. Individuals with NAFLD may develop liver fibrosis of varying degrees, with an annual fibrosis progression rate of nearly 10%, while individuals with advanced fibrosis have a 10-year survival rate of nearly 80%.(1, 2) Previous studies have shown that fibrosis stage is one of the strongest predictors of liver-related morbidity and mortality in NAFLD.(3, 4) Therefore, accurate assessment of liver fibrosis is required for a better stratification and personalized management of patients with NAFLD.

To date, liver biopsy remains the 'gold standard' for staging fibrosis in NAFLD. However, liver biopsy is an invasive technique that is not free from complications.(2) Consequently, there is a need for non-invasive tests to stage the severity of liver fibrosis in NAFLD. In clinical practice, the most widely used non-invasive tests for staging advanced fibrosis are the aspartate aminotransferase (AST) to platelet ratio index (APRI), the NAFLD fibrosis score (NFS), and the fibrosis-4 (FIB-4) index. While the diagnostic performance of these non-invasive tests is satisfactory for ruling out advanced fibrosis, their performance is not adequate for diagnosing clinically significant liver fibrosis.(5, 6) A recent meta-analysis showed that even low levels of liver fibrosis are associated with overall and liver-specific mortality, and these data are supported by the results of a large biopsy study with a mean of 20 years follow-up

showing that advanced fibrosis was associated with liver-specific morbidity and overall mortality.(3) Additionally, distinguishing between NAFLD with or without significant fibrosis is clinically important for determining the clinical prognosis of these patients.(7)

Machine learning refers to uploading large amounts of data into a computer program and choosing a model to "fit" the data so that the computer can make predictions. The way the computer creates the model is via the use of algorithms, which include both simple equations (such as linear equations) and very complex logic/mathematical systems, with the aim of developing the best prediction model (**Supplementary Figure 1**). Machine learning algorithms (MLAs) are a type of artificial intelligence formerly used to extrapolate multifactorial events and behaviors, which have substantial advantages over conventional statistical methods; and MLAs have been implemented with some success in cancer research.(8, 9) Therefore, the aim of our study was to establish a novel prediction model for non-invasively identifying the presence of significant fibrosis (defined by F≥2 fibrosis on histology), by using MLA, in a well-characterized cohort of patients, in whom NAFLD severity has been assessed with histology. We first screened for important variables, then used these variables to construct traditional regression models and MLA, respectively, and compared the diagnostic effectiveness of MLA and other models.

MATERIALS AND METHODS

Patient population

The current analysis was embedded within the well-characterized Prospective Epidemic Research Specifically Of NASH (PERSONS) cohort study.(5) For the purpose of the current analysis, we included a cohort of 553 adult patients with biopsy-proven NAFLD, who were consecutively recruited at the First Affiliated Hospital of Wenzhou Medical University in Wenzhou, China, from December 2016 to December 2018. Subsequently, we randomly subdivided this cohort of 553 patients into a training set (n=278) for logistic regression model (LRM) and MLA development, and a validation set (n=275). Participants were eligible for this study if they met the following inclusion criteria: (1) patients aged 18–75 years; (2) fatty liver diagnosed by imaging and/or abnormal liver function; (3) participants were willing to provide written informed consent. Individuals with significant alcohol intake (≥140 g/week in men or ≥70 g/week in women); use of potentially hepatotoxic drugs, presence of viral hepatitis, autoimmune hepatitis or other known chronic liver diseases; or those with incomplete data were excluded from the study (see details reported in the Results section). The flow chart of the study population is shown in Supplementary Figure 1. The study protocol was approved by the ethics committee of the First Affiliated Hospital of Wenzhou Medical University (2016-246, 1 December 2016). Written informed consent was obtained from each participant.

Clinical and laboratory data

In all participants, demographic data, anthropometry, clinical parameters, as well as concomitant diseases were recorded by standard methods, as reported previously.(6, 10) Hypertension was diagnosed by blood pressure ≥140/90 mmHg and/or use of any antihypertensive drugs. Homeostatic model assessment for insulin resistance (HOMA-IR) was used for assessing insulin resistance. Type 2 diabetes mellitus (T2DM) was diagnosed with at least one of the following criteria: self-reported history of diabetes, use of any hypoglycemic agents, fasting glucose levels ≥7.0 mmol/L or hemoglobin A1c ≥6.5%. Hyperlipidemia was defined as presence of total cholesterol (TC) ≥6.2 mmol/L, low-density lipoprotein (LDL-C) ≥4.1 mmol/L, triglycerides (TG) ≥2.3 mmol/L or use of any lipid-lowering agents. Patients with a body mass index (BMI) ≥25 kg/m^2 were considered as obese. All routine blood tests, including also the circulating values of procollagen III (PC-III), type-IV collagen (IV-C), aspartate aminotransferase (AST) and the ratio of albumin/globulin (A/G), were performed in the Central laboratory of our hospital and were performed using standard laboratory methods. Plasma cytokeratin-18 M30 (CK-18 M30) level was measured using a commercially available ELISA kit from Herui Biomed Company Limited, Suzhou, China. Assays were performed according to the operator's manual. All the tests were double blind and the coefficients of variability were <15%.

Liver histology

Under the guidance of ultrasound, percutaneous biopsy samples were obtained by 16-gauge hepafix needle. Biopsy samples were stained with hematoxylin, Masson's trichrome as well as eosin and subsequently assessed by an experienced liver pathologist, who was blind to patients' clinical and laboratory data. Liver biopsies were required to be greater than 1 cm and the number of portal areas greater than 6. Grading of NAFLD histology was determined by NAFLD activity score (NAS) according the NASH Clinical Research Network scoring system.(11) The NAS score included three histologic features including the presence of steatosis, lobular inflammation, as well as balloon-like hepatocytes. Liver fibrosis was staged from zero to 4 as follows: 0=no fibrosis, 1=perisinusoidal or portal fibrosis; 2=perisinusoidal and portal/periportal fibrosis; 3=bridging fibrosis; and 4=highly suspicious or definite cirrhosis, respectively.(12, 13) Significant fibrosis was defined as fibrosis stage ≥ F2, which is clinically important for determining prognosis,(7, 14) and advanced fibrosis as ≥ F3. Other non-invasive fibrosis scores that have been used to diagnose advanced fibrosis (i.e., the APRI, NFS, and FIB-4 scores) were calculated based upon published studies.

Development of LRM Model

We developed a logistic regression model (LRM) that was predictive for the development of fibrosis ≥F2 using the training set (n=278). The least absolute shrinkage and selection operator (LASSO) method was used to identify the variables

significantly associated with the presence of fibrosis ≥F2 that were used to build the final LRM. The LASSO logistic regression model was used with penalty parameter tuning that was conducted by 10-fold cross-validation based on minimum criteria. LASSO is suitable for the regression of high-dimensional data, to select significant fibrosis features with non-zero coefficients from among the 23 demographic, anthropometric measurements, concomitant diseases, and laboratory parameters.(15) A formula was generated using a linear combination of selected features that were weighted by their respective LASSO coefficients; the formula was then used to calculate an individual risk score for each patient to reflect the risk of having significant liver fibrosis. The predictive accuracy of LRM was quantified by the area under the receiver-operator characteristic curve (AUROC) in both training and validation sets.

Development of MLA model

Random forest analysis was used, which is a machine learning algorithm that can build classification prediction models. This random forest approach divides the initial training set into two groups – "in-bag" and "out-of-bag" samples. The "in-bag" sample is developed using random sampling with replacement from the initial training set, creating a sample equivalent in size to the initial training set. The "out-of-bag" sample is composed of the unsampled data from the initial training set, and comprises about 30% of the initial cohort. This process is repeated 500 times to create multiple

pairings of "in-bag" and "out-of-bag" samples. For each pairing, a decision tree is then constructed using the "in-bag" sample, using a random set of potential predictor variables for each split, and then validated using the "out-of-bag" sample. As each tree is built, only a random subset of the predictor variables is considered as possible splitters for each binary partitioning. The predictions from each tree are used as "votes", and the outcome with the most votes is considered the dichotomous outcome prediction for that sample. Using this method, multiple decision trees are constructed to create the final classification prediction model and to determine the overall variable importance. Accuracies and error rates are computed for each observation using the "out-of-bag" predictions, and then averaged over all observations. Because the "out-of-bag" observations were not used in the fitting of the trees, the "out-of-bag" estimates serve as cross-validated accuracy estimates.(16) Variable importance identifies the most important variables based on their contribution to the predictive accuracy of the model. The most important variables are identified as those that most frequently result in early splitting of the decision trees.

Statistical analyses

Continuous variables were expressed as means ± SD or medians and inter-quartile ranges (IQR), categorical variables were expressed as numbers and percentages. Differences in main clinical, biochemical and histological liver features between the training and validation cohorts were tested by the unpaired Student's *t*-test and the

Mann-Whitney U-test (for either normally or not normally distributed continuous variables) or the chi-square test for categorical variables. The LASSO method was used to identify the variables associated with presence of fibrosis ≥F2, which were used to build both MLA and LRM using the random forest method, and logistic regression, respectively. LRM was calibrated by repeated sampling and assessed by the area under the receiver operator characteristic curve (AUROC) for performance. The diagnostic accuracy of both models was evaluated utilizing the AUROC. After establishing and verifying the diagnostic efficacy of MLA for significant fibrosis, we explored the diagnostic value of MLA for advanced fibrosis in NAFLD. Analytical software used consisted of SPSS version 22.0 (SPSS, Chicago, IL, USA), and R3.3.1 (R Advancement Core Group, http://www.r-project.org).

RESULTS

Patient characteristics

Of the 596 adults with suspected NAFLD, who were initially enrolled in the study, we excluded 43 persons for the following reasons: 15 for alcoholic fatty liver, 6 for autoimmune hepatitis, 3 for drug-induced liver injury and 19 for missing data. As a consequence, 553 adults with biopsy-proven NAFLD were included in the final analysis. These patients were randomly assigned to a training set (n=278) and a validation set (n=275). **Table 1** summarizes the main clinical, laboratory and histological characteristics of participants. As expected due to the randomization,

there were no significant differences in any of these variables between the training and validation sets.

Selection of main predictors of significant fibrosis

Among the 23 demographic, anthropometric measurements, concomitant diseases, and laboratory parameters, five significant predictors of significant fibrosis were selected in the training set (**Figure 1A and B**). These predictors were BMI, serum procollagen type III (PC-III), type IV collagen (IV-C), AST and A/G ratio.

Development of LRM

In LASSO regression algorithm, a LRM formula was generated using a linear combination of selected features that were weighted by their respective LASSO coefficients. The following LRM formula (0.00265 *AST + 0.00426 * PC-III + 0.00669 * IV-C + 0.01893 * BMI - 0.53322 * A/G) was used to calculate an individual risk score (defined as the LAM score) for each patient to reflect the risk of significant fibrosis. The LRM for the prediction of significant fibrosis, which included the five aforementioned variables, showed excellent diagnostic accuracy in the training set (as shown in **Supplementary Figure 2**).

Comparison of the diagnostic performance of LRM and non-invasive fibrosis markers in the training set

Supplementary Figure 3A shows the AUROCs of the LRM, NFS, FIB-4 and APRI scores for identifying significant fibrosis. **Table 2** shows the diagnostic performance of all of these predictive models in the training set (n=278). LRM showed the highest AUROC (0.764, 95%CI 0.710-0.816) in predicting significant fibrosis compared with APRI (0.684, 95%CI 0.605-0.762, P = 0.047 vs. LRM's AUROC), NFS (0.557, 95%CI 0.470-0.664, P < 0.001), and FIB-4 (0.594, 95%CI 0.503-0.685, P <0.001), respectively. The overall sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of LRM for significant fibrosis were 75.1%, 80.1%, 66.4%, and 90.3%, respectively. In contrast, the overall sensitivity, specificity, PPV, and NPV for: a) APRI, b) NFS and c) FIB-4 scores were: a) 69.8%, 60.0%, 29.1%, 79.4%; b) 74.3%, 38.7%, 23.3%, 85.8%; and c) 32.1%, 79.9%, 44.7%, 85.0%, respectively.

Validation of LRM

Supplementary **Figure 3B** shows the ROC curves for the LRM, NFS, FIB-4 and APRI scores. **Table 2** shows the performances of all these predictive models in the validation set (n=275). LRM showed the highest AUROC (0.786, 95%CI 0.719-0.852) in predicting significant fibrosis compared with APRI (0.615, 95%CI 0.532-0.698, P=0.031 vs. LRM's AUROC), NFS (0.575, 95% CI 0.486-0.664, P<0.001), and FIB-4

(0.578, 95%CI 0.498-0.659, P<0.001), respectively. The overall sensitivity, specificity, PPV, and NPV of LRM for predicting significant fibrosis were 72.6%, 82.5%, 71.1%, and 85.4%, respectively. The overall sensitivity, specificity, PPV, and NPV for: a) APRI, b) NFS, and c) FIB-4 scores were: a) 38.9%, 80.1%, 33.9%, 80.4%, b) 38.9%, 79.1%, 31.3%, 84.1%; and c) 87.0%, 32.1%, 23.9%, 82.5%, respectively. As shown in **Table 3**, subgroup analyses performed among NAFLD patients stratified by obesity, T2DM or serum ALT levels confirmed that the diagnostic performance of LRM was good in both training and validation sets.

Development of MLA model

We developed a MLA model that included the following five variables (BMI, PC-III, IV-C, AST and A/G ratio), which were found to be strongly predictive of significant fibrosis in the validation set. The relative importance of the above-mentioned variables in MLA were 100% for PC-III, 87% for IV-C, 78% for BMI, 68% for AST and 60% for A/G ratio, respectively.

Comparison of the diagnostic performance of LRM and MLA in the training and validation sets

In the training set, the diagnostic accuracy of MLA (AUROC 0.902, 95%CI 0.869-0.904) for identifying significant fibrosis was much better than that of LRM

(AUROC 0.764, 95%CI 0.710-0.816) ($p<0.05$, **Figure 2A**). Similarly, in the validation set, the diagnostic accuracy of MLA (AUROC 0.893, 95%CI 0.864-0.901) was also superior to LRM (AUROC 0.786, 95%CI 0.719-0.852; $p<0.05$, **Figure 2B**). As shown in **Table 4**, subgroup analyses stratifying NAFLD patients by obesity, T2DM or serum ALT levels confirmed that the diagnostic performance of MLA was better than that of LRM in both training and validation sets.

Diagnostic performance of MLA for predicting advanced fibrosis and cirrhosis

Of the 553 patients included, 22 had advanced fibrosis (F ≥3 stage). We also tested the diagnostic performance of MLA for predicting advanced fibrosis. These preliminary data showed that MLA showed excellent performance (AUROC 0.996, 95%CI 0.967-0.998) for predicting the presence of advanced fibrosis (**Figure 3**). Furthermore, we also tested the diagnostic performance of MLA for predicting cirrhosis in NAFLD. These data showed that MLA had excellent performance (AUROC 0.989, 95%CI 0.977-0.996) for predicting the presence of cirrhosis.

DISCUSSION

Our data shows that the newly developed MLA algorithm had excellent diagnostic performance for predicting fibrosis F≥2 in patients with biopsy-confirmed NAFLD. To our knowledge, this is the first observational study that used MLA for developing a

prediction model (involving clinical parameters and serum biomarkers) for identifying non-invasively the presence of significant fibrosis in a large cohort of Chinese adults with biopsy-proven NAFLD.

Through the LASSO regression analysis, we showed that BMI, PC-III, IV-C, AST and A/G ratio were the five strongest predictors of significant fibrosis in our cohort of NAFLD patients. PC-III is a byproduct of type III collagen synthesis, which occurs during hepatocellular injury repair by extracellular matrix remodeling. In addition, several reports have shown a close association between severity of liver fibrosis and increasing serum levels of PC-III.(17-20) Similarly, IV-C is a marker of enhanced basement membrane synthesis during accelerated extracellular matrix remodeling in liver fibrosis and has been demonstrated to be indicative of liver fibrotic injury.(21-23) BMI is a known risk factor for the development and progression of NAFLD. AST is also known to closely correlate to advanced fibrosis and is considered to be an important factor in other widely used non-invasive tests of advanced fibrosis, including the FIB-4, APRI, NFS and BARD scores. By combining PC-III, IV-C, BMI, AST and A/G, we developed a MLA for predicting significant fibrosis and our newly developed prediction model resulted in remarkably better diagnostic performance, compared to that of LRM, FIB-4, APRI and NFS scores.

In the training set, we compared the performances of LRM, FIB-4, APRI, and NFS

scores in identifying the presence of significant fibrosis and found that these scores were moderately accurate in our patient cohort, as compared to other previously published studies;(5, 24, 25) although this finding could be partly affected by differences in the prevalence of significant fibrosis between the studied populations. Similar findings were also noted in the validation set where the diagnostic performances in identifying significant fibrosis were, respectively, 0.786 (95%CI 0.719-0.852) for LRM; 0.578 (95%CI 0.498-0.659) for FIB-4; 0.615 (95%CI 0.532-0.698) for APRI; and 0.575 (95%CI 0.486-0.664) for the NFS score. Notably, MLA performed much better than LRM with an AUROC of 0.902 (95%CI 0.869-0.904) in the training set and an AUROC of 0.893 (95%CI 0.864-0.901) in the validation set, respectively.

Patients with NAFLD are likely to be obese and to have coexisting metabolic diseases such as T2DM. Elevated ALT levels remain the most widely used laboratory parameter in primary clinics for referral to hepatologists of those with suspected NAFLD. However, patients with NAFLD and normal ALT levels are often neglected from further assessment of their liver disease. Therefore, we compared the diagnostic performance of MLA and LRM in subgroups stratified by obesity, T2DM and serum ALT levels. Interestingly, MLA performed better in patients who are non-obese, non-diabetic or having serum ALT <40 U/L compared to their counterparts with obesity, T2DM or elevated ALT levels in both training and validation sets. However,

in training and validation sets, MLA outperformed LRM in every subgroup of patients. There are many reasons why the prediction efficiency of random forest model is better than other models. For example, in various current data sets, the random forest has a great advantage over other algorithms. The introduction of two randomness makes the random forest have a good anti-noise capability. Random forest may process data of high dimensions without feature selection, and it has a strong adaptability to data sets. It may process both discrete data and continuous data, and data sets need not be normalized.

In other machine learning studies, subjects were either from animal models or non-liver biopsy patients with NAFLD, or the sample size of liver biopsy patients with NAFLD of previous studies was smaller than that of our study and the results lacked validation.(26-29) In addition, previous MLA studies have mostly focused on NASH, while few have focused on liver fibrosis. It should be noted that the degree of fibrosis is the most important risk factor of the prognosis of NAFLD patients.

Our study has some important limitations that should be mentioned. Firstly, our data are derived from a single-center study that includes patients with baseline characteristics that may differ compared with other cohorts. Additionally, our results require further verification in other cohorts and in other ethnic groups. The heterogeneity between studies may contribute to differing diagnostic performances of

existing non-invasive tests for fibrosis (i.e. FIB-4, APRI and NFS scores) when comparing results from our cohort and other studies. Secondly, our patients are of Chinese Han ethnicity and, therefore, may not be generalizable to other ethnic groups. Thirdly, the lack of adequate comparators such as the ELF test or elastographic methods and the relatively few patients (n=22) with advanced fibrosis (F⩾3 stage) are another limitation. The study cohort predominantly comprised of patients with the early stages of liver fibrosis. There were few patients with advanced fibrosis (F⩾3 stage) because liver biopsy is less frequently required as a diagnostic method in this patient group. Therefore, there may be a degree of selection bias affecting our study results towards patients with lower levels of liver fibrosis. The lack of a suitable comparison object may to some extent expand the diagnostic effectiveness of MLA. Thus, with this small number of patients, we were able only to perform exploratory analyses for testing the diagnostic ability of the MLA for predicting advanced fibrosis and cirrhosis.

In conclusion, our newly developed MLA algorithm showed excellent performance for predicting fibrosis F≥2 in adults with biopsy-confirmed NAFLD. The MLA may aid clinicians identify patients with a high probability of significant liver fibrosis who need further management and advice.

REFERENCES

1. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology (Baltimore, Md). 2016;64(1):73-84.

2. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the American Gastroenterological Association, American Association for the Study of Liver Diseases, and American College of Gastroenterology. Gastroenterology. 2012;142(7):1592-609.

3. Hagstrom H, Nasr P, Ekstedt M, Hammar U, Stal P, Hultcrantz R, et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. Journal of hepatology. 2017;67(6):1265-73.

4. Vilar-Gomez E, Calzadilla-Bertot L, Wai-Sun Wong V, Castellanos M, Aller-de la Fuente R, Metwally M, et al. Fibrosis Severity as a Determinant of Cause-Specific Mortality in Patients With Advanced Nonalcoholic Fatty Liver Disease: A Multi-National Cohort Study. Gastroenterology. 2018;155(2):443-57.e17.

5. Zhou YJ, Ye FZ, Li YY, Pan XY, Chen YX, Wu XX, et al. Individualized risk prediction of significant fibrosis in non-alcoholic fatty liver disease using a novel nomogram. United European Gastroenterol J. 2019;7(8):1124-34.

6. Vilar-Gomez E, Chalasani N. Non-invasive assessment of non-alcoholic fatty liver disease: Clinical prediction rules and blood-based biomarkers. Journal of hepatology. 2018;68(2):305-15.

7. Yu SJ, Kim W, Kim D, Yoon JH, Lee K, Kim JH, et al. Visceral Obesity Predicts Significant Fibrosis in Patients With Nonalcoholic Fatty Liver Disease. Medicine. 2015;94(48):e2159.

8. Ho WH, Lee KT, Chen HY, Ho TW, Chiu HC. Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network. PloS one. 2012;7(1):e29179.

9. Maiellaro PA, Cozzolongo R, Marino P. Artificial neural networks for the prediction of response to interferon plus ribavirin treatment in patients with chronic hepatitis C. Curr Pharm Des. 2004;10(17):2101-9.

10. Ye FZ, Liu WY, Zheng KI, Pan XY, Ma HL, Wang XD, et al. Homeostatic model assessment of insulin resistance closely related to lobular inflammation in nonalcoholic fatty liver disease. Eur J Gastroenterol Hepatol. 2020;32(1):80-6.

11. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology (Baltimore, Md). 2005;41(6):1313-21.

12. Park CC, Nguyen P, Hernandez C, Bettencourt R, Ramirez K, Fortney L, 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. 2017;152(3):598-607.e2.

13. Bedossa P, Patel K. Biopsy and Noninvasive Methods to Assess Progression of Nonalcoholic Fatty Liver Disease. Gastroenterology. 2016;150(8):1811-22.e4.

14. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, 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(2):389-97 e10.

15. Wu S, Zheng J, Li Y, Yu H, Shi S, Xie W, et al. A Radiomics Nomogram for the Preoperative Prediction of Lymph Node Metastasis in Bladder Cancer. Clin Cancer Res. 2017;23(22):6904-11.

16. Singal AG, Mukherjee A, Elmunzer BJ, Higgins PD, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol. 2013;108(11):1723-30.

17. Bril F, McPhaul MJ, Caulfield MP, Clark VC, Soldevilla-Pico C, Firpi-Morell RJ, et al. Performance of Plasma Biomarkers and Diagnostic Panels for Nonalcoholic Steatohepatitis and Advanced Fibrosis in Patients With Type 2 Diabetes. Diabetes Care. 2020;43(2):290-7.

18. Mosca A, Comparcola D, Romito I, Mantovani A, Nobili V, Byrne CD, et al. Plasma N-terminal propeptide of type III procollagen accurately predicts liver fibrosis severity in children with non-alcoholic fatty liver disease. Liver international : official journal of the International Association for the Study of the Liver. 2019;39(12):2317-29.

19. Nielsen MJ, Veidal SS, Karsdal MA, Orsnes-Leeming DJ, Vainer B, Gardner SD, et al. Plasma Pro-C3 (N-terminal type III collagen propeptide) predicts fibrosis

progression in patients with chronic hepatitis C. Liver international : official journal of the International Association for the Study of the Liver. 2015;35(2):429-37.

20. Daniels SJ, Leeming DJ, Eslam M, Hashem AM, Nielsen MJ, Krag A, et al. ADAPT: An Algorithm Incorporating PRO-C3 Accurately Identifies Patients With NAFLD and Advanced Fibrosis. Hepatology (Baltimore, Md). 2019;69(3):1075-86.

21. Nielsen MJ, Villesen IF, Gudmann NS, Leeming DJ, Krag A, Karsdal MA, et al. Serum markers of type III and IV procollagen processing predict recurrence of fibrosis in liver transplanted patients. Scientific reports. 2019;9(1):14857.

22. Nielsen MJ, Karsdal MA, Kazankov K, Gronbaek H, Krag A, Leeming DJ, et al. Fibrosis is not just fibrosis - basement membrane modelling and collagen metabolism differs between hepatitis B- and C-induced injury. Aliment Pharmacol Ther. 2016;44(11-12):1242-52.

23. Nielsen MJ, Kazankov K, Leeming DJ, Karsdal MA, Krag A, Barrera F, et al. Markers of Collagen Remodeling Detect Clinically Significant Fibrosis in Chronic Hepatitis C Patients. PloS one. 2015;10(9):e0137302.

24. Staufer K, Halilbasic E, Spindelboeck W, Eilenberg M, Prager G, Stadlbauer V, et al. Evaluation and comparison of six noninvasive tests for prediction of significant or advanced fibrosis in nonalcoholic fatty liver disease. United European Gastroenterol J. 2019;7(8):1113-23.

25. Angulo P, Hui JM, Marchesini G, Bugianesi E, George J, Farrell GC, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology (Baltimore, Md). 2007;45(4):846-54.

26. Ramot Y, Zandani G, Madar Z, Deshmukh S, Nyska A. Utilization of a Deep Learning Algorithm for Microscope-Based Fatty Vacuole Quantification in a Fatty Liver Model in Mice. Toxicol Pathol. 2020;48(5):702-7.

27. Garcia-Carretero R, Vigil-Medina L, Barquero-Perez O, Ramos-Lopez J. Relevant Features in Nonalcoholic Steatohepatitis Determined Using Machine Learning for Feature Selection. Metab Syndr Relat Disord. 2019;17(9):444-51.

28. Ma H, Xu CF, Shen Z, Yu CH, Li YM. Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China. BioMed research international. 2018;2018:4304376.

29. Perakakis N, Polyzos SA, Yazdani A, Sala-Vila A, Kountouras J, Anastasilakis AD, 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. 2019;101:154005.

TABLE LEGENDS

Table 1. Baseline characteristics of adult patients with biopsy-proven NAFLD.

Table 2. Diagnostic performances of LRM, FIB-4, APRI, and NFS scores in identifying the presence of significant liver fibrosis.

Table 3. Diagnostic performance of LRM in identifying the presence of significant liver fibrosis different in subgroups of patients.

Table 4. Diagnostic performances of the machine learning algorithm (MLA) and LRM in identifying the presence of significant liver fibrosis different in subgroups of patients.

FIGURE LEGENDS

Figure 1. Texture feature selection using the LASSO model.

(A) Selection of the tuning parameter (ג) in the LASSO model via 10-fold cross-validation based on minimum criteria. Binomial deviance from the LASSO regression cross-validation procedure were plotted as a function of log(ג). The y-axis indicates binomial deviance. The x-axis indicates the log(ג). Numbers along the upper x-axis represent the average number of predictors. Red dots indicate average deviance values for each model with a given ג, and vertical bars through the red dots show the upper and lower values of the deviance. The vertical black lines define the optimal values of ג, where the model provides its best fit to the data. The optimal ג value of 0.057 with log(ג) = - 2.87 was selected.

(B) The LASSO coefficient profiles of clinical features. The dotted vertical line was plotted at the value selected using 10-fold cross-validation in A. The five resulting features with non-zero coefficients are indicated in the plot.

Figure 2. Diagnostic performances of machine learning algorithm (marked as "a") and LRM (marked as "b") for identifying the presence of significant liver fibrosis both in the training set (A) and the validation set (B).

Figure 3. Diagnostic performances of the machine learning algorithm for identifying

the presence of advanced liver fibrosis in the whole cohort.

Supplementary Figure 1. A simpler schematic for machine learning

Supplementary Figure 2. Study flow chart.

From an initial sample of 966 individuals with suspected NAFLD (based on imaging techniques and/or elevated serum liver enzyme levels), we subsequently excluded 413 subjects for the following reasons: (1) alcohol consumption in excess of 140 g per week for men and 70 g per week for women (n = 108); (2) viral hepatitis (n = 196); (3) drug induced hepatitis (n=3), autoimmune hepatitis (n = 17) or liver cancers and other extra-hepatic tumors (n = 10); (4) missing biochemical data (n = 14); (5) refuse liver biopsy (n = 26); (6) hepatic steatosis < 5% on liver histology (n = 39).

Supplementary Figure 3. (A) Receiver operating characteristics (ROC) curve of LRM for predicting significant fibrosis. (**B**) Calibration curves for LRM. The calibration curve of the LRM to identify the presence of significant fibrosis showed good agreement with the predicted presence of significant fibrosis. An ideal model would result in a plot where the actual and predicted probabilities fall along the 45 degree line.

Supplementary Figure 4. Diagnostic performances of LRM, FIB-4, APRI, and NFS scores for identifying the presence of significant liver fibrosis both in the training set (A) and validation set (B).