Clinical and Molecular Hepatology - Review Article in the Special Issue on NAFLD

Noninvasive biomarkers for liver inflammation in NAFLD: current and future

Short title: Inflammation biomarkers for NAFLD

Terry Cheuk-Fung Yip*1,2,3, Fei Lyu*4, Huapeng Lin1,2,3, Guanlin Li1,2,3, Pong-Chi Yuen4, Vincent Wai-Sun Wong1,2,3, Grace Lai-Hung Wong1,2,3

1Medical Data Analytic Centre, 2Department of Medicine and Therapeutics, 3Institute of Digestive Disease, The Chinese University of Hong Kong, Hong Kong
4Department of Computer Science, Hong Kong Baptist University, Hong Kong
*TCF Yip and F Lyu have equal contribution.

Grant support: This work was supported by the Health and Medical Research Fund (HMRF) of the Food and Health Bureau (Reference no.: 07180216) awarded to Grace Wong.

Abbreviations: ALT = alanine aminotransferase, CI = confidence intervals, DM = diabetes mellitus, HCC = hepatocellular carcinoma.

Keywords (not in title): Cytokeratin-18, deep learning, fatty liver, hepatic steatosis, liver cancer, machine learning, necroinflammation.

Abstract word count: 147       Total word count: 4,224       No. of table: 4

Correspondence:
1. Grace Lai-Hung Wong, MD - Department of Medicine and Therapeutics, 9/F Prince of Wales Hospital, 30-32 Ngan Shing Street, Shatin, Hong Kong.
   Telephone: 852-3505-1205, Fax: 852-2637-3852, Email: wonglaihung@cuhk.edu.hk
2. Pong-Chi Yuen, PhD - Department of Computer Science, Room 634, David C Lam Building (DLB634), Hong Kong Baptist University, Kowloon Tong, Hong Kong.
   Telephone: 852-3411-7091, Fax: 852-3411-7892, Email: pcyuen@comp.hkbu.edu.hk

Disclosures:
Terry Yip has served as a speaker and an advisory committee member for Gilead Sciences.
Vincent Wong has served as an advisory committee member for AbbVie, Allergan, Echosens, Gilead Sciences, Janssen, Perspectum Diagnostics, Pfizer and Terns, and a speaker for Bristol-Myers Squibb, Echosens, Gilead Sciences and Merck.
Grace Wong has served as an advisory committee member for Gilead Sciences and Janssen, as a speaker for Abbott, Abbvie, Bristol-Myers Squibb, Echosens, Furui, Gilead Sciences, Janssen and Roche, and received research grant from Gilead Sciences.
The other authors declare that they have no competing interests.

Author Contributions: All authors were responsible for the writing plan, content, drafting and critical revision of the manuscript for important intellectual content.
Abstract

Inflammation is the key driver of liver fibrosis progression in nonalcoholic fatty liver disease (NAFLD).

Unfortunately, it is often challenging to assess inflammation in NAFLD due to its dynamic nature and poor correlation with liver biochemical markers. Liver histology keeps its role as the standard tool, yet it is well-known for substantial sampling, intraobserver and interobserver variability. Serum proinflammatory cytokines and apoptotic markers namely cytokeratin-18 (CK-18) are well studied with reasonable accuracy; whereas serum metabolomics and lipidomics have been adopted in some commercially available diagnostic models. Ultrasound and computed tomography (CT) imaging techniques are attractive because of their wide availability; yet their accuracies may not be comparable with magnetic resonance imaging (MRI)-based tools. Machine learning and deep learning models, be they supervised or unsupervised learning, are promising tools to identify various subtypes of NAFLD, including those with dominating liver inflammation, contributing to sustainable care pathways for NAFLD.
Introduction

Nonalcoholic fatty liver disease (NAFLD) affects over 30% of the general adult population and is emerging as an important cause of cirrhosis and hepatocellular carcinoma. Its more active form nonalcoholic steatohepatitis (NASH) is characterized by the presence of hepatic steatosis, inflammation (both lobular and portal) and hepatocyte ballooning. Assessment of inflammation is important. Although studies have consistently shown that the fibrosis stage has a stronger correlation with adverse liver-related outcomes than features of NASH, inflammation is after all the driver of fibrosis progression. Moreover, the United States Food and Drug Administration and the European Medicines Agency accept NASH resolution with no worsening of fibrosis and/or fibrosis improvement with no worsening of NASH as key histological endpoints for conditional approval of new drugs for NASH. Until the regulators accept the use of noninvasive surrogate biomarkers in place of liver biopsy, assessment of inflammation will remain central in the drug development process.

That being said, assessment of inflammation is difficult. Above all, there is substantial sampling, intraobserver and interobserver variability in the histological assessment of inflammation and diagnosis of NASH. When paired biopsies are performed to assess treatment response, errors at each biopsy add up. If the histological reference standard is unreliable, this would underestimate the performance of even an excellent biomarker. Moreover, compared with fibrosis, inflammation changes more rapidly. Therefore, the time interval between liver biopsy and noninvasive test assessment would have a greater impact on the evaluation of inflammation than fibrosis biomarkers. For the same reason, one may expect inflammatory markers to vary over time, and a single-point assessment may not mean much.

In this article, we review blood and imaging biomarkers of inflammation in NAFLD. We also highlight the emerging role of artificial intelligence and machine learning in diagnostics.

Liver Histology

Liver histology remains the standard to assess inflammation and diagnose NASH. Pathologists diagnose NASH based on a global picture that takes into account the degree and pattern of steatosis, inflammation and hepatocyte ballooning and/or the presence of Mallory-Denk bodies. In 2005, Kleiner and colleagues from the NASH Clinical Research Network proposed the NAFLD activity score, which is the numerical sum of the steatosis grade (0-3), lobular inflammation (0-3), and ballooning (0-2). Later, it was apparent that it is inappropriate to use the score to diagnose NASH, mainly because of the heavy weighting assigned to steatosis. Thus, a patient can have severe steatosis but
mild inflammation, resulting in a high NAFLD activity score but not meeting the pathological diagnosis of NASH. Currently, the NAFLD activity score is mainly used in early-phase clinical trials to evaluate treatment response.

In contrast, Bedossa and colleagues proposed the Steatosis-Activity-Fibrosis (SAF) score in 2012, thus separating the assessment of steatosis and inflammation.\textsuperscript{11} They also developed the Fatty Liver Inhibition of Progression (FLIP) algorithm, which essentially means that one can diagnose NASH when a patient scores 1 or more in steatosis, lobular inflammation and ballooning.\textsuperscript{12} The algorithm has demonstrated a higher degree of interobserver agreement.

One main limitation of the original scores is the relative underweighting of ballooning, which experts agree should be the defining feature of NASH. Besides, complete disappearance of ballooning is uncommon. This explains the very low percentage of patients with NASH resolution in clinical trials, rendering this histological endpoint often useless.\textsuperscript{13} Recently, Pai and colleagues proposed to expand the scale of ballooning scoring from 0-2 to 0-4 to increase granularity and reliability of the assessment of NASH.\textsuperscript{14}

Other than assessment variability, liver biopsy is also limited by its invasiveness nature, poor patient acceptance, cost, pain and potential complications.\textsuperscript{15} It is therefore important to develop noninvasive tests for routine clinical use.

**Serum markers**

Traditionally, alanine aminotransferase (ALT) and aspartate aminotransferase (AST) are used in routine clinical practice as biochemical markers of inflammatory damage in hepatocytes, or hepatitis in a simpler term. Unfortunately, a more active form of disease like NASH and advanced fibrosis is often found in NAFLD patients exhibiting normal aminotransferase levels; such levels may even paradoxically decrease in patients with progressive fibrosis,\textsuperscript{16} suggesting that ALT or AST levels are not reliable in establishing active inflammation in NAFLD. Combining routine clinical parameters is another popular approach; a handful of diagnostic panels were proposed and validated to identify liver inflammation in NASH (Table 1). Most of these models have the beauty of wide availability of parameters included and reasonably good diagnostic accuracy, yet specific cut-offs need to be further optimized.\textsuperscript{17}

Proinflammatory cytokines and apoptotic markers are possible diagnostic biomarkers for patients with NASH. The most evaluated NASH serum biomarker is cytokeratin-18 (CK-18), which is a well-recognized hepatocyte apoptosis product and accounts for about 5% of liver proteins.\textsuperscript{18} Two antigens of CK-18, M30 and M65, are of the same protein
yet distinctive mechanisms - M30 measures the caspase-cleaved CK-18 revealed during apoptosis, while M65 measures the full-length protein, including both caspase-cleaved and intact CK-18, which is released from cells undergoing necrosis. In general, models with CK-18 perform better than those with solely routine laboratory parameters (Table 1).

Serum metabolomics and lipidomics are also widely studied; pyroglutamic acid, phosphatidylcholine, sphingomyelin, fatty acids, hydroxyeicosatetraenoic acid (HETE), glycyrrhetic acid, taurocholate and various subtypes of triglycerides levels were incorporated in different models (Table 1). Some diagnostic models have been commercially available, e.g. by OWL Metabolomics.

While most of the biomarkers and models were derived and validated in a cross-sectional fashion, dedicated studies to evaluate the dynamic change, in particular, the reduction of score after treatment which correlates with inflammation improvement, are much warranted in the era of active development of novel therapeutics for NASH.

Ultrasound imaging (Table 2)

Transabdominal ultrasonography

Conventional B-model ultrasound (US) is the most widely used imaging technique for the non-invasive assessment of NAFLD. US is currently the first-line diagnostic approach for NAFLD suggested by clinical practice guidelines of the European Association for the Study of the Liver due to its low cost, wide availability and repeatability. However, several studies found no correlation between the US characteristics and liver histologic features including inflammation and ballooning. Hamaguchi scoring system was developed based on US findings including bright liver and hepatorenal echo contrast (0-3), deep attenuation (0-2), vessel blurring (0-1). The scoring system further improved the diagnostic performance of NAFLD in obese patients, with an AUROC of 0.98. Ultrasonographic fatty liver indicator (US-FLI) is another scoring system ranging from 2-8 based on the intensity of liver or kidney contrast, attenuation of ultrasound beam, vessel blurring, and visualization of gallbladder wall, diaphragm and areas of focal sparing. The AUROC of US-FLI for predicting NASH was 0.80 (0.68-0.92) and US-FLI was correlated with lobular inflammation according to Kleiner’s criteria. Hamaguchi score and US-FLI score lack validation in large series of patients, and whether the dynamic change of scores correlates with inflammation progression or improvement need to be validated in the future.
Vibration-controlled transient elastography

Vibration-controlled transient elastography (VCTE) technique measures the velocity of shear wave through the liver parenchyma and the velocity is related to the degree of liver tissue stiffness. Controlled attenuation parameter (CAP) captures the attenuation in the amplitude of ultrasound waves to estimate the degree of hepatic steatosis and it was available for clinical practice since 2010. Fibroscan 502 Touch was the first VCTE device commercially available with CAP. An examination is considered valid if ≥10 valid liver stiffness measurement (LSM) and CAP and an interquartile range-to-median ratio of the measurements of ≤0.3 of LSM and CAP. According to previous studies, Fibroscan has high accuracy, simplicity, and reproducibility to assess hepatic steatosis and fibrosis. Series studies focus on the discriminative ability of CAP and LSM for NASH patients. HW Lee et al. conducted a prospective Korean study based on 183 patients with biopsy-proven NAFLD patients and showed that a cutoff value of 7 kPa for liver stiffness by VCTE can achieve an AUROC of 0.75 (95%CI 0.68-0.82), a sensitivity of 73.4% and a specificity of 78.7%. Based on VCTE, they developed a scoring system named “CLA score” using three independent predictors including CAP value, liver stiffness by VCTE and ALT level to identify NASH patients. CLA score had a significantly higher diagnostic performance than NAFLD fibrosis score (NFS) (AUROC 0.81 vs 0.62).

FAST score

FibroScan-AST (FAST) score was a logistic regression-based scoring system for detecting fibrotic NASH, which includes liver stiffness by VCTE, CAP and AST. The diagnostic performance of FAST score was validated in multiple large global cohorts. AUROC's were ranged from 0.74 to 0.95, with sensitivity and specificity up to 1 and 0.86, and NPV ranged from 0.73 to 1. Compared to FIB4, NFS, and APRI, FAST score had a significantly higher diagnostic performance for fibrotic NASH. FAST can be used as a noninvasive tool to screen fibrotic NASH to reduce the number of unnecessary liver biopsies. The relationship between dynamic changes of FAST score and liver inflammation should be explored in the future.

Computed tomography

Computed tomography (CT) uses computer processing of X-ray data of the body to produce images created from detection of X-rays traversing tissues. While CT is widely used to characterize focal liver lesions, in NAFLD patients CT is more often studied to assess steatosis and fibrosis but not as much for inflammation. Only one retrospective study of 88 NAFLD patients found that non-contrast-enhanced CT (NECT) texture analysis with a 2-mm filter predicted NASH with accuracy above 90%; yet the accuracy dropped to 60% if a 4-mm filter was used.
emerging CT techniques including dual-energy CT (DECT), post-processing software, perfusion CT (pCT), and photon-counting detector CT (PCD-CT) are promising tools and potentially more accurate to detect inflammation. Currently, CT is not the preferred primary modality to measure liver inflammation given its lack of sensitivity for steatohepatitis and the need for exposure of the subjects to radiation.

Magnetic resonance imaging (Table 3)

LiverMultiScan

LiverMultiScan (LMS) is an emerging diagnostic tool using multiparametric magnetic resonance imaging (MRI) to quantify liver disease. The technology is comprised of corrected T1 (cT1), T2, and liver fat assessment by advanced MRI. LMS measures the amount of iron in the liver to correct for its effect on T1-cT1 as excess iron in the liver reduces T1 relaxation time and leads to underestimation of liver disease. cT1 correlates with necroinflammation and fibrosis and may serve as a non-invasive method in NASH. LMS had fewer technical failures, especially compared with ultrasound-based techniques which were less reliable in patients with a higher body mass index. The success rate exceeded 95% in previous clinical studies. One recent pooled study examined the utility of cT1 and proton density fat fraction (PDFF) for identifying NASH and fibrotic NASH. The diagnostic accuracy (AUROC) of cT1 to identify patients with NASH was 0.78 (95% CI, 0.74–0.82), for MRI liver fat was 0.78 (95% CI, 0.73–0.82), and when combined cT1 with MRI liver fat was 0.82 (95% CI, 0.78–0.85). The diagnostic accuracy of cT1 to identify patients with fibrotic NASH (AUROC [0.78; 95% CI, 0.74–0.82]) was superior to that of MRI liver fat (AUROC [0.69; 95% CI, 0.64–0.74]). There is one ongoing study (NCT03743272) which aims to investigate the repeatability and reproducibility of LMS. Multiparametric MRI has been evaluated to be associated with liver-related clinical outcomes in a cohort of patients with chronic liver disease. Longitudinal change of MRI-PDFF correlated well with the biopsy results, and there was one study evaluated that a 30% relative decline in MRI-PDFF predicted fibrosis regression in NAFLD patients.

MEFIB

MEFIB index was a combination of MR elastography and fibrosis-4 (FIB-4) used for identification of fibrotic NASH. The positive predictive value (PPV) exceeded 90% with AUROC of 0.84 (95% CI, 0.78-0.89) in the validation cohort of the study by Jung et al. MEFIB was evaluated to have a higher diagnostic accuracy than MAST and FAST score for significant fibrosis and also fibrotic NASH. The MEFIB index had a robust association with liver-related outcome with a hazard ratio of 20.6 (95% CI, 10.4–40.8) and the negative predictive value (NPV) for the
outcome reached 99.1% at 5 years. Future studies should explore if the dynamic change of MEFIB index correlated with liver-related outcomes.

MAST

Given that MRI-PDFF has been shown to be more accurate than VCTE-based CAP in identifying all grades of steatosis in patients with NAFLD, and MR elastography is more accurate than VCTE in detecting liver fibrosis, Noureddin et al. proposed the MAST score based on MRI-PDFF, MR elastography and AST value. In their validation cohort, the MAST score demonstrated high performance and discrimination (AUROC 0.93, 95% CI 0.88-0.97) which was significantly better than NAFLD fibrosis score, FIB-4 index and FAST score. However, MEFIB index showed a higher AUROC with the PPV and NPV reached 95.3% and 90.1% for ruling in and ruling out fibrotic NASH compared with MAST in a head-to-head comparison study. There is still a lack of published studies on prognostication and also the dynamic change in fibrosis progression or regression by MAST score.

3D MR elastography

Recently, several studies by Allen et al from Mayo Clinic evaluated the role of three-dimensional (3D) MR elastography in identifying NASH in patients undergoing bariatric surgery. By combing the 3D MR elastography with MRI-PDFF, the AUROC was 0.73 for the diagnosis of NASH. Additionally, they demonstrated that the 3D MR elastography and MRI-PDFF could detect histologic changes in NASH resolution after bariatric surgery. There are limited studies on the association between 3D MR elastography with liver-related outcomes.

Machine learning models

In the past decade, the advancement of artificial intelligence has led to its numerous applications in hepatology. Artificial intelligence, machine learning, and deep learning can be considered three overlapping domains that use computer programs to mimic functions of human intelligence, including learning, problem solving, classification, and decision making. Particularly, machine learning methods are usually applied for developing diagnostic or predictive models. Machine learning and deep learning algorithms can be supervised or unsupervised. Supervised learning methods occur when a label for the outcome is given in the training data. For example, if we aim to predict the presence of NASH among patients with biopsy-proven NAFLD, the information of whether the patients had NASH needs to be provided to the learning algorithms during training so that the model can distinguish patients with and without NASH based on that. As a result, the learning algorithm can identify combinations and interactions of factors
that best separate the two groups of patients and yield an accurate prediction. In contrast, information on the presence and absence of NASH is not provided in unsupervised learning. The purpose of unsupervised learning is to identify several clusters of patients who are similar in terms of data distribution. In other words, patients within the same cluster have similar clinical characteristics, which may represent a certain disease phenotype or subtype.

Common supervised machine learning algorithms examined in identifying inflammation in NAFLD patients including logistic regression with penalization, decision tree, random forest, support vector machine, and different boosting methods. Regarding the use of covariates, existing literature usually includes laboratory parameters or histological features from liver biopsy for the prediction. Fialoke and colleagues utilized electronic health records from the Optum administrative claim dataset to develop machine learning models for identifying NASH patients from NAFLD patients or healthy patients without NAFLD. In this study, NAFLD and NASH were identified based on diagnosis codes. Supervised machine learning algorithms including logistic regression, decision tree, random forest, and eXtreme Gradient Boosting (XGBoost) were examined. Temporal mean of laboratory parameters including ALT, AST, and platelets, together with age, gender, race, and the presence of type 2 diabetes were included as covariates. The four models yielded satisfactory classification performance with AUROC over 0.83 in internal validation (Table 4). This study demonstrated the possibility of using machine learning in identifying NASH in a large group of patients, while the good performance may be due to a more obvious separation between healthy individuals and NASH patients.

NASHmap is another example of machine learning model for predicting NASH. Docherty and colleagues utilized a biopsy cohort to derive the machine learning models. Similarly, logistic regression, classification and regression trees (CART) (aka decision tree), random forest, and XGBoost were considered. Fourteen clinical and laboratory parameters were included in the models, which yielded AUROCs of around 0.7-0.8. Hemoglobin A1c (HbA1c) was found to be the most predictive covariate, followed by AST and ALT. The models were then externally validated in the Optum dataset and demonstrated comparable AUROC. Slightly reduced performance was observed in reduced models using 5 parameters including HbA1c, AST, ALT, total protein, and triglycerides. Moreover, Canbay et al. developed a logistic regression model to distinguish NASH from NAFLD in obese patients, with an AUROC of 0.70 in an independent validation cohort. The logistic model included age, gamma-glutamyl transferase, CK-18 M30, adiponectin, and HbA1c. All these laboratory-based machine learning models highlighted the importance of HbA1c, AST, and ALT in identifying NASH patients. On the other hand, there is emerging evidence of the difference in the characteristics of lipidomic, glycomic, and hormonal features in patients with NAFLD and NASH due to their strong
relationship with metabolic syndrome. Perakakis and colleagues incorporated these omics features into machine learning models including support vector machine, k-nearest neighbor classifier, and random forest. Using 29 features, the machine learning models achieved AUROCs of over 0.95 in selecting patients with NASH from patients with NAFLD or healthy individuals in internal validation (Table 4).56

Unsupervised learning can be useful to identify clinically relevant subtypes of NAFLD patients including those with significant liver inflammation. Using a hierarchical clustering algorithm based on Manhattan distance of similarity, Vandromme and colleagues identified 5 disease subtypes among NAFLD patients.57 Some of the subtypes showed evidence of liver inflammation such as a high proportion of elevated ALT, as well as notable comorbidities such as diabetes and hypertension.

The presence of lobular inflammation is one of the key histological characteristics of NAFLD activity score (NAS) besides the presence of hepatocyte ballooning and steatosis. Traditional scoring system such as NAFLD activity score only offers a non-linear and categorical assessment of the disease. Machine learning thus has a role here to provide quantification of the assessment.58 Liu and colleagues developed an algorithm to analyze the liver biopsy and quantify different components of the NASH Clinical Research Network (CRN) scoring system. They used special microscopy and image analysis to visualize and quantify inflammation in liver biopsy.58 The algorithms performed well in a three-center study to predict lobular inflammation and other components of the NASH CRN scoring system (Table 4).

Deep Learning Methods

Deep learning methods attempt to train deep neural networks for solving complex problems and show more promising prediction results compared to traditional methods based on handcrafted features. Recent deep learning techniques have led to wide applications in healthcare areas,59 and they have been increasingly applied for the prediction and diagnosis of NASH. Popular deep learning approaches include Convolutional Neural Network (CNN), Graph Neural Network (GNN) and Recurrent Neural Network (RNN). Besides developing sophisticated network architectures to improve prediction accuracy, other important questions in deep learning methods are also explored, such as model interpretability and annotation-efficient learning.

CNN is the most widely used technique of deep learning and has been proved effective in solving many medical problems. CNN achieves better performance when dealing with image-related tasks, such as analyzing CT, MRI, and
pathology data. A typical model based on CNN contains a series of layers, including convolution layers, pooling layers, and fully connected layers. In convolution layers, each convolutional neuron only processes data within its receptive field, thus the architecture is ideal for large-scale data such as high-resolution images. NAS is important for diagnosing NASH, and liver biopsy is used for calculating NAS. CNN can be used for quantitative measurement of liver histology and disease monitoring in NASH, and CNN-based methods are proven accurate with strong correlations with expert pathologists and good risk stratification of patients with NASH.\textsuperscript{60,61} CT is non-invasive and less expensive compared with liver biopsy, recent works proposed to combine the information from CT and pathology data for predicting NAS and fibrosis stage.\textsuperscript{52} CNN is first used for feature extraction, and different fusion strategies are proposed to combine these two pieces of information for better prediction performance. Their results showed that combining data from different modalities is beneficial for improving the prediction performance of NAS. To conclude, existing studies demonstrate that CNNs can automatically learn better features for NASH diagnosis compared with traditional approaches based on manually designed features.

GNN is a rapidly growing field of deep learning, which is suitable for processing graph data which contains rich relation information among elements.\textsuperscript{63} GNN is able to extract multi-scale localized spatial features by exchanging information between the nodes of graphs, and its key element is pairwise message passing. There is an increasing number of GNN applications, such as electrical health records modeling and synthesizing chemical compounds. GNN is also attracting more attention in pathology data analysis,\textsuperscript{64} since they learn features which can well represent the tissue spatial structure. A recent work proposes to study liver biopsy on two histological stains namely Trichrome (TC) and hematoxylin and eosin (H&E) with GNN.\textsuperscript{65} The latent embeddings extracted from the graphs are concatenated to predict NAS, and their results show superiority over competing methods. Graph representation is able to integrate the tissue features from the whole slide image, and deserves further study in the evaluation of tissue biopsies for NASH diagnosis.

RNN can process data with any length, and is a good choice for sequential data processing.\textsuperscript{66,67} Electronic health records (EHRs) contain medical time series of laboratory tests, and RNN-based methods can analyze the conditions of patients using these records. Long short-term memory (LSTM) is a representative method of RNN, and its gating mechanism within each LSTM cell is effective to avoid the long-term dependency problem in standard RNNs. Deep learning approaches based on LSTM are utilized to identify patients at risk of developing NASH, and show better performance compared with other competing methods like XGBoost.\textsuperscript{68} Considering there is a large amount of EHRs
available in hospitals, RNN-based methods can work as powerful tools to analyze these existing valuable data for NASH diagnosis.

Even though deep learning methods have achieved great success in solving many medical problems, applying them to clinical practice remains skeptical. Deep learning methods are often described as ‘black boxes’, however, interpretability is especially important in the medical domain. Some recent works attempted to deal with the interpretability problem of deep learning methods. One promising solution is to incorporate domain knowledge into model design. For example, clinically interpretable features (e.g., nuclei and fat droplets) can be incorporated into NAS prediction. Pathologists normally focus on the nuclei and fat droplet regions for evaluating a liver biopsy image and developing models to mimic the diagnosis process of pathologists is proven effective. Moreover, the success of deep learning models depends on large-scale training data, while collecting such datasets is extremely difficult in the medical domain. Developing data-efficient deep learning models is thus important and requires further study for NASH diagnosis, and one possible solution is to fully utilize free-text reports stored in hospital archiving and communication systems.

Conclusions and perspectives

This review summarizes the latest developments in histological and noninvasive assessments of inflammation in NAFLD. In routine clinical practice, noninvasive tests have already largely replaced liver biopsy in the evaluation of patients with NAFLD. However, liver biopsy remains valuable in case of diagnostic uncertainty, such as uncertain etiology or indeterminate or conflicting noninvasive test results. At present, liver biopsy is still required in late phase clinical trials for NASH. The limitation of serial liver biopsies to determine NASH resolution is well documented. Artificial intelligence-aided assessment of key histological features including ballooning and fibrosis has made much progress and should be incorporated into future clinical trials, subject to agreement by the regulators. To the least, artificial intelligence has consistently demonstrated a much higher reproducibility than traditional pathological assessments. Eventually, the aim should be to use noninvasive tests in both clinical trials and routine clinical practice. With a disease that affects over 30% of the population, noninvasive tests are simply the only feasible option if we are to build robust and sustainable clinical care pathways and improve NAFLD management.
References


12. Bedossa P, Consortium FP. Utility and appropriateness of the fatty liver inhibition of progression (FLIP) algorithm and steatosis, activity, and fibrosis (SAF) score in the evaluation of biopsies of nonalcoholic fatty liver disease. *Hepatology* 2014;60(2):565-75.


Table 1. Diagnostic models for liver inflammation in nonalcoholic steatohepatitis (NASH) (adapted from Zeng et al).16

<table>
<thead>
<tr>
<th>Models</th>
<th>Variables</th>
<th>AUROC</th>
<th>Cutoff</th>
<th>Sn</th>
<th>Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLI</td>
<td>BMI, waist, TG, GGT</td>
<td>0.84</td>
<td>&lt;30 and ≥60</td>
<td>87%</td>
<td>86%</td>
</tr>
<tr>
<td>HAIR score</td>
<td>HT, ALT, insulin, Glu</td>
<td>0.68</td>
<td>3</td>
<td>57%</td>
<td>77%</td>
</tr>
<tr>
<td>NASHTest-2</td>
<td>A2M, ApoA1, Hapt, TBil, GGT, TC, TG</td>
<td>0.59</td>
<td>0.5</td>
<td>83.3%</td>
<td>37.5%</td>
</tr>
<tr>
<td>MACK-3</td>
<td>CK-18 M30, AST, HOMA</td>
<td>0.81</td>
<td>≤0.167 and ≥0.551</td>
<td>84.2%</td>
<td>81.4%</td>
</tr>
<tr>
<td>G-NASH</td>
<td>CK-18 M30, GP73</td>
<td>0.85</td>
<td>N.A.</td>
<td>82.1%</td>
<td>80.5%</td>
</tr>
<tr>
<td>Nice model</td>
<td>CK-18, ALT, MS</td>
<td>0.88</td>
<td>0.14</td>
<td>84%</td>
<td>86%</td>
</tr>
<tr>
<td>FIC-22</td>
<td>CK-18 M30, FIB-4</td>
<td>0.82</td>
<td>1</td>
<td>89.1%</td>
<td>62.5%</td>
</tr>
<tr>
<td>NASH diagnostic™</td>
<td>CK-18 M30, adiponectin, resistin</td>
<td>0.91</td>
<td>0.2272</td>
<td>94.45%</td>
<td>70.21%</td>
</tr>
<tr>
<td>CHeK</td>
<td>CK-18 M30, GGT, age, HbA1c, adiponectin</td>
<td>0.73</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>NASH score</td>
<td>PNPLA3, insulin, AST</td>
<td>0.77</td>
<td>-1.054</td>
<td>75%</td>
<td>74%</td>
</tr>
<tr>
<td>NASH PT score</td>
<td>PNPLA3, TM6SF2, diabetes, AST, HOMA-IR, hsCRP</td>
<td>0.86</td>
<td>-0.785</td>
<td>91%</td>
<td>58.1%</td>
</tr>
<tr>
<td>NIS4</td>
<td>miRNA-34a, A2M, YKL-40, HbA1c</td>
<td>0.80</td>
<td>&lt;0.36</td>
<td>80.8%</td>
<td>65.2%</td>
</tr>
<tr>
<td>GlycoNASHTest</td>
<td>Log(NGA2F/NA2)</td>
<td>0.74</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

A2M: alpha-2 macroglobulin; ALT = alanine aminotransferase; ApoA1 = apolipoprotein A1; AUC = area under the receiver operating characteristic curve; BMI = body mass index; CK-18 = cytokeratin-18; FGF-21 = fibroblast growth factor-21; FIB-4 = fibrosis-4; GGT = γ-glutamyl transpeptidase; Glu = glucose; GP73 = golgi protein 73; HT = hypertension; Hapt = haptoglobin; HbA1c = glycosylated hemoglobin; HOMA = homeostasis model assessment; hsCRP = high-sensitivity C-reactive protein; IL-6 = interleukin-6; miRNA = MicroRNA; MS = metabolic syndrome; N.A. = not available; PLT = platelet; Sn = sensitivity; Sp = specificity; TBil = total bilirubin; TC = total cholesterol; TG = triglycerides.
Table 2. Diagnostic performance of ultrasound imaging for liver inflammation in nonalcoholic steatohepatitis (NASH).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Variables</th>
<th>Outcome</th>
<th>AUROC</th>
<th>Cutoff</th>
<th>Sn</th>
<th>Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>NA</td>
<td>Severe NAFLD</td>
<td>0.93</td>
<td>NA</td>
<td>84.8%</td>
<td>93.6%</td>
</tr>
<tr>
<td>US-FLI</td>
<td>US findings</td>
<td>NASH</td>
<td>0.80</td>
<td>5</td>
<td>83.3%</td>
<td>62.9%</td>
</tr>
<tr>
<td>VCTE</td>
<td>NA</td>
<td>NASH</td>
<td>0.75</td>
<td>7</td>
<td>73.4%</td>
<td>78.7%</td>
</tr>
<tr>
<td>FAST score</td>
<td>Liver stiffness by VCTE, CAP and AST</td>
<td>Fibrotic NASH</td>
<td>0.74- 0.95</td>
<td>≤0.35 and ≥0.67</td>
<td>64%- 100%</td>
<td>35%-86%</td>
</tr>
</tbody>
</table>

AUROC, area under the receiver operating characteristic curve; NA, not available; Sn, sensitivity; Sp, specificity; US, Conventional B-model ultrasound; US-FLI, Ultrasonographic fatty liver indicator; VCTE, vibration-controlled transient elastography.
## Table 3. Diagnostic performance of magnetic resonance imaging for nonalcoholic steatohepatitis

<table>
<thead>
<tr>
<th>Models</th>
<th>Variables</th>
<th>Outcome</th>
<th>AUROC</th>
<th>Sn</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiverMultiScan</td>
<td>cT1, T2 and PDFF</td>
<td>Fibrotic NASH</td>
<td>0.69-0.79</td>
<td>0.39-</td>
<td>0.56-</td>
<td>0.45-</td>
<td>0.78-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.86</td>
<td>0.90</td>
<td>0.60</td>
<td>0.91</td>
</tr>
<tr>
<td>MEFIB</td>
<td>MRE and FIB-4</td>
<td>Fibrotic NASH</td>
<td>0.84-0.90</td>
<td>0.85-</td>
<td>0.94-</td>
<td>0.91-</td>
<td>0.85-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.94</td>
<td>0.98</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>MAST</td>
<td>MRE, PDFF and AST</td>
<td>Fibrotic NASH</td>
<td>0.86-0.93</td>
<td>0.89-</td>
<td>0.89-</td>
<td>0.50-</td>
<td>0.91-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.94</td>
<td>0.90</td>
<td>0.55</td>
<td>0.98</td>
</tr>
<tr>
<td>3D MRE</td>
<td></td>
<td>NASH</td>
<td>0.73</td>
<td>0.67</td>
<td>0.80</td>
<td>0.73</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Abbreviation: AUROC, the area under the receiver operating characteristic curve; Sn, sensitivity; Sp, specificity; PPV, positive predictive value; NPV, negative predictive value; NASH, nonalcoholic steatohepatitis; MRE, MR elastography; FIB-4, fibrosis-4; AST, aspartate aminotransferase.
<table>
<thead>
<tr>
<th>Study</th>
<th>Machine learning algorithms</th>
<th>Predicted variable</th>
<th>AUROC</th>
<th>Cutoff</th>
<th>Sn</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fialoke et al.</strong></td>
<td>DT with 3 temporal laboratory and 3 demographic variables</td>
<td>NASH vs. healthy individuals</td>
<td>0.842*</td>
<td>0.5</td>
<td>74.5%</td>
<td>N.A.</td>
<td>78.6%</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td>LR with 3 temporal laboratory and 3 demographic variables</td>
<td>NASH vs. healthy individuals</td>
<td>0.835*</td>
<td>0.5</td>
<td>74.3%</td>
<td>N.A.</td>
<td>77.0%</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td>RF with 3 temporal laboratory and 3 demographic variables</td>
<td>NASH vs. healthy individuals</td>
<td>0.870*</td>
<td>0.5</td>
<td>76.8%</td>
<td>N.A.</td>
<td>80.4%</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td>XGB with 3 temporal laboratory and 3 demographic variables</td>
<td>NASH vs. healthy individuals</td>
<td>0.876*</td>
<td>0.5</td>
<td>77.4%</td>
<td>N.A.</td>
<td>80.8%</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Docherty et al.</strong></td>
<td>DT with 14 clinical and laboratory variables</td>
<td>NASH vs. NAFLD</td>
<td>0.72</td>
<td>N.A.</td>
<td>78%</td>
<td>N.A.</td>
<td>76%</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td>LR with 14 clinical and laboratory variables</td>
<td>NASH vs. NAFLD</td>
<td>0.77</td>
<td>N.A.</td>
<td>79%</td>
<td>N.A.</td>
<td>79%</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td>RF with 14 clinical and laboratory variables</td>
<td>NASH vs. NAFLD</td>
<td>0.82</td>
<td>N.A.</td>
<td>82%</td>
<td>N.A.</td>
<td>80%</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td>XGB with 14 clinical and laboratory variables</td>
<td>NASH vs. NAFLD</td>
<td>0.82</td>
<td>N.A.</td>
<td>81%</td>
<td>N.A.</td>
<td>81%</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Canbay et al.</strong></td>
<td>LR with 5 clinical and laboratory variables</td>
<td>NASH vs NAFLD among obese patients</td>
<td>0.70</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Perakakis et al.</strong></td>
<td>SVM using 29 lipidomic features</td>
<td>NSAH vs. healthy individuals or NAFLD patients</td>
<td>0.96</td>
<td>N.A.</td>
<td>92%</td>
<td>93%</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Study</td>
<td>Algorithms</td>
<td>Predicted variable</td>
<td>AUROC</td>
<td>Cutoff</td>
<td>Sn</td>
<td>Sp</td>
<td>PPV</td>
<td>NPV</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>--------------------------------------------</td>
<td>-------</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
</tbody>
</table>
| Liu et al.  | qInflammation | Lobular inflammation^  
0 vs. ≥1 | 0.838 | 1.251  | 83%  | 100%| 100%| 14% |
|             | qInflammation | Lobular inflammation^  
≤1 vs. ≥2 | 0.820 | 1.357  | 93%  | 58% | 58% | 93% |
|             | qInflammation | Lobular inflammation^  
≤2 vs. 3 | 0.831 | 1.503  | 100%| 79% | 12% | 100%|

AUROC = area under the receiver operating characteristic curve, DT = decision tree, LR = logistic regression, NPV = negative predictive value, PPV = positive predictive value, RF = random forest, Sn = sensitivity, Sp = specificity, SVM = support vector machine, XGB = XGBoost.  
* AUROC by internal validation with cross validation.  
& AUROC in an independent validation cohort.  
^ Lobular inflammation based on NASH CRN scoring system.