

ORIGINAL ARTICLE

# Machine learning to predict progression of non-alcoholic fatty liver to non-alcoholic steatohepatitis or fibrosis

Sina Ghandian,\* Rahul Thapa,\* Anurag Garikipati,\* Gina Barnes,† Abigail Green-Saxena,†  Jacob Calvert,\* Qingqing Mao‡ and Ritankar Das‡

Departments of \*Data Science, †Research and Writing and ‡Leadership, Houston, Texas, USA

## Key words

fibrosis, machine learning, non-alcoholic steatohepatitis.

Accepted for publication 6 February 2022.

## Correspondence

Abigail Green-Saxena, Dascena, Inc., 12333 Sowden Road, Ste B, PMB 65148, Houston, TX 77080-2059, USA.  
Email: abigail@dascena.com

Sina Ghandian and Rahul Thapa contributed equally to this work.

**Declaration of conflict of interest:** All authors who have affiliations listed with Dascena (Houston, Texas, U.S.A) are employees or contractors of Dascena.

**Author contribution:** Sina Ghandian, Rahul Thapa, Anurag Garikipati, contributed to the data analysis of this work; Gina Barnes, Abigail Green-Saxena, contributed to the drafting and revision of this work; and Jacob Calvert, Ritankar Das, and Qingqing Mao contributed to the conception and revision of this work.

## Introduction

Non-alcoholic fatty liver (NAFL) disease consists of a spectrum of liver disorders from isolated steatosis, termed NAFL, to non-alcoholic steatohepatitis (NASH). NASH is considered to be progressive and is a more severe form of NAFL. While a minority (10–20%) of NAFL patients develop NASH, they are at increased risk for progression to fibrosis, cirrhosis, hepatocellular carcinoma (HCC), and liver-related mortality.<sup>1</sup> Although previously under-recognized, NAFL patients may also progress to advanced fibrosis.<sup>2,3</sup> NAFL and NASH have a large economic impact on health care use.<sup>4,5</sup> Estimates of total NAFL-associated costs in the United States are approximately 103 billion dollars per year.<sup>6</sup> The prevalence of risk factors for NAFL, such as diet westernization, sedentary lifestyle, obesity, and type 2 diabetes (T2D) mellitus has increased, resulting in increased prevalence of NAFL. It is estimated that up to 34% of the United States and

## Abstract

**Background:** Non-alcoholic fatty liver (NAFL) can progress to the severe subtype non-alcoholic steatohepatitis (NASH) and/or fibrosis, which are associated with increased morbidity, mortality, and healthcare costs. Current machine learning studies detect NASH; however, this study is unique in predicting the progression of NAFL patients to NASH or fibrosis.

**Aim:** To utilize clinical information from NAFL-diagnosed patients to predict the likelihood of progression to NASH or fibrosis.

**Methods:** Data were collected from electronic health records of patients receiving a first-time NAFL diagnosis. A gradient boosted machine learning algorithm (XGBoost) as well as logistic regression (LR) and multi-layer perceptron (MLP) models were developed. A five-fold cross-validation grid search was utilized for hyperparameter optimization of variables, including maximum tree depth, learning rate, and number of estimators. Predictions of patients likely to progress to NASH or fibrosis within 4 years of initial NAFL diagnosis were made using demographic features, vital signs, and laboratory measurements.

**Results:** The XGBoost algorithm achieved area under the receiver operating characteristic (AUROC) values of 0.79 for prediction of progression to NASH and 0.87 for fibrosis on both hold-out and external validation test sets. The XGBoost algorithm outperformed the LR and MLP models for both NASH and fibrosis prediction on all metrics.

**Conclusion:** It is possible to accurately identify newly diagnosed NAFL patients at high risk of progression to NASH or fibrosis. Early identification of these patients may allow for increased clinical monitoring, more aggressive preventative measures to slow the progression of NAFL and fibrosis, and efficient clinical trial enrollment.

25.2% of the worldwide population may have NAFL.<sup>7–9</sup> Similarly, increasing global rates of T2D and obesity are expected to lead to an increase in the burden of NASH and associated complications.<sup>4</sup>

Currently, there are insufficient fully validated noninvasive diagnostic tests for NAFL.<sup>10</sup> The gold standard for NASH diagnosis and staging is liver biopsy, which is invasive, costly, and incurs procedure-related risk for patients.<sup>11</sup> Noninvasive diagnostic procedures using ultrasound and MRI have been developed. However, these methods remain costly and time-consuming, require expert use and interpretation, cannot capture all histopathological features assessed during evaluation for NASH, and are not fully validated for diagnosis of NASH. Serum biomarkers and biomarker-based indexes have also been explored as diagnostic aids, but are not widely in use, as they are not perceived to provide additional actionable information relative to standard clinical evaluation.<sup>12,13</sup> Disease activity scores, such as the

NAFL Activity Score (NAS) utilize liver biopsies yet lack utility in predicting disease progression and are not considered a clinical standard at this time.<sup>9</sup> An urgent need remains for tools to predict specific clinical outcomes in NAFL.<sup>9</sup>

Early detection is a significant challenge in NASH and fibrosis diagnosis and management.<sup>14,15</sup> Improved screening methods for patients at high risk of progression will enable clinicians to identify high-risk NAFL patients requiring early intensive intervention. These methods may also assist investigators and sponsors of clinical trials in identifying appropriate patients for NAFL, NASH, and fibrosis drug clinical trials.<sup>16,17</sup> This study utilized clinical information from NAFL-diagnosed patients to predict the likelihood of progression to NASH or fibrosis.

## Methods

**Data sources.** Data were collected from a proprietary national longitudinal electronic health record (EHR) repository that incorporates clinical, claims, and other medical administrative data obtained from over 700 inpatient and ambulatory care sites. The data were extracted from January 2016 to June 2020, and resulted from aggregation across several different EHR systems. All data were de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA) and thus, this study did not require institutional review board approval as per 45 Code of Federal Regulations 46.102. The dataset included lab results, vital sign measurements, medication orders, patient diagnoses, and demographic information, as well as natural language processing (NLP)-extracted features derived from clinicians' notes (Table S1, Supporting information).

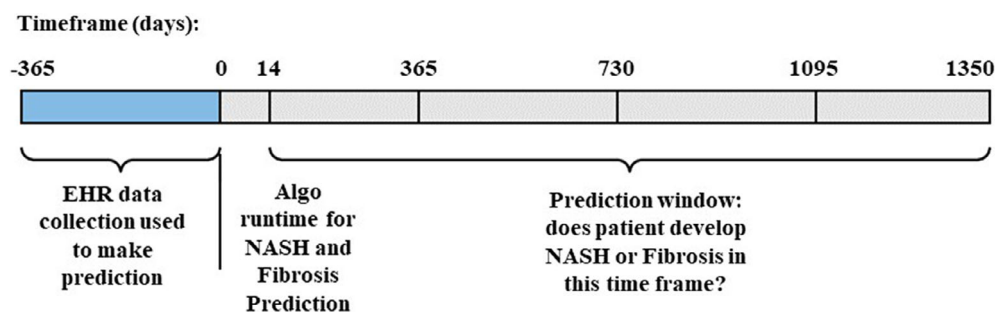
**Gold standard.** NAFL, NASH, and fibrosis were defined using International Classification of Disease, Tenth Revision (ICD-10) codes (Table S2, Supporting information). To meet the gold standard definition of the study, progression to NASH or fibrosis from NAFL had to occur within 4 years—but no earlier than 14 days—from the first diagnosis of NAFL. The majority of patients in the dataset developed NASH or fibrosis within this period (Fig. 1).

**Patient measurements and inclusion criteria.** Minimal inclusion criteria were applied in order to increase the likelihood of our studied population having sufficient data. Only patients that had at least one of their vital signs measured at the

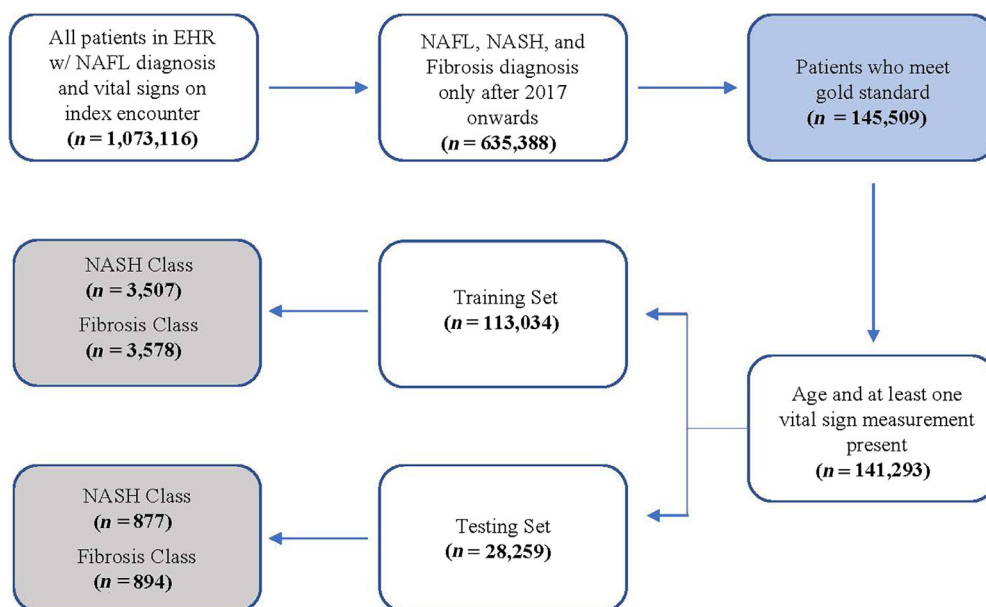
index encounter in which they were diagnosed with NAFL were included in the study. Further, patients were required to have had data available in the database for at least 2 years after their first NAFL diagnosis to be included, with at least one additional healthcare encounter within 2 years from their index encounter (Fig. 2).

Inputs for the machine learning algorithm (MLA) included demographic features, vital signs, and laboratory results that were reported in encounters 1 year prior to the index encounter when a patient was diagnosed with NAFL (Table 1). Vital signs and laboratory results were first classified into measurements taken at the current or from any previous hospital encounter. A previous hospital encounter was defined as any visit, outpatient or inpatient, that happened >24 h prior to the current encounter. The measurements for each included feature were then summarized by their mean, standard deviation, minimum, maximum, and total number of observations. As a result, there were distinct summary statistics for prior and current measurements for every lab or vital sign feature included in the model. This method of incorporating clinical observations was chosen because their distributions across an encounter and over time may indicate underlying health and illness severity, and are therefore helpful to the model. Including all measurements taken over a patient's current or previous encounter would allow for a more direct time-series analysis. However, that representation is generally more valuable for the inpatient setting, as acute conditions can change rapidly. Because NASH and fibrosis develop at a slower rate, the use of a more explicit feature set was chosen in order to represent the data more simply and with greater explainability.

**Machine learning model.** XGBoost (XGB) was the primary model architecture employed in this study. XGB is a gradient boosting algorithm implemented in Python.<sup>18</sup> The XGB algorithm combined results from various decision trees to give prediction scores. Within each decision tree, the patient population was split into successively smaller groups, as each tree branch divided patients who entered it into one of two groups according to the variable value and a predetermined threshold. NASH and fibrosis patient encounters were represented at the end of the decision tree, which was a set of leaf nodes. As the XGB model was trained, successive trees were developed in order to improve the accuracy of the model. Successive iterations



**Figure 1** Study design timeline. At time 0, the patient was diagnosed with NAFL using ICD-10 codes. EHR data were collected 365 days prior to NAFL diagnosis. The prediction window for risk of progression to either NASH or fibrosis was from 14 days post NAFL diagnosis and spanned up to ~4 years (1350 days).



**Figure 2** Patient encounter inclusion diagram. EHR, electronic health record; NAFL, non-alcoholic fatty liver; NASH, non-alcoholic steatohepatitis.

**Table 1** Best features as measured by an XGB model

Demographics	Laboratory measures
Age	Alanine transaminase (ALT)
Sex	Aspartate aminotransferase (AST)
Race-ethnicity	Bilirubin (direct)
Vital signs	Ferritin
Diastolic blood pressure	Platelets
Heart rate	Body mass index (BMI)
Respiratory rate	Prothrombin time or international normalized ratio (INR)
Systolic blood pressure	Gamma-glutamyl transferase (GGT)
Temperature	

These features were then used to train the final XGBoost (XGB), logistic regression (LR), and multi-layered perceptron (MLP) models.

of trees used gradient descent on the prior trees in order to minimize the error of the next tree that is formed.

XGB model performance was compared to that of logistic regression (LR) and multi-layer perceptron (MLP) models. Because LR and MLP models are unable to incorporate missing data, the median observation was used for imputation of that feature. Data were standardized for both the LR and MLP models to a standard Gaussian distribution. The XGB model was trained using the same 139 inputs as the MLP and LR models.

The dataset was partitioned into a train: hold-out test ratio of 80:20 with stratified sampling, because the positive class was relatively small with respect to the negative class. As our dataset comprises multiple distinct facilities from throughout the United States, a second external validation experiment was performed in which the dataset was partitioned into a training set composed of many separate sites and a validation set composed of one distinct facility not seen by the algorithm during training (see Table S3, Supporting

information for demographic information). Inputs for the LR and MLP models were standardized, and missing data were imputed within the training and test datasets independently. All of the models underwent hyperparameter selection with a five-fold cross-validated grid search. The optimization of the hyperparameters was confirmed by evaluating the area under the receiver operating characteristic (AUROC) for different combinations of hyperparameters included in the grid search. For XGB, optimization parameters included maximum tree depth, regularization term (lambda), scale positive weight, learning rate, and number of estimators. Similarly, for LR, optimization parameters included penalty term, optimization problem solver, and inverse of regularization strength (C). For MLP, optimization parameters included maximum iteration, hidden layer size, and learning rate. Performance was assessed against the hold-out test set and external validation cohort with respect to the AUROC curve, sensitivity, and specificity. Confidence intervals for these metrics were constructed using 1000 bootstrapped resamples. A SHAP (SHapley Additive exPlanations)<sup>19</sup> analysis was performed to evaluate how important different features were to the performance of each model.

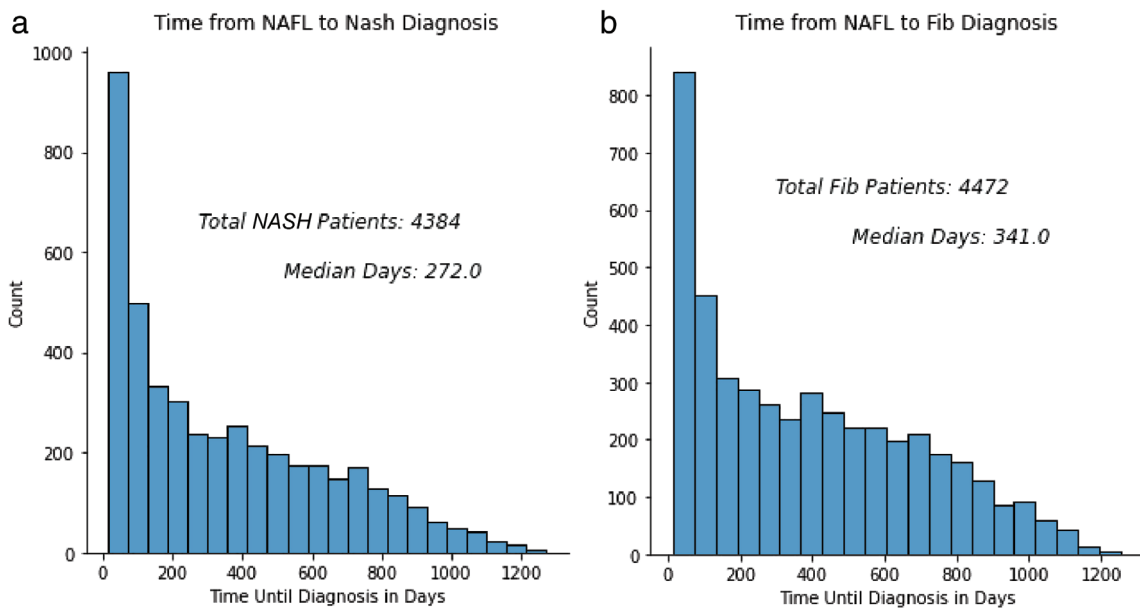
## Results

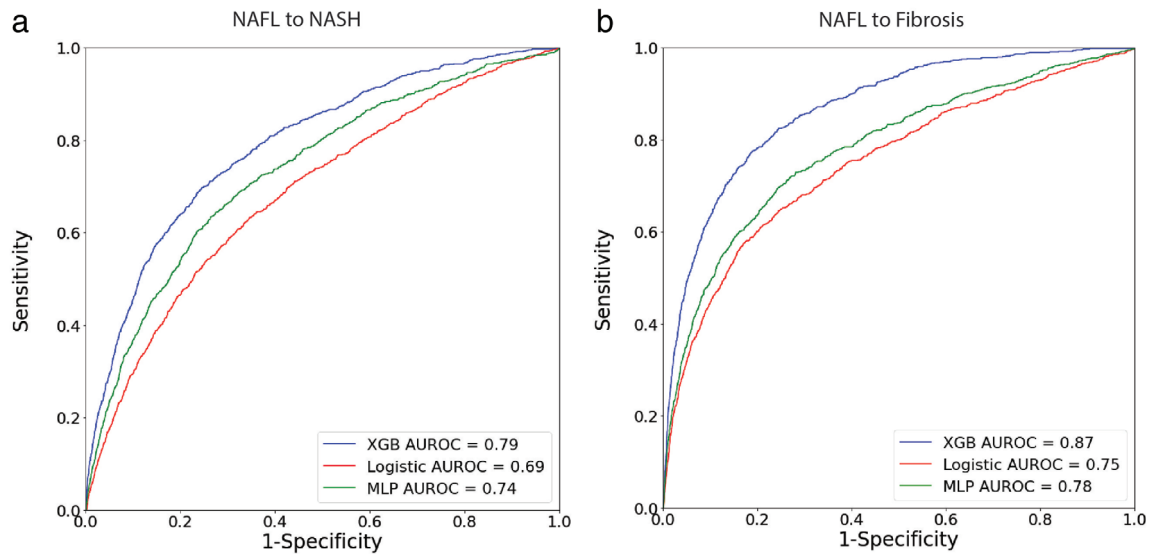
In total, 141,293 patients were included in the experiments, 4384 and 4472 of whom were eventually diagnosed with NASH or fibrosis, respectively (Table 2). *P*-values were calculated using an exact binomial test for noncontinuous observations and Welch's *t*-test for continuous observations (such as age) to handle the unequal variance associated with our positive and negative classes. Median time to NASH or fibrosis diagnosis was 272 and 341 days after NAFL diagnosis, respectively, with a range of 15 to 1250 days (Fig. 3).

For the three models trained using 139 features, constructed as outlined above through the use of summary statistics

**Table 2** Characteristics of the study sample

Characteristic	Non-NASH patients (%) <i>n</i> = 136 909	NASH patients (%) <i>n</i> = 4384	<i>P</i> -value	Nonfibrosis patients (%) <i>n</i> = 136 821	Fibrosis patients (%) <i>n</i> = 4472	<i>P</i> -value
Age						
<30	9690 (7.08)	343 (7.82)	0.0583	9825 (7.18)	207 (4.55)	<0.001
30–49	44 044 (32.17)	1444 (32.94)	0.2842	44 253 (32.33)	1235 (27.24)	<0.001
50–59	38 997 (28.48)	1345 (30.68)	0.0015	38 876 (28.44)	1466 (32.83)	<0.001
60–69	29 935 (21.86)	946 (21.58)	0.6515	29 749 (21.75)	1132 (25.90)	<0.001
70–79	12 271 (8.96)	276 (6.30)	<0.001	12 169 (8.89)	378 (8.38)	0.3071
80+	1971 (1.44)	30 (0.68)	<0.001	1948 (1.42)	53 (1.10)	0.1839
Sex						
Male	60 978 (44.53)	1692 (38.59)	<0.001	60 595 (44.31)	2075 (45.91)	0.0054
Female	75 885 (55.42)	2689 (61.34)	<0.001	76 178 (55.65)	2396 (54.05)	0.0054
Ethnicity						
Hispanic	11 821 (8.63)	459 (10.47)	<0.001	11 842 (8.66)	438 (9.66)	0.0078
Not Hispanic	114 805 (83.85)	3651 (83.30)	0.3088	114 735 (83.84)	3721 (83.59)	0.2444
Unknown	10 283 (7.51)	274 (6.25)	0.0018	10 244 (7.49)	313 (6.74)	0.2444
Race						
Caucasian	112 602 (82.24)	3658 (83.45)	0.0415	112 632 (82.31)	3628 (81.32)	0.0397
African American	9898 (7.22)	212 (4.83)	<0.001	9787 (7.15)	323 (7.17)	0.8590
Asian	2717 (1.98)	98 (2.23)	0.2419	2716 (1.98)	99 (2.19)	0.2815
Other/unknown	11 692 (8.54)	416 (9.48)	0.0271	11 686 (8.55)	422 (9.31)	0.0353
Comorbidities						
Obesity	21 612 (15.78)	848 (19.34)	<0.001	21 781 (15.92)	679 (15.12)	0.1853
Hypertension	61 145 (44.66)	2042 (46.57)	0.120	61 122 (44.68)	2065 (46.32)	0.0466
Dyslipidemia	45 034 (32.89)	1596 (36.40)	<0.001	45 231 (33.05)	1399 (31.30)	0.0130
Obstructive sleep apnea	15 287 (11.16)	645 (14.71)	<0.001	15 430 (11.27)	502 (11.39)	0.9137

**Figure 3** Distribution of patients diagnosed with either NASH or fibrosis over time (the positive class). (a) the number of patients diagnosed with NASH, indicating the progression from NAFL to NASH and (b) the number of patients diagnosed with fibrosis (Fib) over time, indicating the progression from NAFL to fibrosis.

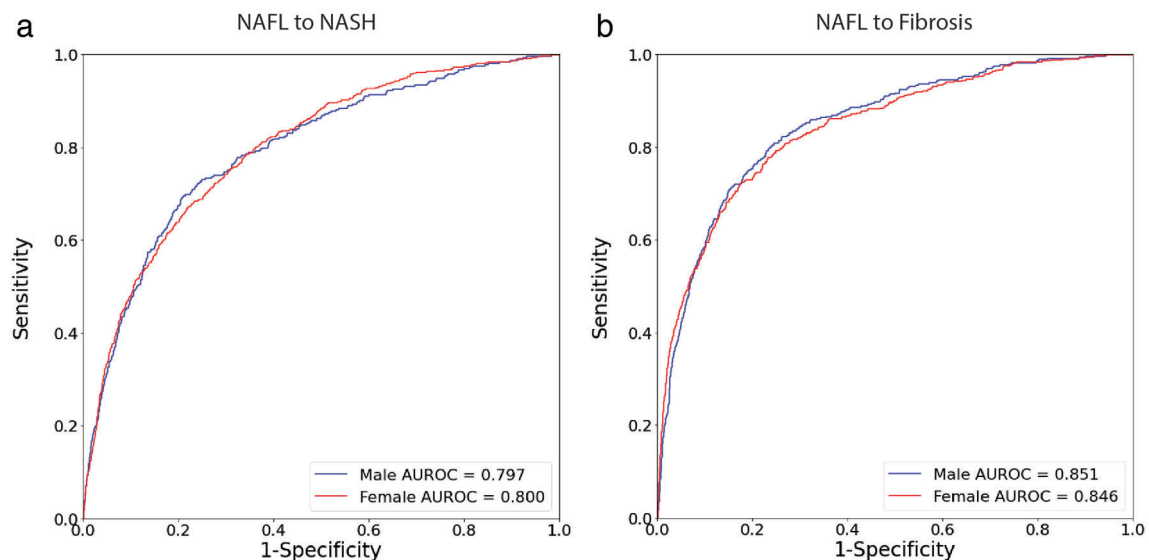


**Figure 4** Receiver operating characteristic curves for the prediction of progression to NASH (a) and fibrosis (b) within 4 years using models trained on 139 features on the hold-out test set. AUROC, area under the receiver operating characteristic; MLP, multi-layer perceptron; XGB, XGBoost.

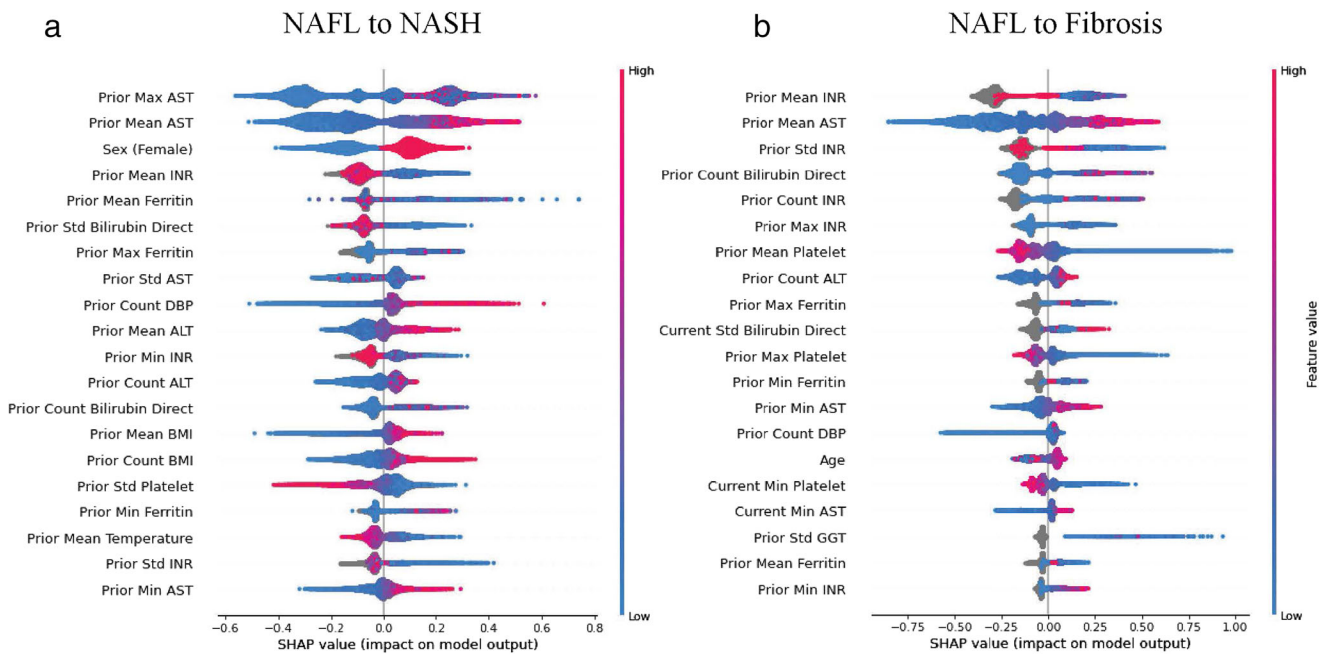
for observations collected during current and previous encounters, the XGB model demonstrated the highest performance with an AUROC of 0.792 (95% CI [0.777–0.808]) for prediction of progression to NASH and of 0.871 (95% CI [0.859–0.882]) for prediction of progression to fibrosis within 4 years on the hold-out test set. The performance of the XGB model on the external validation cohort was similar with AUROC values of 0.795 (95% CI [0.770–0.817]) for progression to NASH and 0.871 (95% CI [0.843–0.897]) for progression to fibrosis (Fig. S1, Supporting information). The LR and MLP models demonstrated AUROC values on the hold-out test set of 0.689 (95% CI [0.669–0.731]) and 0.737 (95% CI [0.720–0.755]), respectively,

for prediction of progression to NASH and of 0.753 (95% CI [0.735–0.771]) and 0.784 (95% CI [0.768–0.799]), respectively, for prediction of progression to fibrosis (Fig. 4). Similarly, the AUROC values for the LR and MLP models on the external validation set for prediction of progression to NASH were 0.680 (95% CI [0.653–0.706]) and 0.740 (95% CI [0.716–0.765]), respectively, and for progression to fibrosis were 0.795 (95% CI [0.762–0.828]) and 0.829 (95% CI [0.801–0.859]), respectively.

Using 139 features for training, the XGB model demonstrated similar performance in both the hold-out test set (results shown) and external validation cohort (Fig. S2, Supporting information) in determining the risk of progression from NAFL to



**Figure 5** Receiver operating characteristic curves for prediction of progression to NASH (a) and fibrosis (b) on male and female populations within 4 years using XGBoost models trained on 139 features on the hold-out test set. AUROC, area under the receiver operator characteristic.



**Figure 6** Feature correlations and distribution of feature importance for each patient for the XGBoost model on the hold-out test set. (a) Feature importance plots for model predicting risk progression from NAFL to NASH and (b) NAFL to fibrosis. Features are ranked in descending order of importance as measured by SHAP values. Red indicates a high feature value; blue indicates a low feature value. Dots to the right resulted in a higher score; dots to the left resulted in a lower score. Min, max, mean, count, and std. represent minimum, maximum, average, number of data points for a certain measurement, and standard deviation, respectively. ALT, alanine transaminase; AST, aspartate aminotransferase; BMI, body mass index; DBP, diastolic blood pressure; GGT, gamma-glutamyl transferase; INR, international normalized ratio; SBP, systolic blood pressure

NASH for males and females with AUROCs of 0.797 and 0.800, respectively, and progression from NAFL to fibrosis of 0.851 and 0.846, respectively (Fig. 5).

Feature importance plots for the XGB model trained to predict the progression of NAFL to NASH show the top three most important features on the hold-out test set were prior max and mean aspartate aminotransferase (AST) and sex. For the XGB model trained to predict the progression of NAFL to fibrosis, the three most important features were prior mean and standard deviation of prothrombin time or international normalized ratio (INR) and prior mean AST (Fig. 6). Similar results for the external validation cohort are presented in Figure S3, Supporting information.

## Discussion

In this study, we retrospectively validated an MLA to predict progression to NASH or fibrosis in newly NAFL-diagnosed patients. The XGB algorithm demonstrated an AUROC of 0.79 and 0.87 for prediction of progression to NASH and fibrosis, respectively, and outperformed LR and MLP models in both the hold-out test set and external validation cohort. The results indicate that the XGB algorithm is capable of accurately identifying newly diagnosed NAFL patients at high risk of progression to NASH or fibrosis.

An exploration of feature importance for the 139 feature XGB model revealed that lab values collected at previous visits were the most important features, with previous mean and maximum values of AST and sex forming the three most important features for prediction of NASH development (Fig. 6). In

addition to previous mean AST values, the previous mean and standard INR values were the top three features for the prediction of the development of fibrosis. AST and alanine aminotransferase (ALT) were important features in both sets. Although both may be elevated in NAFL, ALT elevations are generally greater and the AST:ALT ratio is typically less than 1.<sup>20</sup> The importance of historical AST measurements for predicting NAFL progression to NASH in the XGB model, therefore, suggests that historical trends in AST levels may be an important indicator for likely progression at the time of NAFL diagnosis. This is consistent with the observation that AST increases as hepatic steatosis and fibrosis occurs.<sup>21</sup> The role of prothrombin time–INR as an indirect marker of liver fibrosis is well studied,<sup>22,23</sup> and the predictive importance of INR for progression of fibrosis in the feature sets is well correlated with its diagnostic accuracy. INR has also been identified as significantly associated with histologically proven NASH<sup>24</sup> and thus it is congruent that it is an important feature in both feature sets of the model.

Several additional features were also identified as ranking high in predictive value for risk progression in both sets. The importance of these features is consistent with previous studies and literature. Ferritin expression, which can be elevated due to systemic inflammation as well as increased iron stores, often shows elevated levels in patients with NAFL disease.<sup>25</sup> Ferritin stores iron, which is a putative element that interacts with oxygen radicals.<sup>26</sup> Increased iron stores are associated with increased liver damage in patients with NASH.<sup>26</sup> It is also associated with an increased risk of fibrosis.<sup>27</sup> Serum ferritin has been identified as a predictor in

NASH,<sup>28,29</sup> although some studies have disputed the veracity of its predictive value.<sup>30</sup> Other studies have found value in the use of ferritin as part of a diagnostic panel in evaluating individual patients, but not as an independent predictor.<sup>31</sup> Our analysis showed previous mean and maximum levels of ferritin were highly valuable for predicting risk of progression in both feature sets. Platelet count is another important feature identified in both risk prediction models. Peripheral platelet production is regulated via a hormone primarily produced in the liver; decreased stimulation of platelet production and increased platelet destruction due to liver-disease associated hypersplenism can result in decreased platelet counts, or thrombocytopenia.<sup>32,33</sup> Low platelet count has been documented in subsets of patients with NAFL<sup>34</sup> and was found to be an important predictor in determining risk of progression to NASH and fibrosis.<sup>35</sup>

There is a need to separately study risk of progression to NASH and fibrosis for men and women diagnosed with NAFL.<sup>36</sup> Sex differences in the prevalence, risk factors, fibrosis, and clinical outcomes of NAFLD<sup>36–39</sup> should be included when analyzing disease progression. The National Institutes of Health recommend that experiments address sex differences in preclinical studies,<sup>40</sup> as well as studies involving algorithm development and interpretation.<sup>41</sup> The XGB model demonstrated an AUROC of 0.797 and 0.800 for the risk of progression from NAFL to NASH and 0.851 and 0.846 for the risk of progression to fibrosis for males and females, respectively, on the hold-out test set and similar results on the external validation cohort. This indicates that the algorithm predicts the risk development of NASH and fibrosis equally well in both males and females.

Despite the significant disease burden posed by NAFL, effective targeted treatments for NAFL, NASH, and fibrosis are limited.<sup>10,17</sup> Newly diagnosed NAFL patients are encouraged to make a number of lifestyle modifications, including weight loss, dietary modifications, decreased alcohol consumption, and increased exercise.<sup>42</sup> In addition, pharmacological therapies, including pioglitazone and omega-3 polyunsaturated fatty acids, have been investigated for off-label use in treating NASH in certain patient populations, though their use is generally recommended only for those with advanced stages or at high risk of disease progression.<sup>43</sup> While such treatments are not recommended for the general NAFL population, their use could be appropriate for those at high risk of progression to NASH, as could more aggressive weight-loss interventions such as bariatric surgery.<sup>44</sup> Although pharmacologic efforts to treat fibrosis due to NASH have recently been set back with the rejection of obeticholic acid by the Food and Drug Administration (FDA),<sup>45</sup> there are currently more than 20 agents in investigational trials for treatment of various etiologies of fibrosis.<sup>46</sup>

Despite the potential clinical benefits of accurately identifying NAFL patients most at risk of progression to advanced NASH, few methods exist to readily identify these individuals. The majority of research into machine learning (ML) methods for NASH has explored ways to assist in diagnosis via detecting existing NASH using various features, for example, biopsy specimens and biochemical parameters in peripheral blood.<sup>47–49</sup> One study also demonstrated the superior performance of the XGB algorithm for NAFL and NASH detection, where it outperformed other ML approaches, including LR, decision tree, and random forest.<sup>49</sup> While these studies provide context and demonstrate comparably superior performance of the XGB model as seen in

this study, it is notable that the prediction tasks, input features, and population cohorts differed from our study. To the authors' knowledge, only one study to date has explicitly used ML to assist in prognosis, specifically to identify predictors of NASH among a hypertensive patient population.<sup>50</sup> However, this study analyzed features in a specific subpopulation and so algorithm performance is not directly comparable. This research adds to the literature by demonstrating that an MLA can be trained to accurately predict progression to NASH or fibrosis in a general NAFL patient population with no other filtering of the patient population by the presence of other comorbidities.

In addition to improving prognostic predictions, our methods may also help identify and enrich clinical trial populations. The algorithm developed in this study could improve study power by enrolling NAFL patients most likely to experience rapid disease progression, increasing study efficiency. Slow disease progression has been noted as a major barrier to clinical trials of NAFL, given the long follow-up required for most patients to see endpoints of disease progression and mortality.<sup>51</sup> As such, researchers have recommended focusing enrollment on patients most likely to benefit from the trial protocol.<sup>55</sup> Researchers have also noted difficulties in enrolling NASH patients in clinical trials due to low detection and diagnosis rates, a lack of formal diagnosis, and/or discrepancies between initial pathological readings of liver biopsies and pathology readings performed by study clinicians.<sup>16,52,53</sup> These difficulties have prompted suggestions that other methods of screening patients may improve trial enrollment by reducing screening failures. The algorithm presented in this study is capable of defining a population at high risk of progression to NASH, potentially enabling prevention as well as treatment trials.

There are several limitations to this study. First, the timing of NAFL, NASH, and fibrosis onset was determined by the first application of ICD codes. NAFL may be clinically silent for a prolonged period prior to diagnosis, therefore the observed onset may differ from the pathophysiological onset of disease. Second, although the gold standard for diagnosis of NASH and fibrosis is liver biopsy, in the absence of these data we instead used clinician diagnosis via ICD codes as a proxy. Third, the retrospective nature of this study means that we cannot make inferences about the impact this algorithm may have on clinical care or patient outcomes. Future prospective research will be critical in elucidating the role that prognostic MLAs may play in treatment planning and clinical trial design for NASH.

## Conclusion

In this study, we have developed an MLA capable of accurately identifying newly diagnosed NAFL patients at high risk of progression to NASH or fibrosis. Identification of patients at high risk of progression to NASH or fibrosis may serve to improve patient outcomes through increased patient monitoring and aggressive preventative measures.

## Acknowledgments

We would like to thank Logan Ryan and Samson Mataraso for their assistance with study design; Jana Hoffman, Anna Siefkas, and Megan Handley for their contributions to the writing of this manuscript, and Nicole Zellin for assistance with study design and manuscript review.

## References

- 1 de Alwis NMW, Day CP. Non-alcoholic fatty liver disease: the mist gradually clears. *J. Hepatol.* 2008; **48**: S104–12.
- 2 Singh S, Allen AM, Wang Z, Prokop LJ, Murad MH, Loomba R. Fibrosis progression in nonalcoholic fatty liver versus nonalcoholic steatohepatitis: a systematic review and meta-analysis of paired-biopsy studies. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* 2015; **13**: 643–654.e9.
- 3 De A, Duseja A. Natural history of simple steatosis or nonalcoholic fatty liver. *J. Clin. Exp. Hepatol.* 2020; **10**: 255–62.
- 4 Younossi ZM. Patient-reported outcomes and the economic effects of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis: the value proposition. *Hepatology.* 2018; **68**: 2405–12.
- 5 Sayiner M, Otgonsuren M, Cable R *et al.* Variables associated with inpatient and outpatient resource utilization among Medicare beneficiaries with nonalcoholic fatty liver disease with or without cirrhosis. *J. Clin. Gastroenterol.* 2016; **51**: 254–60.
- 6 Younossi ZM, Blissett D, Blissett R *et al.* The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. *Hepatology.* 2016; **64**: 1577–86.
- 7 Kim D, Kim WR, Kim HJ, Therneau TM. Association between non-invasive fibrosis markers and mortality among adults with non-alcoholic fatty liver disease in the United States. *Hepatology (Baltimore Md.)*. 2013; **57**: 1357–65.
- 8 Wong RJ, Aguilar M, Cheung R *et al.* Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. *Gastroenterology.* 2015; **148**: 547–55.
- 9 Sanyal AJ. Past, present and future perspectives in nonalcoholic fatty liver disease. *Nat. Rev. Gastroenterol. Hepatol.* 2019; **16**: 377–86.
- 10 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.* 2016; **64**: 73–84.
- 11 Chalasani N, Younossi Z, Lavine JE *et al.* The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. *Hepatology (Baltimore Md.)*. 2012; **55**: 2005–23.
- 12 Castera L, Friedrich-Rust M, Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. *Gastroenterology.* 2019; **156**: 1264–1281.e4.
- 13 Nallagangula KS, Nagaraj SK, Venkataswamy L, Chandrappa M. Liver fibrosis: a compilation on the biomarkers status and their significance during disease progression. *Future Sci. OA.* 2017; **4**: FSO250.
- 14 Polis S, Fernandez R. Impact of physical and psychological factors on health-related quality of life in adult patients with liver cirrhosis: a systematic review protocol. *JBI Database Syst. Rev. Implement. Rep.* 2015; **13**: 39–51.
- 15 Li C, Li R, Zhang W. Progress in non-invasive detection of liver fibrosis. *Cancer Biol. Med.* 2018; **15**: 124–36.
- 16 Schulz M, Tacke F. Identifying high-risk NASH patients: what we know so far. *Hepat. Med.: Evid. Res.* 2020; **12**: 125–38.
- 17 Schuppan D, Kim YO. Evolving therapies for liver fibrosis. *J. Clin. Invest.* 2013; **123**: 1887–901.
- 18 Chen T, Guestrin C. *XGBoost: a scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13 Aug 2016; 785–94.
- 19 Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. ArXiv170507874 Cs Stat [Internet]. 24 Nov 2017. Cited 3 May 2021. Available from URL: <http://arxiv.org/abs/1705.07874>
- 20 Hall P, Cash J. What is the real function of the liver ‘function’ tests? *Ulster Med. J.* 2012; **81**: 30–6.
- 21 Sattar N, Forrest E, Preiss D. Non-alcoholic fatty liver disease. *BMJ.* 2014; **349**: g4596.
- 22 Croquet V, Vuillemin E, Ternisien C *et al.* Prothrombin index is an indirect marker of severe liver fibrosis. *Eur. J. Gastroenterol. Hepatol.* 2002; **14**: 1133–41.
- 23 Komp DM, Selden RF. Coagulation abnormalities in cystic fibrosis. *Chest.* 1970; **58**: 501–3.
- 24 Fierbinteanu-Braticevici C, Baicus C, Tribus L, Papacocea R. Predictive factors for nonalcoholic steatohepatitis (NASH) in patients with nonalcoholic fatty liver disease (NAFLD). *J. Gastrointest. Liver Dis.* 2011; **20**: 153–9.
- 25 Dever JB, Mallory MA, Mallory JE, Wallace D, Kowdley KV. Phenotypic characteristics and diagnoses of patients referred to an iron overload clinic. *Dig. Dis. Sci.* 2010; **55**: 803–7.
- 26 George DK, Goldwurm S, MacDonald GA *et al.* Increased hepatic iron concentration in nonalcoholic steatohepatitis is associated with increased fibrosis. *Gastroenterology.* 1998; **114**: 311–8.
- 27 Kowdley KV, Belt P, Wilson LA *et al.* Elevated serum ferritin is an independent predictor of histologic severity and advanced fibrosis among patients with nonalcoholic fatty liver disease. *Hepatology (Baltimore Md.)*. 2012; **55**: 77–85.
- 28 Yoneda M, Nozaki Y, Endo H *et al.* Serum ferritin is a clinical biomarker in Japanese patients with nonalcoholic steatohepatitis (NASH) independent of HFE gene mutation. *Dig. Dis. Sci.* 2010; **55**: 808–14.
- 29 Kim YS, Jung ES, Hur W *et al.* Noninvasive predictors of non-alcoholic steatohepatitis in Korean patients with histologically proven nonalcoholic fatty liver disease. *Clin. Mol. Hepatol.* 2013; **19**: 120–30.
- 30 Chandok N, Minuk G, Wengiel M, Uhanova J. Serum ferritin levels do not predict the stage of underlying non-alcoholic fatty liver disease. *J. Gastrointest. Liver Dis.* 2012; **21**: 53–8.
- 31 Buzzetti E, Petta S, Manuguerra R *et al.* Evaluating the association of serum ferritin and hepatic iron with disease severity in non-alcoholic fatty liver disease. *Liver Int. Off. J. Int. Assoc. Study Liver.* 2019; **39**: 1325–34.
- 32 Aster RH. Pooling of platelets in the spleen: role in the pathogenesis of “hypersplenic” thrombocytopenia. *J. Clin. Invest.* 1966; **45**: 645–57.
- 33 Mitchell O, Feldman DM, Diakow M, Sigal SH. The pathophysiology of thrombocytopenia in chronic liver disease. *Hepat. Med. Evid. Res.* 2016; **8**: 39–50.
- 34 López-Trujillo MA, Olivares-Gazca JM, Cantero-Fortiz Y *et al.* Non-alcoholic fatty liver disease and thrombocytopenia iii: its association with insulin resistance. *Clin. Appl. Thromb. Hemost.* 2019; **25**: 1076029619888694.
- 35 Angulo P. Trial design in NASH—realities and challenges. *Nat. Rev. Gastroenterol. Hepatol.* 2011; **8**: 424–5.
- 36 Lonardo A, Nascimbeni F, Ballestri S *et al.* Sex differences in NAFLD: state of the art and identification of research gaps. *Hepatology (Baltimore Md.)*. 2019; **70**: 1457–69.
- 37 Balakrishnan M, Patel P, Dunn-Valadez S *et al.* Women have a lower risk of nonalcoholic fatty liver disease but a higher risk of progression vs men: a systematic review and meta-analysis. *Clin. Gastroenterol. Hepatol.* 2021; **19**: 61–71.e15.
- 38 Lonardo A, Suzuki A. Sexual dimorphism of NAFLD in adults. focus on clinical aspects and implications for practice and translational research. *J. Clin. Med.* 2020; **9**: 1278.
- 39 Dufour J-F, Scherer R, Balp M-M *et al.* The global epidemiology of nonalcoholic steatohepatitis (NASH) and associated risk factors—A targeted literature review. *Endocr. Metab. Sci.* 2021; **3**: 100089.
- 40 Clayton JA, Collins FS. NIH to balance sex in cell and animal studies. *Nature.* 2014; **509**: 282–3.
- 41 Cirillo D, Catuara-Solarz S, Morey C *et al.* Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit. Med.* 2020; **3**: 81.



- 42 Rinella ME, Sanyal AJ. Management of NAFLD: a stage-based approach. *Nat. Rev. Gastroenterol. Hepatol.* 2016; **13**: 196–205.
- 43 Leoni S, Tovoli F, Napoli L, Serio I, Ferri S, Bolondi L. Current guidelines for the management of non-alcoholic fatty liver disease: A systematic review with comparative analysis. *World J. Gastroenterol.* 2018; **24**: 3361–73.
- 44 Lassailly G, Caiazzo R, Buob D *et al.* Bariatric surgery reduces features of nonalcoholic steatohepatitis in morbidly obese patients. *Gastroenterology.* 2015; **149**: 379–88.
- 45 Mullard A. FDA rejects NASH drug. *Nat. Rev. Drug Discov.* 2020; **19**: 501–1.
- 46 Damiris K, Tafesh ZH, Pysopoulos N. Efficacy and safety of anti-hepatic fibrosis drugs. *World J. Gastroenterol.* 2020; **26**: 6304–21.
- 47 Forlano R, Mullish BH, Giannakeas N. *et al.* High-Throughput, Machine Learning-Based Quantification of Steatosis, Inflammation, Ballooning, and Fibrosis in Biopsies from Patients with Nonalcoholic Fatty Liver Disease [published online ahead of print]. Vol. 2019; S1542-3565(19)31505-8.
- 48 Perakakis N, Polyzos SA, Yazdani A *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.
- 49 Fialoke S, Malarstig A, Miller MR, Dumitriu A. Application of machine learning methods to predict non-alcoholic steatohepatitis (NASH) in non-alcoholic fatty liver (NAFL) patients. *AMIA Annu. Symp. Proc.* 2018; **2018**: 430–9.
- 50 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**: 444–51.
- 51 Ofosu A, Ramai D, Reddy M. Non-alcoholic fatty liver disease: controlling an emerging epidemic, challenges, and future directions. *Ann. Gastroenterol.* 2018; **31**: 288–95.
- 52 Ratziu V, Friedman SL. *Why Do So Many NASH Trials Fail?* Gastroenterology [Internet]. 18 May 2020. Cited 17 Jun 2021. Available from URL: <https://www.sciencedirect.com/science/article/pii/S0016508520306806>
- 53 Rinella ME, Noureddin M. STELLAR 3 and STELLAR 4: lessons from the fall of Icarus. *J. Hepatol.* 2020; **73**: 9–11.

## Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website:

**Appendix S1.** Supporting Information.