Non-Invasive Prescreening for Fatty Liver Disease

1. Summary

- With the first treatments for nonalcoholic steatohepatitis (NASH) coming to market, developing a cheap, widely distributable, non-invasive method for diagnosing or even prescreening for fibrosis staging can reduce costs and expand patient care.
- Non-invasive blood tests, including ALBI score, APRI, and FIB-4, have been explored for diagnosing fibrosis staging with moderate success.
- Machine learning models can improve upon existing methods by considering many correlative variables as opposed to a few blood tests.
- Using patient data coming from the Mayo Clinic, a logistic regression model was shown to outperform ALBI, APRI, FIB-4, and a baseline random model for diagnosing fibrosis staging.
- Machine learning models can reduce costs and expand patient care, but additional data is required to further evaluate the methods.

2. Introduction

A. Problem Statement

Nonalcoholic steatohepatitis (NASH) affects approximately 5% of US adults and 8% globally. Prescreening for NASH can reduce costs and expand patient care. The goal is to test machine learning methods for prescreening NASH.

B. Background

NASH, also commonly referred to as fatty liver disease, affects approximately 5% of US adults and 8% of the global population. NASH commonly leads to development of fibrosis through the continual stress and inflammation placed on the liver. NASH-associated fibrosis is categorized into five stages: F0, F1, F2, F3, and F4 with F4 referring to cirrhosis. Those with late stages of fibrosis, both F3 and F4, are at a greater risk of liver-related complications and liver-related death. Liver damage up to F3 stage of NASH is reversible, but F4, is irreversible and those with cirrhosis are at greater risk for developing liver cancer.

Recently, the first ever therapy was approved for treating NASH was granted for Madrigal Pharmaceutical's Rezdiffra, but only for patients with noncirrhotic NASH with moderate to advanced fibrosis (consistent with F2 and F3 staging). Following Madrigal's success, similar therapies, such as those being developed by Viking Therapeutics, Pfizer, and Novo Nordisk, are likely to come to market. To be eligible for such therapies, NASH staging must be confirmed by either liver biopsy, an

invasive procedure and costly procedure, or by MRI-proton density fat fraction (MRI-PDFF), a non-invasive but specialized and potentially costly procedure.

Cheap, widely distributable, non-invasive methods for diagnosing, or even prescreening for, NASH staging can reduce costs and expand patient care. Several non-invasive measures using only blood test results, including the albumin-bilirubin (ALBI) score, aspartate aminotransferase-platelet ratio index (APRI), and fibrosis-4 (FIB-4) index, have been explored for diagnosing NASH staging with moderate success. By considering additional correlative measures, machine learning methods can improve upon ALBI score, APRI, and FIB-4, and potentially reduce cost and expand patient care for NASH patients.

C. Goal

The purpose of this project is to develop a cheap, widely distributable, non-invasive prescreening method for fibrosis staging by using machine learning to improve upon existing non-invasive blood tests, such as ALBI score, APRI, and FIB-4. The data used for this project comes from the Mayo Clinic and contains real records of 306 patients with varying stages of fibrosis. Based on the results, even simple machine learning methods, such as logistic regression, can provide more accurate predictions for fibrosis staging compared to ALBI score, APRI, and FIB-4. However, additional data is required to thoroughly assess the utility of using logistic regression models to prescreen for fibrosis staging.

3. Dataset

The data used to train and assess the machine learning models comes from the Mayo Clinic and can be found on Kaggle. After cleaning, the dataset contains 306 records (418 records prior to cleaning) of patient data with 18 clinical measures, including fibrosis stage, age, platelet count, cholesterol level, bilirubin concentration, albumin concentration, triglyceride concentration, and other correlative measures that are linked to liver damage. Because the goal is to prescreen for fibrosis stage, fibrosis stage will be the target variable with the remaining columns being our feature variables. As fibrosis stage is a categorical variable, the machine learning models employed will be ones suited for classification problems.

4. Data Preparation

A. Some Features Contain Outliers or are Uninformative

Preparing the raw data for analysis and learning primarily involved two steps. The first step was removing outliers and removing uninformative features. Figure 1 shows that bilirubin concentration, cholesterol level, copper levels, alkaline phosphatase concentration, alanine aminotransferase concentration, triglyceride levels, and platelet counts are all skewed with tails toward high values. Furthermore, each variable contains outliers at the high end. Because outliers are likely to skew results, the outliers were removed.

Additionally, other columns including Status, Drug, and N_Days, which are not shown for expediency, were completely removed from the data as they do not contain meaningful information for diagnosing fibrosis stage. Finally, records that did not contain values for fibrosis stage were removed.

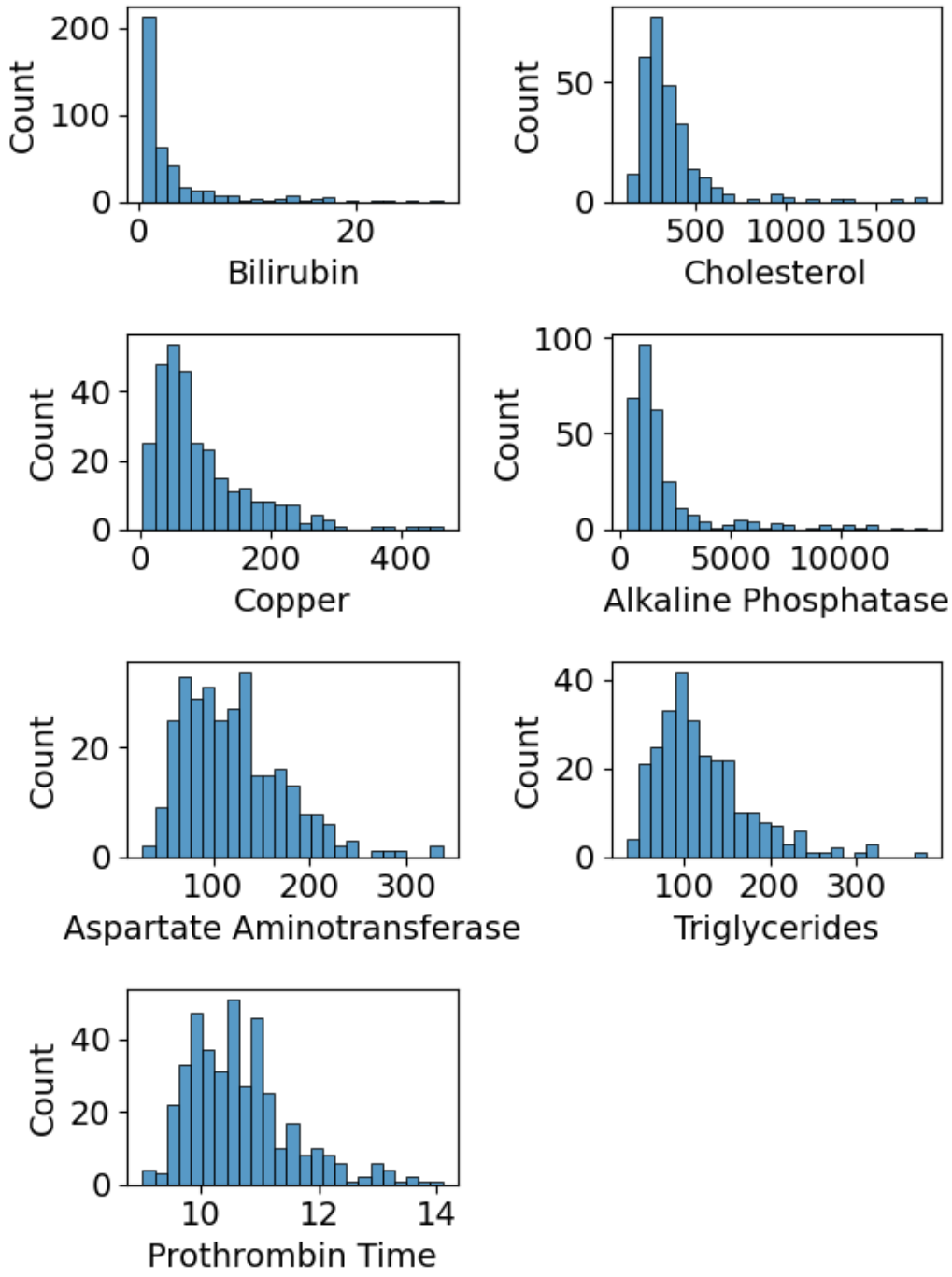

Figure 1: Clinical measures bilirubin, cholesterol, copper, alkaline phosphatase, aspartate aminotransferase, triglycerides, and prothrombin time all contain outliers with bilirubin concentration being the most widely spread. In all cases, outliers were removed.

B. Records From Separate Study Are Removed

According to the data source, the first 312 records come from a single clinical trial with an additional 106 records added later from patients that did not participate in the same clinical trial but were included for a separate study. These same 106 records also contain 8 or 9 missing values while the first 312 records contain no more than 3 missing values. Therefore, the additional 106 records were removed from the dataset. After removing outliers, records with 8 or more missing values, and records with fibrosis stage not recorded, there were 306 patient records.

5. Exploratory Data Analysis

A. Cannot Reliably Classify F2 and Below

Because fibrosis stage is the target variable, it is imperative to explore any patterns within this column. The first step is to explore how fibrosis stage is distributed, which is shown in Figure 2. Most notably, there are only 15 records for F1, or mild NASH-associated fibrosis. With only 15 records, it is very unlikely that any model could accurately classify F1 individuals. Furthermore, as is shown in the next section, the feature variables are not well separated by staging, further suggesting that the dataset is insufficient for classifying F1 individuals.

With few cases of mild fibrosis (F1) and no cases of healthy individuals, F2 individuals are also unlikely to be reliably classified as any model trained using this data is unlikely to learn a well-defined lower bound for moderate fibrosis: F2 individuals could be classified as not having severe fibrosis (F3) or cirrhosis (F4) instead of having moderate fibrosis.

Therefore, only F3 and F4 individuals may be reliably classified. Fortunately, because F3 individuals are the most in need for treatment, foregoing accurate diagnosis of F1 and F2 does not significantly undermine the results. Furthermore, by focusing on screening for F3, F1 and F2 can be combined to improve statistics. However, with only 120 and 105 cases for F3 and F4, respectively, any model attempting to classify F3 and F4 will need to be tested against a larger dataset.

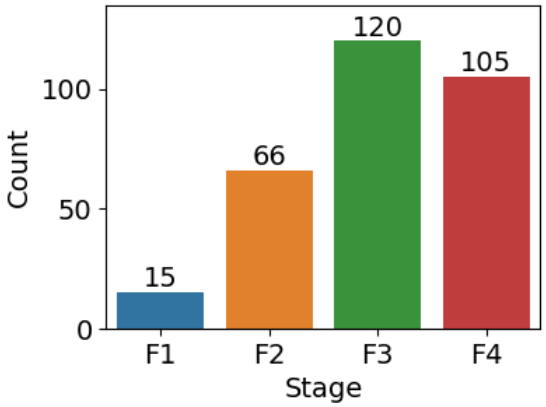

Figure 2: Fibrosis staging is asymmetric with F1 being in the minority and F0 and healthy individuals not included. F3 and F4 are roughly equally represented, although F3 contains a few more records. Since healthy and F0 individuals are not included and F1 individuals are not well represented, it is unlikely that any model could reliably classify F2 or below.

B. Most Features Not Separable by Staging

After grouping F1 and F2 into F1/F2, clinical measures and fibrosis scores were grouped according to fibrosis staging to better understand intrinsic patterns in the data with their respective distributions shown in Figure 3. Although not all clinical measures are shown, the ones shown are, the distributions show that individual features are not separable by fibrosis staging. Furthermore, even the fibrosis scores, which aggregate quantities including Albumin, Bilirubin, Aspartate Aminotransferase, Alanine Aminotransferase, and Platelets, are not separable by fibrosis stage. And of the three fibrosis scores, ALBI, APRI, and FIB-4, ALBI is best separable by fibrosis stage while APRI and FIB-4 are nearly identical. Therefore, any diagnostic test that uses a single signal is likely to be inaccurate. That said, a combination of signals may be more accurate if the data are better separated in a higher dimensional space.

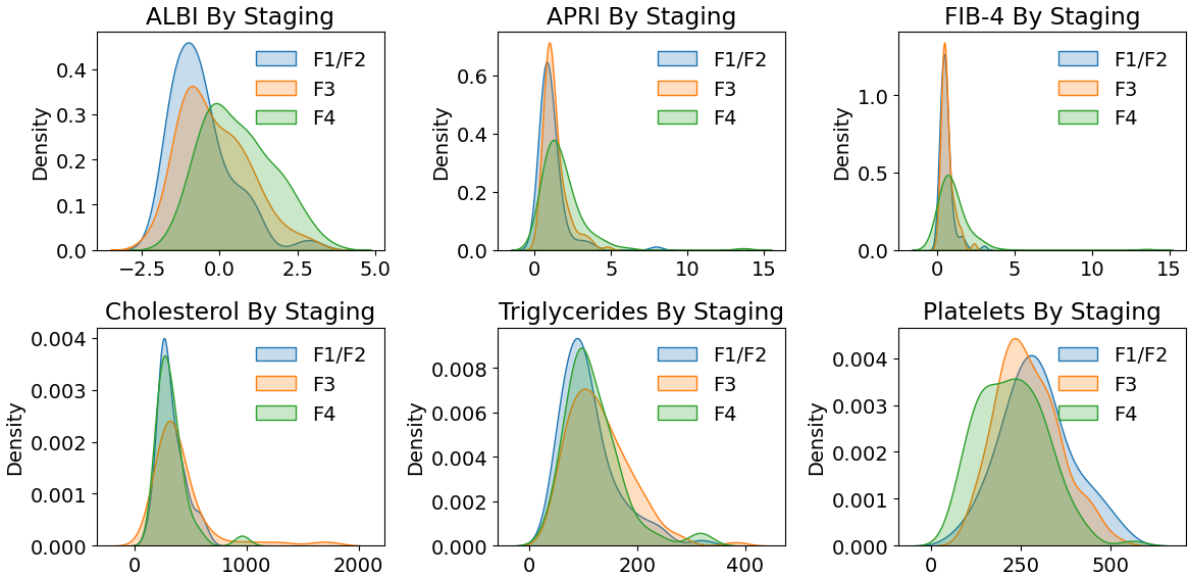

Figure 3: Clinical measures and fibrosis scores, ALBI, APRI, and FIB-4, are not separable by staging. Diagnosing fibrosis staging using a single fibrosis score or clinical measure will be inaccurate. However, a combination of measures may be more accurate if staging is more separable in a higher dimensional space. For brevity, not all clinical measures are shown, however, the set that are shown are more than representative.

6. Modeling Liver Disease Staging

A. Benchmarks

To contextualize machine learning models, benchmarks were examined. At the very minimum, any meaningful model should out-perform a randomized model in which patients are diagnosed randomly according to population statistics. Furthermore, any machine learning model that takes advantage of high dimensional data should out-perform models that use fewer dimensions. For such comparisons, ALBI, APRI, and FIB-4 scores were considered.

Given the nature of the problem, the primary performance metric is F1-score with false negative rate as the secondary metric. The following tables report on each model's performance via a confusion matrix. Precision, recall, and F1-score are also presented.

a. Random Model

At the very least, any diagnostic model needs to be more predictive than a random guess. With 120 patients with F3 stage of liver fibrosis out of 306 total patients, a random model would guess F3 stage fibrosis at a rate of 39.2% and with 39.2% of the population having F3 stage fibrosis, the precision, recall, and f1-score are 39.2%. Also, such a model would predict F1/F2 at a rate of 26.5% and with 73.5% of the population being either F3 or F4, the random model predicts severe fibrosis or cirrhosis with a false negative rate of 78.6%.

Random		Predicted			
		F1/F2	F3	F4	Total
True	F1/F2	21.4	31.8	27.8	81
	F3	31.8	47.0	41.2	120
	F4	27.8	41.2	36.0	105
	Total	81	120	105	306
	Precision:	0.26	0.39	0.34	
	Recall:	0.26	0.39	0.34	
	F1-Score:	0.26	0.39	0.34	

ALBI		Predicted			
		F1/F2	F3	F4	Total
True	F1/F2	20.0	38.0	23.0	81
	F3	9.0	52.0	59.0	120
	F4	1.0	19.0	85.0	105
	Total	30	109	167	306
	Precision:	0.67	0.48	0.51	
	Recall:	0.25	0.43	0.81	
	F1-Score:	0.36	0.45	0.63	

APRI		Predicted			
		F1/F2	F3	F4	Total
True	F1/F2	1.0	60.0	20.0	81
	F3	0.0	71.0	49.0	120
	F4	0.0	38.0	67.0	105
	Total	1	169	136	306
	Precision:	1.00	0.42	0.49	
	Recall:	0.01	0.59	0.64	
	F1-Score:	0.02	0.49	0.56	

FIB-4		Predicted			
		F1/F2	F3	F4	Total
True	F1/F2	0.0	59.0	22.0	81
	F3	0.0	79.0	41.0	120
	F4	0.0	43.0	62.0	105
	Total	0	181	125	306
	Precision:	0.00	0.44	0.50	
	Recall:	0.00	0.66	0.59	
	F1-Score:	0.00	0.52	0.54	

Table 1: Confusion matrix, precision, recall, and f1-score for random guess, ALBI, APRI, and FIB-4 as a prescreening method in diagnosing fibrosis stage.

b. ALBI

ALBI score has been used to diagnose fibrosis by choosing a threshold value which best separates the patient population according to fibrosis stage. Depending on the training data, the threshold can vary from -2.125 to -2.95 for classifying advanced fibrosis. With the data used in this report, the best cutoff value was -1.18 for classifying advanced fibrosis and -0.3 for cirrhosis. Using these values, a diagnostic model trained on the data presented using only ALBI score would result in a precision of 0.48, a recall of 0.43, a f1-score of 0.45, and a false negative rate of 0.33. Compared to the random model, the ALBI model scores better according to all metrics. However, the ALBI model

does predict F1/F2 at a significantly lower rate than the random model (9.8% vs 26.5%), which is not surprising as F1/F2 patients significantly overlap with F3 and F4 patients according to ALBI score.

c. APRI and FIB-4

Both APRI and FIB-4 are similarly distributed according to fibrosis staging as can be seen in Figure 3, which is not surprising as FIB-4 is equivalent to APRI excluding a factor for age and aspartate aminotransferase count. Not only are they similarly distributed, but for both measures, F1/F2 is inseparable from F3. With both being qualitatively similar, discussion for APRI and FIB-4 are combined.

Like ALBI score, both APRI and FIB-4 have also been used to diagnose fibrosis by choosing a threshold value which best separates the patient population according to fibrosis stage. Again, depending on the training data, the threshold value can vary with some reports using an APRI cutoff value of 0.7 for advanced fibrosis and a FIB-4 cutoff value of 3.25 for advanced fibrosis. With the data used in this report, the best APRI cutoff value was -1.06 for classifying advanced fibrosis and -0.22 for classifying cirrhosis. Similarly, the best FIB-4 cutoff values for classifying advanced fibrosis and cirrhosis were -2.00 and -0.16, respectively. Using these values, precision, recall, f1-score, and false negative rate for classifying advanced fibrosis using APRI score are 0.42, 0.59, 0.49, and 0, respectively, shown in Table 1. Similarly, precision, recall, f1-score, and false negative rate for classifying advanced fibrosis using FIB-4 score was 0.44, 0.66, 0.52, and NaN, respectively, also shown in Table 1.

Compared to the random model and ALBI model, precision is lower but recall and f1-score are both improved. While nominally better, both APRI and FIB-4 significantly under predict F1/F2 with APRI predicting F1/F2 only once and FIB-4 not predicting F1/F2 at all. Therefore, both are not informative despite the improved recall and f1-score relative to the ALBI model. Going forward, the ALBI model will serve as the benchmark to beat for machine learning models.

B. Random Forest

An initial random forest model was trained using all available features to determine which features are informative and which are not. After calculating feature importances using this initial random forest model, a second model was trained. The specifics of training are included in the accompanying Jupyter notebook labeled “03_av_preprocessing_and_training.ipynb.” The results of this refined model are shown in Table 2, as well as the results of a logistic regression model. It is worth mentioning that other models were trained and tested, and their results can be found in the notebook. However, their results are excluded for brevity as they all, excluding a feedforward neural network, underperformed compared to the random forest model and logistic regression model. The feedforward neural network was excluded because such a model would be more difficult to implement for minimal gain in performance.

Using a random forest model, the precision, recall, f1-score, and false negative rate for predicting advanced fibrosis are 0.44, 0.64, 0.52, and 0.55, respectively. Compared to the random model, the random forest model scores higher according to all performance metrics. Compared to the ALBI model, recall and f1-score are improved but precision and false negative rate are worse. Looking at

the rates in which F1/F2, F3, and F4 are predicted, the random forest model predicts F1/F2 more often (14.6% for random forest vs. 9.8% for ALBI) which contributes to the increased false negative rate of the random forest model: although the random forest model predicts F1/F2 at a higher rate than the ALBI model, both models have the same recall for F1/F2. Additionally, the random forest model predicts F3 at a higher rate than the ALBI model (57.5% vs. 35.6%) at the expense of predicting F4 (28.0% vs. 54.6%). With the increased rate in predicting F3, F3 recall is much higher for the random forest model compared to the ALBI model with only a small decrease in F3 precision. Overall, the random forest model better classifies F3 compared to the ALBI model.

Random Forest		Predicted			
		F1/F2	F3	F4	Total
True	F1/F2	20.1	52.7	8.2	81
	F3	20.5	77.3	22.2	120
	F4	4.0	45.8	55.2	105
	Total	44.6	175.8	85.6	306
	Precision:	0.45	0.44	0.64	
	Recall:	0.25	0.64	0.53	
	F1-Score:	0.32	0.52	0.58	

Log. Regression		Predicted			
		F1/F2	F3	F4	Total
True	F1/F2	10.8	65.3	4.9	81
	F3	6.8	96.5	16.7	120
	F4	1.4	57.5	46.1	105
	Total	19	219.3	67.7	306
	Precision:	0.57	0.44	0.68	
	Recall:	0.13	0.80	0.44	
	F1-Score:	0.22	0.57	0.53	

Table 2: Confusion matrix, precision, recall, and f1-score for using a random forest model and logistic regression model as a prescreening method in diagnosing fibrosis stage.

C. Logistic Regression

Like the random forest model, an initial logistic regression model was trained to identify important features using permutation importance. Again, specifics of training are included in "03_av_preprocessing_and_training.ipynb." And like the random forest model, results for the logistic regression model are shown in Table 2.

With the logistic regression model, precision, recall, f1-score, and false negative rate for classifying F3 are 0.44, 0.80, 0.57, and 0.43, respectively. Like the random forest model, logistic regression improves upon all performance metrics compared to the random model. Compared to the ALBI model, the logistic regression model improves on recall and f1-score at the cost of precision and false negative rate. Looking at the rates at which F1/F2, F3, and F4 are predicted, the logistic regression model heavily predicts F3 (71.7% vs. 35.6%) mostly at the expense of predicting F4 (22.1% vs. 54.6%). The increased rate of predicting F3 significantly improves recall (0.80 vs. 0.43) with only a minor decrease in precision (0.44 vs. 0.48). As expected, the logistic model also incorrectly classifies F4 individuals as F3 more often than the ALBI model. However, because F3 and F4 both require medical attention, this type of error may be outweighed by correctly classifying more F3 individuals. At the same time, the logistic regression model is worse at classifying F1/F2 compared to the ALBI model. Overall, because classifying F3 is more important than classifying F1/F2 or F4, the logistic regression model is preferred over both benchmark models.

Compared to the random forest model, the logistic regression model improves upon recall, f1-score, and false negative rate. Looking at the rates of predicting F1/F2, F3, and F4, the logistic regression model predicts F3 more often at the expense of predicting F1/F2 and F4. However,

because F3 precision is 0.44 for both models, the logistic regression model is not worse at classifying F3 for predicting it more frequently. Between the two machine learning models, random forest and logistic regression, the logistic regression model is better at classifying F3.

7. Conclusion

Given the rising ability to treat advanced liver fibrosis and the need for treating patients, prescreening for liver fibrosis can potentially improve patient care and reduce costs. Cheap, non-invasive tests are ideal for a prescreening test and prior blood tests including ALBI score, APRI, and FIB-4 have been tested for the same purpose. Rather than prescreening or diagnosing with a single measure, machine learning models can classify individuals using trends in higher dimensional spaces and can potentially better prescreen or diagnose liver fibrosis. With data provided by the Mayo Clinic, a logistic regression model does better identify F3 individuals, although additional data and testing is required.