Iran Red Crescent Med J. 2016 November; 18(11):e32858. doi: 10.5812/ircmj.32858.

Published online 2016 August 9. **Research Article**

Prediction and Diagnosis of Non-Alcoholic Fatty Liver Disease (NAFLD) and Identification of Its Associated Factors Using the Classification Tree Method

Mehdi Birjandi,[1] Seyyed Mohammad Taghi Ayatollahi,[1,*] Saeedeh Pourahmad,[1] and Ali Reza Safarpour[2]

[1]Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, IR Iran
[2]Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, IR Iran

[*]*Corresponding author*: Seyyed Mohammad Taghi Ayatollahi, Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, IR Iran. Tel: +98-7132349330, Fax: +98-7132349330, E-mail: ayatolahim@sums.ac.ir

Received 2015 September 01; **Revised** 2015 October 25; **Accepted** 2015 November 22.

Abstract

Background: Non-alcoholic fatty liver disease (NAFLD) is the most common form of liver disease in many parts of the world.
Objectives: The aim of the present study was to identify the most important factors influencing NAFLD using a classification tree (CT) to predict the probability of NAFLD.
Patients and Methods: This cross-sectional study was conducted in Kavar, a town in the south of Fars province, Iran. A total of 1,600 individuals were selected for the study via the stratified method and multiple-stage cluster random sampling. A total of 30 demographic and clinical variables were measured for each individual. Participants were divided into two datasets: testing and training. We used the training dataset (1,120 individuals) to build the CT and the testing dataset (480 individuals) to assess the CT. The CT was also used to estimate class and to predict fatty liver occurrence.
Results: NAFLD was diagnosed in 22% of the individuals in the sample. Our findings revealed that the following variables, based on univariate analysis, had a significant association with NAFLD: marital status, history of hepatitis B vaccine, history of surgery, body mass index (BMI), waist-hip ratio (WHR), systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL), triglycerides (TG), alanine aminotransferase (ALT), cholesterol (CHO0, aspartate aminotransferase (AST), glucose (GLU), albumin (AL), and age ($P < 0.05$). The main affecting variables for predicting NAFLD based on the CT and in order of importance were as follows: BMI, WHR, triglycerides, glucose, SBP, and alanine aminotransferase. The goodness of fit model based on the training and testing datasets were as follows: prediction accuracy (80%, 75%), sensitivity (74%, 73%), specificity (83%, 77%), and the area under the receiver operating characteristic (ROC) curve (78%, 75%), respectively.
Conclusions: The CT is a suitable and easy-to-interpret approach for decision-making and predicting NAFLD.

Keywords: Classification Tree, Decision Tree, Non-Alcoholic Fatty Liver Disease, Prediction

1. Background

Non-alcoholic fatty liver disease (NAFLD), or fat accumulation in the liver parenchyma not due to the consumption of alcohol, is the most common form of liver disease in many parts of the world (1). NAFLD encompasses a spectrum of diseases ranging from simple steatosis to inflammatory steatohepatitis (NASH) with increasing levels of fibrosis and ultimately cirrhosis (2). The prevalence of NAFLD is quickly increasing such that it is a worldwide public health problem (3). According to reports, about 14% - 30% of the general population is affected by the disease (4, 5). There is also a global trend in obesity and type II diabetes (6). The prevalence of NAFLD in Western and Asian countries is estimated to be 20% - 30%, and it is approximately 15% among the adult population (4, 7, 8). Based on the Dallas Heart Study report, 30% of US adults are affected by NAFLD (9). In Iran, reports of its prevalence differ greatly (2.8% - 24%) depending on age group (10, 11) and geographical location. In southern Iran, the percentages reach 21.5% and 15.3% (12, 13).

Due to the undesirable outcomes of a delayed NAFLD diagnosis, the correct and timely diagnosis of NAFLD cases would yield important benefits. Accordingly, a common goal of many clinical studies is to develop a reliable clinical decision-making guide for its diagnosis (14). Early identification of high-risk individuals not only prevents damage to hepatocytes, but also NAFLD's side effects, such as heart failure, which is one cause of mortality in fatty liver patients (15-18). In order to determine the most appropriate treatment for each patient, major risk factors and their

relationships should be identified (14, 19).

So far, no attempts have been made to achieve these goals through a classification tree (CT) approach. A CT is a non-parametric statistical learning approach that can be used to develop predictive models (14). This method screens a large number of variables and classifies them based on their relations and importance. The data can be recursively partitioned into subsets using predictors based on a set of decision criteria. The flexibility and hierarchical nature of CTs are two important features that make them a common method used to solve practical decision-making problems (20, 21).

2. Objectives

The aim of the present study was to identify the major factors influencing NAFLD as well as to use the CT to predict the probability of NAFLD.

3. Patients and Methods

3.1. Materials

The present research was comprised of cross-sectional study conducted from January to August 2013 in the city of Kavar, Fars province, southern Iran. The city has a Mediterranean climate and a population of 75,000; it is located 35 km from Shiraz city. A total of 1,600 individuals were selected using a stratified method and a multiple-stage cluster random sampling method. This means that firstly, towns and villages were considered as stratified. The town was partitioned based on health care center (as a cluster). Some of these centers were randomly selected, while the subjects were randomly selected based on family registration data, which is available at these centers. From the selected villages, the samples were randomly selected based on data from health houses.

Thirty attributes, including demographic and clinical characteristics, were studied in order to predict the existence of NAFLD (with/without NAFLD). Individuals with a history of liver cirrhosis, underlying liver disease, hepatobilliary cancers, those with $>$ 20g/day alcohol consumption, and individuals receiving anti-thyroid medications were excluded from the study. A questionnaire was designed for this study that gathered demographic information (age, sex, marital status, education, etc.), medical history, and health-relevant behaviors such as smoking habits. Participants' waist-hip ratio (WHR) was calculated by measuring waist circumference and dividing it by hip circumference; body mass index (BMI) was calculated by dividing weight in kilograms by height in squared meters. The subjects were placed in the following categories based on the World Health Organization's (WHO) criteria for BMI: $<$ 18.5 (underweight), 18.5 - 24.9 (normal), 25 - 29.9 (overweight), and $\geq$ 30 (obese) in both the male and female groups (22). All of the participants were asked to attend the clinic after fasting overnight. A team of nurses and physicians performed interviews, filled out the questionnaires, and took intravenous blood samples to measure each subject's triglycerides (TG), cholesterol (CHO), high-density lipoprotein (HDL), alanine aminotransferase (ALT), aspartate aminotransferase (AST), glucose (GLU), HBSAG, HBSAB , and albumin (AL).

In addition, each participant's systolic blood pressure (SBP) and diastolic blood pressure (DBP) were measured at three different times; their mean was recorded as a reference. We measured liver enzyme levels for our clinical variables, as they are sensitive in diagnosing NAFLD. We found that the elevations in ALT and AST were typically four times greater than normal (23); further, gamma-glutamyltransferase (GGT) in the serum was frequently elevated in patients with NAFLD and was associated with advanced fibrosis and increased mortality in NAFLD patients (24). Using a cutoff serum GGT value of 96.5 U/L, GGT predicted advanced fibrosis with 83% sensitivity and 69% specificity (25). The process of calibrating and subsequently verifying the calibration of the spectrophotometers for low-level measurements was very sensitive to the user's technique and the surrounding environment. As measured activity levels dropped below 1.0 nephelometric turbidity units (NTU), interferences caused by bubbles, particulate contamination, and stray light became major factors. There are several ways to minimize these errors, the most common of which are one-point and two-point calibrations.

With the criteria we used to measure the variables for GGT, the test's sensitivity increased to 53%, but its specificity dropped to 75%. An elevated transaminase level had a positive predictive value of 90% for NAFLD (26). Although measuring transaminases as a population-based screening test has its pitfalls, it continues to be used widely in clinical practice to stratify patients with risk factors for NAFLD (27).

NAFLD was diagnosed according to increased echogenicity of the liver parenchyma and attenuation of the portal vein or echogenicity of the diaphragmatic area measured via a trans-abdominal-sonography-calibrated sonography machine. The sonographers were trained before the study (28).

Shiraz University of Medical Sciences' ethics committee evaluated and approved the study (Code: 92-6869, Date: May 11, 2012). All of the participants read the study objectives and signed informed consent forms.

3.2. Methods

CTs are non-parametric classifiers that construct hierarchical decision trees. A CT structure is made of root, internal, and leaf nodes. Each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision (Figure 1) (21); thus, CT can be used to model the relationships among the predictors and their interactions in determining the outcome (29). Three major elements support the development of a CT: 1) choosing a sampling-splitting rule that defines the tree branch connected to the classification nodes; 2) evaluating the classification, which is produced by the splitting rule at each node; and 3) choosing an optimal or final tree by using the criteria for classification purposes (30).

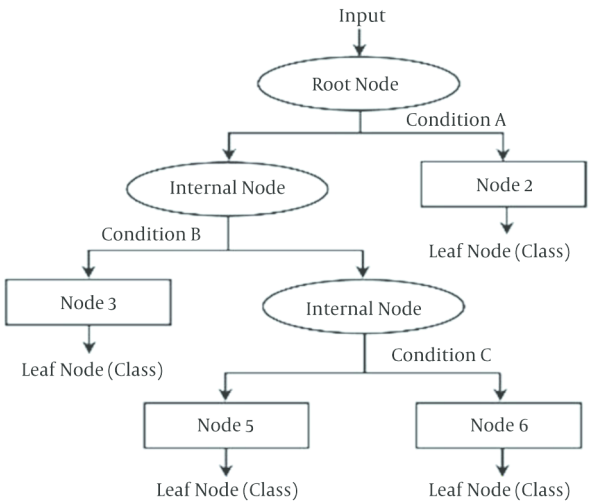


Figure 1. CT Structure

With the GUIDE method, chi-square tests were used to measure the degree of association between each predictor variable and each response variable (if the predictor variable was deemed continuous by evaluating its sample quartiles, it was divided into four groups). Then, the most significant predictor variable was chosen for partitioning in the first step. If the selected variable was continuous, the method searched for the best cut score in order to make the subsets as homogeneous as possible. If the selected variable was categorical, the best split was chosen by evaluating the subset of the values. This step was used recursively in each partition, and the whole process was described by the tree structure. When the sample size was less than the pre-specified sample, the partitioning was stopped. Since the resulting CT model was very complex, a sequence of smaller tree models was obtained by sequentially pruning the tree structure using a 10-fold cross-validation until only one node was left. Finally, the tree model with the smallest cross-validated error rate was chosen (31). The whole set was then divided into a training set (almost 70% of all cases), which was used for the induction of a CT that classified the individual into "with" or "without" risk of NAFLD as well as a testing set (30%) that was used to check the accuracy of the obtained solution.

The CT was built using NAFLD as a response variable with the following steps: From each of the predictor variables of interest, the variable that split the data into the two most pure response groups (or nodes) using prespecified criteria was chosen. These criteria included specification of the minimum number of observations to enter each node (10 observations), the minimum number in a node before attempting to split (10 observations), and the "costs" assigned to misclassifying the items. Cost was measured in terms of proportion of misclassified cases. In order to better predict the classification of patients who actually had NAFLD, different costs were applied to the classification of the two groups (14). Regarding this point, the cost of misclassifying an individual with a high risk of NAFLD as having a low risk NAFLD was two times that of the opposite scenario.

Finally, the validity of the model was examined by the area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, accuracy rate in prediction, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), and negative likelihood ratio (NLR). Pearson's chi-square test and Fisher's exact test were applied to identify any associations between qualitative risk factors and fatty liver data. Independent t-tests were used to assess differences in the mean value of the continuous variables between subjects with and without NAFLD. Statistical significance was set as $P < 0.05$.

The GUIDE program (www.stat.wisc.edu/~ loh/guide.html) version 13 was used to construct the CT. STATA 10 was used to analyze the data.

4. Results

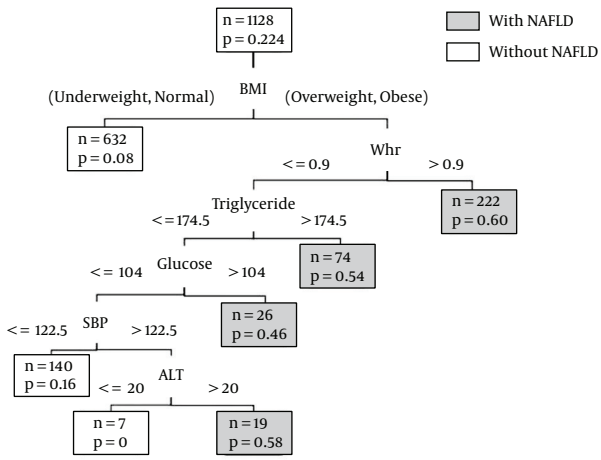
The data of 1,600 subjects, comprised of 471 (29.4%) males and 1,129 (70.6%) females, was analyzed. The individuals' mean age was 37.3 $\pm$ 17.2 (range: 16 - 88). A total of 12% of the individuals were underweight, 44% were normal, 32% were overweight, and 12% were obese. NAFLD was diagnosed in (22.4%) subjects; this rate was 23.4% in males and 22% in females.

Table 1 summarizes the demographic and clinical characteristics of the participants. The mean age of the participants with NAFLD (45.9 $\pm$ 13.34) was significantly higher than those without NAFLD (34.85 $\pm$ 17.45) ($P < 0.001$).

The anthropometric indices BMI and WHR were associated with presence of NAFLD ($P < 0.001$). The mean HDL, AL, SBP, and DBP were associated with the presence of NAFLD ($P < 0.001$). The mean TG, ALT, CHO, and AST of the participants with NAFLD were significantly higher than those without NAFLD ($P < 0.001$). Also, there was a significant association between marital status, history of hepatitis B vaccine, history of surgery, and NAFLD ($P < 0.001$). The other variables did not have a significant association with NAFLD ($P > 0.05$) (Table 1).

Thirty attributes were used as the inputs of the algorithm to predict the binary outcome representing the issue of NAFLD. To begin the modeling process, 1,120 individuals were randomly assigned to the training set, while the rest were assigned to the testing set. Figure 2 shows the classification tree of the NAFLD subjects derived from the training dataset.

Figure 2. CT for Predicting NAFLD



The number of patients (n) and the probability of NAFLD are given inside each node. Predicted class is given beneath each terminal node.

Each terminal node in Figure 2 indicates the presence or absence of NAFLD (with its related probability). For each subject, one branch was followed based on his or her circumstances until we reached the terminal node. At each intermediate node, an observation went to the branches if and only if the condition was satisfied. This tree can be helpful in exploring high-order interaction between the attributes. For instance, the GUIDE analysis in our dataset showed that the BMI variable was the most important factor affecting NAFLD, as it appeared on the top of the tree in the first step. Furthermore, it had a relationship with WHR. In obese and overweight subjects whose WHR mg/dL was greater than 0.9, the probability of NAFLD was 60% and the predicted class was "with NAFLD," while for the others with a WHR less than 0.9 mg/dL, the TG was important. Based on the branches of the derived tree, seven rules were extracted related to seven leaf nodes. To simplify the decision-making process for clinical experts, these rules are summarized in Table 2.

The rules consist of a sequence of logical if then else statements about the patients' attributes (23) with a simple interpretation. For example, rule six expressed that the probability of NAFLD was 58% and that the predicted class was "with NAFLD" for a person who was overweight or obese with a WHR less than 0.9, a TG less than 74.5 mg/dL, a GLU less than 104 mg/dL, an SBP greater than 122.5 mmHg, and an ALT greater than 20.5 IU/L.

To test the validity of the training tree, the testing dataset, which did not contribute to the construction of the tree, was applied. The evaluation measures of the CT were calculated for both the training and testing samples (Table 3). Regarding the sensitivity of 73% of the test set, 73% of the individuals actually had NAFLD and therefore did not contribute to creating the tree; therefore, they were diagnosed correctly by this method. The PLR was 3.04 and 4.35 for the testing and training sets, respectively. The NLR, on the other hand, was 0.35 and 0.31 for the testing and training sets, respectively, which is near zero.

In addition, the area under the ROC curve, diagnostic accuracy, PPT, NPT, PLR, and NLR were obtained for both the training and testing datasets. Youden's index equaled 60%.

Dark nodes represent the predicted class "with NAFLD" and white nodes represent the predicted class "without NAFLD.

5. Discussion

We found that approximately 22.4% of the subjects had NAFLD. Further, the following variables, based on univariate analysis, were significantly associated with NAFLD: marital status, history of hepatitis B vaccine, history of surgery, BMI, WHR, SBP, DBP, HDL, TG, ALT, CHO, AST, GLU, AL, and age ($P < 0.05$).

The main variables for predicting NAFLD based on the CT, and in order of importance, they were BMI, WHR, TG, GLU, SBP, and ALT. We predicted the probability of having NAFLD based on the decision rules in the CT.

According to our findings on sensitivity, specificity, area under the ROC curve, NPV, accuracy, and NLR, there was no perceptible difference between the testing and training datasets, indicating the capability of this model to estimate and predict NAFLD. The only difference we found was that PLR was overestimated in the model for the training dataset; it was large and significant based on the testing dataset due to its difference of zero (3.04). Using the testing dataset to diagnose the disease revealed that the

Table 1. Demographic and Clinical Characteristics of Participants According to Groups (No. (%) or Mean ± SD)

Risk Factor	Abbreviation	Level	Without NAFLD (n = 1241)	With NAFLD (n = 359)	P Value
Sex	SEX	Male	361 (29.1)	110 (30.7)	0.55
		Female	880 (70.9)	249 (69.3)	
		Single	447 (36)	27 (7.5)	
Marital status	MS	Married	726 (58.5)	297 (83)	< 0.001[a]
		Other	68 (5.5)	35 (9.5)	
		Yes	538 (43.4)	70 (19.3)	
History of hepatitis B vaccine	HEP	No	703 (56.6)	289 (80.7)	< 0.001[a]
History of blood transfusion	BT	Yes	22 (1.8)	11 (3.1)	
		No	1,219 (98.2)	348 (96.9)	0.139
Thalassemia	THAL	Yes	2 (.2)	1 (.3)	0.533
		No	1,239 (99.8)	358 (99.7)	
Hemophilia	HEMO	Yes	3 (.2)	0 (.0)	0.99
		No	1,238 (99.8)	359 (100)	
Dialysis	DI	Yes	3 (.2)	1 (.3)	0.99
		No	1,238 (99.8)	358 (99.7)	
Surgery	SU	Yes	3 (.2)	1 (.3)	0.99
		No	1,238 (99.8)	358 (99.7)	
History of surgery	HS	Yes	356 (28.7)	141 (39.4)	< 0.001[a]
		No	885 (71.3)	218 (60.4)	
History of dental surgery	DE	Yes	1,002 (80.7)	303 (84.6)	0.104
		No	239 (19.3)	56 (15.4)	
History of phlebotomy	PH	Yes	94 (7.6)	35 (9.8)	0.186
		No	1,147 (92.4)	324 (90.2)	
Tattoos	TA	Yes	38 (3.1)	19 (5.3)	0.052
		No	1,203 (96.9)	340 (94.7)	
History of unsanitary ear piercing	UPE	Yes	541 (43.6)	141 (39.4)	0.147
		No	700 (56.4)	218 (60.6)	
Use of hookah	HOO	Yes	83 (6.7)	28 (7.8)	0.479
		No	1,158 (93.3)	331 (92.2)	
Current smoker (tobacco)	SMOK	Yes	39 (3.1)	19 (5.3)	0.076
		No	1,202 (96.9)	340 (94.7)	
History of drug use	HDU	Yes	28 (2.3)	6 (1.7)	0.677
		No	1,213 (97.7)	353 (98.3)	
HBSAG	HBSAG	Negative	1,215 (98.1)	353 (98.5)	0.83
		Positive	26 (1.9)	6 (1.5)	
HBSAB	HBSAB	Negative	1,079 (88.5)	307 (87.0)	0.142
		Positive	162 (11.5)	52 (13.0)	
		Underweight (UW)	197 (15.9)	1 (.3)	
Body mass index	BMI	Normal (N)	633 (51)	62 (17.3)	< 0.001[a]
		Overweight (OW)	320 (25.8)	186 (51.7)	
		Obese (OB)	87 (7)	110 (30.7)	
Waist-hip ratio	WHR		0.83 ± 0.09	0.92 ± 0.09	<0.001[a]
Systolic blood pressure	SBP		100.05 ± 26.1	108.42 ± 31.86	<0.001[a]
Diastolic blood pressure	DBP		82.14 ± 20.01	93.37 ± 23.85	<0.001[a]
High-density lipoprotein	HDL		50.95 ± 11.5	48.9 ± 9.73	0.009[a]
Triglycerides	TG		120.3 ± 68.52	193.89 ± 113.5	< 0.001[a]
Alanine aminotransferase	ALT		15.56 ± 10.92	19.11 ± 12.5	< 0.001[a]
Cholesterol	CHO		184.94 ± 42.58	207.62 ± 41.79	< 0.001[a]
Aspartate aminotransferase	AST		24.84 ± 11.66	28.06 ± 17.84	< 0.001[a]
Glucose	GLU		96.68 ± 26.86	108.45 ± 39.56	< 0.001[a]
Albumin	AL		4.32 ± 0.37	4.23 ± 0.4	< 0.001[a]
Age	AGE		34.85 ± 17.45	45.9 ± 13.34	< 0.001[a]

[a] Significant at the 0.05 level.

Table 2. The Rules for the CT

Variable	Response					Class	
	Fatty Liver					Predicted Probability (%)	Predicted Class
BMI	WHR	TG	GLU	SBP	ALT		
UW. N						8	Without NAFLD
OB. OW	> 0.9					60	With NAFLD
OB. OW	< 0.9	> 174.5				54	With NAFLD
OB. OW	< 0.9	< 174.5	> 104			46	With NAFLD
OB. OW	< 0.9	< 174.5	< 104	< 122.5		16	Without NAFLD
OB. OW	< 0.9	< 174.5	< 104	> 122.5	>20.5	58	With NAFLD
OB. OW	< 0.9	< 174.5	< 104	> 122.5	<20.5	0	Without NAFLD

Abbreviations: UW, underweight, N, normal, OW, overweight, O, obese.

Table 3. The Measures of the CT for the Training and Testing Samples[a]

Value	Training Sample	Testing Sample
Sensitivity	74	73
Specificity	83	77
Area under the ROC curve	78	75
Prediction accuracy	80	75
Positive predictive value	58	57
Negative predictive value	91	93
Positive likelihood ratios	4.35	3.04
Negative likelihood ratios	0.31	0.35

[a]Values are expressed as No. (%).

PPV values had clinical importance. Therefore, our findings showed that the calculated and predicted PPV values based on the training and testing datasets did not have perceptible differences from one another; thus, we could rely on the CT in estimating the PPV values. Low levels of PPV are not relevant to the sensitivity and specificity of the tests; instead, they vary according to the disease prevalence in each population. Thus, low levels of PPV should not be considered as a failure of the CT method. In this study, almost 82% of the patients with NAFLD were overweight or obese. The results showed that BMI was the most important factor in NAFLD. It was mentioned as a base factor in all of the previous studies on fatty liver (1, 7, 10, 12, 32-35).

WHR was identified as another important factor in diagnosing NAFLD. In many studies, waist circumference is introduced as a significant factor in diagnosing NAFLD (1, 7, 10, 12, 32-35). However, in the study performed by Eshraghiyan et al. (13), WHR, in comparison with waist circumference, was a more precise risk factor for NAFLD.

High TG was another significant factor that affected diagnosis of NAFLD. This result is in line with other studies (7, 10, 12, 32, 34-37). Bedongi et al. (7) indicated that TG, BMI, and waist circumference were the three most important factors in diagnosing NAFLD similarly, these variables were major factors at the top of the tree in the present study.

Fotbolcu et al. (33) and Bedongi et al. (7) compared patients with NAFLD and a control group. Both found a significant difference in the SBP between the groups, which was also considered to be a risk factor in the present study. Marcheini et al. (38) and Paschos et al. (39) obtained similar results.

We found that NAFLD was associated with BMI, WHR, high SBP, and high TG. These variables were also used to predict NAFLD based on the CT. These determinants of NAFLD are metabolic and anthropometric features of metabolic syndrome (40). In our study, there was a very close association between NAFLD and metabolic syndrome. Lankarani et al. (12) and Lonardo et al. (41), beside obtaining the same results as we did, showed that an additional feature of metabolic syndrome could be NAFLD. The notable point in this study is that the CT predicted NAFLD based on the components identified in metabolic syndrome.

In this study, GLU and ALT were the other factors that affected NAFLD. In a case-control study performed by Dey et al. (37), the mean fasting GLU and ALT levels in the group with NAFLD were significantly higher than in the control group. Moreover, GLU was strongly associated with the development and progression of NAFLD (41, 42).

Estakhri et al. (43) evaluated the effect of NAFLD on the outcome of chronic hepatitis B disease. They found a relationship between NAFLD and increased ALT levels. A study performed on an Iranian population by Eshraghiyan et al. (13) obtained similar results. Hosseini et al. (44) used a logistic regression model in order to determine the risk fac-

tors for NAFLD. They found that BMI, waist circumference, and serum TG were significant.

While the previous research on NAFLD has not focused on interaction effects between the variables, even in cases where logistic regression was used to determine the effect of risk factors, these interactions are highly informative. This suggests that there may be better ways to predict NAFLD in individuals.

Lankarani et al. (12) showed that high CHO, unlike HDL, had a significant association with NAFLD .In contrast, Tomizawa et al. (45) showed that HDL and NAFLD were associated. Eshraghian et al. (13) argued that there was no relation between high CHO and HDL with NAFLD. Of course, these disagreements may be due to environmental and genetic differences, the differences in methodology, and the type of diagnostic criteria. Although in our study, there were significant associations between HDL and high CHO with NAFLD based on our univariate analysis; these variables were not used to structure the tree. One major limitation with a CT is that the variables it uses might not be the only important ones that exist. If there were a high correlation between two or more independent variables as equally important in prediction, at most, one would be selected. Of course, this problem exists in other statistical models, such as logistic regression, in which multicollinearity makes estimated regression problematic. Another limitation of CT is that we were unable to use a more traditional P value (19).

Individuals' fatty liver diagnoses were done based on sonography, which does not have high sensitivity and specificity for fatty liver diagnosis. No detection tools with high accuracy for diagnosing fatty liver currently exist. The golden standard for diagnosing fatty liver is biopsy, which is not often used because of its dangers and the ethical problems of invasive diagnostic tests in apparently healthy individuals (46). Another weakness of our study is the probable inter-observer bias due to the absence of periodic estimation of kappa statistics. Intermittent assessments of expert examiners from the cohort personnel were used to overcome this problem. However, regarding this limitation, the results of the GUIDE classification tree are valuable and can be used to predict NAFLD. Our study offers just one alternative diagnostic method to safely and accurately diagnose NAFLD.

In order to evaluate the accuracy and reliability of the predicted classes, we relied on repetition such as Bootstrap and other data mining methods; repetition is particularly important, and we suggest that it be used in future studies. Our attempt to model the grade of fatty liver disease based on related risk factors may yield valuable results for its prevention and treatment.

Acknowledgments

We would like to thank the gastroenterohepatology research center at Shiraz University of Medical Sciences for gathering data and for its collaboration in this study. We would also like to thank Dr. Nasrin Shokrpour at the center for development of clinical research, Nemazee hospital for her editorial assistance.

Footnotes

Authors' Contribution: Mehdi Birjandi: study concept, design analysis, data interpretation, statistical analysis, manuscript drafting; Seyyed Mohammad Taghi Ayatollahi: study supervision, manuscript drafting, critical revision of the manuscript for important intellectual content; Saeedeh Pourahmad: study supervision, manuscript drafting, critical revision of the manuscript for important intellectual content, administration; Ali Reza Safarpour: acquisition of data, critical revision of the manuscript for important intellectual content.

Funding/Support: This study was extracted from the PhD thesis of the first author and was financially supported by grant number 92-6869, Shiraz University of Medical Sciences.

References

1. Bellentani S, Marino M. Epidemiology and natural history of non-alcoholic fatty liver disease (NAFLD). *Ann Hepatol.* 2009;**8 Suppl 1**:S4–8. [PubMed: 19381118].
2. Dowman JK, Tomlinson JW, Newsome PN. Systematic review: the diagnosis and staging of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis. *Aliment Pharmacol Ther.* 2011;**33**(5):525–40. doi: 10.1111/j.1365-2036.2010.04556.x. [PubMed: 21198708].
3. Kelishadi R, Poursafa P. Obesity and air pollution: global risk factors for pediatric non-alcoholic fatty liver disease. *Hepat Mon.* 2011;**11**(10):794–802. doi: 10.5812/kowsar.1735143X.746. [PubMed: 22224077].
4. Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, et al. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. *Hepatology.* 2004;**40**(6):1387–95. doi: 10.1002/hep.20466. [PubMed: 15565570].
5. Kelishadi R, Farajian S, Mirlohi M. Probiotics as a novel treatment for non-alcoholic Fatty liver disease; a systematic review on the current evidences. *Hepat Mon.* 2013;**13**(4):ee7233. doi: 10.5812/hepatmon.7233. [PubMed: 23885277].
6. Bjornsson E, Angulo P. Non-alcoholic fatty liver disease. *Scand J Gastroenterol.* 2007;**42**(9):1023–30. doi: 10.1080/00365520701514529. [PubMed: 17710666].
7. Bedogni G, Miglioli L, Masutti F, Tiribelli C, Marchesini G, Bellentani S. Prevalence of and risk factors for nonalcoholic fatty liver disease: the Dionysos nutrition and liver study. *Hepatology.* 2005;**42**(1):44–52. doi: 10.1002/hep.20734. [PubMed: 15895401].
8. Chitturi S, Wong VW, Farrell G. Nonalcoholic fatty liver in Asia: Firmly entrenched and rapidly gaining ground. *J Gastroenterol Hepatol.* 2011;**26 Suppl 1**:163–72. doi: 10.1111/j.1440-1746.2010.06548.x. [PubMed: 21199528].

9. Browning JD. Statins and hepatic steatosis: perspectives from the Dallas Heart Study. *Hepatology.* 2006;**44**(2):466–71. doi: 10.1002/hep.21248. [PubMed: 16871575].
10. Alavian SM, Mohammad-Alizadeh AH, Esna-Ashari F, Ardalan G, Hajarizadeh B. Non-alcoholic fatty liver disease prevalence among school-aged children and adolescents in Iran and its association with biochemical and anthropometric measures. *Liver Int.* 2009;**29**(2):159–63. doi: 10.1111/j.1478-3231.2008.01790.x. [PubMed: 18492015].
11. Jamali R, Khonsari M, Merat S, Khoshnia M, Jafari E, Bahram Kalhori A, et al. Persistent alanine aminotransferase elevation among the general Iranian population: prevalence and causes. *World J Gastroenterol.* 2008;**14**(18):2867–71. [PubMed: 18473412].
12. Lankarani KB, Ghaffarpasand F, Mahmoodi M, Lotfi M, Zamiri N, Heydari ST. Non alcoholic fatty liver disease in southern Iran: a population based study. *Hepatitis Mon.* 2013;**13**(5).
13. Eshraghian A, Dabbaghmanesh MH, Eshraghian H, Fattahi MR, Omrani GR. Nonalcoholic fatty liver disease in a cluster of Iranian population: thyroid status and metabolic risk factors. *Arch Iran Med.* 2013;**16**(10):584–9. [PubMed: 24093139].
14. Breiman L. Classification and regression trees. USA: Chapman & Hall; 1984.
15. Marchesini G, Moscatiello S, Di Domizio S, Forlani G. Obesity-associated liver disease. *J Clin Endocrinol Metab.* 2008;**93**(11 Suppl 1):S74–80. doi: 10.1210/jc.2008-1399. [PubMed: 18987273].
16. Targher G, Arcaro G. Non-alcoholic fatty liver disease and increased risk of cardiovascular disease. *Atherosclerosis.* 2007;**191**(2):235–40. doi: 10.1016/j.atherosclerosis.2006.08.021. [PubMed: 16970951].
17. Targher G, Marra F, Marchesini G. Increased risk of cardiovascular disease in non-alcoholic fatty liver disease: causal effect or epiphenomenon?. *Diabetologia.* 2008;**51**(11):1947–53. doi: 10.1007/s00125-008-1135-4. [PubMed: 18762907].
18. Ballestri S, Lonardo A, Bonapace S, Byrne CD, Loria P, Targher G. Risk of cardiovascular, cardiac and arrhythmic complications in patients with non-alcoholic fatty liver disease. *World J Gastroenterol.* 2014;**20**(7):1724–45. doi: 10.3748/wjg.v20.i7.1724. [PubMed: 24587651].
19. Piper ME, Loh WY, Smith SS, Japuntich SJ, Baker TB. Using decision tree analysis to identify risk factors for relapse to smoking. *Subst Use Misuse.* 2011;**46**(4):492–510. doi: 10.3109/10826081003682222. [PubMed: 20397871].
20. Rokach L. Data Mining with Decision Trees. Incorporated; 2008.
21. Gorunescu F. Data Mining: Concepts, Models and Techniques. USA: Springer; 2011.
22. Anuurad E, Shiwaku K, Nogi A, Kitajima K, Enkhmaa B, Shimono K, et al. The new BMI criteria for asians by the regional office for the western pacific region of WHO are suitable for screening of overweight to prevent metabolic syndrome in elder Japanese workers. *J Occup Health.* 2003;**45**(6):335–43. [PubMed: 14676412].
23. Torres DM, Harrison SA. Diagnosis and therapy of nonalcoholic steatohepatitis. *Gastroenterology.* 2008;**134**(6):1682–98. doi: 10.1053/j.gastro.2008.02.077. [PubMed: 18471547].
24. Ruhl CE, Everhart JE. Elevated serum alanine aminotransferase and γ-glutamyltransferase and mortality in the United States population. *Gastroenterology.* 2009;**136**(2):477–85.
25. Tahan V, Canbakan B, Balci H, Dane F, Akin H, Can G. Serum gamma-glutamyltranspeptidase distinguishes non-alcoholic fatty liver disease at high risk. *Hepato-Gastroenterol.* 2007;**55**(85):1433–8.
26. Choudhury J, Sanyal AJ. Clinical aspects of fatty liver disease. Seminars in liver disease. .
27. Clark JM, Brancati FL, Diehl AM. Nonalcoholic fatty liver disease. *Gastroenterology.* 2002;**122**(6):1649–57. [PubMed: 12016429].
28. Saverymuttu S, Joseph A, Maxwell J. Ultrasound scanning in the detection of hepatic fibrosis and steatosis. *British Med J.* 1986;**292**(13).
29. Anyanwu MN, Shiva SG. Comparative analysis of serial decision tree classification algorithms. *Int J Computer Science Security.* 2009;**3**(3):230–40.
30. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes.* 2011;**4**:299. doi: 10.1186/1756-0500-4-299. [PubMed: 21849043].
31. Loh WY. Improving the precision of classification trees. *Annals Applied Statistics.* 2009:1710–37.
32. Chan DF, Li AM, Chu WC, Chan MH, Wong EM, Liu EK, et al. Hepatic steatosis in obese Chinese children. *Int J Obes Relat Metab Disord.* 2004;**28**(10):1257–63. doi: 10.1038/sj.ijo.0802734. [PubMed: 15278103].
33. Fotbolcu H, Yakar T, Duman D, Karaahmet T, Tigen K, Cevik C, et al. Impairment of the left ventricular systolic and diastolic function in patients with non-alcoholic fatty liver disease. *Cardiol J.* 2010;**17**(5):457–63. [PubMed: 20865675].
34. Fu CC, Chen MC, Li YM, Liu TT, Wang LY. The risk factors for ultrasound-diagnosed non-alcoholic fatty liver disease among adolescents. *Ann Acad Med Singapore.* 2009;**38**(1):15–7. [PubMed: 19221666].
35. Gelpi Mendez JA, Castellanos Fillot A, Sainz Gutierrez JC, Quevedo Aguado L, Martin Barallat J. [Prevalence of non-alcoholic fatty liver disease and associated risk factors among managers from the community of Madrid]. *Arch Prev Riesgos Labor.* 2014;**17**(2):84–90. doi: 10.12961/aprl.2014.17.2.03. [PubMed: 24718630].
36. Shannon A, Alkhouri N, Carter-Kent C, Monti L, Devito R, Lopez R, et al. Ultrasonographic quantitative estimation of hepatic steatosis in children With NAFLD. *J Pediatr Gastroenterol Nutr.* 2011;**53**(2):190–5. doi: 10.1097/MPG.0b013e31821b4b61. [PubMed: 21788761].
37. Dey PK, Sutradhar SR, Barman TK, Khan NA, Hasan I, Haque MF, et al. Risk factors of non-alcoholic fatty liver disease. *Mymensingh Med J.* 2013;**22**(4):649–54. [PubMed: 24292291].
38. Marchesini G, Bugianesi E, Forlani G, Cerrelli F, Lenzi M, Manini R, et al. Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology.* 2003;**37**(4):917–23. doi: 10.1053/jhep.2003.50161. [PubMed: 12668987].
39. Paschos P, Paletas K. Non alcoholic fatty liver disease and metabolic syndrome. *Hippokratia.* 2009;**13**(1):9–19. [PubMed: 19240815].
40. Loria P, Lonardo A, Carulli L, Verrone AM, Ricchi M, Lombardini S, et al. Review article: the metabolic syndrome and non-alcoholic fatty liver disease. *Aliment Pharmacol Ther.* 2005;**22 Suppl 2**:31–6. doi: 10.1111/j.1365-2036.2005.02592.x. [PubMed: 16225469].
41. Lonardo A, Ballestri S, Marchesini G, Angulo P, Loria P. Nonalcoholic fatty liver disease: a precursor of the metabolic syndrome. *Dig Liver Dis.* 2015;**47**(3):181–90. doi: 10.1016/j.dld.2014.09.020. [PubMed: 25739820].
42. Anstee QM, Targher G, Day CP. Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis. *Nat Rev Gastroenterol Hepatol.* 2013;**10**(6):330–44. doi: 10.1038/nrgastro.2013.41. [PubMed: 23507799].
43. Estakhri A, Sari AA, Nedjat S, Rohban M, Rakhshani N, Tavangar S. The effect of NAFLD (non-alcoholic fatty liver disease) on long-term outcome of chronic hepatitis B in Iranian patients. *Open J Gastroenterol.* 2012;**2012**(2):18–21.
44. Hosseini SM, Mousavi S, Poursafa P, Kelishadi R. Risk Score Model for Predicting Sonographic Non-alcoholic Fatty Liver Disease in Children and Adolescents. *Iran J Pediatr.* 2011;**21**(2):181–7. [PubMed: 23056785].
45. Tomizawa M, Kawanabe Y, Shinozaki F, Sato S, Motoyoshi Y, Sugiyama T, et al. Triglyceride is strongly associated with nonalcoholic fatty liver disease among markers of hyperlipidemia and diabetes. *Biomed Rep.* 2014;**2**(5):633–6. doi: 10.3892/br.2014.309. [PubMed: 25054002].
46. Vernon G, Baranova A, Younossi ZM. Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. *Aliment Pharmacol Ther.* 2011;**34**(3):274–85. doi: 10.1111/j.1365-2036.2011.04724.x. [PubMed: 21623852].