Original Article

Data Mining and Machine Learning Algorithms Using IL28B Genotype and Biochemical Markers Best Predicted Advanced Liver Fibrosis in Chronic Hepatitis C

Hend Ibrahim Shousha1*, Abubakr Hussein Awad2, Dalia Abdelhamid Omran1, Mayada Mohamed Elnegouly1, and Mahasen Mabrouk1

1Endemic Medicine Department, Faculty of Medicine; and
2Computer Science Department, Faculty of Computers and Information, Cairo University, Cairo, Egypt

SUMMARY: IL28B single nucleotide polymorphism (rs12979860) is an etiology-independent predictor of hepatitis C virus (HCV)-related hepatic fibrosis. Data mining is a method of predictive analysis which can explore tremendous volumes of information from health records to discover hidden patterns and relationships. The current study aims to evaluate and compare the prediction accuracy of scoring system like aspartate aminotransferase-to-platelet ratio index (APRI) and fibrosis-4 (FIB-4) index versus data mining for the prediction of HCV-related advanced fibrosis. This retrospective study included 427 patients with chronic hepatitis C. We used data mining analysis to construct a decision tree by reduced error (REP) technique, followed by Auto-WEKA tool to select the best classifier out of 39 algorithms to predict advanced fibrosis. APRI and FIB-4 had sensitivity-specificity parameters of 0.523–0.831 and 0.415–0.917, respectively. REPTree algorithm was able to predict advanced fibrosis with sensitivity of 0.749, specificity of 0.729, and receiver operating characteristic (ROC) area of 0.796. Out of the 16 attributes, IL28B genotype was selected by the REPTree as the best predictor for advanced fibrosis. Using Auto-WEKA, the multilayer perceptron (MLP) neural model was selected as the best predictive algorithm with sensitivity of 0.825, specificity of 0.811, and ROC area of 0.880. Thus, MLP is better than APRI, FIB-4, and REPTree for predicting advanced fibrosis for patients with chronic hepatitis C.

INTRODUCTION

Hepatic fibrosis is the core pathological process in chronic hepatitis C virus (HCV) infection with the subsequent of development of liver cirrhosis, thereby leading to mortality. Liver cirrhosis leads to hepatocellular dysfunction, portal hypertension, and the development of hepatocellular carcinoma (HCC) (1,2). Egypt has the highest prevalence of HCV, with genotype 4 (G4) being responsible for over 90% of the cases and the remaining due to genotype 1 (G1) (3). The progression of hepatic fibrosis is influenced by environmental, viral, and host factors such as duration of infection, age, sex, alcohol consumption, immunosuppression, obesity, insulin resistance, co-infection with other viruses, and genetic factors (4).

Host genetics has a great impact on liver fibrosis, especially the variants in genes controlling immune responses and inflammation pathways. The discovery of 2 single nucleotide polymorphisms (SNP) (rs12979860 and rs8099917) in the interferon-lambda (IFN-λ) region on chromosome 19 made a breakthrough in HCV research. It appeared to be the strongest factor associated with viral clearance, either spontaneously or after IFN-based therapy in genotypes 1, 2, 3, and 4 (5–10) and across different ethnic groups (11). In patient with G4, rs8099917 genotype in the carriers of the heterozygous C allele of rs12979860 was found to improve the prediction of sustained virologic response (SVR)(12). The IL28B genotype neither predicts nor affects the viral load in Egyptians with chronic HCV infections (13). Therefore, the importance of IL28B genetic variants in the setting of direct antiviral therapy is controversial (8,14,15). However, a study reported that the cost-effectiveness of sofosbuvir improved in IL28B CT/TT genotype (16).

Non-invasive methods for diagnosing the stages of liver fibrosis, either radiological (as in FibroScan; Echosens, Paris, France) or serological, were found to be more reliable diagnostic tools compared to liver biopsy (17). The aspartate aminotransferase-to-platelet ratio index (APRI) and the fibrosis-4 (FIB-4) score are simple, noninvasive, easy to perform, inexpensive, and reproducible algorithms that were found to have satisfactory performance in the diagnosis of moderate to severe stages of fibrosis (17–19).

Data mining is a method of predictive analysis which can explore tremendous volumes of data to discover hidden patterns and relationships in highly complex dataset, thus enabling the development of predictive models. Decision trees are preferred algorithm for building understandable predictive models, that are simple yet fast-to-build, with good accuracy, easily converted to classification rules, does not require any domain knowledge and easy to assimilate by physicians (20).
Analysis of such large amounts of data is a difficult and computationally intensive task for most of the existing data mining algorithms. One of the "soft" approaches for reducing the need for computer power in data analysis, is the introduction of meta-learning systems for selection and ranking of the best-suited algorithms for different datasets (21). This study aimed to evaluate and compare the prediction accuracy of APRI and FIB-4 versus data mining techniques (selecting the best performing algorithm), in the prediction of advanced fibrosis in patients with chronic HCV infection, using clinical information, serum biomarkers, in addition to IL28B rs12979860 SNP genotype as input data.

**MATERIALS AND METHODS**

**Patients:** Our retrospective study included 427 Egyptian patients with chronic hepatitis C who were naïve candidates for an antiviral therapy. A local ethical committee approval was obtained before starting the data collection. To respect the patients’ confidentiality, all patients were represented in the study by code numbers, concealing their personal data. The protocol followed in the study conformed to the ethical guidelines of the Declaration of Helsinki (1975). Full personal and clinical information were retrieved, routine pre-treatment including complete blood count, liver biochemical tests (alanine aminotransferase [ALT], aspartate aminotransferase [AST], alkaline phosphatase, serum bilirubin, albumin, and international normalized ratio [INR]), urea and creatinine levels, serum α-fetoprotein [AFP], thyroid stimulating hormone [TSH], blood sugar, α-bilirubin, albumin, and international normalized ratio (INR), were included. Liver biochemical tests were performed in the laboratory of the Egypt for Advance Medical Services (EAMS), Egypt.

**Methods:**

**IL28B genotyping:** Genomic DNA extraction: Five milliliter of blood was collected. Genomic DNA was extracted and purified from peripheral blood leukocytes using the Whole Blood DNA Purification Kit (Bio Vision, Milpitas, CA, USA).

 Detection of IL28B rs12979860 C/T polymorphism: PCR-restriction fragment length polymorphism (RFLP) was performed as proposed by El Awady et al. (25) for the analysis of IL28B rs12979860 C/T polymorphism with minor modifications using primer pair (Thermo Fisher Scientific, Waltham, MA, USA); forward primer 5’-CGAGGGCCCCCTAACCTCTGCA-3’ and the reverse primer 5’-GGGAGCCGGATGCAATTCA-3’ (23). Enzymatic amplification was performed by PCR using Master Taq polymerase (Quanta Bio, Beverly, MA, USA) and TProfessional Standard 96-well Gradient Thermal Cycler (Biometa, Göttingen, Germany). The PCR reaction mixture (20 µL) contained 10 µL 2 × PCR Master Mix (10 × PCR buffer, 4 mM MgCl2, 0.5 µL Taq DNA polymerase, 0.4 mM dNTPs [dATP, dCTP, dGTP, dTTP]), 1 µL of each primer (25 pmol), 2 µL of genomic DNA, and 6 µL sterilized nuclelease-free water. The thermal cycler protocol was as follows: 35 cycles of denaturation (95°C for 60 s), annealing (65°C for 60 s), and polymerization (72°C for 60 s). There was a 5-min pre-incubation at 95°C before the first cycle, and a 5-min additional extension at 72°C after completion of the cycles. The amplified band (139 bp) was detected by electrophoresis using 1.5% agarose gel containing ethidium bromide and visualized by UV transilluminator (ProEM; Princeton Instruments, Trenton, NJ, USA).

The amplification products were then digested with the addition of 2 U BsrUI FastDigest restriction endonuclease (Thermo Fisher Scientific). The digested samples were separated by electrophoresis on a 4.0% agarose gel containing ethidium bromide. The bands were visualized using UV transillumination. Restriction fragments at 139 bp indicated the TT genotype, those at 109 bp indicated the CC genotype; those at 139 and 109 bp indicated the CT genotype.

**Calculated scores:** Aspartate aminotransferase-to-platelet ratio index (APRI): APRI was calculated using Wai’s formula (26): (AST/upper limit of normal)/(platelet count [PLT, × 10^9/L] × 100). FIB-4 score: The score was calculated using Sterling’s formula (27): (Age [yr] × AST [IU/L])/(PLT [10^9/L] × ALT [IU/L])^1/2.

**Data mining techniques:** Data collection, processing, and cleansing: A standardized enrollment questionnaire was completed by the patients’ physicians. Each feature was classified as numerical or categorical and was checked for correctness, to setup high quality data characterized by accuracy, integrity, completeness, validity, consistency, uniformity, and uniqueness.

Developing computational algorithms: Data mining analysis using WEKA implementation of reduced error pruning tree (REPtree) has been performed based on routine pre-treatment data and IL28B genotypes to build the predictive algorithm for fibrosis. WEKA is available at <http://www.cs.waikato.ac.nz/ml/weka>. WEKA aids the implementation of learning algorithms that are easily applicable to a dataset. The calculated

- ≥ 9.5 kPa for advanced fibrosis (≥ F3) and ≥ 12.5 kPa for cirrhosis (F4).

**IL28B genotyping:** Genomic DNA extraction: Five milliliter of blood was collected. Genomic DNA was extracted and purified from peripheral blood leukocytes using the Whole Blood DNA Purification Kit (Bio Vision, Milpitas, CA, USA).

Detection of IL28B rs12979860 C/T polymorphism: PCR-restriction fragment length polymorphism (RFLP) was performed as proposed by El Awady et al. (25) for the analysis of IL28B rs12979860 C/T polymorphism with minor modifications using primer pair (Thermo Fisher Scientific, Waltham, MA, USA); forward primer 5’-CGAGGGCCCCCTAACCTCTGCA-3’ and the reverse primer 5’-GGGAGCCGGATGCAATTCA-3’ (23). Enzymatic amplification was performed by PCR using Master Taq polymerase (Quanta Bio, Beverly, MA, USA) and TProfessional Standard 96-well Gradient Thermal Cycler (Biometa, Göttingen, Germany). The PCR reaction mixture (20 µL) contained 10 µL 2 × PCR Master Mix (10 × PCR buffer, 4 mM MgCl2, 0.5 µL Taq DNA polymerase, 0.4 mM dNTPs [dATP, dCTP, dGTP, dTTP]), 1 µL of each primer (25 pmol), 2 µL of genomic DNA, and 6 µL sterilized nuclelease-free water. The thermal cycler protocol was as follows: 35 cycles of denaturation (95°C for 60 s), annealing (65°C for 60 s), and polymerization (72°C for 60 s). There was a 5-min pre-incubation at 95°C before the first cycle, and a 5-min additional extension at 72°C after completion of the cycles. The amplified band (139 bp) was detected by electrophoresis using 1.5% agarose gel containing ethidium bromide and visualized by UV transilluminator (ProEM; Princeton Instruments, Trenton, NJ, USA).

The amplification products were then digested with the addition of 2 U BsrUI FastDigest restriction endonuclease (Thermo Fisher Scientific). The digested samples were separated by electrophoresis on a 4.0% agarose gel containing ethidium bromide. The bands were visualized using UV transillumination. Restriction fragments at 139 bp indicated the TT genotype, those at 109 bp indicated the CC genotype; those at 139 and 109 bp indicated the CT genotype.

**Calculated scores:** Aspartate aminotransferase-to-platelet ratio index (APRI): APRI was calculated using Wai’s formula (26): (AST/upper limit of normal)/(platelet count [PLT, × 10^9/L] × 100). FIB-4 score: The score was calculated using Sterling’s formula (27): (Age [yr] × AST [IU/L])/(PLT [10^9/L] × ALT [IU/L])^1/2.

**Data mining techniques:** Data collection, processing, and cleansing: A standardized enrollment questionnaire was completed by the patients’ physicians. Each feature was classified as numerical or categorical and was checked for correctness, to setup high quality data characterized by accuracy, integrity, completeness, validity, consistency, uniformity, and uniqueness.

Developing computational algorithms: Data mining analysis using WEKA implementation of reduced error pruning tree (REPtree) has been performed based on routine pre-treatment data and IL28B genotypes to build the predictive algorithm for fibrosis. WEKA is available at <http://www.cs.waikato.ac.nz/ml/weka>. WEKA aids the implementation of learning algorithms that are easily applicable to a dataset. The calculated

- ≥ 9.5 kPa for advanced fibrosis (≥ F3) and ≥ 12.5 kPa for cirrhosis (F4).
algorithm was validated using the k-fold cross-validation approach. Briefly, the original sample was divided into k sub-samples. This is considered to be a powerful methodology to overcome data over-fitting. The cross-validation process was repeated k times, and each of the k sub-samples was used once as the validation data.

Selecting the best computational algorithm (meta-learning): Performance of the algorithm was assessed according to an evaluation matrix based on the values for the correctly classified instance, precision (specificity), recall (sensitivity), and receiver operating characteristic (ROC) curve. We propose a system that includes open source data mining environments and services based on meta-learning approach. This system integrates meta-learning framework for ranking and selection of the best predictive algorithms for data.

Automatic selection methods for machine learning algorithms and hyper-parameter values: Auto-WEKA considers all the 39 machine learning classification algorithms and hyper-parameter values: Auto-WEKA predicts all the algorithms for data.

**RESULTS**

This study includes 427 patients with HCV-related liver fibrosis. The patients were divided into 2 groups based on the results of their FibroScan. Group 1, 204 patients (47.8%) with no, minimal, or moderate fibrosis stages (F0-F2) and Group 2, 223 patients (52.2%) with advanced fibrosis stages (F3-F4). The demographic and laboratory features of the studied groups are shown in Table 1. There was no statistically significant difference between the 2 groups in terms of their age, body mass index (BMI), liver biochemical profile, AFP, or baseline quantitative HCV PCR. Male patients represented 51.8% (221 patients) of the study population. The results of combined pegylated interferon and ribavirin therapy

<table>
<thead>
<tr>
<th>Table 1. Clinical and laboratory features of studied patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F0-F2</strong> (n = 204)</td>
</tr>
<tr>
<td>Age (yr)</td>
</tr>
<tr>
<td>BMI</td>
</tr>
<tr>
<td>ALT (IU/L)</td>
</tr>
<tr>
<td>AST (IU/L)</td>
</tr>
<tr>
<td>Albumin (g/dL)</td>
</tr>
<tr>
<td>Total bilirubin (mg/dL)</td>
</tr>
<tr>
<td>Fasting blood glucose (mg/dL)</td>
</tr>
<tr>
<td>Hemoglobin (g/dL)</td>
</tr>
<tr>
<td>WBC (× 10^3/μL)</td>
</tr>
<tr>
<td>PLT (× 10^3/μL)</td>
</tr>
<tr>
<td>AFP (ng/mL)</td>
</tr>
<tr>
<td>TSH (μU/mL)</td>
</tr>
<tr>
<td>Quantitative HCV PCR (IU)</td>
</tr>
</tbody>
</table>

BMI, body mass index; ALT, alanine aminotransferase; AST, aspartate aminotransferase; WBC, white blood cell; PLT, platelet; AFP, α-fetoprotein; TSH, thyroid stimulating hormone.

* = statistically significant.

<table>
<thead>
<tr>
<th>Table 2. <em>IL28B</em> genotype distribution within the 2 groups of fibrosis stages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fibrosis stage (No., %)</strong></td>
</tr>
<tr>
<td><strong>IL28B</strong> genotype</td>
</tr>
<tr>
<td>CC</td>
</tr>
<tr>
<td>ratio (%)</td>
</tr>
<tr>
<td>within CC genotype of the same stage</td>
</tr>
<tr>
<td>CT</td>
</tr>
<tr>
<td>ratio (%)</td>
</tr>
<tr>
<td>within CT genotype</td>
</tr>
<tr>
<td>within CT genotype of the same stage</td>
</tr>
<tr>
<td>within the total population</td>
</tr>
<tr>
<td>TT</td>
</tr>
<tr>
<td>ratio (%)</td>
</tr>
<tr>
<td>within TT genotype</td>
</tr>
<tr>
<td>within TT genotype of the same stage</td>
</tr>
</tbody>
</table>
were available for only 263 patients; non-responders represented 31.6% (83 patients), relapers were 28.9% (76 patients), and finally patients who achieved SVR were 39.5% (104 patients).

**IL28B** genotyping results revealed that only 85 (19.9%) patients of the total population carried the CC allele, and the majority of them (88.2%) had fibrosis stages between F0 and F2. CT heterozygosity was the most prevalent genotype (68.9%) in the total population. It was present in 82.5% of patients with advanced fibrosis, and 53.9% of patients between F0 and F2. Thus, among the 223 patients with advanced fibrosis, the genotype distribution was 4.5% CC, 82.5% CT, and 13.0% TT. Among the 204 patients, between early to moderate fibrosis stages, the genotype distribution was 36.8% CC, 53.9% CT, and 9.3% TT (Table 2).

With an APRI score of < 1 or ≥ 1, the sensitivity and specificity of predicting advanced fibrosis reached 0.523 and 0.831, respectively. In FIB-4 scoring system, the sensitivity and specificity reached 0.415 and 0.917 for the cut-off value < 1.45, whereas 0.459 and 0.929 for the cut-off value ≥ 3.25 (Table 3).

When applying data mining techniques, we used REPTree as the predictive algorithm that showed sensitivity of 0.749, specificity of 0.729, negative predictive value of 0.304, positive predictive value of 0.749, and ROC area of 0.796 (Table 3, Fig.1). The REPTree model revealed **IL28B** genotype as the initial splitting variable (most decisive), the second split differs according to the genotype. Patients with the favorable CC genotype have a higher probability for early fibrosis stages (F0-F2, 88.2%). For patients with the heterozygous CT genotype (the most prevalent genotype), PLT, albumin level, and age were the decisive attributes. PLT was the second important splitting attribute. According to REPTree predictions, patients with CT genotype, having PLT less than 152.5 × 10^3 μL and albumin level less than 4.2 mg/dL have 89.1% probability of advanced fibrosis. Moreover, patients with PLT above 152.5 × 10^3 μL, older than 49.5 years, and AFP less than 65 ng/mL have 62.3% probability of advanced fibrosis. For patients with TT genotype, white blood cell counts (WBC) was the decisive attribute. Patients carrying TT genotype and WBCs less than 8.63 × 10^3/μL have 68.3% probability of having advanced fibrosis.

In contrast, other attributes, such as BMI, ALT, AST, WBC, AFP, fasting blood glucose, hemoglobin level, TSH, and quantitative HCV PCR had less decisive role in the prediction of response (Fig. 1). To assess the performance of this REPTree-based model, ROC curve was generated, and area under the curve (AUC) showed that it could be used for the prediction of advanced fibrosis stage (AUC = 0.796).

By applying WEKA-implemented classifiers to choose the best classifying model, the multilayer perceptron (MLP) neural model was chosen as the best classifying model with sensitivity of 0.825, specificity of 0.811, positive predictive value of 0.825, and negative predictive value of 0.211. Regarding significant fibrosis, the MLP neural network model showed the highest AUC of 0.880, followed by the decision tree with a value of 0.796 (Table 3).

Age and PLT were considered as independent factors for advanced fibrosis by multivariate regression analysis (Table 4). Our REPTree and the MLP model were more accurate regarding the factors associated with advanced fibrosis. The presence of TT genotype was

---

**Table 3. APRI, FIB-4, decision tree, and multilayer perceptron (MLP) neural model prediction accuracy**

<table>
<thead>
<tr>
<th></th>
<th>F0-F2</th>
<th>F3-F4</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>APRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 1 (n = 145)</td>
<td>31</td>
<td>114</td>
<td>0.523</td>
<td>0.831</td>
<td>0.786</td>
<td>0.594</td>
<td>0.734</td>
</tr>
<tr>
<td>&lt; 1 (n = 256)</td>
<td>152</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIB-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1.45 (n = 94)</td>
<td>76</td>
<td>(107)</td>
<td>0.415</td>
<td>0.917</td>
<td>0.806</td>
<td>0.651</td>
<td></td>
</tr>
<tr>
<td>≥ 3.25 (n = 113)</td>
<td>13</td>
<td>(170)</td>
<td>0.459</td>
<td>0.929</td>
<td>0.885</td>
<td>0.590</td>
<td>0.799</td>
</tr>
<tr>
<td>Decision tree</td>
<td></td>
<td></td>
<td>0.749</td>
<td>0.729</td>
<td>0.749</td>
<td>0.304</td>
<td>0.796</td>
</tr>
<tr>
<td>MLP</td>
<td></td>
<td></td>
<td>0.825</td>
<td>0.811</td>
<td>0.825</td>
<td>0.211</td>
<td>0.880</td>
</tr>
</tbody>
</table>

PPV, positive predictive value; NPV, negative predictive value.

**Table 4. Multivariate analysis for predictors of advanced fibrosis**

<table>
<thead>
<tr>
<th></th>
<th>unstandardized coefficient</th>
<th>standardized coefficient</th>
<th>t</th>
<th>P-value</th>
<th>95% confidence interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>SE</td>
<td>Beta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.019</td>
<td>0.004</td>
<td>0.407</td>
<td>4.906</td>
<td>0.0001</td>
</tr>
<tr>
<td>PLT</td>
<td>-0.001</td>
<td>0.001</td>
<td>-0.178</td>
<td>-2.140</td>
<td>0.034</td>
</tr>
</tbody>
</table>

1: Not applicable.
significantly correlated with the failure of treatment response, especially early failure, as presented in the non-responder group, the CC genotype was significantly correlated with favorable response, as presented in the SVR group (P = 0.03, Table 5).

DISCUSSION

Staging of liver fibrosis is an important component in the appropriate management of chronic hepatitis C (CHC) and the prediction of the risk of hepatocarcinogenensis and patient prognosis. Patients with minimal or no liver fibrosis progress slowly over a long period of time, while those with advanced fibrosis will eventually progress to cirrhosis within 10 years, carrying the risk of liver-related complications and eventually death. Whilst there have been many non-invasive biomarkers investigated and proposed, an ideal serological test for liver fibrosis that reduces the need to perform liver biopsy is still awaited (29).

The current study included 427 Egyptian patients with CHC at different fibrosis stages who are naïve to treatment. The aim of this study was to use data mining techniques, while incorporating IL28B rs12979860 SNP, to diagnose advanced liver fibrosis by the best performing algorithm, with respect to APRI and FIB-4 scores.

The results of this study revealed IL28B rs12979860 C/T genotype as the commonest genotype (68.9%) and is in agreement with previous studies done on Egyptian patients (19,30–32), suggesting a possible role of this genotype in the transition of a normal hepatocyte into an abnormal one. This genotype distribution was suggested to be a consequence of environmental or ethnic variation (30,32).

Several studies reported that the C/C genotype is frequently associated with viral clearance, while both the C/T and T/T genotypes have less chance to spontaneously clear the virus. Egyptian patients showed higher frequencies of the C allele, as observed by Thomas et al. (33), but the frequency of distribution of the CC genotype among chronically infected patients was different across the studies: El-Bendary et al. reported 27.9%, while Abdelwahab et al. reported 31.4% (13), and Knapp et al. reported 27% (36). The frequency of CC genotype in our samples was smaller than all those mentioned above. This may be explained by the results obtained by El-Awady et al. that showed a sharp decline in the C/C genotype along the progression of chronic HCV infection towards the end-stage liver disease (31), and the majority of our patients had advanced fibrosis stages (F3-F4).

When applying our REPTree, IL28B genotype was chosen as the most decisive variable, occupying the root node, to predict advanced fibrosis stages. IL28B gene plays a role in the outcome of chronic hepatitis C (30). The SNP (rs12979860) is an etiology-independent predictor of hepatic inflammation and fibrosis (2).

IL28B CC genotype was found to be associated with enhanced immunity (increased Th1 cytokine production), with enhanced hepatic necro-inflammatory changes, higher ALT, worse clinical course, but double rates of spontaneous and treatment-associated clearance in individuals of both African and European ancestry (5). The T allele was higher in HCV-related cirrhosis than in mild hepatitis and after development of HCC (5). A meta-analysis by Sato et al. revealed that the genotype CC was significantly associated with an increased possibility of severe fibrosis and a higher possibility of severe inflammatory activity (37). On the other hand, Youssef et al. studied over 124 patients and found no association between IL28B rs12979860 polymorphism and liver fibrosis or inflammation in Egyptian patients with HCV genotype 4 (38). This difference from our results may be attributed to the larger number of patients considered in our study.

Our REPTree considered IL28B genotype, in addition to age, sex, serum albumin, WBC, PLT, and AFP as the most decisive variables. Older ages were associated with advanced fibrosis; this could be due to longer duration of viral exposure (18). The severity of hepatic fibrosis was associated with elevated serum AFP level in accordance with other studies (17,18). PLT and serum albumin level were also found to be independent predictors of fibrosis (17,39), however, multivariate analysis in our study revealed only age and PLT as the independent predictors of advanced fibrosis.

Comparing APRI and FIB-4 results with our data, the REPTree demonstrated sensitivity, specificity, and ROC better than FIB-4 and APRI. FIB-4 was first described in HCV/HIV co-infection, and later validated as an accurate marker for the diagnosis of fibrosis stages in patients with HCV mono-infection (17,40), particularly advanced fibrosis and cirrhosis. At different cut-off values the accuracy of FIB-4 approached 70% (41) with high sensitivity and moderate specificity (32). For predicting significant fibrosis, FIB-4 had an AUC value of 0.645, as reported by Alboraiie et al. (17).

Previous studies on the APRI score regarding its ability to diagnose advanced fibrosis were contradictory. This could be due to the inclusion of PLT in the score that may in turn be affected by several factors other than HCV, such as presence of portal hypertension and alcohol intake (17,42).

Decision tree is a powerful data mining tool, the root node of the tree being the most influential piece of data that affects the response variable in the model (43–46).
An alternative way to build a decision tree is to grow a multilayer feed-forward artificial neural network learning by back-propagation. It belongs to the class of supervised neural networks (41). A single previous study that used MLP to predict liver cirrhosis in patients with chronic hepatitis B reported the tree built with 7 routine clinical parameters. The model had a sensitivity of 95.2%, a specificity of 84.2%, and an overall accuracy of 89.9%. Thus it was superior to those of APRi and FIB-4 with an AUC of 0.942 (48).

In conclusion, the comparison of APRi, FIB-4, REPTree, and MLP neural prediction model revealed MLP to be significantly more accurate in predicting the stage of advanced fibrosis than the others. It could potentially be used clinically for predicting of HCV-induced advanced stages of fibrosis.

Conflict of interest None to declare.

REFERENCES


44. Song Y, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27:130-5.


