Original Paper

# Machine-Learning Monitoring System for Predicting Mortality Among Patients With Noncancer End-Stage Liver Disease: Retrospective Study

Yu-Jiun Lin[1,2], MD; Ray-Jade Chen[3,4], MD, MSc; Jui-Hsiang Tang[5], MD; Cheng-Sheng Yu[1,2*], PhD; Jenny L Wu[1,2], BSc; Li-Chuan Chen[6,7], MSN; Shy-Shin Chang[1,2,6*], MD, PhD

[1]Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[2]Department of Family Medicine, Taipei Medical University Hospital, Taipei, Taiwan

[3]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[4]Division of General Surgery, Department of Surgery, Taipei Medical University Hospital, Taipei, Taiwan

[5]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Taipei Medical University Hospital, Taipei, Taiwan

[6]Department of Community and Preventive Medicine, Taipei Medical University Hospital, Taipei, Taiwan

[7]School of Gerontology Health Management, College of Nursing, Taipei Medical University, Taipei, Taiwan

[*]these authors contributed equally

**Corresponding Author:**
Shy-Shin Chang, MD, PhD
Department of Family Medicine
School of Medicine, College of Medicine
Taipei Medical University
250 Wuxing Street
Taipei, 11031
Taiwan
Phone: 886 2 23565926
Email: sschang0529@gmail.com

## Abstract

**Background:** Patients with end-stage liver disease (ESLD) have limited treatment options and have a deteriorated quality of life with an uncertain prognosis. Early identification of ESLD patients with a poor prognosis is valuable, especially for palliative care. However, it is difficult to predict ESLD patients that require either acute care or palliative care.

**Objective:** We sought to create a machine-learning monitoring system that can predict mortality or classify ESLD patients. Several machine-learning models with visualized graphs, decision trees, ensemble learning, and clustering were assessed.

**Methods:** A retrospective cohort study was conducted using electronic medical records of patients from Wan Fang Hospital and Taipei Medical University Hospital. A total of 1214 patients from Wan Fang Hospital were used to establish a dataset for training and 689 patients from Taipei Medical University Hospital were used as a validation set.

**Results:** The overall mortality rate of patients in the training set and validation set was 28.3% (257/907) and 22.6% (145/643), respectively. In traditional clinical scoring models, prothrombin time-international normalized ratio, which was significant in the Cox regression ($P<.001$, hazard ratio 1.288), had a prominent influence on predicting mortality, and the area under the receiver operating characteristic (ROC) curve reached approximately 0.75. In supervised machine-learning models, the concordance statistic of ROC curves reached 0.852 for the random forest model and reached 0.833 for the adaptive boosting model. Blood urea nitrogen, bilirubin, and sodium were regarded as critical factors for predicting mortality. Creatinine, hemoglobin, and albumin were also significant mortality predictors. In unsupervised learning models, hierarchical clustering analysis could accurately group acute death patients and palliative care patients into different clusters from patients in the survival group.

**Conclusions:** Medical artificial intelligence has become a cutting-edge tool in clinical medicine, as it has been found to have predictive ability in several diseases. The machine-learning monitoring system developed in this study involves multifaceted analyses, which include various aspects for evaluation and diagnosis. This strength makes the clinical results more objective and reliable. Moreover, the visualized interface in this system offers more intelligible outcomes. Therefore, this machine-learning monitoring system provides a comprehensive approach for assessing patient condition, and may help to classify acute death

patients and palliative care patients. Upon further validation and improvement, the system may be used to help physicians in the management of ESLD patients.

**KEYWORDS**

visualized clustering heatmap; machine learning; ensemble learning; noncancer-related end-stage liver disease; data analysis; medical information system

## Introduction

End-stage liver disease (ESLD) is a major public health problem. It is estimated that 1 million patients died from ESLD globally in 2010, accounting for approximately 2% of all deaths [1-6]. Despite improvements in health care, mortality due to ESLD increased by 65% from 1999 to 2016 [7]. Patients with ESLD have limited treatment options and have a deteriorated quality of life with an uncertain prognosis [8]. Early identification of patients with ESLD who have a poor prognosis is fundamental for palliative care.

Several ESLD risk prediction models have been developed using traditional statistical modeling, including the Child-Pugh score [9], model for end-stage liver disease (MELD) [9,10], adjusted MELD scores (eg, MELD-Na score and integrated MELD score) [11-13], albumin-bilirubin score [14], Chronic Liver Failure Consortium (CLIF) Acute Decompensation Score [15], CLIF Sequential Organ Failure Score [16], CLIF Consortium Acute-on-Chronic Liver Failure Score [17], and a novel score recently developed by our group [18]. Unfortunately, these prediction scores were all found to have poor discrimination between survival and death [19-22]. In addition, these traditional risk scores cannot differentiate patients that need acute care or palliative care.

Machine learning, which is the use of computer algorithms that improve automatically through experience, has recently been utilized in disease diagnosis and prediction. In fact, several studies found that machine-learning models have either better or similar performances as traditional statistical modeling approaches [23-26]. Supervised machine-learning models can predict binary disease outcomes, but the prediction accuracy drops when the disease outcome involves several stages. Unsupervised machine-learning models have been successfully utilized to classify diseases that have several stages, such as chronic kidney diseases [27,28]. ESLD is a progressive disease that requires either acute or palliative care. Therefore, the goal of this study was to utilize both supervised and unsupervised machine learning to improve the care of ESLD patients. Specifically, we aimed to create a machine-learning monitoring system that combines several machine-learning models with visualized graphs, including decision trees, ensemble learning methods, and clustering, to predict the mortality of ESLD patients.

## Methods

### Study Participants and Data Collection

We conducted a retrospective cohort study using the electronic medical records (EMRs) of patients from Wan Fang Hospital and Taipei Medical University (TMU) Hospital (Figure 1). The training dataset comprised patients from Wan Fang Hospital only, whereas the validation set comprised patients from TMU Hospital. By validating our results in different settings, we tried to ensure that the models developed remained valid and robust in different hospitals.

**Figure 1.** Study flowchart depicting the series of procedures from enrollment to outcome for data collection from patients with noncancer-related end-stage liver disease. WFH: Wan Fang Hospital; TMUH: Taipei Medical University Hospital; EMR: electronic medical record; ICD: International Classification of Diseases.
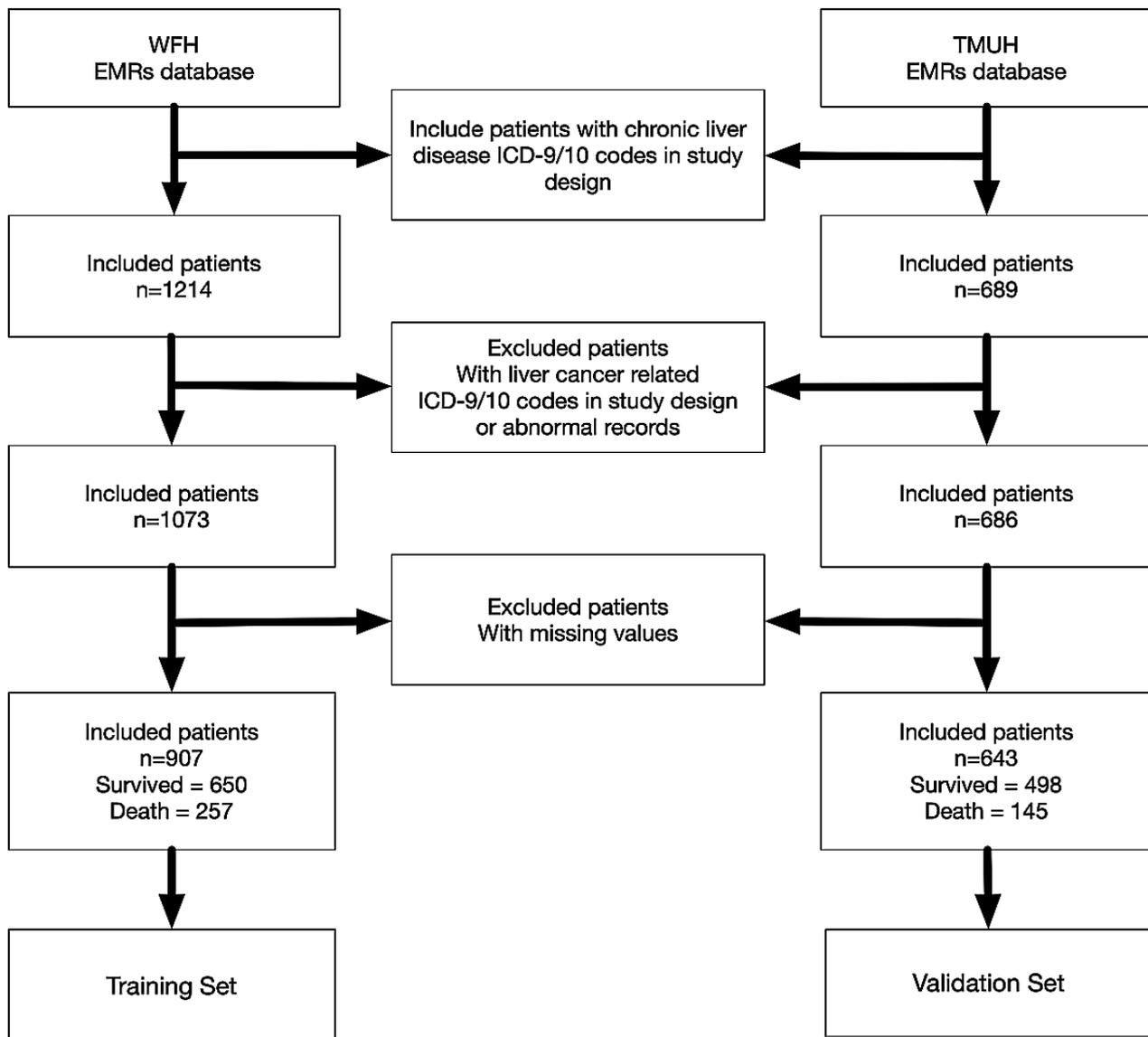


The study included all adults (aged>18 years) who were diagnosed as having chronic liver diseases with or without related complications of spontaneous bacterial peritonitis, hepatic coma, and esophageal varices (Table 1). In addition, included patients needed to have laboratory EMR data available within 24 hours of admission. Exclusion criteria included pregnancy, cancer, or had a liver transplantation.

Wan Fang Hospital and TMU Hospital are both managed by TMU. The clinical database of TMU includes the EMRs of the two hospitals. The study was approved by the TMU Institutional Review Board (approval number: N202002023) and was conducted in accordance with the Helsinki Declaration.

**Table 1.** International Classification of Diseases (ICD)-9-Clinical Modification (CM) and ICD-10-CM codes for noncancer end-stage liver disease (ESLD).

| Diseases | Included in noncancer ESLD | ICD-9-CM code | ICD-10-CM |
|---|---|---|---|
| Cirrhosis | Yes | 571.xx | K74.xx |
| Hepatic coma | Yes | 070.xx; 572.xx | K70.xx |
| Spontaneous bacterial peritonitis | Yes | 567.xx | K65.2 |
| Esophageal varices | Yes | 456.xx | I85.xx |
| Malignant neoplasm of the liver | No | 155.xx | C22.xx; Z51.12 |
| Liver transplant | No | 996.82 | Z94.4;T86.4x |

## Study Overview and Design

The aim of this study was to develop noncancerous liver disease survival prediction models using both traditional statistical modeling approaches and machine-learning approaches (Figure 2). Both supervised and unsupervised machine-learning models were investigated in parallel. For supervised machine learning, the main output was to identify the model with the best survival prediction performance via comparison of the concordance statistic (c-statistic). For unsupervised learning, the main output was the dynamic visualization of ESLD patients to aid in the palliative care of patients. Therefore, ESLD patients were classified into acute death, palliative care, and survived. Acute death was defined as death within 30 days and palliative care was defined as death within 1-9 months from the date of first admission. Mortality was defined using EMR codes related to patient death or critical illness and discharge against medical advice.

**Figure 2.** Flowchart depicting the structure of the machine learning medical system. CART: classification and regression tree; Adaboost: adaptive boosting.
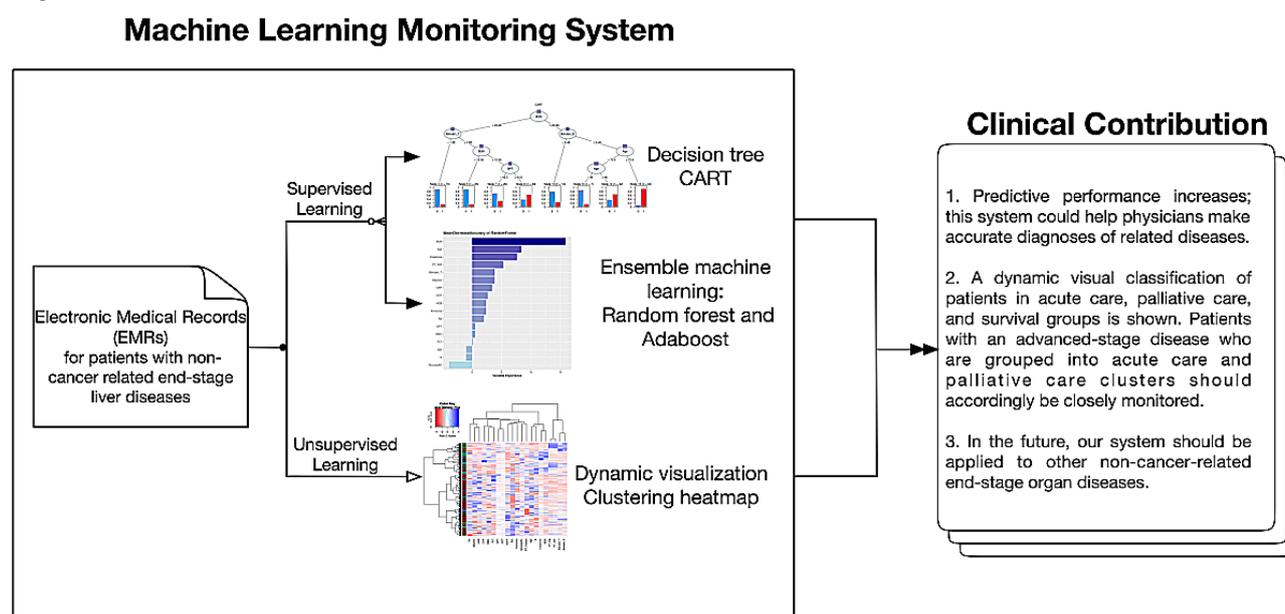


Data input was based on the literature and the physician's clinical judgment. For example, the following biochemical parameters associated with chronic liver disease were recorded: ammonia, albumin, blood urea nitrogen (BUN), complete blood count, C-reactive protein (CRP), creatinine, glutamic-pyruvic transaminase, prothrombin time (PT) and international normalized ratio (INR), glutamic-oxaloacetic transaminase, serum sodium, serum potassium, and total bilirubin.

## Statistical Analysis

Continuous variables were compared by the nonparametric Wilcoxon rank-sum test and categorical variables were compared by the chi-square test.

An initial bivariate analysis was performed to identify significant associations between mortality and all variables available in the study. Significant variables ($P<.10$) were subsequently tested in a stepwise multivariate logistic regression and stepwise Cox proportional hazards regression to identify independent predictors of mortality ($P<.05$). The final model for the stepwise regressions was selected as that with the lowest Akaike information criterion.

The validation dataset was used to compare the performances among all models. Performance was assessed according to comparison of receiver operating characteristic (ROC) curves for the different machine-learning models, including random forest with the MELD score, MELD-Na score, and our novel score [18].

All statistical analyses were performed using R (version 3.6.1) and SAS Enterprise Guide (version 7.1) software. For all analyses, $P<.05$ represented statistical significance.

## Machine-Learning Techniques

Machine learning is a statistical-based model that computer systems use to perform a task without using explicit instructions or inferences [29]. In general, machine-learning algorithms can be subdivided into either supervised or unsupervised learning algorithms. Supervised learning involves building a mathematical model of a dataset, termed training data, that contains the inputs and desired outputs known as a supervisory signal. The model is then tested using a validation set. Supervised learning algorithms involve classification and regression. The supervised machine-learning tools utilized in this study included linear discriminant analysis (LDA), support vector machine (SVM), naive Bayes classifier, decision tree, random forest, and adaptive boosting. By contrast, for unsupervised learning, a dataset is taken that contains only inputs and the structure is identified in the data, such as through grouping and clustering.

### LDA

LDA is commonly used in multivariate statistical analysis, as it can find a linear combination of features that separates two groups of objects. Hence, LDA is usually used in classification and dimensionality reduction. In this study, LDA was applied to predict the mortality of patients with chronic liver diseases using the "MASS" package in R [30].

### SVM

SVM constructs a hyperplane in a high-dimensional space for classification and regression. The ideal hyperplane will have the largest distance of margins that separates the two groups of objects. SVM is a nonprobabilistic binary classifier, as it can divide two groups of subjects and can assign new events to one group or the other [31].

### Naïve Bayes Classifiers

Naïve Bayes classifier is based on the Bayes' theorem, with an independence assumption among these features as probabilistic classifiers. Naïve Bayes can be considered a conditional probability model, which assigns a class label according to the maximum a posteriori decision rule [32].

### Decision Tree

A decision tree model is a nonparametric and effective machine-learning model. Classification and regression tree (CART) is a typical tree-based model that can predict either a continuous (regression tree) or categorical (classification tree) outcome, and visualizes the decision rule [33]. In decision tree, the Gini index (Equation 1) is used to decide the nodes on a decision branch where $p_i$ represents the relative frequency of the class that is being observed in the dataset and $c$ represents the number of classes. The process of the CART algorithm at each node for classification is as follows: (1) construct a split condition, (2) select a split condition, (3) calculate the impurity by the Gini index (Equation 1), (4) execute steps 2 to 4 until the minimum impurity is selection, and (5) construct the classification in the node.

The Gini index is calculated as:

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \qquad \textbf{(1)}$$

where $p_i$ is the probability of an object being classified to a particular class $i$.

In this study, the tree depth of CART was controlled at 4 (ie, maxdepth=4) in the R package to avoid overfitting based on a previous study [26].

### Ensemble Learning

Ensemble learning uses multiple learning algorithms to improve machine-learning results, and has generally been found to have better predictive performance than a single model. This is achieved by combining several decision classification and regression tree models [34]. Two types of ensemble learning (random forest and adaptive boosting) were used in this study.

### Random Decision Tree

Random forest, a random decision tree model, can extract the most relevant variables by performing classification, regression, or other applications based on a decision tree structure. Parallel methods were used to exploit the independence between the base learners because the error can be minimized by averaging. By creating multiple decision trees and combining the output generated by each tree, the model increases predictive power and reduces bias.

The basic single tree model in random forest is a CART using the Gini index as the selection criterion, and the random forest algorithm applies the bagging technique to implement the teamwork of numerous decision tree models, thereby improving the performance of a single model. The bagging procedure is as follows:

(1) Given a training set $X = x_1, x_2, \ldots, x_n$, with response $Y = y_1, y_2, \ldots, y_n$;

(2) For $b = 1, 2, \ldots, B$, as the repeated bagging time;

(3) Bagging select a random sample $X_b, Y_b$ with replacement of the training set;

(4) Generate a classification tree from $X_b, Y_b$;

(5) Prediction for unseen or testing samples $z$ by taking the majority vote from all of the individual classification trees.

The variable importance is determined by the decrease in node impurity, which is weighted by the probability of reaching the node. We determined the node probability by the number of samples that reached the node divided by the total number of samples. Thus, the variable becomes more significant as the value gets higher. The feature importance was implemented by Scikit-learn according to Equations (2) and (3). Assuming a binary tree, Scikit-learn calculates a node's importance using the Gini index.

$$importance\ (n_i) = w_i G_i - w_{left(i)} G_{left(i)} - w_{right(i)} G_{right\ (i)}$$
$$\textbf{(2)}$$

$$fi_j = \frac{\sum_{l:node\ l\ splits\ on\ feature\ j} importance\ (n_l)}{\sum_{k \in all\ nodes} importance\ (n_k)} \qquad \textbf{(3)}$$

Where *importance (n_i)* is the importance of node *i*, $w_i$ is the weighted number of samples reaching node *i*, $G_i$ is the impurity value of node *j*, *left(i)* is the left child node from node *i*, *right(i)* is the right child node from node *i*, and $fi_j$ is the importance of feature *j*.

The final feature importance at the random forest is the average over all CART tree models after normalization. That is, the sum of the feature's importance values on each tree is divided by the total number of trees [35]. We used the R package randomForest in this study [36].

### Adaptive Boosting

Adaptive boosting is an ensemble learning method in which base learners are generated sequentially. It is also used in conjunction with many weak learners (ie, those with poor-performance classifiers) to improve performance. Improving weak learners and creating an aggregated model to improve model accuracy is crucial for boosting algorithm performance. The output of weak learners is combined into a weighted sum that represents the final output of the boosted classifier. Adaptive boosting is adaptive because the motivation for using sequential methods is exploiting the dependence between the base learners. In addition, the predictive ability can be boosted by weighing previously mislabeled examples with a higher weight. In addition, bagging, a method that combines bootstrapping and aggregating, was used. Because the bootstrap estimate of the data distribution parameters is more accurate and robust, after combining them, a similar method can be used to obtain a classifier with superior properties [37,38]. This study used the "adabag" package for implementing adaptive boosting in R.

### ROC

We used ROC curves to compare the mortality predictive performances based on the c-statistic, which is equivalent to the area under the curve (AUC) value. The false positive rate (related to specificity) and the true positive rate (also called sensitivity or recall) were calculated for comparison.

### Heatmap and Clustering

A heatmap was used to visualize the pattern of the clinical variables. The clinical and laboratory data of patients are represented as grids of colors with hierarchical clustering analysis applied for both rows and columns [39]. Patients were separated by Euclidean distance (Equation 4) and clustered using the Ward hierarchical clustering algorithm (Equation 5). Clustering can be upgraded using different similarity measures and clustering algorithms [40]. The heatmap was constructed using the "ggplot" package in R. The Euclidean distance between points *p* and *q* is the length in multidimensional *n*-space calculated as:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad \textbf{(4)}$$

We followed the general agglomerative hierarchical clustering procedure suggested by the Ward method. The criterion for choosing a pair of clusters to merge at each step is based on the Ward minimum variance method, which can be defined and implemented recursively by a Lance–Williams algorithm [41]. The recursive formula gives the updated cluster distances following the pending merge of clusters. We used the following formula to compute the updated cluster distance:

$$d(C_i \cup C_j, C_k) = a_i d(C_i, C_k) + a_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma \mid d(C_i, C_k) - d(C_j, C_k) \mid \quad \textbf{(5)}$$

where $d(C_i, C_j)$ is the distance defined between cluster *i* and cluster *j*; thus, for each of the metrics we can compute the parameters $\alpha_i$, $\alpha_j$, $\beta$, and $\gamma$.

The Ward minimum variance method can be implemented by the Lance–Williams formula as follows:

$$d(C_i \cup C_j, C_k) = n_i + n_k / n_i + n_j + n_k d(C_i, C_k) + n_j + n_k / n_i + n_j + n_k \, d(C_j, C_k) - n_k / n_i + n_j + n_k \, d(C_i, C_j) \quad \textbf{(6)},$$

where $n_i$, $n_j$, and $n_k$ is the size of each cluster, $a_i$ is $n_i + n_k / n_i + n_j + n_k$, $a_j$ is $n_j + n_k / n_i + n_j + n_k$, $\beta$ is $- n_k / n_i + n_j + n_k$, and $\gamma$ is 0.

The "ggplot" package provides the function to apply heatmap and hierarchical clustering in R. In the function, "scale" was subject to normalization, and "RowSideColors" were set according to the death outcomes.

## Results

Figure 1 shows an overview of the study participants and Figure 2 gives an overview of the study. Initially, a total of 1214 patients from Wan Fang Hospital were used to establish a dataset for training and 689 patients from TMU Hospital were used for validation. After data preprocessing (ie, excluding cases with abnormal records and liver cancer cases), the overall mortality rate of patients in the training set at Wan Fang Hospital was 28.3% (257/907) and that at TMU Hospital was 22.6% (145/643). Table 2 and Table 3 summarize the baseline characteristics of all patients according to survival status and separated by the training and validation datasets, respectively.

**Table 2.** Demographic and laboratory characteristics of patients with noncancerous liver diseases according to survival status.

| Demographic variables[a] | Died (n=257) | Survived (n=650) | *P* value |
|---|---|---|---|
| **Sex, n (%)** | | | .25 |
| Male | 155 (60.3) | 419 (64.5) | |
| Female | 102 (39.7) | 231 (35.5) | |
| Age (years) | 69 (56-82) | 60 (50-72) | <.001 |
| Albumin (g/dL) | 2.8 (2.5-3.1) | 3.1 (2.8-3.6) | <.001 |
| Ammonia (μg/dL) | 55 (34-82) | 43 (31-68) | .003 |
| Blood urea nitrogen (mg/dL) | 29 (18-52) | 16 (12-26) | <.001 |
| Total bilirubin (mg/dL) | 1.7 (0.9-4.3) | 1.3 (0.7-2.5) | <.001 |
| Direct bilirubin (mg/dL) | 1 (0.4-2.9) | 0.5 (0.2-1.2) | <.001 |
| Creatinine (mg/dL) | 1.2 (0.8-2.2) | 0.9 (0.7-1.3) | <.001 |
| C-reactive protein (mg/dL) | 5.2 (2.8-8.9) | 4.1 (1.2-6.3) | <.001 |
| eGFR[b] (mL/min/1.73 m$^2$) | 54.6 (33.4-60.5) | 62 (57-80) | <.001 |
| Glucose ante cibum (mg/dL) | 111 (96-139) | 107 (94-139) | .31 |
| Serum GOT[c] (U/L) | 54 (32-94) | 40 (26-72) | <.001 |
| Serum GPT[d] (U/L) | 35 (22-59) | 33 (21-58) | .42 |
| Hemoglobin (g/dL) | 10 (9-11) | 12 (10-13) | <.001 |
| Potassium (mEq/L) | 4 (3.7-4.4) | 3.9 (3.7-4.2) | .001 |
| Sodium (mEq/L) | 138 (134-141) | 138 (136-139) | .43 |
| Platelets (10$^3$/μL) | 130 (86-177) | 162 (110-217) | <.001 |
| PT[e] Control (seconds) | 10.8 (10.8-12.5) | 11.8 (11.8-12.6) | <.001 |
| PT fibrinogen (seconds) | 14.8 (12.7-17.1) | 13.5 (12.2-15.1) | <.001 |
| PT international normalized ratio | 1.3 (1.13-1.54) | 1.15 (1.04-1.29) | <.001 |
| Leukocyte count (10$^3$/μL) | 8.10 (5.99-10.82) | 7.02 (5.43-9.28) | <.001 |

[a]Continuous variables are presented as median (IQR).

[b]eGFR: estimated glomerular filtration rate.

[c]GOT: glutamic-oxaloacetic transaminase.

[d]GPT: glutamic-pyruvic transaminase.

[e]PT: prothrombin time.

**Table 3.** Demographic and laboratory characteristics of patients with noncancerous liver diseases in the training and validation datasets.

| Demographic variables[a] | Training (n=907) | Validation (n=643) | P value |
|---|---|---|---|
| **Sex, n (%)** | | | .51 |
| Male | 574 (63.3) | 420 (65.3) | |
| Female | 333 (36.7) | 223 (34.7) | |
| Age (years) | 62 (52-75) | 61 (51-73) | .11 |
| Albumin (g/dL) | 3 (2.7-3.5) | 3.3 (3.1-3.7) | <.001 |
| Ammonia (μg/dL) | 48 (31-75) | 83 (49-116) | <.001 |
| Blood urea nitrogen (mg/dL) | 18 (13-33) | 16 (12-27) | .003 |
| Total bilirubin (mg/dL) | 1.4 (0.8-2.8) | 1.5 (0.7-3.1) | .77 |
| Direct bilirubin (mg/dL) | 0.6 (0.2-1.7) | 1.1 (0.5-2.7) | <.001 |
| Creatinine (mg/dL) | 0.9 (0.7-1.5) | 0.9 (0.7-1.2) | <.001 |
| C-reactive protein (mg/dL) | 4.4 (1.6-7) | 3.3 (1.3-4.9) | <.001 |
| eGFR[b] (mL/min/1.73 m$^2$) | 60.5 (47.6-73.5) | 94.1 (65.1-123.6) | <.001 |
| Glucose ante cibum (mg/dL) | 108 (94-139) | 121(104-151) | <.001 |
| Serum GOT[c] (U/L) | 43 (27-80) | 53 (34-91) | <.001 |
| Serum GPT[d] (U/L) | 34 (21-59) | 39 (25-64) | .001 |
| Hemoglobin (g/dL) | 11 (10-13) | 11 (10-13) | .50 |
| Potassium (mEq/L) | 4 (3.7-4.2) | 3.9 (3.6-4.2) | .003 |
| Sodium (mEq/L) | 138 (135-140) | 137 (135-139) | .049 |
| Platelets (10$^3$/μL) | 154 (102-209) | 138 (87-197) | <.001 |
| PT[e] control (seconds) | 11.7 (10.8-12.6) | 13.3 (13.2-13.4) | <.001 |
| PT fibrinogen (seconds) | 13.8 (12.3-15.6) | 15 (13.7-17.4) | <0.001 |
| PT international normalized ratio | 1.19 (1.05-1.37) | 1.23 (1.08-1.48) | <.001 |
| Leukocyte count (10$^3$/μL) | 7.38 (5.56-9.71) | 6.8 (5.28-8.82) | <.001 |

[a]Continuous variables are presented as median (IQR).

[b]eGFR: estimated glomerular filtration rate.

[c]GOT: glutamic-oxaloacetic transaminase.

[d]GPT: glutamic-pyruvic transaminase.

[e]PT: prothrombin time.

Table 4 shows the risk factors of mortality-based stepwise multivariate logistic and Cox regression analyses for the training dataset. PT-INR, which was significant in the Cox regression, had a prominent influence on predicting mortality. Moreover, BUN and CRP had significant effects on mortality.

**Table 4.** Significant factors in stepwise multivariate logistic and Cox regression analyses.

| Factors | P value | Odds ratio/hazard ratio[a] (95% CI) |
| --- | --- | --- |
| **Stepwise multivariate logistic regression** | | |
| Age | <.001 | 1.029 (1.017-1.042) |
| Albumin | .002 | 0.590 (0.421-0.827) |
| Blood urea nitrogen | .04 | 1.009 (1.000-1.018) |
| C-reactive protein | <.001 | 1.101 (1.056-1.147) |
| Hemoglobin | <.001 | 0.795 (0.717-0.882) |
| Sodium | .02 | 1.053 (1.007-1.101) |
| Platelets | <.001 | 0.995 (0.992-0.997) |
| Total bilirubin | <.001 | 1.149 (1.087-1.216) |
| Leukocyte count | .01 | 1.075 (1.016-1.137) |
| **Stepwise Cox regression** | | |
| Age | .03 | 1.005 (1.001-1.010) |
| Blood urea nitrogen | <.001 | 1.013 (1.009-1.018) |
| Creatinine | .002 | 0.920 (0.873-0.969) |
| C-reactive protein | <.001 | 1.027 (1.013-1.042) |
| PT[b] international normalized ratio | <.001 | 1.288 (1.131-1.468) |
| Total bilirubin | <.001 | 1.036 (1.018-1.053) |

[a]Odds ratios are reported for logistic regression and hazard ratios are reported for Cox regression.

[b]PT: prothrombin time.

Similar results were obtained using machine-learning methods. Figure 3 shows the variable of importance for random forest and adaptive boosting, which had better performances among all of the supervised machine-learning methods tested (Table 5, Figure 4). BUN was regarded as the primary factor for predicting mortality by both random forest and adaptive boosting models. Creatinine, PT-INR, and bilirubin also emerged as remarkable factors in prediction.

**Figure 3.** Variable importance ordered by the accuracy of mean decrease in random forest, adaptive boosting (AdaBoost), and AdaBoost + bootstrap. The order of variables is followed by the rank of leading variables in the random forest.
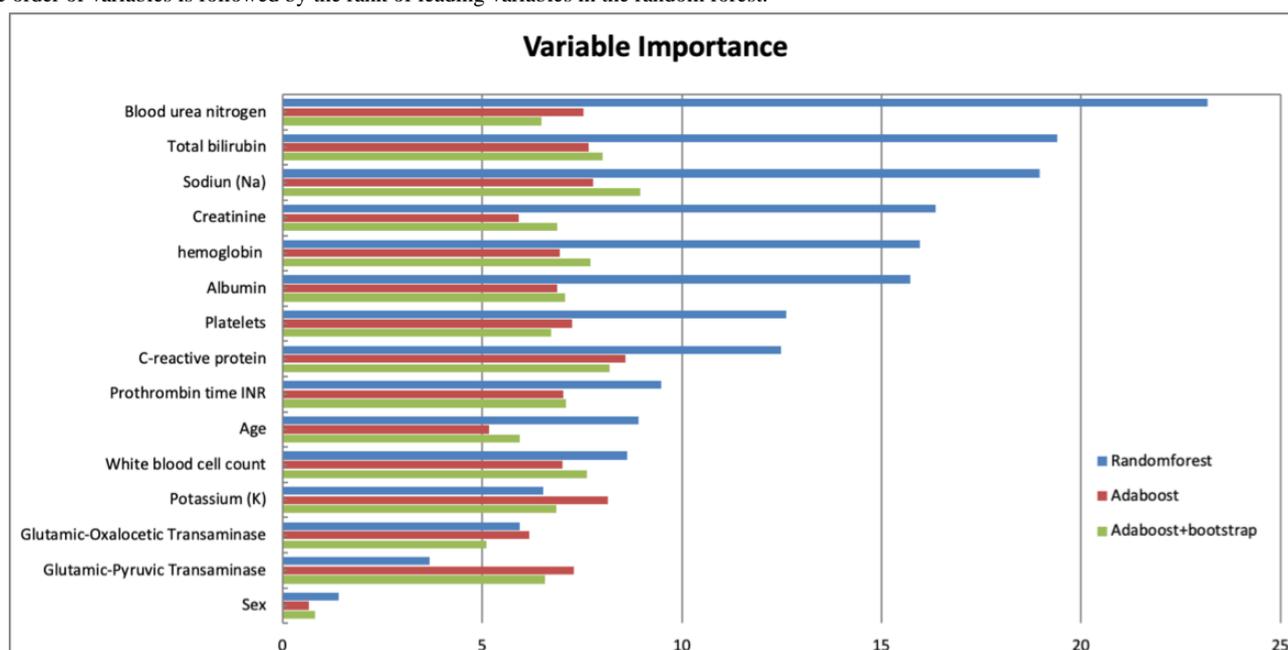
**Table 5.** Performance of different machine-learning models on predicting mortality of patients with noncancer end-stage liver diseases using the validation dataset.

| Model | Accuracy | Sensitivity | Specificity | c-statistic[a] |
|---|---|---|---|---|
| LDA[b] | 0.823 | 0.701 | 0.839 | 0.829 |
| SVM[c] | 0.818 | 0.310 | 0.966 | 0.817 |
| Naïve Bayes | 0.784 | 0.290 | 0.928 | 0.824 |
| CART[d] | 0.790 | 0.379 | 0.910 | 0.744 |
| Random Forest | 0.824 | 0.372 | 0.956 | 0.852 |
| Adaboost[e] | 0.813 | 0.455 | 0.918 | 0.833 |

[a]c-statistic: concordance statistic of the receiver operating characteristic curve.

[b]LDA: linear discriminant analysis.

[c]SVM: support vector machine.

[d]CART: classification and regression tree.

[e]AdaBoost: adaptive boosting.

**Figure 4.** Receiver operating characteristic (ROC) curves with area under the curve (AUC) statistics of classification and regression tree (CART), supervised machine learning (SVM), linear discriminant analysis (LDA), random forest, naive Bayes, and adaptive boosting (Adaboost).
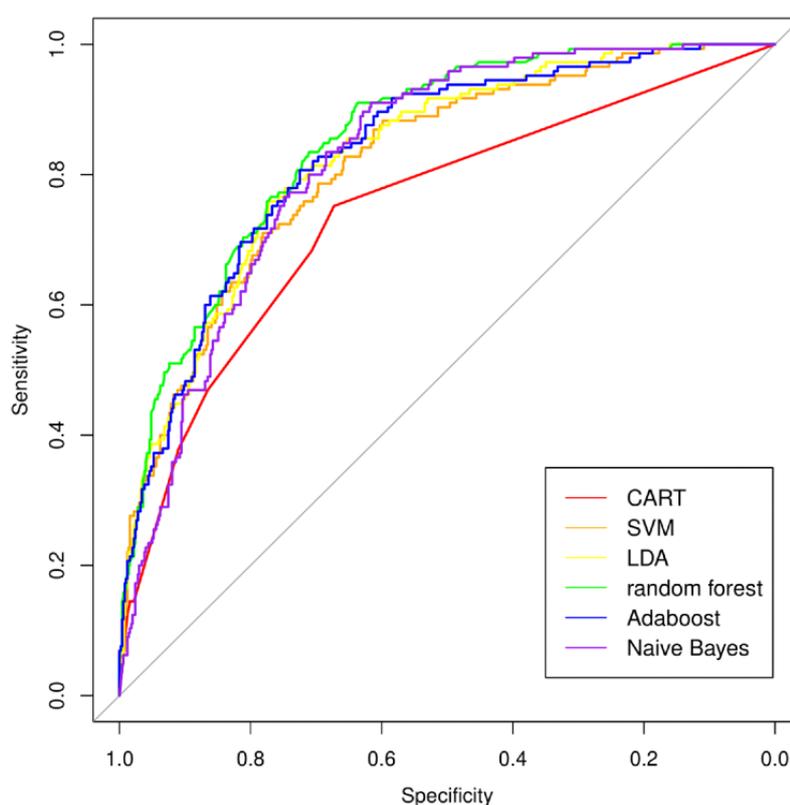


Figure 5 compares the ROC curves for mortality prediction between random forest, as the top-performing machine-learning model, with traditional risk scores. It is clear that random forest (blue curve) had better predictability than all traditional risk scores. However, there were overlaps among traditional risk scores, and it is difficult to differentiate the predictive ability of the MELD score (red, AUC=0.76), MELD-Na (orange, AUC=0.79), and novel score (green, AUC=0.75). Figure 6 shows the calibration plots for the different machine-learning models. The calibration plot is divided into 5 risk strata to match the MELD score. In general, most of the points are close to the diagonal, and the random forest model was found to be better calibrated than other machine-learning techniques. Therefore, the majority of machine-learning models showed better performance (according to the c-statistic in Table 5) than the traditional scoring models. The specificity of each machine-learning model was also above 0.80.

**Figure 5.** Receiver operator characteristic (ROC) curves with area under the curve (AUC) statistics of random forest, model for end-stage liver disease (MELD) score, MELD-NA score, and novel score.
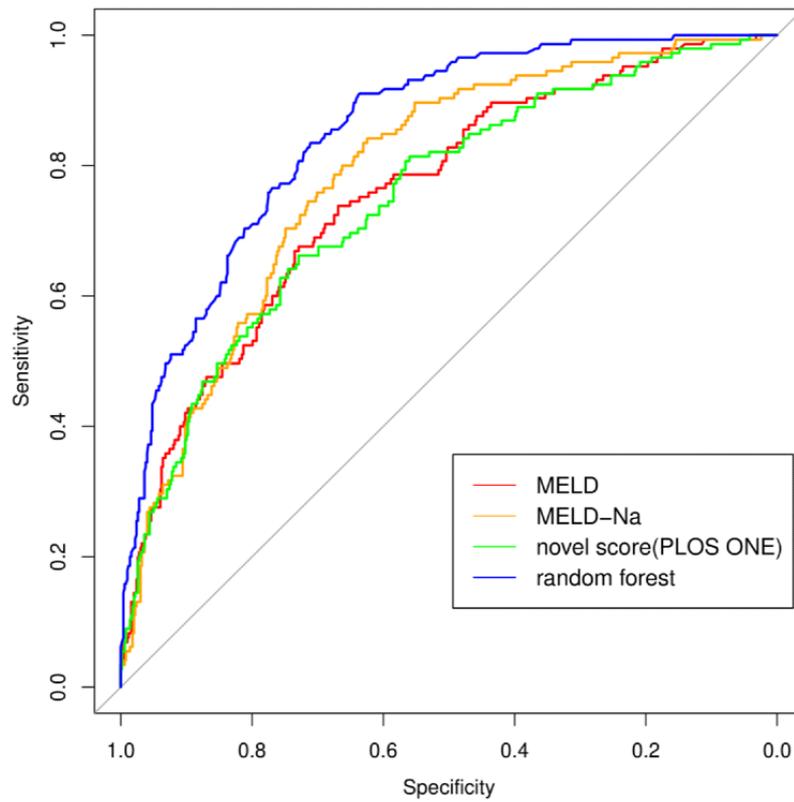


**Figure 6.** Calibration plots of classification and regression tree (CART), supervised machine learning (SVM), linear discriminant analysis (LDA), random forest, naive Bayes (NB), and adaptive boosting (Adaboost).
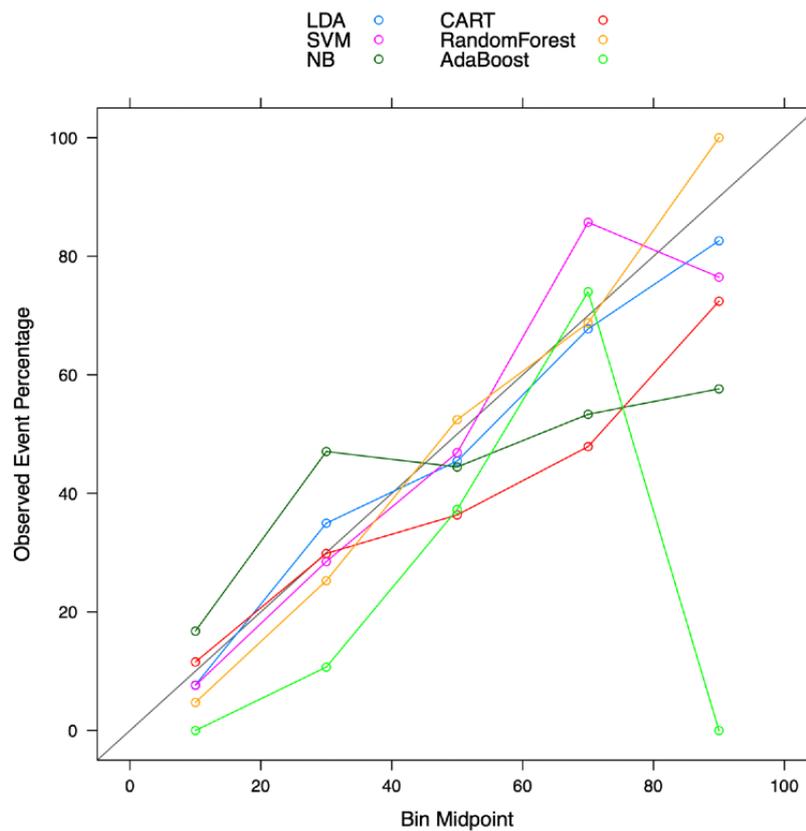
In unsupervised machine learning using the heatmap, patients were grouped into death within 30 days (red), death within 1-9 months (yellow), and survival (green) (Figure 7). We found that different clusters had specific color patterns related to laboratory outcomes.

**Figure 7.** Heatmap showing the classification of acute care (death within 30 days), palliative care (death within 1-9 months), and survival groups in Wan Fang Hospital (WFH) cohort (A and C) and Taipei Medical University Hospital (TMUH) cohort (B and D). BUN: blood urea nitrogen; Bilirubin_T: total bilirubin; CRP: C-reactive protein; eGFR: estimated glomerular filtration rate; GlucoseAC: glucose ante cibum; GOT: serum glutamic-oxaloacetic transaminase; GPT: serum glutamic-pyruvic transaminase; HGB: hemoglobin; K: potassium; Na: sodium; PLT: platelets; PT: prothrombin time; INR: international normalized ratio; WBC: leukocyte count.
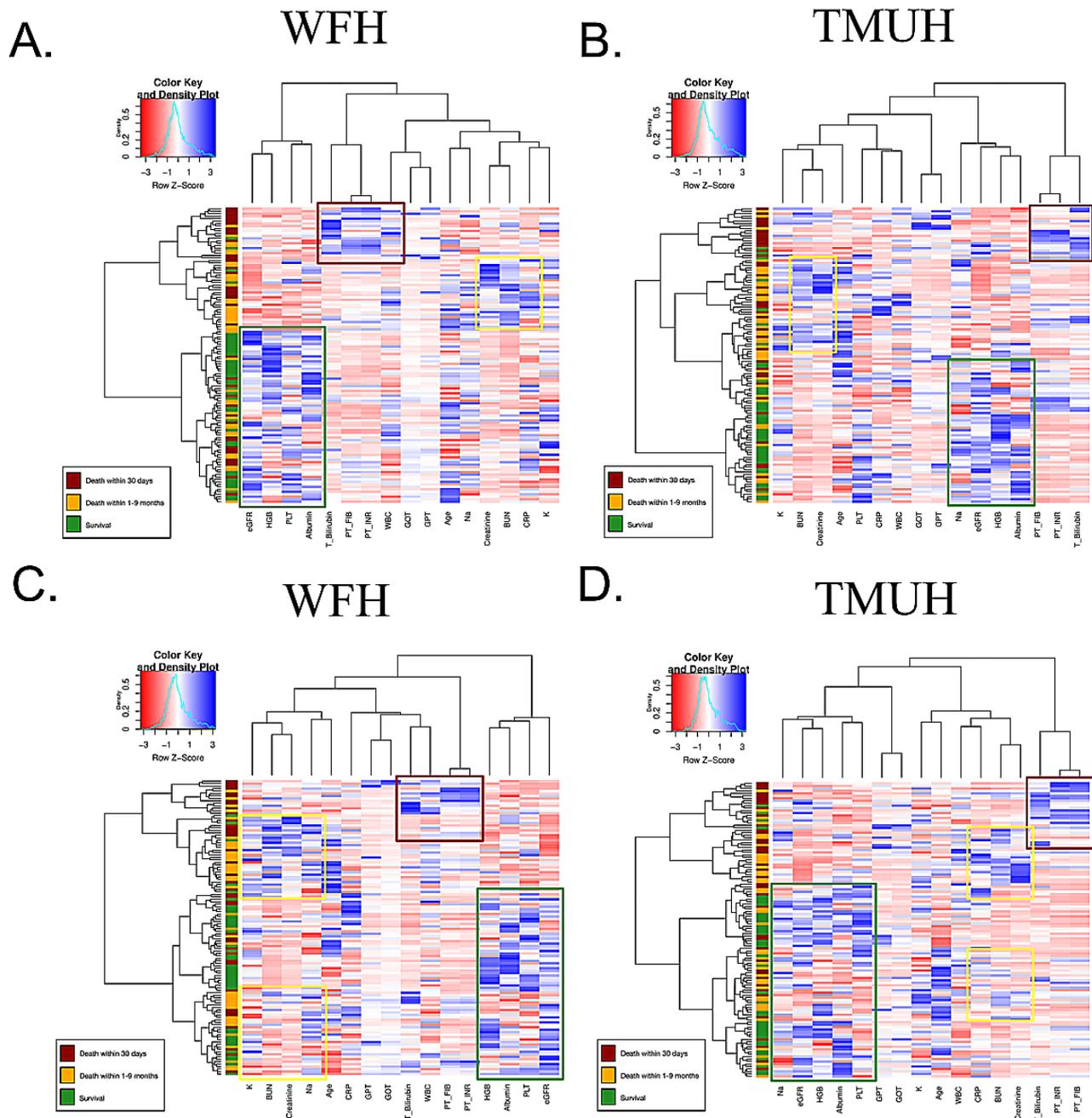


## Discussion

### Principal Findings

A major limitation in traditional statistical modeling is poor predictive ability, especially in nonhomogeneous patients representing several different disease stages. Supervised and unsupervised machine-learning methods are data-driven techniques that have been shown to have either better or similar performances as traditional statistical modeling approaches. In this study, we found that supervised ensemble learning models have better predictive performance than traditional statistical modeling. The AUC of traditional statistical modeling techniques was around 0.75, whereas that of machine-learning techniques was around 0.80. The AUC of the machine-learning technique with the best performance (random forest) was 0.85. In unsupervised learning analysis using hierarchical clustering,

ESLD patients were separated into three clusters: acute death, palliative care, and survived.

Traditional regression analysis showed that PT-INR had the highest odds ratio among all of the significant variables in predicting mortality. This is likely because critically ill patients develop hemostatic abnormalities, and PT-INR has been associated with early death among patients with sepsis-associated coagulation disorders [42]. Similar to previous studies, we also found that BUN and CRP can predict mortality in critically ill patients and for those receiving palliative care [43,44]. A prior study also found that total bilirubin is an excellent predictor of short-term (1-week) mortality in patients with chronic liver failure [45]. High bilirubin levels combined with low albumin levels may be used to predict the severity and progression of liver injury [46,47]. Hyperkalemia (high potassium) and hyponatremia (low sodium) have also been found to increase the mortality risk of ESLD patients [48,49].

In the variable of importance analysis using supervised machine-learning models, BUN was regarded as the primary factor for predicting mortality. This result is in line with a recent study showing that a high BUN concentration is robustly associated with adverse outcomes in critically ill patients, and the results remained robust after correction for renal failure [43]. Interestingly, our variable of importance analysis suggested that BUN might be a more crucial parameter for risk stratification than creatinine level in critically ill patients. We hypothesize that BUN could be an independent risk factor for renal failure, which might indicate neurohumoral activation and disturbed protein metabolism.

In the unsupervised learning analysis, ESLD patients were successfully separated into three clusters. We found that leukocyte count, PT, and bilirubin had specific and similar patterns in the acute death cluster when compared with the palliative care and survival clusters. This is likely related to the fact that these parameters are excellent predictors of short-term mortality and were therefore classified with the acute patient group [42,45]. Acute‐on‐chronic liver failure (ACLF) is one of the main causes of mortality of ESLD patients. One of the marked pathophysiological features of ACLF is excessive systemic inflammation, which is mainly manifested by a significant increase in the levels of plasma proinflammatory factors, leukocyte count, and CRP [50,51], as observed in our study.

ESLD patients with hepatorenal syndrome typically have the worst prognosis. There are two types of hepatorenal syndrome: type 1 progresses quickly to renal failure, whereas type 2 evolves slowly. Type 2 hepatorenal syndrome is typically associated with refractory ascites and the 3-month survival is 70% [52]. Although BUN, creatinine, sodium, and potassium are indicators of renal function, considering the progression of hepatorenal syndrome, the clustering heatmap classified these parameters in the palliative care group. Thus, visualization of the monitoring system using machine-learning techniques may furnish health care personnel with sufficient relevant information to manage the treatment of patients with chronic liver diseases.

## Strengths and Limitations

Medical artificial intelligence has become a cutting-edge tool in clinical medicine, as it has been found to have predictive ability in several diseases. The machine-learning monitoring system developed in this study involves multifaceted analyses, which provide various aspects for evaluation and diagnosis. This strength makes the clinical results more objective and reliable. Moreover, the visualized interface in this system offers more intelligible outcomes.

However, this study has several limitations. First, although this study enrolled thousands of ESLD patients, the numbers of ESLD patients who received palliative care or who experienced acute death were small relative to the number of ESLD patients that have survived. Including data from a larger sample of ESLD patients who received palliative care or who died from acute disease will further improve the accuracy of the machine-learning model in differentiating these three types of ESLD patients. Second, this study enrolled only patients in the Taiwanese population, and the external validity of this study with a cohort of different ethnicity remains to be tested. Third, this was a retrospective study, and a cohort study with prospectively enrolled patients is required to determine the usefulness of our system in clinical practice.

## Conclusions and Implications

Our machine-learning monitoring system provides a comprehensive approach for evaluating the condition of patients with ESLD. We found that supervised machine-learning models have better predictive performance than traditional statistical modeling, and the random forest model had the best performance of all models investigated. In addition, our unsupervised machine-learning model may help to differentiate patients that require either acute or palliative care, and may help physicians in their decision in patient treatment. In the future, it will be beneficial to apply our model to several other end-stage organ diseases without the involvement of cancer.

## Conflicts of Interest

None declared.

## References

1.   Bhala N, Aithal G, Ferguson J. How to tackle rising rates of liver disease in the UK. BMJ 2013 Feb 08;346:f807. [doi: 10.1136/bmj.f807] [Medline: 23396387]

2.   Cox-North P, Doorenbos A, Shannon SE, Scott J, Curtis JR. The Transition to End-of-Life Care in End-Stage Liver Disease. J Hosp Palliat Nurs 2013;15(4):209-215. [doi: 10.1097/njh.0b013e318289f4b0]

3.   da Rocha MC, Marinho RT, Rodrigues T. Mortality Associated with Hepatobiliary Disease in Portugal between 2006 and 2012. GE Port J Gastroenterol 2018 Apr 6;25(3):123-131 [FREE Full text] [doi: 10.1159/000484868] [Medline: 29761148]

4.   Mokdad A, Lopez A, Shahraz S, Lozano R, Mokdad A, Stanaway J, et al. Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med 2014 Sep 18;12:145 [FREE Full text] [doi: 10.1186/s12916-014-0145-y] [Medline: 25242656]

5.   Liver cirrhosis, age-standardized death rates (15+), per 100,000 population. The Global Health Observatory.: World Health Organization URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/liver-cirrhosis-age-standardized-death-rates-(15-)-per-100-000-population [accessed 2020-07-17]

6.   Eurostat (European Commission). European social statistics 2013 edition. Luxembourg: Publications Office of the European Union; 2013. URL: https://ec.europa.eu/eurostat/documents/3930297/5968986/KS-FP-13-001-EN.PDF/6952d836-7125-4ff5-a153-6ab1778bd4da

7.   Tapper EB, Parikh ND. Mortality due to cirrhosis and liver cancer in the United States, 1999-2016: observational study. BMJ 2018 Jul 18;362:k2817. [doi: 10.1136/bmj.k2817] [Medline: 30021785]

8.   Verma M, Tapper EB, Singal AG, Navarro V. Nonhospice Palliative Care Within the Treatment of End-Stage Liver Disease. Hepatology 2020 Jun 30;71(6):2149-2159. [doi: 10.1002/hep.31226] [Medline: 32167615]

9.   Peng Y, Qi X, Guo X. Child-Pugh Versus MELD Score for the Assessment of Prognosis in Liver Cirrhosis: A Systematic Review and Meta-Analysis of Observational Studies. Medicine 2016 Feb;95(8):e2877. [doi: 10.1097/MD.0000000000002877] [Medline: 26937922]

10.  Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, et al. A model to predict survival in patients with end-stage liver disease. Hepatology 2001 Feb;33(2):464-470. [doi: 10.1053/jhep.2001.22172] [Medline: 11172350]

11.  Kim WR, Biggins SW, Kremers WK, Wiesner RH, Kamath PS, Benson JT, et al. Hyponatremia and mortality among patients on the liver-transplant waiting list. N Engl J Med 2008 Sep 04;359(10):1018-1026 [FREE Full text] [doi: 10.1056/NEJMoa0801209] [Medline: 18768945]

12.  Lai JC, Covinsky KE, Dodge JL, Boscardin WJ, Segev DL, Roberts JP, et al. Development of a novel frailty index to predict mortality in patients with end-stage liver disease. Hepatology 2017 Aug;66(2):564-574 [FREE Full text] [doi: 10.1002/hep.29219] [Medline: 28422306]

13.  Luca A, Angermayr B, Bertolini G, Koenig F, Vizzini G, Ploner M, et al. An integrated MELD model including serum sodium and age improves the prediction of early mortality in patients with cirrhosis. Liver Transpl 2007 Aug;13(8):1174-1180. [doi: 10.1002/lt.21197] [Medline: 17663415]

14.  Chen RC, Cai YJ, Wu JM, Wang XD, Song M, Wang YQ, et al. Usefulness of albumin-bilirubin grade for evaluation of long-term prognosis for hepatitis B-related cirrhosis. J Viral Hepat 2017 Mar;24(3):238-245. [doi: 10.1111/jvh.12638] [Medline: 27862671]

15.  Moreau R, Jalan R, Gines P, Pavesi M, Angeli P, Cordoba J, CANONIC Study Investigators of the EASL–CLIF Consortium. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology 2013 Jun;144(7):1426-1437. [doi: 10.1053/j.gastro.2013.02.042] [Medline: 23474284]

16.  Jalan R, Pavesi M, Saliba F, Amorós A, Fernandez J, Holland-Fischer P, CANONIC Study Investigators; EASL-CLIF Consortium. The CLIF Consortium Acute Decompensation score (CLIF-C ADs) for prognosis of hospitalised cirrhotic patients without acute-on-chronic liver failure. J Hepatol 2015 Apr;62(4):831-840. [doi: 10.1016/j.jhep.2014.11.012] [Medline: 25463539]

17.  Jalan R, Saliba F, Pavesi M, Amoros A, Moreau R, Ginès P, CANONIC study investigators of the EASL-CLIF Consortium. Development and validation of a prognostic score to predict mortality in patients with acute-on-chronic liver failure. J Hepatol 2014 Nov;61(5):1038-1047. [doi: 10.1016/j.jhep.2014.06.012] [Medline: 24950482]

18.  Tsai YW, Tzeng IS, Chen YC, Hsieh TH, Chang SS. Survival prediction among patients with non-cancer-related end-stage liver disease. PLoS One 2018;13(9):e0202692 [FREE Full text] [doi: 10.1371/journal.pone.0202692] [Medline: 30240398]

19.  Kim HJ, Lee HW. Important predictor of mortality in patients with end-stage liver disease. Clin Mol Hepatol 2013 Jun;19(2):105-115 [FREE Full text] [doi: 10.3350/cmh.2013.19.2.105] [Medline: 23837134]

20.  Said A, Williams J, Holden J, Remington P, Gangnon R, Musat A, et al. Model for end stage liver disease score predicts mortality across a broad spectrum of liver disease. J Hepatol 2004 Jun;40(6):897-903. [doi: 10.1016/j.jhep.2004.02.010] [Medline: 15158328]

21.  Moraes ACOD, Oliveira PCD, Fonseca-Neto OCLD. The Impact of the Meld Score on Liver Transplant Allocation and Results: an Integrative Review. Arq Bras Cir Dig 2017;30(1):65-68. [doi: 10.1590/0102-6720201700010018] [Medline: 28489174]

22. Tsai YW, Chan YL, Chen YC, Cheng YH, Chang SS. Association of elevated blood serum high-sensitivity C-reactive protein levels and body composition with chronic kidney disease: A population-based study in Taiwan. Medicine 2018 Sep;97(36):e11896. [doi: 10.1097/MD.0000000000011896] [Medline: 30200074]

23. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]

24. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509 [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]

25. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814 [FREE Full text] [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]

26. Yu CS, Lin YJ, Lin CH, Wang ST, Lin SY, Lin SH, et al. Predicting Metabolic Syndrome With Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study. JMIR Med Inform 2020 Mar 23;8(3):e17110 [FREE Full text] [doi: 10.2196/17110] [Medline: 32202504]

27. Yu CS, Lin CH, Lin YJ, Lin SY, Wang ST, Wu JL, et al. Clustering Heatmap for Visualizing and Exploring Complex and High-dimensional Data Related to Chronic Kidney Disease. J Clin Med 2020 Feb 02;9(2):403 [FREE Full text] [doi: 10.3390/jcm9020403] [Medline: 32024311]

28. Yu CS, Lin YJ, Lin CH, Lin SY, Wu JL, Chang SS. Development of an Online Health Care Assessment for Preventive Medicine: A Machine Learning Approach. J Med Internet Res 2020 Jun 05;22(6):e18585 [FREE Full text] [doi: 10.2196/18585] [Medline: 32501272]

29. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

30. Abdi H. Discriminant correspondence analysis. In: Salkind N, editor. Encyclopedia of Measurement and Statistics. Thousand Oaks: Sage; 2007:270-275.

31. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]

32. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. Upper Saddle River: Prentice Hall; 2009.

33. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton: CRC Press; 2017.

34. Rokach L. Ensemble-based classifiers. Artif Intell Rev 2009 Nov 19;33(1-2):1-39. [doi: 10.1007/s10462-009-9124-7]

35. Breiman L. Bagging predictors. Mach Learn 1996 Aug;24(2):123-140. [doi: 10.1007/bf00058655]

36. Breiman L. Random forests. Mach Learn 2001;45:5-32. [doi: 10.1023/A:1010933404324]

37. Freund Y, Schapire R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci 1997 Aug;55(1):119-139. [doi: 10.1006/jcss.1997.1504]

38. Alfaro E, Gámez M, García N. ADABAG: An R package for classification with boosting and bagging. J Stat Soft 2013;54(2). [doi: 10.18637/jss.v054.i02]

39. Perrot A, Bourqui R, Hanusse N, Lalanne F, Auber D. Large interactive visualization of density functions on big data infrastructure. 2015 Presented at: 2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV); October 25-26, 2015; Chicago, IL. [doi: 10.1109/LDAV.2015.7348077]

40. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 1963 Mar;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]

41. Gordon AD. Classification, 2nd Edition. Boca Raton: Chapman & Hall/CRC; 2019.

42. Iba T, Arakawa M, Ohchi Y, Arai T, Sato K, Wada H, et al. Prediction of Early Death in Patients With Sepsis-Associated Coagulation Disorder Treated With Antithrombin Supplementation. Clin Appl Thromb Hemost 2018 Dec;24(9_suppl):145S-149S. [doi: 10.1177/1076029618797474] [Medline: 30198317]

43. Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig J, et al. Blood Urea Nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU. PLoS One 2018;13(1):e0191697 [FREE Full text] [doi: 10.1371/journal.pone.0191697] [Medline: 29370259]

44. Amano K, Maeda I, Morita T, Miura T, Inoue S, Ikenaga M, et al. Clinical Implications of C-Reactive Protein as a Prognostic Marker in Advanced Cancer Patients in Palliative Care Settings. J Pain Symptom Manage 2016 May;51(5):860-867 [FREE Full text] [doi: 10.1016/j.jpainsymman.2015.11.025] [Medline: 26826676]

45. López-Velázquez JA, Chávez-Tapia NC, Ponciano-Rodríguez G, Sánchez-Valle V, Caldwell SH, Uribe M, et al. Bilirubin alone as a biomarker for short-term mortality in acute-on-chronic liver failure: an important prognostic indicator. Ann Hepatol 2013;13(1):98-104 [FREE Full text] [Medline: 24378272]

46. Chen B, Lin S. Albumin-bilirubin (ALBI) score at admission predicts possible outcomes in patients with acute-on-chronic liver failure. Medicine 2017 Jun;96(24):e7142. [doi: 10.1097/MD.0000000000007142] [Medline: 28614241]

47. Ma T, Li QS, Wang Y, Wang B, Wu Z, Lv Y, et al. Value of pretransplant albumin-bilirubin score in predicting outcomes after liver transplantation. World J Gastroenterol 2019 Apr 21;25(15):1879-1889 [FREE Full text] [doi: 10.3748/wjg.v25.i15.1879] [Medline: 31057301]

48. Fernández-Esparrach G, Sánchez-Fueyo A, Ginès P, Uriz J, Quintó L, Ventura PJ, et al. A prognostic model for predicting survival in cirrhosis with ascites. J Hepatol 2001 Jan;34(1):46-52. [doi: 10.1016/s0168-8278(00)00011-8] [Medline: 11211907]

XSL•FO
RenderX

49.   Borroni G, Maggi A, Sangiovanni A, Cazzaniga M, Salerno F. Clinical relevance of hyponatraemia for the hospital outcome of cirrhotic patients. Digest Liver Dis 2000 Oct;32(7):605-610 [FREE Full text] [doi: 10.1016/S1590-8658(00)80844-0]

50.   Garcia-Martinez R, Caraceni P, Bernardi M, Gines P, Arroyo V, Jalan R. Albumin: pathophysiologic basis of its role in the treatment of cirrhosis and its complications. Hepatology 2013 Nov;58(5):1836-1846. [doi: 10.1002/hep.26338] [Medline: 23423799]

51.   Mahmud N, Kaplan DE, Taddei TH, Goldberg DS. Incidence and Mortality of Acute-on-Chronic Liver Failure Using Two Definitions in Patients with Compensated Cirrhosis. Hepatology 2019 May;69(5):2150-2163 [FREE Full text] [doi: 10.1002/hep.30494] [Medline: 30615211]

52.   de Mattos ÂZ, de Mattos AA, Méndez-Sánchez N.  Hepatorenal syndrome: Current concepts related to diagnosis and management. Ann Hepatol 2016;15(4):474-481 [FREE Full text] [Medline: 27236146]

## Abbreviations

**ACLF:** acute-on-chronic liver failure
**AUC:** area under the curve
**BUN:** blood urea nitrogen
**CART:** classification and regression tree
**CLIF:** Chronic Liver Failure Consortium
**CRP:** C-reactive protein
**c-statistic:** concordance statistic of the receiver operating characteristic curve
**EMR:** electronic medical record
**ESLD:** end-stage liver disease
**INR:** international normalized ratio
**LDA:** linear discriminant analysis
**MELD:** model for end-stage liver disease
**PT:** prothrombin time
**ROC:** receiver operating characteristic
**SVM:** support vector machine
**TMU:** Taipei Medical University

XSL•FO
RenderX