JGH Open Open Access An open access journal of gastroenterology and hepatology JGHF Journal of Gastroenterology and Hepatology Foundation

doi:10.1002/jgh3.12915

ORIGINAL ARTICLE

# Machine learning prediction model for treatment responders in patients with primary biliary cholangitis

Naruhiro Kimura, Kazuya Takahashi, Toru Setsu, Shu Goto, Suguru Miida, Nobutaka Takeda, Yuichi Kojima, Yoshihisa Arao, Kazunao Hayashi, Norihiro Sakai, Yusuke Watanabe, Hiroyuki Abe, Hiroteru Kamimura, Akira Sakamaki, Takeshi Yokoo, Kenya Kamimura, Atsunori Tsuchiya and Shuji Terai

Division of Gastroenterology and Hepatology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan

**Key words**
machine learning, primary biliary cholangitis, treatment response, ursodeoxycholic acid.

Accepted for publication 11 May 2023.

**Correspondence**
Naruhiro Kimura, Division of Gastroenterology and Hepatology, Niigata University Graduate School of Medical and Dental Sciences, 1-757 Asahimachi-dori, Niigata, Japan.
Email: naruhiro-kimura@live.jp

**Declaration of conflict of interest:** The authors have no conflicts of interest to declare.
**Author contribution:** Naruhiro Kimura, Kazuya Takahashi, and Shuji Terai drafted the manuscript. Naruhiro Kimura, Yoshihisa Arao, and Toru Setsu collected the data and created the figures. All the authors critically reviewed the manuscript, approved the final version to be published, and agreed to be accountable for all aspects of the work.
**Financial support:** This research received no external funding.

## Abstract

**Background and Aim:** Treatment response to ursodeoxycholic acid may predict the prognosis of patients with primary biliary cholangitis (PBC). Recent studies have suggested the benefits of using machine learning (ML) to forecast complex medical predictions. We aimed to predict treatment response in patients with PBC using ML and pretreatment data.
**Methods:** We conducted a single-center retrospective study and collected data from 194 patients with PBC who were followed up for at least 12 months after treatment initiation. Patient data were analyzed with five ML models, namely random forest, extreme gradient boosting (XGB), decision tree, naïve Bayes, or logistic regression, to predict treatment response using the Paris II criteria. The established models were assessed using an out-of-sample validation. The area under the curve (AUC) was used to evaluate the efficacy of each algorithm. Overall survival and liver-related deaths were analyzed using Kaplan–Meier analysis.
**Results:** Compared to logistic regression (AUC = 0.595, $P = 0.0219$, 0.031 models), ML analyses showed significantly high AUC in the random forest (AUC = 0.84) and XGB (AUC = 0.83) models; however, the AUC was not significantly high for decision tree (AUC = 0.633) or naïve Bayes (AUC = 0.584) models. Kaplan–Meier analysis showed significantly improved prognoses in patients predicted to achieve the Paris II criteria by XGB (log-rank = 0.005 and 0.007).
**Conclusion:** ML algorithms could improve treatment response prediction using pretreatment data, which could lead to better prognoses. In addition, the ML model using XGB could predict the prognosis of patients before treatment initiation.

## Introduction

Primary biliary cholangitis (PBC) is a cholestatic disease with an autoimmune pathophysiology. Guidelines recommend treatment with weight-based ursodeoxycholic acid (UDCA) as a standard first-line therapy for this chronic inflammatory liver disease.[9,11,17,23–25] Treatment response at 12 months after initial UDCA treatment is recognized as a surrogate marker for early detection of high-risk patients as per the guidelines. Well-known methods to detect good treatment response are the Paris I, Paris II, Barcelona, and Ehime criteria.[2,8,16,27] We have reported previously that the Paris II criteria may be able to predict overall survival (OS), liver-related deaths (LRD), and newly developed symptoms.[17,31] As drugs such as bezafibrate and obeticholic acid have been reported to be effective in patients with PBC,[3,10,13,18] UDCA nonresponders could be predicted in advance, and additional treatment with these drugs may improve outcomes. However, to date, there has been no method to predict UDCA response.

Recently, machine learning (ML) algorithms for treatment response have been reported in various fields, including oncology,[29,30] psychiatry,[7,15] cardiovascular diseases,[32] orthopedics,[21] and gastroenterology,[19] showing the clinical usefulness of ML. Additionally, ML has expanded into hepatology to assess liver cirrhosis based on multiple factors.[19] In the past, linear and logistic regressions have been used for forecasting. However, more complex algorithms for ML have become available in recent years, and these models can easily incorporate many variables,[22] as all calculations are performed using a computer.[1] Describing the complex and unpredictable natural progression of human disease may be improved by ML algorithms. Therefore, we hypothesized that predictive models using ML algorithms would successfully predict the treatment response in patients with PBC.

This study aimed to develop models using ML to predict which patients were likely to develop a good treatment response and to compare the results of models established using ML.

## Methods

***Participants.*** This study was reviewed and approved by the institutional review board of the University of Niigata (approval No. 2021-0385) and was conducted in accordance with the Declaration of Helsinki. The diagnosis of PBC was based on the criteria established by the Intractable Hepato-Biliary Disease Study Group of Japan.[36] Patients diagnosed between 1 January 1990 and 31 December 2020 were enrolled. Participants aged <18 years who had other chronic liver diseases or any malignancy, who were followed up for <1 year, or who lacked data were excluded from this study. Patients treated with prednisolone were excluded to avoid the potential prevalence of autoimmune hepatitis (Fig. 1). All participants in this study had been previously treated with UDCA and/or bezafibrate. Prior to and after 12 months of UDCA treatment (range: 1 January 1990–31 December 2020), blood samples were biochemically analyzed (serum white blood cells, red blood cells [RBCs], hemoglobin, hematocrit [Ht], platelet [Plt], prothrombin time [PT], total bilirubin [T-Bil], direct bilirubin [D-Bil], albumin, alkaline phosphatase [ALP], gamma-glutamyl transpeptidase, aspartate aminotransferase [AST], alanine aminotransferase [ALT], blood urea nitrogen [BUN], creatinine [Cre], immunoglobulin G [IgG], Ig A, IgM, and anti-mitochondria [AMA] M2 antibody). After data collection, information that could identify individual participants was deleted.

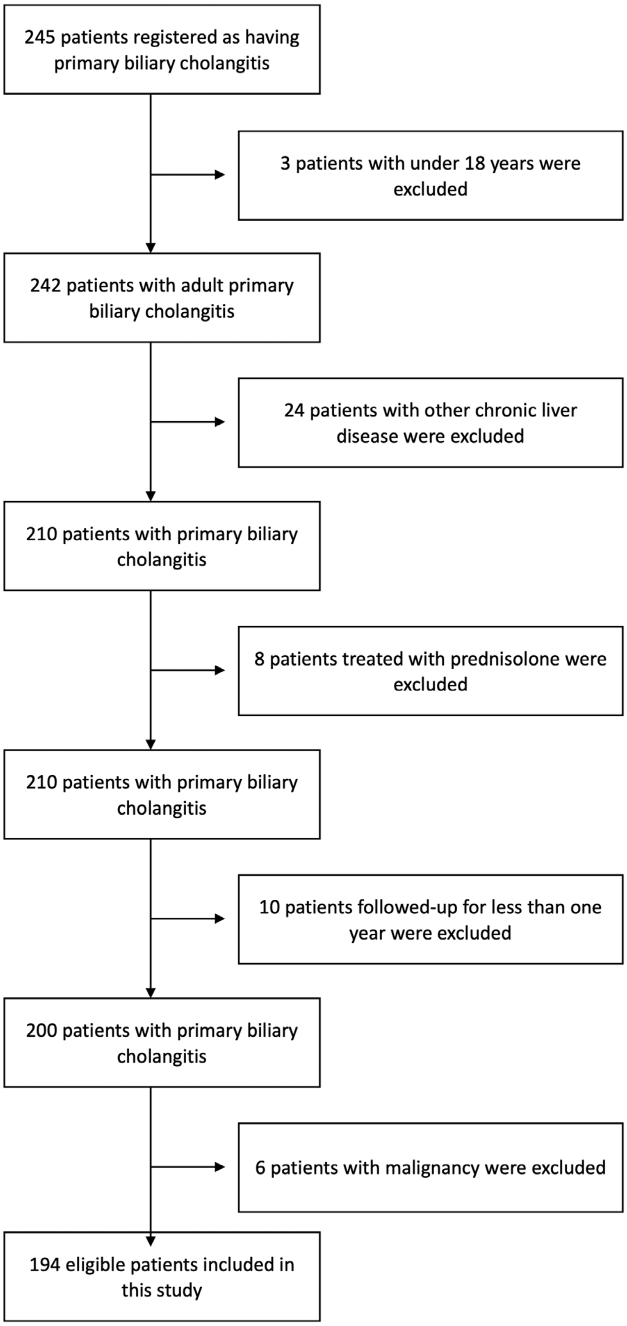
245 patients registered as having primary biliary cholangitis
3 patients with under 18 years were excluded
242 patients with adult primary biliary cholangitis
24 patients with other chronic liver disease were excluded
210 patients with primary biliary cholangitis
8 patients treated with prednisolone were excluded
210 patients with primary biliary cholangitis
10 patients followed-up for less than one year were excluded
200 patients with primary biliary cholangitis
6 patients with malignancy were excluded
194 eligible patients included in this study

**Figure 1** Strategy for inclusion and exclusion of patients and their data.

***Statistical analysis.*** The Paris II criteria were set as the primary analysis in this study. Categorical data are expressed as numbers and percentages, and continuous data are expressed as means ± standard deviations (SDs). Clinical differences between patients divided according to the Paris II criteria were analyzed using either the Mann–Whitney U test or the Chi-square test. Results of the receiver operating characteristic (ROC) and area under the curve (AUC) analyses were compared using the DeLong test. The accumulation ratios of OS and LRD were estimated using the Kaplan–Meier methods. Cox proportional hazard models were used for detecting the predictors of OS and LRD. Statistical significance was set at $P < 0.05$. EZR (Saitama Medical Center, Jichi Medical University, Saitama, Japan), a graphical user interface for R (The R Foundation for Statistical Computing, Vienna, Austria), was used for statistical analysis.[14]

The Paris II criteria define a decrease in ALP level ≤1.5 × normal limits, AST level ≤1.5 × normal limits, or bilirubin level <1 mg/dL after 1 year of treatment.[27]

The GLOBE and UK-PBC scores were calculated as in previous reports.[6,20]

***Supervised predictive model.*** During the ML process, Python (version 3.7.12) and *scikit-learn* (version 1.0.2) were used to develop supervised predictive models.[28] The median values were used to generate the final estimates for clinically important candidate variables with missing data, and continuous variables were standardized. Training and test sets were obtained by randomly splitting the data into subsets of 70 and 30% data, respectively.

We developed five ML models to predict patients who would achieve the Paris II criteria with data collected before treatment. Random forest (RF), extreme gradient boosting tree (XGB) classifier,[28] decision tree (DT), naïve Bayes (NB), and logistic regression (LR) were used. Tuning hyperparameters could be a way to avoid overfitting which is a major problem in ML. The hyperparameters in these ML methods were tuned by grid search for high predictive accuracy using the training dataset. The hyperparameters in each algorithm were as follows: n_estimators, max_features, and max_depth were used in RF; colsample_bytree, min_child_weight, max_depth, and n_estimators in XGB, and model_depth were used in DT. No hyperparameter was used in naïve Bayes. The least absolute shrinkage and selection operator was used in the LR analysis. The feature importance in the XGB and RF models was calculated based on the Gini impurity to evaluate the contribution of each feature to the models. The scale_pos_weigh method was adopted to balance the positive and negative weights. Owing to the imbalance in our dataset, 75.5% of patients met the Paris II criteria, and these parameters were tuned to three.

The performance of each model was evaluated and compared with that of the test dataset in terms of sensitivity, specificity, and accuracy. In addition, the ROC curve and AUC were used to evaluate predictive accuracy. Calibration in AUC curve was measured using calibration_curve and brier_score_loss analyses.

Kaplan–Meier analysis for OS and LRD was performed to evaluate the usefulness of ML prediction using test samples.

## Results

***Baseline characteristics and outcomes.*** Data from 194 patients with PBC were analyzed in this study (Table 1). The median age ± SD at the time of diagnosis was 64 ± 12 years, with 86.1% of patients being female (167/194). A total of 43% of patients had other autoimmune diseases, including Sjogren syndrome (17.4%), chronic thyroiditis (6.7%), and rheumatoid arthritis (3.3%). Regarding treatment, 80.9% of patients (157/194) received UDCA and 8% (16/194) received bezafibrate. The percentage of patients who met the Paris II criteria at the end of the follow-up period was 77.3% (151/194) and the median follow-up period was 2480 ± 3103 days.

We compared patients who met the Paris II criteria with those who did not (Table 2). Those complying with the Paris II criteria were significantly older (64 ± 12 *vs* 58 ± 11 years, $P = 0.029$) and had higher serum prothrombin time international normalized ratio (PT INR) (1.05 ± 0.03 *vs* 1.03 ± 0.04, $P = 0.011$), lower serum AST (38 ± 38 *vs* 42 ± 65 IU/L, $P = 0.039$), lower serum ALT (32 ± 52 *vs* 53 ± 78 IU/L, $P = 0.021$), lower T-Bil (0.6 ± 0.2 *vs* 1.0 ± 0.2 g/dL, $P < 0.001$), lower D-Bil (0.1 ± 0.1 *vs* 0.1 ± 0.2 g/dL, $P = 0.035$), higher BUN (14 ± 4.7 *vs* 12 ± 4.3 mg/dL, $P = 0.028$), and higher Cre levels (0.63 ± 0.24

**Table 1** Characteristics of patients

| Parameter | Median ± SD |
|---|---|
| Age at diagnosis (years) (*N* = 194) | 64 ± 12 |
| Sex (M/F) | 27/167 |
| Presence of other autoimmune diseases, *n* (%) (*N* = 194) | 84 (43.3%) |
| Histology, *n* (%) (*N* = 194) | 120 (61.9%) |
| Height (cm) (*N* = 194) | 155 ± 13 |
| Weight (kg) (*N* = 194) | 52.4 ± 10.0 |
| Treatment | |
| UDCA (mg/kg/day) (*N* = 194) | 9.9 ± 5.4 |
| Bezafibrate, *n* (%) (*N* = 194) | 16 (8%) |
| Lab data | |
| Plt ($\times 10^4$/μL) (*N* = 194) | 20.4 ± 8.3 |
| RBC ($\times 10^4$/μL) (*N* = 194) | 428 ± 46 |
| Hb (g/dL) (*N* = 194) | 13.0 ± 1.8 |
| Hct (%) (*N* = 194) | 39.0 ± 4.6 |
| PT INR (*N* = 102) | 1.03 ± 0.04 |
| TP (g/dL) (*N* = 192) | 8.0 ± 0.7 |
| Alb (g/dL) (*N* = 192) | 4.0 ± 0.5 |
| AST (U/L) (*N* = 194) | 39 ± 47 |
| ALT (U/L) (*N* = 194) | 38 ± 61 |
| LDH (U/L) (*N* = 192) | 205 ± 107 |
| ALP ISCC (U/L) (*N* = 194) | 139 ± 138 |
| γGTP (U/L) (*N* = 192) | 142 ± 226 |
| T-Bil (mg/dL) (*N* = 194) | 0.7 ± 0.4 |
| D-Bil (mg/dL) (*N* = 190) | 0.1 ± 0.2 |
| BUN (mg/dL) (*N* = 187) | 13.0 ± 4.6 |
| Cre (mg/dL) (*N* = 187) | 0.6 ± 0.2 |
| IgM (mg/dL) (*N* = 192) | 279 ± 300 |
| Treatment response | |
| Paris II criteria, *n* (%) (*N* = 194) | 151 (77.3%) |

Alb, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; D-Bil, direct bilirubin; F, female; IgM, immunoglobulin M; M, male; Plt, platelet count; PT INR, prothrombin time international normalized ratio; T-Bil, total bilirubin; UDCA, ursodeoxycholic acid; γGTP, γ-glutamyl transferase.

**Table 2** Characteristics of patients with/without the Paris II criteria

| Parameter | With Paris II, *N* = 151 | Without Paris II, *N* = 43 | *P*-value |
|---|---|---|---|
| Age at diagnosis (years) | 64 ± 12 | 58 ± 11 | 0.029 |
| Sex (M/F) | 23/128 | 4/39 | 0.679 |
| Presence of other autoimmune disease (*n*/%) | 62/41.1% | 22/51.2% | 0.637 |
| Histology (*n*/%) | 87/57.6% | 33/76.7% | 0.095 |
| Height (cm) | 155 ± 8 | 154 ± 8 | 0.948 |
| Weight (kg) | 52.0 ± 9.3 | 53.7 ± 9.6 | 0.371 |
| Treatment | | | |
| UDCA (mg/kg/day) | 10.1 ± 5.4 | 9.5 ± 5.3 | 0.339 |
| Bezafibrate (*n*/%) | 10/6.6% | 6/13.9% | 0.227 |
| Lab data | | | |
| Plt ($\times 10^4$/μL) | 20.9 ± 8.2 | 19.5 ± 8.7 | 0.675 |
| RBC ($\times 10^4$/μL) | 428 ± 44.9 | 428 ± 52.4 | 0.849 |
| Hb (g/dL) | 13.0 ± 1.8 | 13.0 ± 2.0 | 0.983 |
| Ht (%) | 39.0 ± 4.7 | 39.0 ± 4.4 | 0.845 |
| PT INR | 1.05 ± 0.03 | 1.03 ± 0.04 | 0.011 |
| TP (g/dL) | 8.0 ± 0.7 | 8.0 ± 0.6 | 0.272 |
| Alb (g/dL) | 4.0 ± 0.5 | 0.4 ± 0.6 | 0.089 |
| AST (U/L) | 38 ± 38 | 42 ± 65 | 0.039 |
| ALT (U/L) | 32 ± 52 | 53 ± 78 | 0.021 |
| LDH (U/L) | 198 ± 107 | 211 ± 106 | 0.653 |
| ALP ISCC (U/L) | 133 ± 132 | 154 ± 154 | 0.136 |
| γGTP (U/L) | 124 ± 218 | 189 ± 247 | 0.224 |
| T-Bil (mg/dL) | 0.6 ± 0.2 | 1.0 ± 0.5 | <0.001 |
| D-Bil (mg/dL) | 0.1 ± 0.1 | 0.1 ± 0.2 | 0.035 |
| BUN (mg/dL) | 14 ± 4.7 | 12 ± 4.3 | 0.028 |
| Cre (mg/dL) | 0.63 ± 0.24 | 0.60 ± 0.24 | 0.027 |
| IgM (mg/dL) | 279 ± 310 | 261 ± 270 | 0.402 |

Abbreviations: Alb, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; D-Bil, direct bilirubin; F, female; IgM, immunoglobulin M; M, male; Plt, platelet count; PT INR, prothrombin time international normalized ratio; T-Bil, total bilirubin; UDCA, ursodeoxycholic acid; γGTP, γ-glutamyl transferase.

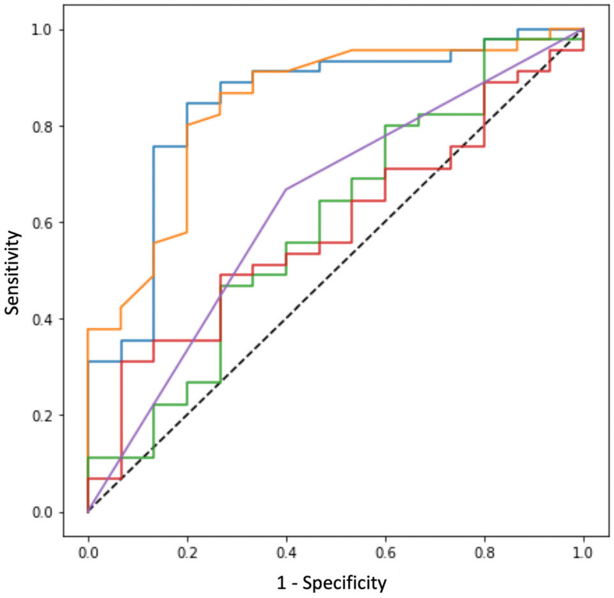

**Figure 2** Receiver operating characteristic (ROC) curve analysis. Area under the curve (AUC) for predicting patients meeting the Paris II criteria using extreme gradient boosting tree (XGB, AUC = 0.830), random forest (RF, AUC = 0.842), logistic regression (LR, AUC = 0.595), Naïve Bayes (NB, AUC = 0.584), and decision tree (DT, AUC = 0.633). 🟦, XGB 0.830; 🟧, RF 0.842; 🟩, LR 0.595; 🟥, NB 0.584; 🟪, DT 0.633.

*vs* 0.60 ± 0.24 mg/dL, $P = 0.027$) compared to patients who did not achieve the Paris II criteria.

***Performance of the predictive models.*** ROC analysis results with AUC of RF, XGB, DT, NB, and LR models for prediction of patients matching the Paris II criteria are shown in Figure 2 and Table 3. RF (AUC: 0.842, 95% confidence interval [CI]: 0.709–0.975) and XGB (AUC: 0.830, 95% CI: 0.702–0.958) had significantly higher AUC than LR (AUC: 0.595, 95% CI: 0.419–0.779, $P = 0.022$ and 0.031, respectively; Table 3). However, the AUCs of DT (AUC: 0.633, 95% CI: 0.487–0.779) and NB (AUC: 0.584, 95% CI: 0.425–0.744) were not significantly different from those of LR ($P = 0.714$ and 0.878, respectively). Calibratiob_curve in both RF and XGB was satisfactory (Fig. S1A,B) and brier_score_loss showed sufficiently low score in both RF (0.154) and XGB (0.130) (Table S1), suggesting that both ML algorisms were good enough for practical use.

***Influence of variables on prediction.*** Feature importance in XGB and RF is shown in Figure 3, following the significance found in the ROC and AUC analysis when compared to LR. The serum ALP level, serum RBC level, serum LDH level, serum AMA M2 index, and serum Ht level were the top five factors for predicting the outcome in RF (Fig. 3a). The serum Plt level, serum Cre level, UDCA dose (mg/kg/day), serum ALP level, and use of fibrate were the top five factors for predicting the Paris II criteria in XGB (Fig. 3b).

***Paris II criteria prediction using XGB could forecast the outcome of patients with PBC.*** To investigate the usefulness of ML algorithms in predicting the outcome of patients with PBC, Kaplan–Meier analysis was performed using test samples ($N = 60$) (Fig. 4). The analysis showed that patients with PBC estimated to match the Paris II criteria with XGB showed significantly better prognoses for both OS (Fig. 4a right) and LRD (Fig. 4b right) (log-rank = 0.005 and 0.007, respectively). Other ML algorithms including RF (Fig. 4a left, b left), DT, and NB did not show significantly different results (data not shown). To compare the efficacy of OS prediction, Cox hazard analysis was performed. The data showed that only the GLOBE score was significantly associated with OS prediction (hazard ratio [HR]: 1.437, 95% CI: 1.006–2.054, $P = 0.047$). XGB had high HR (6.253, 95% CI: 0.831–47.028) but did not reach significance ($P = 0.075$). In terms of LRD, univariate analysis showed that the GLOBE score (HR: 1.534, 95% CI: 1.053–2.235, $P = 0.026$) and XGB (HR: 15.783, 95% CI: 1.427–174.546, $P = 0.024$) were significantly associated with OS prediction, but multivariate analysis showed that neither factor reached significance.

## Discussion

In this study, we were able to use an ML method to predict which patients with PBC would respond well to UDCA treatment. Furthermore, ROC analysis with ML using pretreatment data showed a high AUC. Recent reports have suggested that the response rates to UDCA treatment are 48–71%,[4,25,34,35] which was nearly the same as that observed in our study. From these data, we consider our results to be in line with real-world reports.

As patients who met the Paris II criteria showed significantly better liver function and liver functional reserve at the time of diagnosis, it is understandable that these patients could

**Table 3** Comparison of prediction performance for biochemical response

| | Sensitivity, (95% CI) | Specificity, (95% CI) | Accuracy, (95% CI) | AUC, (95% CI) | *P*-value against LR |
|---|---|---|---|---|---|
| RF | 0.833 (0.516–0.979) | 0.896 (0.773–0.965) | 0.883 (0.774–0.952) | 0.842 (0.709–0.975) | 0.0219 |
| XGB | 0.667 (0.223–0.957) | 0.796 (0.665–0.894) | 0.783 (0.658–0.879) | 0.830 (0.702–0.958) | 0.0305 |
| NB | 0.250 (0.006–0.806) | 0.750 (0.616–0.856) | 0.717 (0.586–0.825) | 0.584 (0.425–0.744) | 0.878 |
| DT | 0.375 (0.188–0.594) | 0.833 (0.672–0.936) | 0.650 (0.516–0.769) | 0.633 (0.487–0.779) | 0.714 |
| LR | 0.375 (0.188–0.594) | 0.833 (0.672–0.936) | 0.650 (0.516–0.769) | 0.595 (0.419–0.770) | |

AUC, area under the curve; CI, confidence interval; DT, decision tree; LR, logistic regression; NB, naïve Bayes; RF, random forest; XGB, extreme gradient boosting tree.

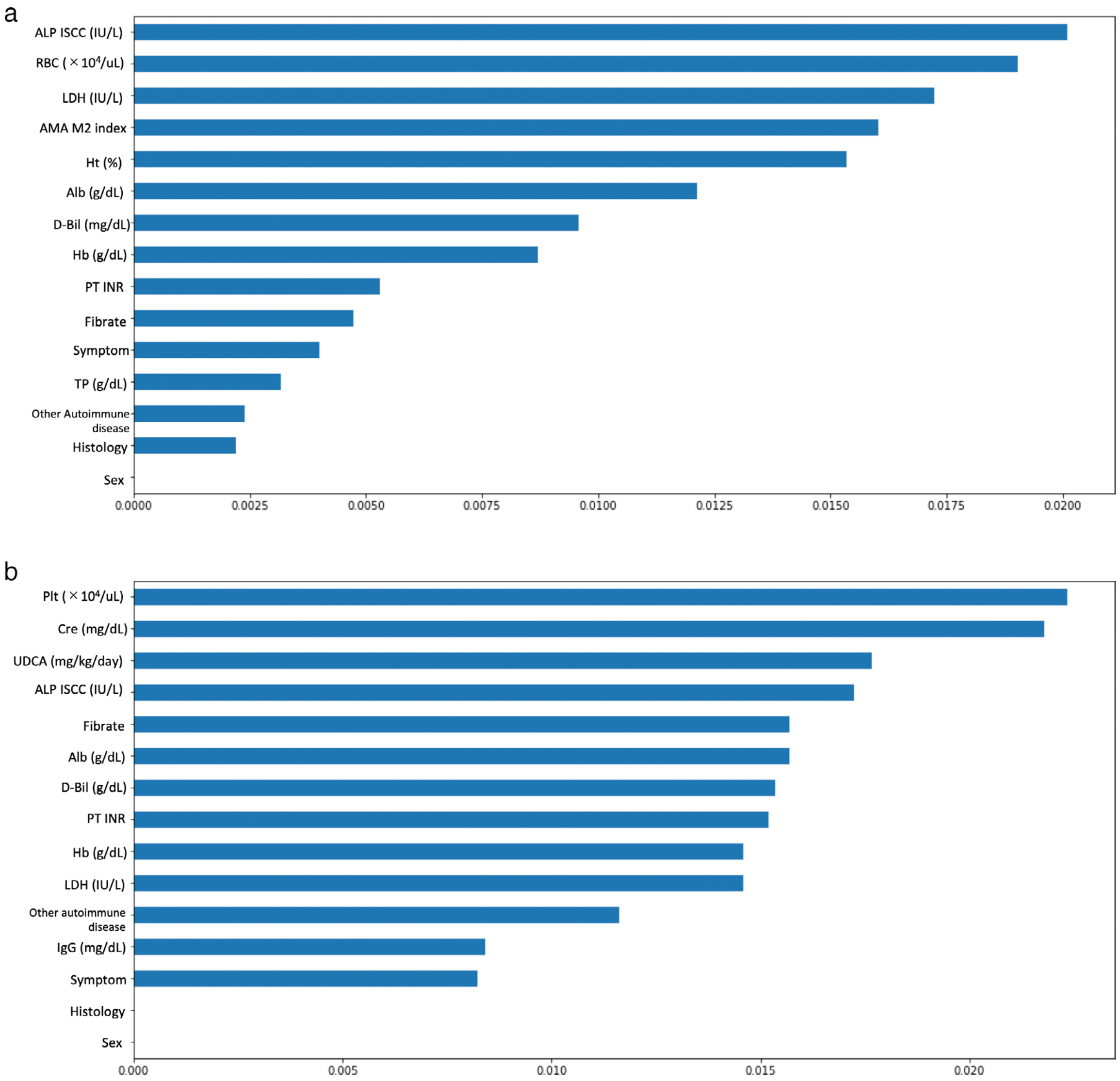

**Figure 3** Feature importance derived from random forest and extreme gradient boosting tree. Feature importance is displayed from the top for random forest (a) and extreme gradient boosting tree (b). Alb, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AMA M2, antimitochondrial M2 antibody; AST, aspartate aminotransferase; BUN, blood urea nitrogen; Cre, creatinine; D-Bil, direct bilirubin; HCC, hepatocellular carcinoma; Ht, hematocrit; IgM, immunoglobulin M; LDH, lactate dehydrogenase; Other AID, other autoimmune diseases; Plt, platelet count; PT%, prothrombin time %; RBC, red blood cell count; T-Bil, total bilirubin; T-Cho, total cholesterol; TP, total protein; UDCA, ursodeoxycholic acid; γ-GTP, γ-glutamyl transpeptidase.

achieve the Paris II criteria. In general, it is challenging to predict treatment response in patients with PBC, and recent studies have shown AUCs of 0.79–0.83 for predicted treatment response using a logistic regression model.[5,33] As these studies aimed to predict the Barcelona, Paris I, or Toronto criteria, it is difficult to compare our results with theirs. However, previous reports have shown that Paris II criteria could predict OS, LRD, and newly developed symptoms in patients with PBC.[16,17] From these data, we deduced that our prediction model could be more useful than previous prediction models. In addition, we compared the results of ML analysis with those of a prognosis prediction model. The GLOBE score seemed to be the best for OS and LRD prediction, but the XGB model also estimates patient prognosis with high HR. Furthermore, XGB used only pretreatment data, and this could be a strong indication to use ML. Recently, a novel approach with ML was reported in patients with PBC.[12] This study established an ML model that developed four clusters of patients depending on serum Alb, T-Bil, and ALP levels, suggesting the convenience and usefulness of ML in PBC cases.

Many factors were found to influence the treatment response predictions in this study. To summarize, there were three groups with important features. First, liver enzyme-related factors, including the serum AST, ALT, γ-GTP, ALP, and bilirubin levels, reportedly predict good treatment response.[26] Further,

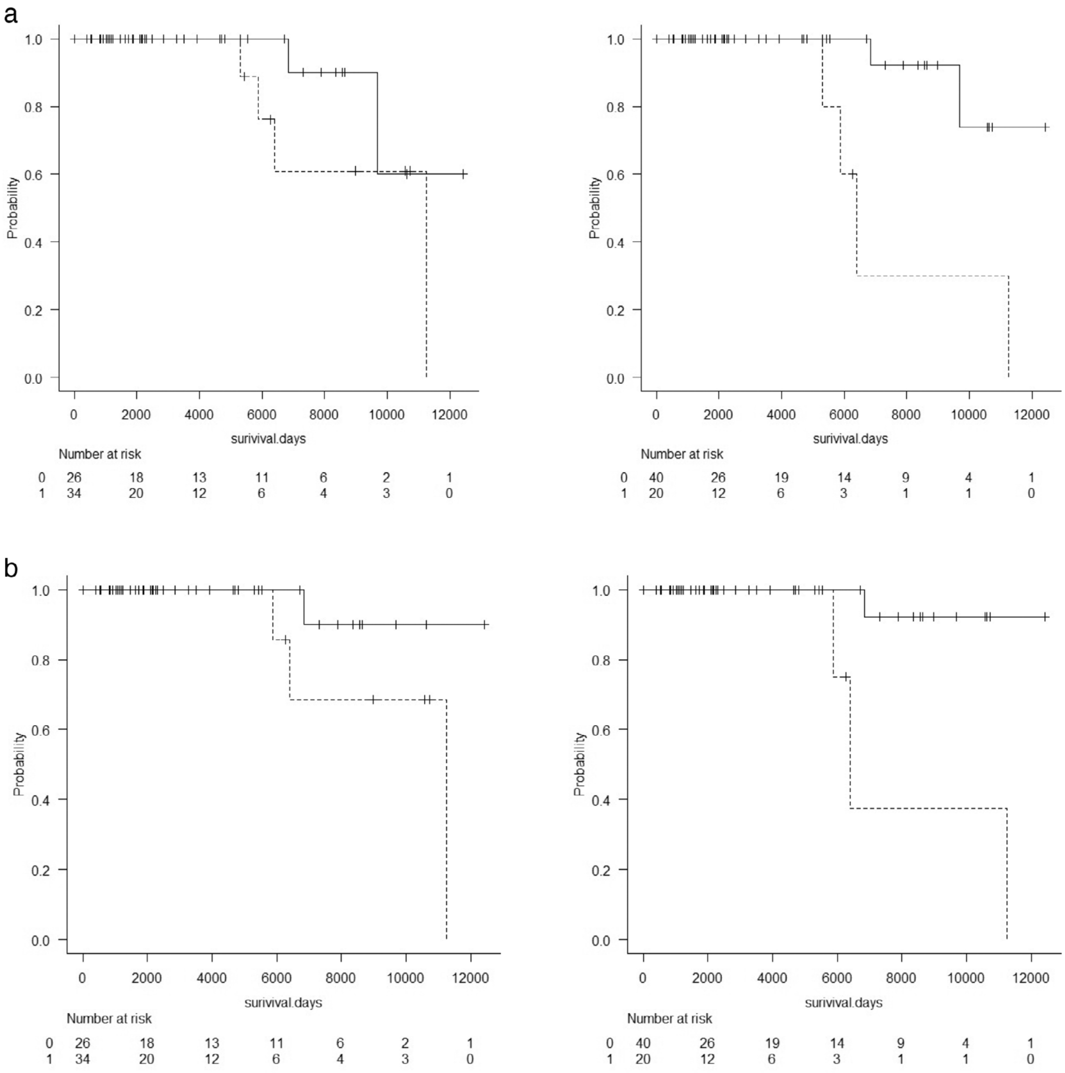

**Figure 4** Kaplan–Meier analysis. The analysis shows that patients who could achieve the Paris II criteria predicted by extreme gradient boosting tree have significantly better prognosis in both overall survival (log-rank = 0.005) (a, right) and liver-related death (log-rank = 0.007) (b, right). Prediction model of Random Forest (a left, and b left) did not show significance. ——, achieve Paris II criteria; ----, not achieve Paris II criteria. Long rank = 01.75.

some of these terms were even included in the Paris II criteria. Second, immune-related data, such as the serum IgM level, AMA M2 antibody index, and prevalence of other autoimmune diseases, were listed. These factors may reflect the severity of autoimmune liver diseases, and patients with elevated levels may have a low probability of a positive treatment response. Patient background, including weight, age, and prevalence of malignancies, was the last group of relevant features. Regarding the age at diagnosis, a previous report suggested that young patients showed more pronounced biochemical hepatic activity and high levels of some hormones, such as estrogen, leading to treatment resistance and poor response to treatment.[26] Malignancies, especially hepatocellular carcinoma, may occur in patients with an advanced stage of liver cirrhosis, and their liver dysfunction treatment may be deprioritized over treatment for the malignancies.

Our study has several limitations. First, this was a single-center retrospective study. As our facility is a university hospital, most of the patients with PBC had a high prevalence of autoimmune diseases. In addition, the 95% CIs in AUC results were wide owing to the small sample size. Therefore, we believe that a multicenter survey is required to establish a stronger evidence base for our conclusions. Second, given the imbalance in the number of patients not meeting the Paris II criteria, a larger population survey is required. The small number of participants could have led to overfitting in ML, resulting in low reproducibility in other facilities. In addition, the possibility of model overfitting and lack of validation cohort were limitations of this study and important for developing ML algorisms with more credibility. Nevertheless, we believe that the usefulness and importance of using ML in clinical fields has been highlighted in this study. These two factors may have influenced the results; however, we consider that our ML forecasting models are adequate for predicting treatment outcomes in patients with PBC and may be useful in clinical practice.

## Conclusion

In conclusion, in this study, we have developed the first ML prediction model based on Paris II criteria using pretreatment data. This algorithm has the potential for application in routine clinical practice to improve PBC outcomes. However, further studies are required to verify the accuracy of these prediction models.

## Acknowledgments

We would like to thank Editage (www.editage.com) for English language editing.

## Ethics approval

This study was reviewed and approved by the institutional review board of the University of Niigata (approval No. 2021-0385).

## Patient consent statement

Patients were given the opportunity to refuse to join this study.

***Data availability statement.*** The datasets used for this study are not publicly available.

## References

1 Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One*. 2014; **9**: e88225.

2 Azemoto N, Abe M, Murata Y *et al*. Early biochemical response to ursodeoxycholic acid predicts symptom development in patients with asymptomatic primary biliary cirrhosis. *J. Gastroenterol.* 2009; **44**: 630–4.

3 Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief. Bioinform.* 2021; **22**: 360–79.

4 Cappon G, Facchinetti A, Sparacino G, Georgiou P, Herrero P. Classification of postprandial glycemic status with application to insulin dosing in type 1 diabetes-an in silico proof-of-concept. *Sensors (Basel)*. 2019; **19**: 1–11.

5 Carbone M, Nardi A, Flack S *et al*. Pretreatment prediction of response to ursodeoxycholic acid in primary biliary cholangitis: development and validation of the UDCA Response Score. *Lancet Gastroenterol. Hepatol.* 2018; **3**: 626–34.

6 Carbone M, Sharp SJ, Flack S *et al*. The UK-PBC risk scores: Derivation and validation of a scoring system for long-term prediction of end-stage liver disease in primary biliary cholangitis. *Hepatology*. 2016; **63**: 930–50.

7 Cikes M, Sanchez-Martinez S, Claggett B *et al*. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur. J. Heart Fail.* 2019; **21**: 74–85.

8 Corpechot C, Abenavoli L, Rabahi N *et al*. Biochemical response to ursodeoxycholic acid and long-term prognosis in primary biliary cirrhosis. *Hepatology*. 2008; **48**: 871–7.

9 Corpechot C, Chazouillères O, Poupon R. Early primary biliary cirrhosis: biochemical response to treatment and prediction of long-term outcome. *J. Hepatol.* 2011; **55**: 1361–7.

10 D'Amato D, De Vincentis A, Malinverno F *et al*. Real-world experience with obeticholic acid in patients with primary biliary cholangitis. *JHEP Rep.* 2021; **3**: 100248.

11 European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu, European Association for the Study of the Liver. EASL Clinical Practice Guidelines: the diagnosis and management of patients with primary biliary cholangitis. *J. Hepatol.* 2017; **67**: 145–72.

12 Gerussi A, Verda D, Bernasconi DP *et al*. Machine learning in primary biliary cholangitis: a novel approach for risk stratification. *Liver Int.* 2022; **42**: 615–27.

13 Honda A, Tanaka A, Kaneko T *et al*. Bezafibrate improves GLOBE and UK-PBC scores and long-term outcomes in patients with primary biliary cholangitis. *Hepatology*. 2019; **70**: 2035–46.

14 Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplant.* 2013; **48**: 452–8.

15 Kautzky A, Möller HJ, Dold M *et al*. Combining machine learning algorithms for prediction of antidepressant treatment response. *Acta Psychiatr. Scand.* 2021; **143**: 36–49.

16 Kimura N, Setsu T, Arao Y *et al*. Cumulative risk of developing a new symptom in patients with primary biliary cholangitis and its impact on prognosis. *JGH Open*. 2022; **6**: 577–86.

17 Kimura N, Takamura M, Takeda N *et al*. Paris II and Rotterdam criteria are the best predictors of outcomes in patients with primary biliary cholangitis in Japan. *Hepatol. Int.* 2021; **15**: 437–43.

18 Kjærgaard K, Frisch K, Sørensen M *et al*. Obeticholic acid improves hepatic bile acid excretion in patients with primary biliary cholangitis. *J. Hepatol.* 2021; **74**: 58–65.

19 Konerman MA, Zhang Y, Zhu J, Higgins PD, Lok AS, Waljee AK. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology*. 2015; **61**: 1832–41.

20 Lammers WJ, Hirschfield GM, Corpechot C *et al*. Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. *Gastroenterology*. 2015; **149**: 1804–12.

21 Le Berre C, Sandborn WJ, Aridhi S *et al*. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology*. 2020; **158**: 76–94.

22 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; **521**: 436–44.

23 Lindor KD, Bowlus CL, Boyer J, Levy C, Mayo M. Primary biliary cholangitis: 2018 practice guidance from the American Association for the Study of Liver Diseases. *J. Hepatol.* 2018; **69**: 394–419.

24 Lindor KD, Bowlus CL, Boyer J, Levy C, Mayo M. Primary biliary cholangitis: 2018 practice guidance from the American Association for the Study of Liver Diseases. *Hepatology*. 2019; **69**: 394–419.

25 Örnolfsson KT, Lund SH, Olafsson S, Bergmann OM, Björnsson ES. Biochemical response to ursodeoxycholic acid among PBC patients: a nationwide population-based study. *Scand. J. Gastroenterol.* 2019; **54**: 609–16.

26 Papastergiou V, Tsochatzis EA, Rodriguez-Peralvarez M *et al*. Biochemical criteria at 1 year are not robust indicators of response to ursodeoxycholic acid in early primary biliary cirrhosis: results from a 29-year cohort study. *Aliment. Pharmacol. Ther.* 2013; **38**: 1354–64.

27 Parés A, Caballería L, Rodés J. Excellent long-term survival in patients with primary biliary cirrhosis and biochemical response to ursodeoxycholic acid. *Gastroenterology*. 2006; **130**: 715–20.

28 Pedregosa F, Varoquaux G, Gramfort A *et al*. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 2011; **12**: 2825–30.

29 Sammut SJ, Crispin-Ortuzar M, Chin SF *et al*. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022; **601**: 623–9.

30 Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. *J. Affect. Disord.* 2021; **281**: 618–22.

31 Tanaka A, Hirohara J, Nakano T *et al*. Association of bezafibrate with transplant-free survival in patients with primary biliary cholangitis. *J. Hepatol.* 2021; **75**: 565–71.

32 Tanphiriyakun T, Rojanasthien S, Khumrin P. Bone mineral density response prediction following osteoporosis treatment using machine learning to aid personalized therapy. *Sci. Rep.* 2021; **11**: 13811.

33 Tian S, Liu Y, Sun K *et al*. A nomogram based on pretreatment clinical parameters for the prediction of inadequate biochemical response in primary biliary cholangitis. *J. Clin. Lab. Anal.* 2020; **34**: e23501.

34 Vespasiani-Gentilucci U, Rosina F, Pace-Palitti V *et al*. Rate of non-response to ursodeoxycholic acid in a large real-world cohort of primary biliary cholangitis patients in Italy. *Scand. J. Gastroenterol.* 2019; **54**: 1274–82.

35 Wilde AB, Lieb C, Leicht E *et al*. Real-world clinical management of patients with primary biliary cholangitis-A retrospective multicenter study from Germany. *J. Clin. Med.* 2021; **10**: 1061.

36 Working Subgroup. English version of Clinical Practice Guidelines for Primary Biliary Cirrhosis. Guidelines for the management of primary biliary cirrhosis. *Hepatol. Res.* 2014; **44**: 71–90.

## Supporting information

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

**Figure S1.**

**Table S1.** Results of Brier score.
**Table S2.** Cox hazard analysis for overall survival.
**Table S3.** Cox hazard analysis for liver-related death.