Gastro Hep Advances 2023;2:778–787

ORIGINAL RESEARCH—CLINICAL

Development and Validation of a Machine Learning–Based Prediction Model for Detection of Biliary Atresia

Ho Jung Choi,[1,*] Yeong Eun Kim,[1,*] Jung-Man Namgoong,[2] Inki Kim,[3] Jun Sung Park,[1] Woo Im Baek,[1] Byong Sop Lee,[1] Hee Mang Yoon,[4] Young Ah Cho,[4] Jin Seong Lee,[4] Jung Ok Shim,[5] Seak Hee Oh,[1] Jin Soo Moon,[6] Jae Sung Ko,[6] Dae Yeon Kim,[2] and Kyung Mo Kim[1]

[1]Department of Pediatrics, Asan Medical Center Children's Hospital, University Ulsan College of Medicine, Seoul, Korea; [2]Division of Pediatric Surgery, Department of Surgery, Asan Medical Center, University Ulsan College of Medicine, Seoul, Korea; [3]Department of Convergence Medicine, Asan Institutes for Life Sciences, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; [4]Department of Radiology, Asan Medical Center, University Ulsan College of Medicine, Seoul, Korea; [5]Department of Pediatrics, Korea University College of Medicine, Seoul, Korea; and [6]Department of Pediatrics, Seoul National University Children's Hospital, Seoul National University College of Medicine, Seoul, Korea

BACKGROUND AND AIMS: Biliary atresia is a rare and devastating bile duct disease that occurs during the neonatal period. Timely identification and prompt surgical intervention is critical for improving the outcome. The aim of the study was to develop a new machine learning–based prediction model for the detection of biliary atresia. **METHODS:** Neonates aged <100 days with cholestasis at least once were retrospectively screened in 2 tertiary referral hospitals between 2015 and 2020. Simple demographic data, routine laboratory indices, and imaging findings of ultrasonography and hepatobiliary scintigraphy were used as features in the multivariate analysis. The extreme gradient boosting (XGBoost) framework was used to develop prediction models according to the diagnostic steps. **RESULTS:** Among 1605 enrolled neonates with all-cause cholestasis, 145 (9%) were included as having biliary atresia. Direct bilirubin, gamma-glutamyl transpeptidase, abdominal sonography, and hepatobiliary scan were the most impactful features in prediction models. The Step II XGBoost model, consisting of nonimaging inputs, showed excellent discriminatory performance (area under the curve = 0.97). The Step III and IV XGBoost models showed near-perfect performances (area under the curve = 0.998 and 0.999, respectively). In external validation (n = 912 with 118 [12.9%] biliary atresia), XGBoost-based prediction models consistently showed acceptable performances. Utilizing shapley additive explanation values also provided visualized insight and explanation of the contribution of features in detecting biliary atresia. The models were integrated into a web-based diagnostic tool for case-level application. **CONCLUSION:** We introduced a new machine learning–based prediction model for detecting biliary atresia in the largest cohorts of neonatal cholestasis.

Keywords: Biliary Atresia; Neonatal Cholestasis; Prediction; Machine Learning; XGBoost

Introduction

Biliary atresia is a serious neonatal bile duct disorder that results in progressive fibrosis of the bile duct and obliteration of bile flow.[1] It rapidly progresses into deteriorating liver cirrhosis without early identification and prompt performance of Kasai portoenterostomy. Restoration of normal bile drainage through early intervention is crucial to avoid devastating consequences and subsequent liver transplantation and surgical intervention is recommended within the first 45 to 60 days of life. Therefore, the early detection of biliary atresia among neonates with cholestasis for prompt intervention is critically important to improving clinical outcomes. However, accurate and prompt identification of biliary atresia from other cholestatic conditions is challenging in clinical practice because vast nonsurgical etiologies are also related to neonatal cholestasis, which presents overlapped phenotypes of biliary atresia.[2–4] Further, there is no single measure to confirm biliary atresia except for intraoperative cholangiography. To overcome diagnostic difficulties for preoperative biliary atresia and lessen its burden of false negative (FN) and false positive (FP) misdiagnosis, clinicians use comprehensive and extensive investigations such as serologic laboratory indices, imaging tools, and invasive acquisition of liver tissue.[5–8] Based on the retrospective dataset from the comprehensive approach, a few diagnostic tools, which were

**Ho Jung Choi and Yeong Eun Kim contributed equally to this article as co-first authors. Seak Hee Oh and Jae Sung Ko are co-corresponding authors.*

Abbreviations used in this paper: **AUC, area under the curve; CI, confidence interval; DB, direct bilirubin; FN, false negative; FP, false positive; GGT, gamma-glutamyl transpeptidase; HBS, hepatobiliary scan; LR, logistic regression; ML, machine learning; MNAR, missing not at random; SHAP, Shapley additive explanation; TB, total bilirubin; TP, true positive; TRIPOD, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; US, ultrasonography; XGBoost, extreme gradient boosting.**

Most current article

2772-5723
https://doi.org/10.1016/j.gastha.2023.05.002

derived mainly from conventional logistic regression (LR)-based prediction, have been developed.[5–9]

In detail, the sensitivity and specificity of high gamma-glutamyl transpeptidase (GGT) level in discriminating biliary atresia from neonatal cholestasis were 40% and 98%, respectively.[10] The pooled sensitivity and specificity of traditional ultrasonography (US) in meta-analysis were known to be 85% and 97%, respectively.[11] Recently introduced elastography of US showed sensitivity of 83% and specificity of 77%.[12] The pooled sensitivity and specificity of hepatobiliary scan (HBS) were 98% and 70%, respectively.[13] A nomogram using clinical and serologic features showed sensitivity of 86% and specificity of 80% and another recent one adding HBS findings showed sensitivity of 89% and specificity of 84% (c-statistics of 0.91). A new scoring system using clinical, laboratory, US, and HBS data showed a high discriminatory performance (c-statistics of 0.981). A scoring system derived from a prospective study using clinical, laboratory, US, and even liver biopsy revealed highest sensitivity of 100% and specificity of 98% and did not repeat its performance on an external study.[14]

Data missingness often occurs in hospitals due to various factors, such as patient compliance and hospital operational reasons. If data missingness is large, conventional list-wise deletion approach (complete-case analysis) in LR may not include substantial numbers, resulting in biased and low precision of prediction models.[15] Herein, missingness should be evaluated in terms of patterns and mechanisms of missing data.[16,17] As the detection of biliary atresia among neonatal cholestasis in general comprises differential diagnostic steps from serologic laboratory tests to invasive tests (Figure 1), nonbiliary atresia patients may systemically drop out by clinical judgment, which includes a strong intention to avoid misdiagnosis due to the nature of its fatal consequence and the possibility of medicolegal issues. Accordingly, the mechanism of missingness may not be at random in a neonatal cholestasis dataset.

Recently, there has been an increase in the use of machine learning (ML) in gastroenterology literature due to its excellent discriminatory ability.[18] Extreme gradient boosting (XGBoost) is an efficient ML classifier,[19,20] superior to random forest and LR in performance. XGBoost framework softly handles missing values without certain types of imputation via a sparsity-aware split finding algorithm. In addition, this ML method using recursive partitioning analysis overcomes limitations of classical LR such as linearity assumption. However, its black-box algorithm's interpretability still limits its application in clinical practice. Furthermore, to provide a decision rule in medical practice, this innovative approach should follow the proper steps of a conventional guideline.[21] This study aimed to develop a new ML-based prediction model for the detection of biliary atresia from a dataset on 2517 patients with neonatal cholestasis from 2 tertiary referral hospital from 2015 to 2020.

Methods

Data processing

We conducted a retrospective study on data collected from neonatal patients who had cholestasis at a tertiary referral hospital, Asan Medical Center, Korea. As inclusion criteria, a dataset of serologic direct bilirubin (DB) tests in a group of neonatal patients aged <100 days from January 2005 to January 2020 was extracted from the medical record database. A cut-off value (serum DB $\geq$ 2.0 mg/dL) from the Childhood Liver Disease Research Network[22] was used to define neonatal cholestasis. Patients who were referred to our hospital after the Kasai operation were excluded. The study was approved by the Institutional Review Board of the Asan Medical Center (#2022-0522). Informed consent was waived as this was a retrospective study.

The primary event (dependent variable) was biliary atresia diagnosis among enrolled patients with neonatal cholestasis. The diagnosis was validated by confirmative findings of intraoperative cholangiography, which showed the absence of contrast filling inside the bile duct, and subsequent compatible pathological findings of bile duct and liver before and after Kasai operation. All medical records and diagnoses of nonbiliary atresia groups in the whole cohort were also manually validated twice by 4 pediatric hepatologists.

For input features (independent variables), we initially included simple inputs—demographic data (age, sex, and body weight) and serologic laboratory tests at the time of enrollment. The serologic laboratory tests were white blood cell ($\times 10^3/\mu$L) and platelet ($\times 10^3/\mu$L) counts, hemoglobin (g/dL), prothrombin time (international normalized ratio), aspartate aminotransferase (IU/L), alanine aminotransferase (IU/L), alkaline phosphatase (IU/L), albumin (g/dL), total bilirubin (TB, mg/dL), DB (mg/dL), creatinine (mg/dL), c-reactive protein (mg/dL), and GGT (IU/L). Binary input features from imaging data such as major compatible findings[11,13] of abdominal ultrasonography (US; gallbladder anomalies and absence, nonvisualization of common bile duct, and triangular sign) and HBS with technetium 99m mebrofenin (HBS; no excretion or severely delayed excretion of bile) were classified as positivity of predicting biliary atresia.

Exploratory data analysis

Descriptive statistics for the variables (features and event) were provided for biliary atresia and nonbiliary atresia groups. Continuous features were expressed as median values with interquartile ranges as most of them did not pass the normality tests. Differences between groups were assessed using Mann-Whitney U test for continuous features and the χ^2 test for categorical features. All statistical calculations were performed using IBM SPSS version 27.0 (SPSS Inc, Armonk, NY) and R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria) with publicly available codes. A P value of $< .05$ was considered statistically significant.

Evaluation of missing data included the identification of mechanisms and patterns of missing values. Missing data pattern refers to the configuration of measured features and their missing values, and missing data mechanism is the possible association of missingness with measured features.[17] The pattern was visualized by a diagram of location of the missing values in the whole dataset. The mechanism of missing values of features was evaluated by classifying them into missing completely at random, missing at random, or missing not at random (MNAR).[16]

Model development and evaluation

XGBoost is a decision tree-based ensemble ML method.[19,20] We used this ML method because it has been shown to be

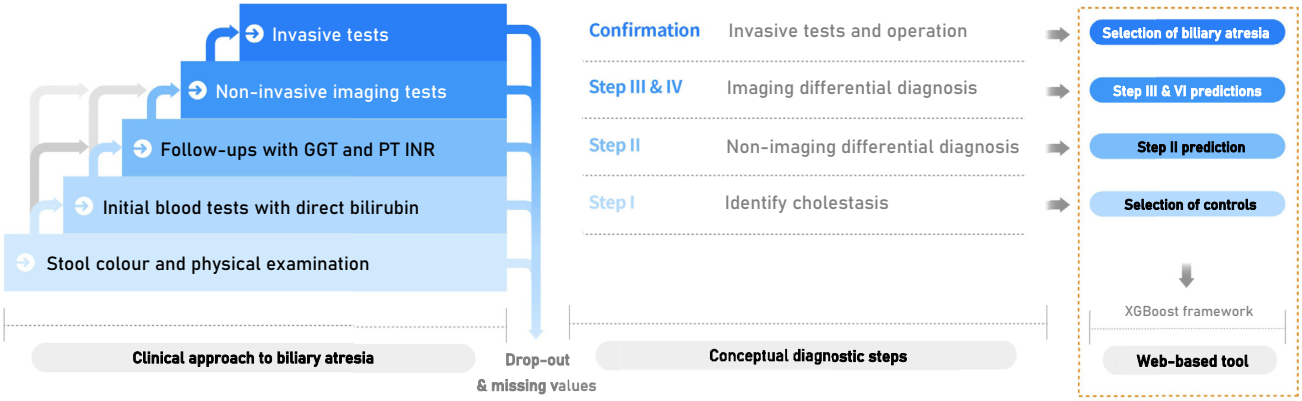

Figure 1. Conceptual diagnostic steps to biliary atresia among all-cause cholestatic neonates. Cholestatic neonates are, in general, initially examined by a primary health care provider. From this perspective, diagnosis of biliary atresia among cholestatic neonates comprises 3 or 4 conceptual steps. After the first diagnostic step of detecting cholestasis, initial differential process (Step II) begins by conducting nonimaging tests including GGT and prothrombin time international normalized ratio. Then, imaging tests such as abdominal sonography and HBS could be added based on the clinical judgment (Steps III and IV, respectively). Finally, clinicians decide whether the patients need invasive confirmative tests such as biopsy and intraoperative cholangiography for the confirmation of biliary atresia. During these diagnostic steps, missing values and drop-outs of nonbiliary atresia controls occur by clinical judgments. GGT, gamma-glutamyl transpeptidase; PT INR, prothrombin time international normalized ratio; XGBoost, extreme gradient boosting.

dominating in ML competitions[20] and has consistently shown high performance in recent medical predictions.[23–25] In addition, XGBoost utilizes subjects with missing values through the sparsity-aware split finding algorithm.[19,26] The performance of the models was evaluated in 2 main aspects: discrimination by measuring the area under the curve (AUC, equivalent to c-statistics in binary classification) in the receiver-operating characteristic curve analysis and reliability by measuring calibration slope and intercept (Supplementary Method 1). For internal validation, 10-fold cross-validation was used. For external validation, a dataset from the Seoul National University Hospital (Institutional Review Board #2204-147-1319) was used and the spatial transportability of the prediction models was evaluated. Clinical usefulness was also evaluated by measuring sensitivity, specificity, positive predictive value, negative predictive value, and diagnostic accuracy on confusion matrix. To assess interpretability of models, we derived the feature contribution represented by the Shapley additive explanations (SHAPs) value of each feature.[27] The SHAP value is estimated via post hoc analysis based on input perturbation and provides features' contribution on a log-odds scale (logarithm of the ratio of high risk to low risk). The SHAP value was calculated using the following equation (Supplementary Method 2):

$$\Phi_i(f,x) = \sum_{S \in N} \frac{|S|!(n-|S|-1)!}{n!}[v(S \cup \{i\}) - v(S)]$$

In this study, the model development adhered to the full steps of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline.[21] Finally, a web-based ML tool was developed using the R shiny package including SHAP plots and Break-down plots (Supplementary Method 3).

Results

Data processing

A total of 17,372 neonates received serologic DB tests of mean 7.4 times during the study period (Figure 2). Of them, 1608 (9.2%) had cholestasis (DB $\geq$ 2.0 mg/dL) at least once during the neonatal period, and biliary atresia was noted in 148 (0.85%) children. Of them, 3 patients referred to our center after Kasai operation elsewhere were excluded. Then, the development dataset (n = 1605 with 145 [9%] biliary atresia) consisted of 923 (57.5%) male and 682 (42.5%) female patients with a median enrollment age of 21 days (interquartile range: 8–48 days) (Table 1). Besides biliary atresia, a variety of specific diagnoses combined with nonbiliary atresia cholestasis (n = 1460, 91%) were noted including hepatic causes (15.8%) and extrahepatic conditions (84.2%). On serologic laboratory tests, the biliary atresia group had higher levels of major laboratory features, such as TB, DB, and GGT compared to those of non-biliary atresia group. Among imaging features, biliary atresia had compatible findings of US with more instances (n = 137/145, 94.4%) compared to that of nonbiliary atresia group (n = 33/1209, 2.7%). HBS, which had more missing instances, also showed a high detection rate of biliary atresia group (n = 135/139, 97.1%). In univariate analysis, discriminatory performances, presented by AUCs, of each laboratory feature ranged from 0.558 to 0.877 (Table A1). Among imaging features, US showed the highest AUC (0.959), while HBS had an AUC of 0.842.

The missingness of input features showed a stepwise monotone pattern in our development dataset (Figure A1A), which is often observed in longitudinal studies where

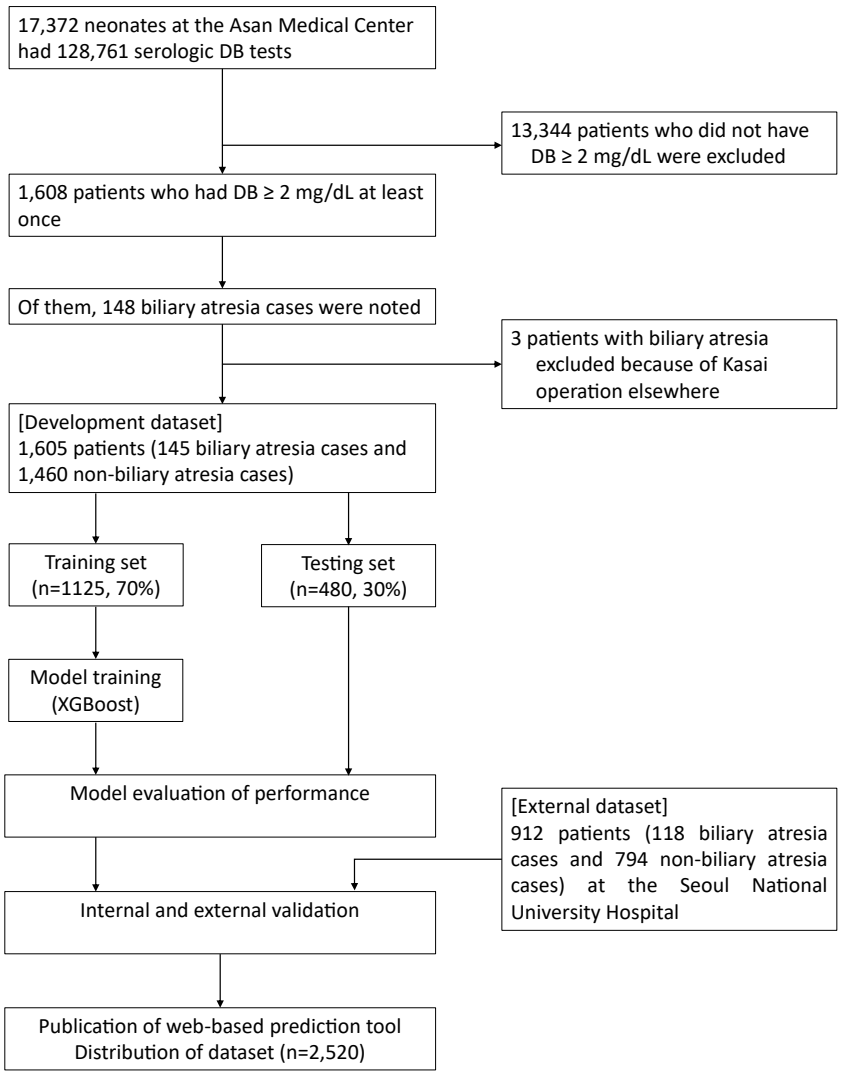

Figure 2. Flow chart of study population. DB, direct bilirubin; XGBoost, extreme gradient boosting.

samples systemically drop out due to protocols or outcomes during the follow-up.[17] In the evaluation of missing mechanism, the 3 well-known impactful features of interest, such as GGT, US, and HBS,[7,11,13] were likely to be MNARs that were closely related not only to other features but also to the primary event (biliary atresia) (Figure A1B). If complete-case analysis with no missing value was conducted, a substantial loss (75%) of nonbiliary atresia samples would occur in our development dataset.

Model development

First, the XGBoost prediction model for Step II prediction, which comprised simple demographic and serologic laboratory data, was built after hyperparameter optimization (Table A2). The train and test sets in the internal dataset were split in a ratio of 0.7 (n = 1125):0.3 (n = 480). The Step II prediction model showed excellent discriminatory performance (AUC = 0.967, 95% confidence interval [CI] = 0.945–0.992) with proper calibration (slope = 0.98, 95% CI = 0.949–1.011) (Table 2 and Figure A2). The contribution of each feature is shown in the SHAP summary plots (Figure 3). In the plots, DB, GGT, TB, and aspartate aminotransferase appeared to be the most impactful top-4 features in Step II prediction, as DB and GGT are features of interest in this field.[8,10] Increase (approaches purple), the SHAP values, which are quantitative metrics of contribution to the risk of biliary atresia, increase. As expected, most MNAR values of GGT (gray dots) remained on the low SHAP values, indicating that missingness is related to a smaller likelihood of a primary event.

Second, categorical input features of US and HBS were included in the development of Step III and IV prediction models. The Step III model consisted of Step II's dataset and additional imaging inputs of US, and the Step IV model consisted of Step III's dataset and additional inputs of HBS. The Step III and IV models showed the AUCs of 0.998 (95% CI = 0.98–1.0) and 0.999 (95% CI = 0.989–1.0), respectively, with acceptable calibration slopes (Table 2 and Figure A2). Overall, Step III and IV prediction models outperformed most previous works that mainly utilized LR analyses (Table A3). In the SHAP summary plot of Step III

Table 1. Characteristics of Study Cohort With Neonatal Cholestasis (n = 1605)

	Total	Biliary atresia	Non-biliary atresia	
Development dataset	n = 1605	n = 145 (9%)	n = 1460 (91%)	*P* value
Demographics at enrollment				
Age, d	21 (8–48)	54 (27–65)	18 (8–45)	<.01
Male	923 (57.5%)	63 (43.4%)	860 (58.9%)	<.01
Weight	2.9 (2–3.7)			<.01
Etiology				
Hepatic	375 (23.4%)	145 (100%)	231 (15.8%)	<.01
Extrahpatic	1229 (76.6%)	0 (0%)	1229 (84.2%)	<.01
Laboratory				
Total bilirubin	6.2 (3.6–9.6)	8.4 (7–9.8)	5.7(3.4–9.5)	<.01
Direct bilirubin	2.4 (2.1–3.5)	5.6 (4.1–6.9)	2.3 (2.1–3)	<.01
GGT	141 (73–266)	396 (217–604)	121 (67–206)	<.01
Imaging				
US, compatible findings	170 (12.6%, available in 1354)	137 (94.4%, available in 145)	33 (2.7%, available in 1209)	<.01
HBS, compatible findings	221 (50.2%, available in 440)	135 (97.1%, available in 139)	86 (28.6%, available in 301)	<.01

Table 2. Performance of Development Models

	Performance					
	Development testing set (n = 480)		Internal validation (n = 1605)		External validation (n = 912)	
Dataset	AUC	Calibration slope	AUC[a]	Calibration slope[a]	AUC	Calibration slope
Step II	0.967	0.98	0.97	0.97	0.949	1.01
Step III	0.998	0.96	0.991	0.98	0.976	0.99
Step IV	0.999	1.01	0.996	1.0	0.978	1.0

[a]Mean values from 10-fold cross validation.

model, the impact of US steeply increased up to a SHAP value of 3–6, suggesting its critical role in detecting biliary atresia. In Step IV plots, the impact of HBS was not strong compared with US, as noted in univariate analyses. As expected, the MNAR values of US and HBS were related to a smaller likelihood of a primary event.

Validation and clinical usefulness

As internal and external validations are essential parts of the TRIPOD guideline,[21] we performed internal validation via 10-fold cross validation, and the performance of each ML model remained acceptable (Table 2). For the external validation, we used the validation dataset from a retrospective cohort of the Seoul National University Hospital, one of the largest tertiary hospitals in Korea, with the same inclusion and exclusion criteria (n = 912 neonatal patients with 118 [12.9%] biliary atresia). On the external validation dataset, the XGBoost-based prediction models also achieved acceptable performances (Table 2, Figure 4A, and Figure A3); AUCs of Step II to IV XGBoost-based prediction models were 0.949, 0.976, and 0.978, respectively. The diagnostic accuracy, sensitivity and specificity were 86.3%, 90.7%, and 85.6% in Step II model; 94.7%, 88.1%, and 85.6%, respectively, in Step III model; 95.5%, 90.6%, and 85.6%, respectively, in Step IV model (Table A4).

For the final updated models, all samples from the internal and external datasets (n = 2517 with 263 [10.4%] biliary atresia) were merged. Using optimal cut-off probabilities determined on maximal Youden's index (Table 3), the clinical usefulness was evaluated on the Step II to IV models. The diagnostic accuracy, sensitivity and specificity were 96.6%, 94.3%, and 96.9%, respectively, in Step II model. Step III model outperformed Step II model. As expected, Step IV model outperformed both Step II and III models. FN rates of Step II, III, and IV models were 5.7%, 2.6%, and 0.8%. SHAP force plots visualize how the Step II models classified patients based on the SHAP values of each feature (Figure 4B). As the optimal cut-off threshold probability (13.5%) was estimated (left upper box), the probability was converted to the corresponding net SHAP value (0.799) (left lower box). On SHAP force plot for Step II model, SHAP values by features were displayed using the merged internal and external datasets and SHAP values of each feature based on its contribution to the risk of biliary atresia were calculated on every patient (right upper box) and net SHAP values of each patient were also displayed (purple dots). True positive (TP), true negative, FN, and FP groups were divided by the cut-off net SHAP value of 0.799. SHAP patterns of FN, TP, and FP groups were

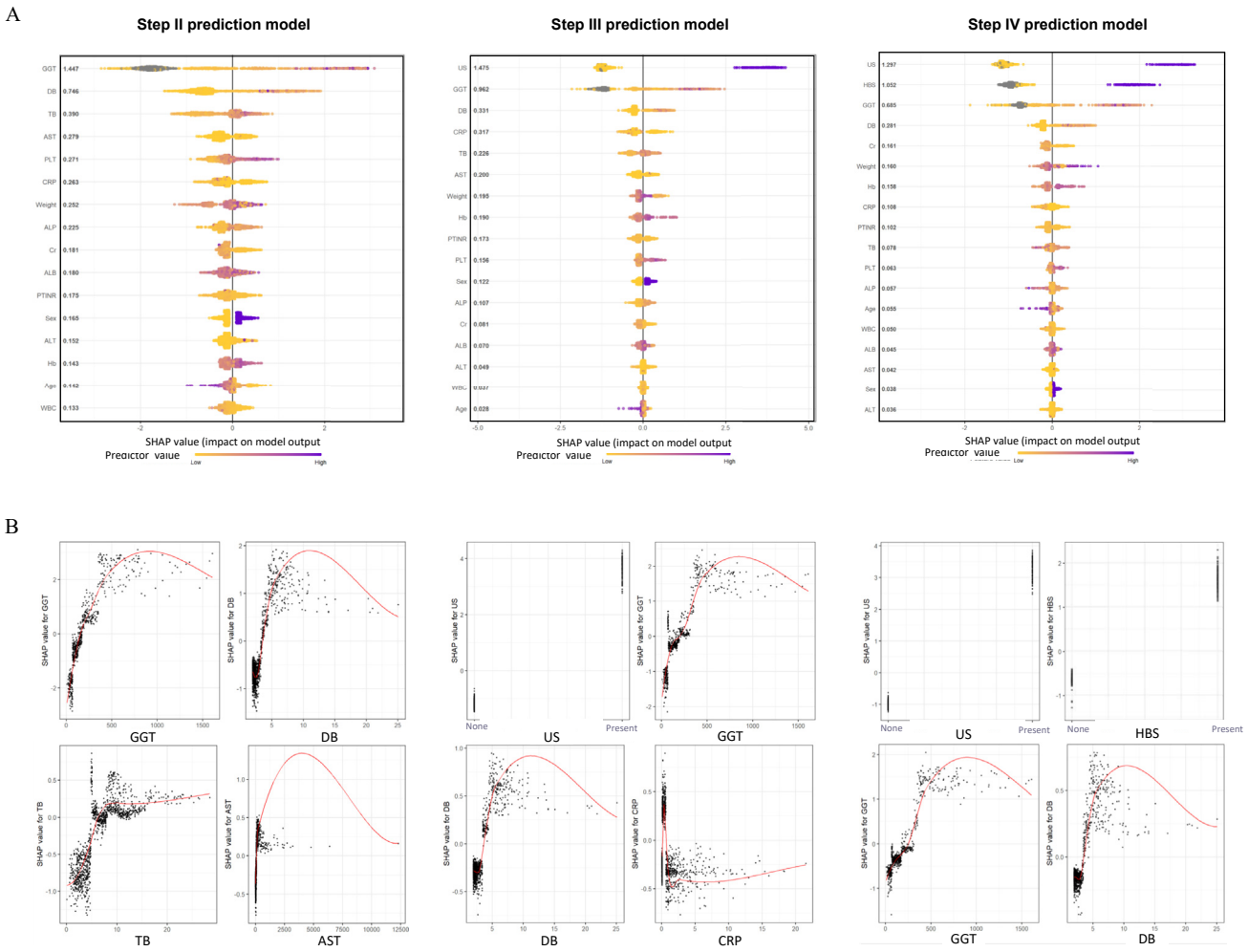

Figure 3. Contributions of features on SHAP summary plots of prediction models. (A) SHAP summary plots. The SHAP value for each patient of the development dataset by input features on the SHAP summary plots. The features on the y-axis are ranked from most important to least important with their mean absolute SHAP value. The x-axis represents the SHAP value contributed by each feature and patient, and a positive value on x-axis indicates a higher impact on the prediction of biliary atresia. The purple color indicates that the individual patients' feature value is high and vice versa (yellow). Gray dots indicate missing values. (B) Partial dependence plots. The plots of the top-4 features show a marginal effect shown by the SHAP values of features in predicting biliary atresia, presented by the average curve (red line for continuous features). ALB, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; Cr, creatinine; CRP, c-reactive protein; DB, direct bilirubin; GGT, gamma-glutamyl transpeptidase; Hb, hemoglobin; HBS, hepatobiliary scan; PLT, platelet; PT INR, prothrombin time; SHAP, shapley additive explanation; TB, total bilirubin; US, ultrasonography; WBC, white blood cell.

zoomed-in (right lower box). Other SHAP force plots for the Step III and IV models were also displayed in Figures A4 and A5.

Publication of prediction models

As using a nomogram or web-based tool is the final step of the TRIPOD guideline,[21] a web-based tool for Step II ∼ IV prediction models for detecting biliary atresia was developed using the R shiny package (URL: https://seakheeoh76.shinyapps.io/XGBoost_BA_prediction/). The user interface of this tool was designed for case-level application of prediction model to a single subject. When input data of a patient are uploaded, the model estimates the probability of biliary atresia. To promote the model's interpretability and user's understanding, we adopted individual SHAP and break-down plots to demonstrate how the model converts feature values of individual patients to risk contribution. Figure 4C shows examples of personalized and interpretable display of TP (biliary atresia) and true negative (nonbiliary atresia) instances on break-down and individual SHAP plots, where visualized patterns of features' contribution to the diagnosis were provided. Finally, the study followed all the steps of the TRIPOD guideline (Table A5).

Discussion

We developed and validated ML-based prediction models based on the conceptual diagnostic steps. The

development models showed excellent performances in predicting biliary atresia from the largest cohorts of all-cause neonatal cholestasis in tertiary hospitals. In addition, these ML-based prediction models maintained the acceptable spatial transportability in external data. Furthermore, the web-based tool provides not only the

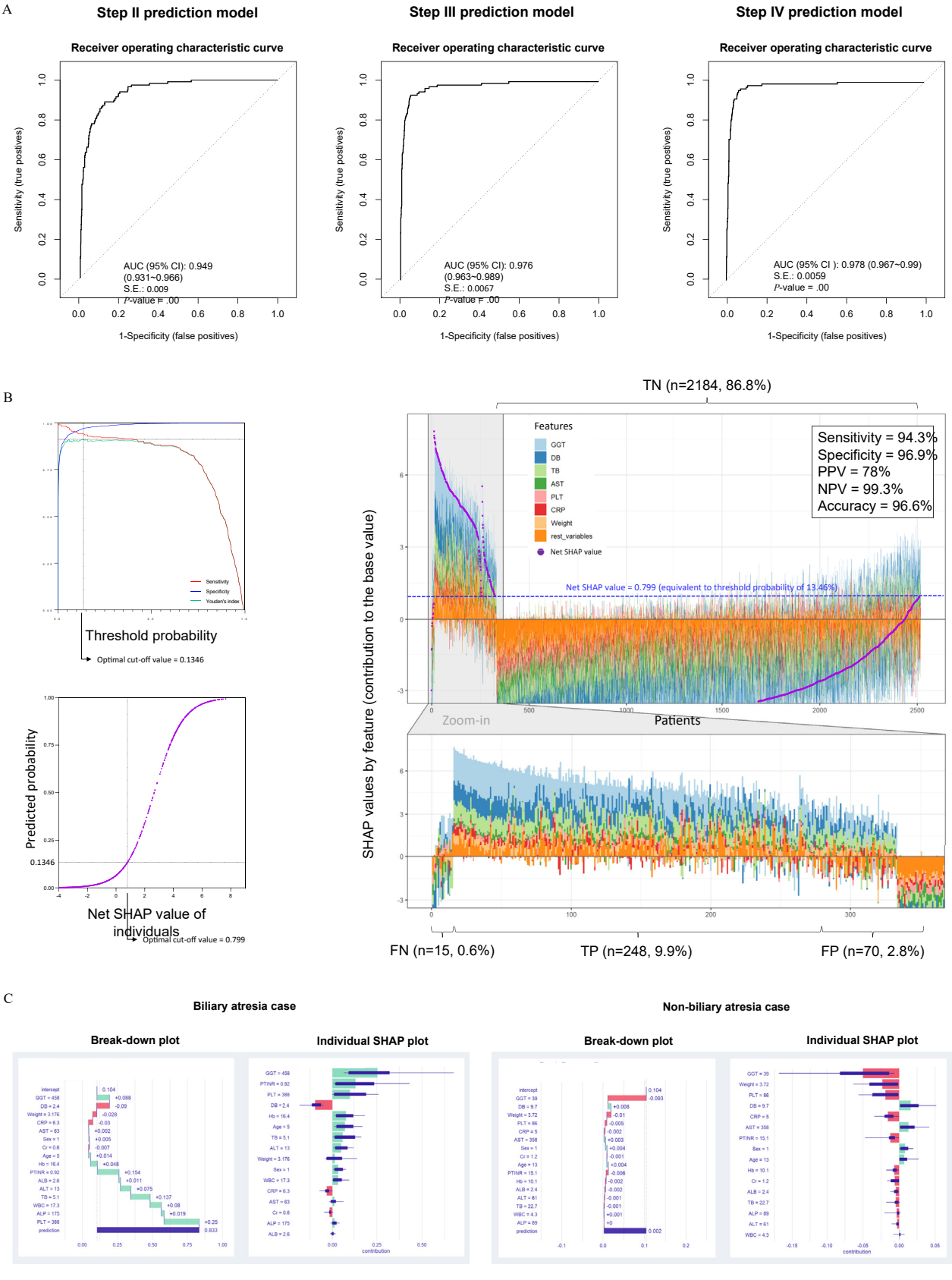

Table 3. Clinical Usefulness of the Final Web-Based Prediction Models

	Clinical usefulness					
	Whole dataset, n = 2517					
Dataset	Cut-off probability	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Diagnostic accuracy
Step II	13.5%	94.3%	96.9%	78%	99.3%	96.6%
Step III	33.7%	97.3%	99.6%	96.6%	99.7%	99.4%
Step IV	10.2%	99.2%	98.3%	87.3%	98.4%	98.4%

probability for the prediction of biliary atresia but also personalized and interpretable prediction by providing visualized plots of impactful features. To the best of our knowledge, this study is the first of its kind to fully utilize ML and properly balanced dataset of neonatal cholestasis reflecting a real-world situation in tertiary hospitals. Overall, the Step II model may be informative when primary health care providers encounter the initial differential diagnosis of biliary atresia without imaging inputs, and the Step III and IV models with near-perfect performances can be applicable for neonates who had advanced imaging tests.

The Step II XGBoost model comprises (1) demographic data (age, weight, and sex at the time of visit) and (2) serologic laboratory tests, which are both easily accessible and essentially objective data. Herein lies the great potential of the Step II model: it is free of human-related judgments or errors and can be easily used by primary health care providers who do not have special medical training in this field. Furthermore, this Step II model may be applicable in low- and middle-income countries, where US and HBS are not easily accessible. In addition, this noninvasive support system would help provide pretest impressions before US and HBS tests, since impactful US and HBS are also not free from FN and FP errors.[11,13]

This study used a large dataset of neonatal cholestasis with balanced event rate of biliary atresia in tertiary hospitals. As most researchers have used LR-based analyses, vast observations of all-cause neonatal cholestasis with missing values could be excluded in their studies. In addition, major reports about the estimation of risks and the development of diagnostic tools were obtained from radiologists and surgeons (Table A3), to whom already selected patients may be referred from primary healthcare providers. In the present study, we selected analytic samples by performing a universal screening of DB on 2 largest hospitals in Korea and utilized whole instances with missing values in the development of the models. Meaningful missingness of the MNAR feature is problematic in medical statistics.[16] Unless missing-at-random assumption is possible, imputation of MNAR may not be valid. However, sparsity-aware split-finding algorithm in XGBoost softly handles missing values.[19,26] Indeed, missing values displayed on the SHAP summary plots (Figure 3) had a negative impact on predicting biliary atresia in our XGBoost models.

XGBoost is a black-box model, in which interpretability of the algorithm between features and event is limited. According to the General Data Protection Regulation by the European Parliament,[28] it is important to obtain explanations about the algorithmic decisions in a ML-based healthcare system. Here, the SHAP value is currently the appropriate one among existing metrics to provide better explanations.[29,30] Utilizing SHAP values, this study provided more suitable explanations by showing various types of plots that yield a collaborated SHAP pattern to biliary atresia risk in both a whole group and individuals. Especially, individual SHAP plots can provide visualized insight into biliary atresia risk in case-level prediction (Figure 4B). Besides, our web-based final models provide break-down plots that utilized probabilities of impactful features (Figure 4C).

This study has inherent limitations; this retrospective observational study used data from electrical health records, resulting in information bias due to missing data and uncontrolled setting of enrollment time (identification of neonatal cholestasis). Bias from missingness was mitigated by the use of sparsity-aware split-finding algorithm in XGBoost. In addition, in defining and labeling diagnoses and

Figure 4. Performances of prediction models and publication of web-based prediction tool. (A) Discriminatory performances of prediction models in the external data. AUC; area under the curve in the receiver-operating characteristic curve analysis. Additional results are listed in Figure A3. (B) SHAP force plot of Step II prediction on all patients (n = 2517 with 263 [10.4%] biliary atresia). (C) Example of personalized and interpretable display of biliary atresia and nonbiliary atresia cases on web-based Step II prediction tool. In addition to providing a predicted probability, the tool also individual plots for user's understanding. Break-down plot and SHAP plot for individual patients illustrate visualized patterns of a feature's contribution. AUC, area under the curve; ALB, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; Cr, creatinine; CRP, c-reactive protein; DB, direct bilirubin; FN, false negative; FP, false positive; GGT, gamma-glutamyl transpeptidase; Hb, hemoglobin; NPV, negative predictive value; PLT, platelet; PPV, positive predictive value; PT INR, prothrombin time; S.E., standard error; SHAP, shapley additive explanation; TB, total bilirubin; TN, true negative; TP, true positive; WBC, white blood cell.

diagnostic indices, evaluating the agreement (concordance) between clinicians by providing some concordance statistics are beyond the scope of this study. Second, data were collected from tertiary referral hospitals, where more severe clinical settings may prevail over community-based settings. Another limitation is the long study period as identification strategies may change over time. Third, the number of study subjects was too small for ML-based models. However, building a large dataset of this rare cholestatic condition from a large population may be a formidable task. In Harpavat's study,[31] only 7 biliary atresia cases were screened among 100,000 neonates who underwent DB tests using population-based screening. Overall, these models were derived from a large cohort data from 2 tertiary hospitals, and therefore, the findings should be generalizable to other centers elsewhere. More external validations are then required for the final web-based models. Fourth, minor image findings of US shown in the literature were not used in the analyses. Fifth, the optimal cut-off value was not estimated based on the cost of FN case. The cost of FN case may be assumed to be very large compared to that of FP case in this field. However, quantitative setting for their costs has not been determined in the literature.

Conclusion

We developed a novel ML-based prediction model for the detection of biliary atresia among all-cause neonatal cholestasis in tertiary hospitals and introduce its web-based tool.

Supplementary Materials

Material associated with this article can be found in the online version at https://doi.org/10.1016/j.gastha.2023.05.002.

References

1. Hartley JL, Davenport M, Kelly DA. Biliary atresia. Lancet 2009;374(9702):1704–1713.
2. Fawaz R, Baumann U, Ekong U, et al. Guideline for the Evaluation of Cholestatic Jaundice in Infants: Joint Recommendations of the North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition and the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition. J Pediatr Gastroenterol Nutr 2017;64(1):154–168.
3. Choi HJ, Kim I, Lee H-J, et al. Clinical characteristics of neonatal cholestasis in a tertiary hospital and the development of a novel prediction model for mortality. EBioMedicine 2022;77:103890.
4. Gottesman LE, Del Vecchio MT, Aronoff SC. Etiologies of conjugated hyperbilirubinemia in infancy: a systematic review of 1692 subjects. BMC Pediatr 2015;15:192.
5. El-Guindi MA, Sira MM, Sira AM, et al. Design and validation of a diagnostic score for biliary atresia. J Hepatol 2014;61(1):116–123.
6. Dong R, Jiang J, Zhang S, et al. Development and validation of novel diagnostic models for biliary atresia in a large cohort of Chinese patients. EBioMedicine 2018;34:223–230.
7. Shneider BL, Moore J, Kerkar N, et al. Initial assessment of the infant with neonatal cholestasis-Is this biliary atresia? PLoS One 2017;12(5):e0176275.
8. Kim JR, Hwang JY, Yoon HM, et al. Risk estimation for biliary atresia in patients with neonatal cholestasis: development and validation of a risk score. Radiology 2018;288(1):262–269.
9. Jiang J, Wang J, Shen Z, et al. Serum MMP-7 in the diagnosis of biliary atresia. Pediatrics 2019;144(5):e20190902.
10. Liu CS, Chin TW, Wei CF. Value of gamma-glutamyl transpeptidase for early diagnosis of biliary atresia. Zhonghua Yi Xue Za Zhi (Taipei) 1998;61(12):716–720.
11. Yoon HM, Suh CH, Kim JR, et al. Diagnostic performance of sonographic features in patients with biliary atresia: a systematic review and meta-analysis. J Ultrasound Med 2017;36(10):2027–2038.
12. Dong B, Weng Z, Lyu G, et al. The diagnostic performance of ultrasound elastography for biliary atresia: A meta-analysis. Front Public Health 2022;10:973125.
13. Kianifar HR, Tehranian S, Shojaei P, et al. Accuracy of hepatobiliary scintigraphy for differentiation of neonatal hepatitis from biliary atresia: systematic review and meta-analysis of the literature. Pediatr Radiol 2013;43(8):905–919.
14. Sciveres M, Milazzo MP, Maggiore G. A scoring system for biliary atresia: is this the right one? J Hepatol 2015;62(4):985–986.
15. Steyerberg E. Clinical prediction models. Stat Biol Health 2008; 127–131.
16. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017;9:157–166.
17. Enders CK. Applied missing data analysis: methodology in the social sciences. New York: Guilford Publications, 2010.
18. Facciorusso A, Licinio R, Di Leo A. Machine learning methods in gastroenterology. Gastroenterology 2015;149(4):1128–1129.
19. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.
20. Morde V. XGBoost algorithm: long may She reign! Accessed June 11, 2022. https://towardsdatasciencecom/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.
21. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1–W73.
22. Hertel PM, Hawthorne K, Kim S, et al. Presentation and outcomes of Infants with Idiopathic cholestasis: a multicenter prospective study. J Pediatr Gastroenterol Nutr 2021;73(4):478–484.
23. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine

learning approach using XGboost. J Transl Med 2020; 18(1):462.
24. Zhang Y, Feng T, Wang S, et al. A novel XGBoost method to identify cancer tissue-of-origin based on copy number variations. Front Genet 2020;11:585029.
25. Li C, Chen L, Chou C, et al. Using machine learning approaches to predict short-term risk of cardiotoxicity among patients with colorectal cancer after starting fluoropyrimidine-based chemotherapy. Cardiovasc Toxicol 2021;22:130–140.
26. Rusdah DA, Murfi H. XGBoost in handling missing values for life insurance risk prediction. SN Appl Sci 2020; 2(8):1336.
27. Shapley LS. 17. A value for n-person games. In: Harold William K, Albert William T, eds. Contributions to the theory of games (AM-28), Volume II. New Jersey: Princeton University Press, 2016:307–318.
28. Goodman B, Flaxman S. European union regulations on algorithmic decision-making and a "right to explanation". AI Mag 2017;38(3):50–57.
29. Baptista ML, Goebel K, Henriques EMP. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artif Intell 2022;306:103667.
30. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: AIES 2020. New York: Society, Association for Computing Machinery, Inc, 2020:180–186.
31. Harpavat S, Garcia-Prats JA, Anaya C, et al. Diagnostic yield of newborn screening for biliary atresia using direct or conjugated bilirubin measurements. JAMA 2020; 323(12):1141–1150.

Received July 15, 2022. Accepted May 12, 2023.

Correspondence:
Address correspondence to: Seak Hee Oh, MD, PhD, Department of Pediatrics, Asan Medical Center Children's Hospital, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-Gu, Seoul 05505, Korea. e-mail: seakhee.oh@amc.seoul.kr. Jae Sung Ko, MD, PhD, Department of Pediatrics, Seoul National University Children's Hospital, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea. e-mail: kojs@snu.ac.kr.

Authors' Contributions:
Ho Jung Choi: Conceived and designed the study, verified and analyzed the data and wrote the paper. Yeong Eun Kim: Conceived and designed the study, verified and analyzed the data and wrote the paper. Jung-Man Namgoong: Conceived and designed the study, verified and analyzed the data and wrote the paper, participated in the management of the patients and analyzed the data. Inki Kim: Verified and analyzed the data and wrote the paper. Jun Sung Park: Reviewed the manuscript. Woo Im Baek: Reviewed the manuscript. Byong Sop Lee: Participated in the management of the patients and analyzed the data. Hee Mang Yoon: Reviewed the manuscript. Young Ah Cho: Reviewed the manuscript. Jin Seong Lee: Reviewed the manuscript. Jung Ok Shim: Reviewed the manuscript. Seak Hee Oh: Conceived and designed the study, verified and analyzed the data and wrote the paper, participated in the management of the patients and analyzed the data. Jae Sung Ko: Verified and analyzed the data and wrote the paper, participated in the management of the patients and analyzed the data. Dae Yeon Kim: Reviewed the manuscript, participated in the management of the patients and analyzed the data. Kyung Mo Kim: Participated in the management of the patients and analyzed the data.

Conflicts of Interest:
The authors disclose no conflicts.

Funding:
This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR21C0198).

Ethical Statement:
The corresponding author, on behalf of all authors, jointly and severally, certifies that their institution has approved the protocol for any investigation involving humans or animals and that all experimentation was conducted in conformity with ethical and humane principles of research.

Data Transparency Statement:
Supplementary data are available at Gastro Hep Advances online. The final model will be available to other reseachers (URL: https://seakheeoh76.shinyapps.io/XGBoost_BA_prediction/). The data underlying this article will be shared on reasonable request to the corresponding author.

Reporting Guidelines:
Helsinki Declaration, TRIPOD.