Retrospective Cohort Study

INTERNATIONAL JOURNAL OF SURGERY

OPEN

Online interpretable dynamic prediction models for clinically significant posthepatectomy liver failure based on machine learning algorithms: a retrospective cohort study

Yuzhan Jin, MSc[a,b], Wanxia Li, MSc[a,b], Yachen Wu, MSc[f], Qian Wang, MD[d], Zhiqiang Xiang, MSc[e], Zhangtao Long, MSc[f], Hao Liang, MSc[f], Jianjun Zou, PhD[b,c,*], Zhu Zhu, MD[f,*], Xiaoming Dai, MD[f,*]

Background: Posthepatectomy liver failure (PHLF) is the leading cause of mortality in patients undergoing hepatectomy. However, practical models for accurately predicting the risk of PHLF are lacking. This study aimed to develop precise prediction models for clinically significant PHLF.

Methods: A total of 226 patients undergoing hepatectomy at a single center were recruited. The study outcome was clinically significant PHLF. Five preoperative and postoperative machine learning (ML) models were developed and compared with four clinical scores, namely, the MELD, FIB-4, ALBI, and APRI scores. The robustness of the developed ML models was internally validated using fivefold cross-validation (CV) by calculating the average of the evaluation metrics and was externally validated on an independent temporal dataset, including the area under the curve (AUC) and the area under the precision–recall curve (AUPRC). SHapley Additive exPlanations analysis was performed to interpret the best performance model.

Results: Clinically significant PHLF was observed in 23 of 226 patients (10.2%). The variables in the preoperative model included creatinine, total bilirubin, and Child–Pugh grade. In addition to the above factors, the extent of resection was also a key variable for the postoperative model. The preoperative and postoperative artificial neural network (ANN) models exhibited excellent performance, with mean AUCs of 0.766 and 0.851, respectively, and mean AUPRC values of 0.441 and 0.645, whereas the MELD, FIB-4, ALBI, and APRI scores reached AUCs of 0.714, 0.498, 0.536, and 0.551, respectively, and AUPRC values of 0.204, 0.111, 0.128, and 0.163, respectively. In addition, the AUCs of the preoperative and postoperative ANN models were 0.720 and 0.731, respectively, and the AUPRC values were 0.380 and 0.408, respectively, on the temporal dataset.

Conclusion: Our online interpretable dynamic ML models outperformed common clinical scores and could function as a clinical decision support tool to identify patients at high risk of PHLF preoperatively and postoperatively.

Keywords: artificial neural network, dynamic prediction, machine learning, posthepatectomy liver failure, preoperative and postoperative models

Introduction

Partial hepatectomy represents an effective treatment for a variety of benign and malignant liver diseases[1–3]. Over the past few decades, although its safety has improved significantly due to tremendous advancements in surgical techniques and perioperative management, posthepatectomy liver failure (PHLF), which is a serious complication, remains unavoidable[4,5]. The incidence

[a]School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, [b]Department of Clinical Pharmacology, Nanjing First Hospital, Nanjing Medical University, Nanjing, [c]Department of Pharmacy, Nanjing First Hospital, China Pharmaceutical University, Nanjing, Jiangsu, [d]Department of Reproductive Medicine, The First Affiliated Hospital, Department of Reproductive Medicine, Hengyang Medical School, University of South China, Hengyang, [e]Department of Hepatobiliary Surgery, Hunan University of Medicine General Hospital, Huaihua and [f]Department of Hepatobiliary Surgery, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, Hunan, People's Republic of China

Yuzhan Jin, Wanxia Li, and Yachen Wu contributed equally to this work. Yuzhan Jin, Wanxia Li, and Yachen Wu are the co-first authors.

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding author. Address: Changle Road No.68, Qinhuai District, Nanjing, Jiangsu 210000, People's Republic of China. Tel.: +86 15380998951. E-mail: Zoujianjun100@126.com (J. Zou); ChuanShan Road No.69, Shigu District, Hengyang, Hunan 421001, People's Republic of China. Tel.: +86 0734 8578563. E-mail: zhuzhu027@gmail.com (Z. Zhu), and Tel.: +86 0734 8578561. E-mail: fydaixiaoming@126.com (X. Dai).

International Journal of Surgery (2024) 110:7047–7057

Received 17 April 2024; Accepted 27 May 2024

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww.com/international-journal-of-surgery.

Published online 18 June 2024

http://dx.doi.org/10.1097/JS9.0000000000001764

of PHLF varies between 1.2 and 32%, with clinically significant PHLF being the leading cause of early postoperative mortality[6,7]. In addition, clinically significant PHLF also leads to prolonged hospitalization, increased healthcare costs, and reduced long-term survival[8]. Therefore, precise prediction through the identification of risk factors for clinically significant PHLF is critical to decrease the incidence of PHLF.

Due to the absence of universally recognized methods, it is still challenging to accurately predict PHLF. Clinical risk scores based on blood tests, including the model for end-stage liver disease (MELD), fibrosis-4 index (FIB-4), albumin-bilirubin (ALBI), and aspartate aminotransferase-to-platelet ratio index (APRI) scoring systems, are widely used to assess liver reserve function[9]. However, owing to the limited number of variables, these methods have several limitations for the prediction of PHLF[10–13] and have unsatisfactory performance, with the area under the receiver operating characteristic curve (AUC) ranging from 0.5 to 0.7[14–16]. In recent years, several nomogram models using the logistic regression (LR) algorithm have been developed to predict PHLF[17–20]. Despite having more included variables, they cannot precisely describe the interactions between different risk factors or address the nonlinear relationship between variables and outcomes[21–23]. Therefore, more efficient methods are necessary to improve the predictive performance of PHLF.

Machine learning (ML), a crucial subfield of artificial intelligence, is capable of generating empirical models through big data to dig deeper into the relationships between risk factors and address nonlinearities arising from data[24,25]. Currently, only a few studies have developed predictive models for PHLF based on ML algorithms using variables such as perioperative clinical features, computed tomography (CT), or MRI-based radiomic features[14–16]. However, these studies still have several limitations, including the use of a model based on a single ML algorithm without comparing the ML algorithm with other algorithms for predictive performance, the use of a model focused exclusively on prediction after surgery, and the applicability of the prediction model, which requires specialized facilities and experienced professionals. Thus, there is an urgent need to establish a PHLF prediction model that is more accurate, applicable, and capable of preoperative and postoperative prediction based on the comparison of multiple ML algorithms.

In this study, we included various important perioperative features and screened key risk factors to develop preoperative and postoperative prediction models for clinically significant PHLF using multiple ML algorithms. Our study aimed to identify patients at high risk of PHLF before and after surgery, which will aid in reducing the incidence of PHLF in clinical practice.

HIGHLIGHTS

- Currently, there is still a lack of an efficient model for precisely predicting the risk of clinically significant post-hepatectomy liver failure (PHLF).
- Five preoperative and postoperative machine learning models were constructed and compared with four clinical scores, namely, the MELD, FIB-4, ALBI, and APRI scores.
- The established preoperative and postoperative artificial neural network (ANN) models, based on creatinine, total bilirubin, Child–Pugh grade, and extent of resection, exhibited an excellent predictive performance that was better than that of the four clinical scores and had the potential to predict clinically significant PHLF.
- The ANN models were internally and externally validated and deployed to the cloud and could be utilized as a clinical-decision support tool to dynamically identify patients at high risk for clinically significant PHLF.

Methods

This study strictly followed the TRIPOD guidelines[26], and our work has been reported in line with the strengthening the reporting of cohort, cross-sectional, and case–control studies in surgery (STROCSS) criteria[27] (Supplemental Digital Content 1, http://links.lww.com/JS9/C788). This study was registered in the Chinese Clinical Trial Registry (ChiCTR2400083151, https://www.chictr.org.cn/showproj.html?proj=222899). The overall workflow is illustrated in Figure 1.

Patient population

The study protocol complied with the Declaration of Helsinki and received approval from the (Anonymised) Ethics (No. Anonymised). The requirement for informed consent was waived by the ethics committee because the study was retrospective.

The patient population originated from a cross-sectional study in the Anonymised Hospital between March 2021 and October 2023. Initially, we recruited 362 patients who underwent therapeutic hepatectomy by a single surgical team. The inclusion criteria were as follows: (1) 18–85 years of age; (2) preoperative Child–Pugh grade A or B; and (3) no cardiopulmonary or renal insufficiency or hepatic encephalopathy preoperatively. The exclusion criteria were as follows: (1) preoperative biliary obstruction and (2) two-stage hepatectomy.

Data collection

In this study, we collected patient demographic, comorbidities, preoperative laboratory, intraoperative, and postoperative data from the electronic medical records. Patient demographic variables included age, sex, and BMI. The comorbidities included diabetes, hypertension, fatty liver, viral hepatitis, cirrhosis, splenomegaly, ascites, esophageal and gastric varices (EGV), preoperative Child–Pugh grade, and the use of hepatoprotective agents. Preoperative laboratory tests included white blood cell (WBC) count, platelet count (PLT), hematocrit (HCT), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin (TBIL), direct bilirubin (DBIL), triglyceride, albumin (ALB), γ glutamyl transpeptidase (GGT), indocyanine green retention rate at 15 min (ICG15), interleukin 6 (IL-6), creatinine (CREA), prothrombin time (PT), fibrinogen (FIB), international normalized ratio (INR), c-reactive protein (CRP), procalcitonin (PCT), erythrocyte sedimentation rate (ESR), hyaluronic acid (HA), laminin (LN), type IV collagen (IV CoI), and type III procollagen peptide (PIIINP). Intraoperative variables included blood loss, operative time, laparoscopy, extrahepatic bile duct resection, and extent of resection (major resection, ≥ 3 segments; minor resection, <3 segments). Postoperative variables included future liver remnant (FLR), standardized future liver remnant (sFLR), and postoperative biliary obstruction.

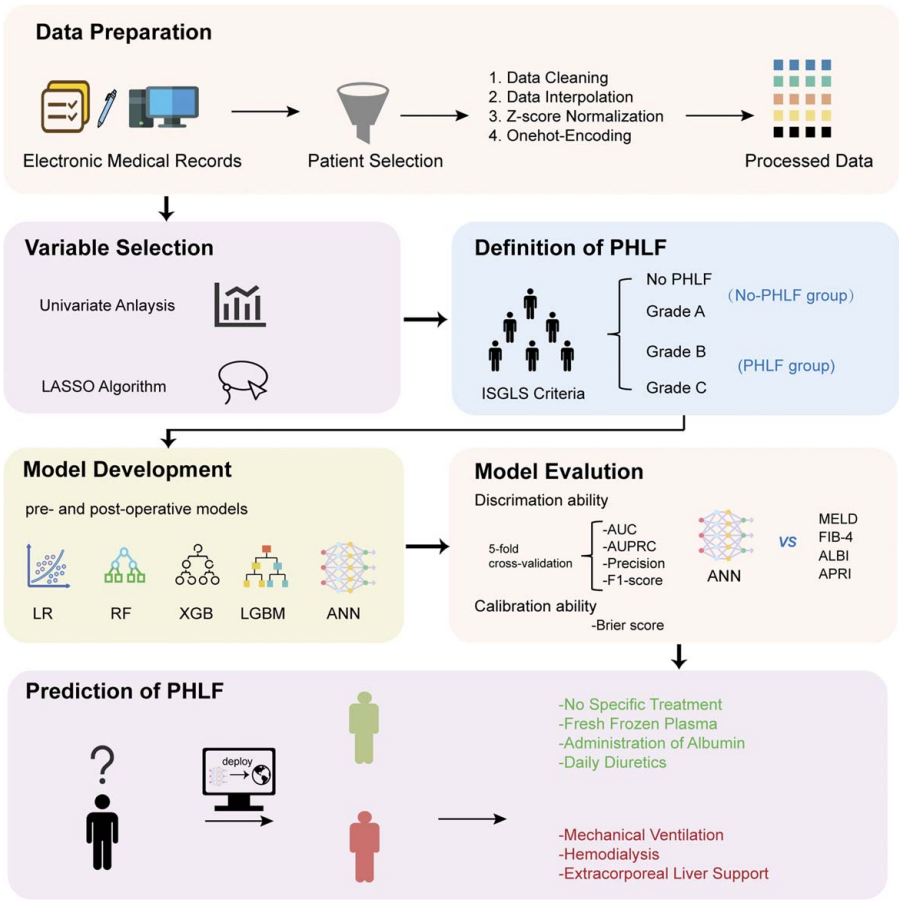

Figure 1. Schematic of the study workflow.

In addition, four clinical scores, namely, the MELD, FIB-4, ALBI, and APRI, were calculated as follows[11,28–30]: $\text{MELD} = 3.78 \times \ln_{\text{TBIL(mg/dL)}} + 11.2 \times \ln_{\text{INR}} + 9.57 \times \ln_{\text{CREA(mg/dL)}} + 6.43$; FIB-4 = [age (years) × AST (IU/L)]/[1/2 × platelets (10^9/L) × ALT (IU/L)]; ALBI = 0.66 × lgTBIL(μmol/L)-0.085 × ALB(g/L); APRI = 100 × AST (IU/L)/PLT (10^9/L).

Sample size calculation

In this research, the sample size of the binary outcome prediction model was calculated according to the following formula[31]: $n = \exp\left(\frac{-0.508 + 0.259\ln(\varphi) + 0.504\ln(P) - \ln(MAPE)}{0.544}\right)$. In this formula, φ is the anticipated outcome proportion (φ = 0.1), P is the number of candidate predictor parameters ($P = 4$), and MAPE is the average absolute error between the observed and true outcome probability (MAPE = 0.05). According to calculations, the minimum sample size required for this research is 260. The sample size of our study is in accordance with the requirements.

Definition of PHLF

PHLF was defined as increased serum TBIL and INR after postoperative day 5 according to the International Study Group of Liver Surgery (ISGLS) diagnostic criteria[32]. PHLF severity was classified into three levels: Grade A, requiring no specific treatment; Grade B, requiring essential noninvasive treatment; and Grade C, requiring invasive treatment. In this study, a positive outcome (PHLF) was defined as PHLF grades B and C. Otherwise, a negative outcome (non-PHLF) was defined as no PHLF or PHLF grade A.

Data preprocessing

First, we interpolated the variables with missing values below the threshold of 25% and then removed those with missing values above the threshold. To interpolate the missing values, we employed the K-nearest neighbor (KNN) method for the continuous variables and the mode interpolation method for the categorical variables. Univariate analysis was conducted to select the variables associated with PHLF ($P < 0.1$). Next, the least absolute shrinkage and selection operator (LASSO) algorithm was used to identify the independent factors of PHLF. The LASSO algorithm, employing the hyperparameter lambda (λ), shrinks the regression coefficients of the redundant variables close to zero and finally selects those variables with nonzero regression coefficients[33]. We performed collinear analysis to evaluate multicollinearity among the selected variables using the variance inflation factor (VIF, <2 indicating no significant multicollinearity). Finally, each continuous variable underwent preprocessing for Z score normalization, while each categorical variable was transformed by one-hot encoding.

Model development and evaluation

In this study, we developed and validated five preoperative and postoperative ML models for predicting PHLF, including LR, random forest classifier (RFC), extreme gradient boosting (XGB), light gradient boosting machine (LGBM), and artificial neural network (ANN) models. A grid search with 10-fold CV was used to determine the optimal hyperparameters for each model. The hyperparameters of each model were selected to maximize the AUC. All ML models were developed using the 'sklearn 1.1.2', 'xgboost 1.6.1', 'lightgbm 3.3.2', and 'Keras 2.9.0' packages in Python 3.9.12.

The performance of the models was evaluated by discrimination and calibration metrics. The discrimination ability was assessed by the AUC, the area under the precision-recall (PR) curve (AUPRC), the precision and the F1-score. The AUPRC is more valuable for evaluating binary classifiers on unbalanced datasets than is the AUC. The F1-score is a comprehensive metric combining precision and recall. These four discriminative indicators were calculated as an average using fivefold CV. CV method is the most practical and flexible method that can be used for model evaluation and selection, and the basic idea is to divide the data repetitively into training data and validation data for estimating model parameters and performance evaluation[34–36]. The final performance evaluation metric results provided for each model are average values using fivefold CV across the dataset. Due to the conservative estimation of performance metrics, it helps to accurately evaluate the performance of our established ML models.

The calibration ability was measured by the Brier score, which ranged from 0 to 1, with lower scores indicating better fitness of the model. Therefore, the AUPRC, F1-score, and Brier score are the main evaluation metrics for selecting the best-performing model. The optimal threshold of each model was determined by the Youden index (Youden index = sensitivity + specificity–1).

Temporal external validation

The generation of the final selected best-performing preoperative and postoperative models was further temporally validated by the later independently collected dataset from December 2023 to January 2024 at the same medical center. The AUC, AUPRC, and Brier score were used to evaluate the performance of the prediction model.

Model interpretation

SHapley Additive exPlanations (SHAP) analysis[37] was used to interpret the output of the best-performing model in our study based on the Shapley values. It provides interpretation to our established model, and can contribute to apply the model in clinical practice[38]. The SHAP value, which was calculated as an average of a variable's contributions through all possible combinations of variables, could be either positive or negative, indicating an increased or decreased probability of the output. SHAP plots were generated using the 'shap 0.42.1' package in Python 3.9.12. The best-performing models were further packaged into an R Shiny-based application for use.

Statistical analysis

The Shapiro–Wilk test was used to assess the normality of continuous variables. Continuous variables are expressed as the mean ± SD or median ± interquartile range (IQR) and were compared using Student's *t*-test or the Mann–Whitney *U* test. Additionally, categorical variables are represented as counts (frequencies) and were compared with Fisher's exact test or the χ^2 test. All tests conducted in this study were two-tailed, and $P < 0.05$ was considered to indicate statistical significance. R software (version 4.2.1) was used to perform all the statistical analyses.

Results

Patient population

A total of 226 patients who met the inclusion criteria were ultimately enrolled in our study. The detailed patient characteristics are described in Table 1. The median patient age was 57 years (range 49–65 years), and 43.4% were female. The incidence of positive outcome (PHLF) was 10.2% (23/226).

Variable selection

In our study, missing situation for the variables with missing values in the original dataset were shown in Supplemental Table S1 (Supplemental Digital Content 2, http://links.lww.com/JS9/C789). Using univariate analysis, we selected seven variables related to the outcome, namely, diabetes, ascites, preoperative Child–Pugh grade, TBIL, DBIL, CREA, and extent of resection. Through the LASSO algorithm, three variables significantly related to the outcome, namely, the preoperative Child–Pugh grade, TBIL, and CREA, were then included to construct the preoperative ML models. To construct postoperative ML models, we ultimately selected four variables significantly related to the outcome, including preoperative Child–Pugh grade, TBIL, CREA, and extent of resection, and there was no significant collinearity among these four variables (see Supplemental Table 2, Supplemental Digital Content 2, http://links.lww.com/JS9/C789).

Model performance

The above variables were entered into our objective to develop five different preoperative and postoperative ML models, including the LR, RFC, XGB, LGBM, and ANN models. The best hyperparameters of each preoperative and postoperative model were detailed (see Supplemental Table S3, Supplemental Digital Content 2, http://links.lww.com/JS9/C789 and Table S4, Supplemental Digital Content 2, http://links.lww.com/JS9/C789).

Among the five preoperative ML models, the ANN model with fivefold CV had the highest AUPRC and F1-score, as well as the second-best AUC of 0.766 (see Table 2, Fig. 2). Additionally, the ANN model showed the second-best calibration ability, with a Brier score of 0.072 (Fig. 3A).

Among the five postoperative ML models, the ANN model with fivefold CV exhibited the highest AUPRC, F1-score, and precision and achieved the second-best AUC of 0.851 (see Table 2, Fig. 2). Furthermore, the ANN model demonstrated the best calibration ability, with a Brier score of 0.034 (Fig. 3B). Thus, we selected the ANN model as having the best performance among the preoperative and postoperative models. Moreover, the MELD, FIB-4, ALBI, and APRI reached AUCs of 0.714, 0.498, 0.536, and 0.551, respectively, and AUPRC values of 0.204, 0.111, 0.128, and 0.163, respectively (Fig. 4A-B).

Table 1

Comparison of the demographics, comorbidities, preoperative laboratory tests, intraoperative, and postoperative variables between patients with or without PHLF in the whole cohort

	Overall ($n=226$)	Non-PHLF ($n=203$)	PHLF ($n=23$)	*P*
Demographics				
Age, year	57 (49–65)	57 (47–65)	56 (52–65)	0.365
BMI, kg/m^2	21.94 (20.56–23.70)	21.91 (20.61–23.69)	22.49 (18.09–23.65)	0.757
Female	98 (43.4)	92 (45.3)	6 (26.1)	0.123
Comorbidities				
Diabetes				0.03*
No	208 (92.0)	190 (93.6)	18 (78.3)	
Yes	18 (8.0)	13 (6.4)	5 (21.7)	
Hypertension				0.235
No	199 (88.1)	181 (89.2)	18 (78.3)	
Yes	27 (11.9)	22 (10.8)	5 (21.7)	
Fatty liver				0.787
No	219 (96.9)	196 (96.6)	23 (100.0)	
Yes	7 (3.1)	7 (3.4)	0 (0.0)	
Cirrhosis				1
No	194 (85.8)	174 (85.7)	20 (87.0)	
Yes	32 (14.2)	29 (14.3)	3 (13.0)	
Viral hepatitis				0.158
None	156 (69)	139 (68.5)	17 (73.9)	
Hepatitis B	63 (27.9)	59 (29.1)	4 (17.4)	
Hepatitis C	7 (3.1)	5 (2.5)	2 (8.7)	
Splenomegaly				0.863
None	202 (89.4)	182 (89.7)	20 (87)	
I	21 (9.3)	18 (8.9)	3 (13)	
II	1 (0.4)	1 (0.5)	0 (0)	
III	2 (0.9)	2 (1.0)	0 (0)	
Ascites				0.011*
None	217 (96.0)	196 (96.6)	21 (91.3)	
Small mount	8 (3.5)	7 (3.4)	1 (4.3)	
Large mount	1 (0.4)	0 (0.0)	1 (4.3)	
EGV				1
No	220 (97.3)	198 (97.5)	22 (95.7)	
Yes	6 (2.7)	5 (2.5)	1 (4.3)	
Preoperative Child-Pugh grade				0.017*
Grade A	169 (74.8)	157 (77.3)	12 (52.2)	
Grade B	57 (25.2)	46 (22.7)	11 (47.8)	
Hepatoprotective agents				0.349
No	83 (36.7)	72 (35.5)	11 (47.8)	
Yes	143 (63.3)	131 (64.5)	12 (52.2)	
Preoperative laboratory tests				
WBC, 10^9/l	5.60 (4.43–7.08)	5.51 (4.46–7.08)	6.58 (4.38–7.30)	0.37
PLT, 10^9/l	188.5 (127.25–239.75)	189.0 (130.5–239.5)	169.0 (125.50–227.0)	0.522
HCT	0.40 (0.36–0.43)	0.40 (0.36–0.43)	0.40 (0.36–0.44)	0.52
ALT, U/l	21.40 (13.25–38.38)	21.40 (13.05–37.90)	22.70 (14.70–50.35)	0.442
AST, U/l	25.00 (17.70–37.85)	25.00 (17.60–36.45)	23.90 (19.75–43.00)	0.493
TBIL, μmol/l	10.40 (8.00–15.55)	10.30 (7.80–14.52)	15.00 (11.95–26.55)	0.003*
DBIL, μmol/l	4.25 (3.40–6.18)	4.20 (3.35–6.10)	5.40 (4.10–11.50)	0.019*
Triglyceride, mmol/l	1.21 (0.95–1.63)	1.20 (0.94–1.63)	1.34 (1.12–1.53)	0.387
ALB, g/l	42.85 (39.40–46.68)	42.80 (39.50–46.55)	43.63 (39.15–47.10)	0.652
CREA, μmol/l	73.00 (62.00–85.75)	73.00 (60.50–85.00)	75.00 (69.00–98.00)	0.063*
PT, s	13.10 (12.20–14.00)	13.10 (12.20–14.00)	12.90 (11.75–13.60)	0.495
FIB, g/l	3.00 (2.50–3.65)	3.00 (2.50–3.68)	2.94 (2.69–3.17)	0.794
$10 \times$ INR	10.3 (9.5–11.0)	10.30 (9.50–11.00)	10.60 (9.95–11.10)	0.415
CRP, mg/l	1.29 (0.40–5.42)	1.20 (0.41–5.39)	2.60 (0.42–5.75)	0.714
Intraoperative variables				
Blood loss, ml	150.0 (50.0–200.0)	150.0 (50.0–200.0)	100.0 (50.00–175.0)	0.301
Operative time, min	270.0 (200.0–348.17)	275.0 (201.5–350.0)	240.0 (186.67–317.5)	0.358
Laparoscopy				0.518
No	35 (15.5)	33 (16.3)	2 (8.7)	
Yes	191 (84.5)	170 (83.7)	21 (91.3)	
Extrahepatic bile duct resection				1

Table 1

(*Continued*)

	Overall ($n=226$)	Non-PHLF ($n=203$)	PHLF ($n=23$)	*P*
No	224 (99.1)	201 (99.0)	23 (100.0)	
Yes	2 (0.9)	2 (1.0)	0 (0)	
Extent of resection				0.044*
Minor	179 (79.2)	165 (61.3)	14 (60.9)	
Major	47 (20.8)	38 (18.7)	9 (39.1)	
Postoperative variables				
FLR, ml	799.98 (736.42–875.47)	798.15 (739.56–870.52)	803.62 (709.08–920.63)	0.881
sFLR	0.79 (0.73–0.85)	0.79 (0.73–0.85)	0.80 (0.74–0.88)	0.58
Biliary obstruction				1
No	225 (99.6)	202 (99.5)	23 (100.0)	
Yes	1 (0.4)	1 (0.5)	0 (0.0)	

Data are presented as *n* (%), mean (SD), or median (IQR). Variables with normal distribution are presented as mean (SD); Variables without normal distribution are presented as median (interquartile range, IQR). ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; CREA, creatinine; CRP, C-reactive protein; DBIL, direct bilirubin; EGV, esophageal and gastric varices; FIB, fibrinofen; FLR, future liver remnant; HCT, hematocrit; INR, international normalized ratio; PHLF, posthepatectomy liver failure; PLT, platelet count; PT, prothrombin time; sFLR, standardized future liver remnant; TBIL, total bilirubin; WBC, white blood cell count.

Model interpretation

To determine the best performing model among the preoperative and postoperative models, SHAP analysis was performed to assess the contribution of each variable. The top three variables for the preoperative model and the top four variables for the postoperative model are depicted in Figure 5A-B. An overview of the positive or negative contributions of variables to the ANN model is shown in Figure 5C-D.

Temporal external validation

The data of 31 patients were included in the temporal external validation cohort. The trained preoperative ANN model achieved an AUC of 0.720, an AUPRC of 0.380, and a Brier score of 0.382. The trained postoperative ANN model reached an AUC of 0.732, an AUPRC of 0.408, and a Brier score of 0.362 (Fig. 6A-C).

Table 2

The average performance of discrimination metrics of the five different preoperative and postoperative ML models by fivefold cross-validation

	Optimal cut-off	AUC	AUPRC	Precision	F1-score
LR					
Preoperative	0.165	0.714	0.335	0.221	0.280
Postoperative	0.103	0.718	0.325	0.194	0.286
RFC					
Preoperative	0.144	0.673	0.247	0.193	0.279
Postoperative	0.149	0.702	0.256	0.215	0.308
XGB					
Preoperative	0.166	0.693	0.271	0.189	0.243
Postoperative	0.151	0.693	0.268	0.172	0.226
LGBM					
Preoperative	0.166	0.841	0.407	0.102	0.185
Postoperative	0.151	0.877	0.516	0.102	0.185
ANN					
Preoperative	0.111	0.766	0.441	0.194	0.297
Postoperative	0.160	0.851	0.645	0.444	0.521

ANN, artificial neural network; AUC, the area under the receiver operating characteristic curve; AUPRC, the area under the precision-recall (PR) curve; LGBM, light gradient boosting machines; LR, logistic regression; ML, machine learning; RFC, random forest classifier; XGB, extreme gradient boosting.

Model deployment

In order to enhance the clinical utility of the developed predictive models, the preoperative and postoperative ANN models were deployed to the cloud (see URLs, https://cheason.shinyapps.io/ANN_pre/; https://cheason.shinyapps.io/ANN_post/). The screenshots of the web-based tools are shown in Supplemental Figure S1 (Supplemental Digital Content 2, http://links.lww.com/JS9/C789).

Discussion

To predict the risk of clinically significant PHLF, we developed preoperative and postoperative prediction models based on comprehensive perioperative factors and multiple ML algorithms with good accuracy and ease of use. After evaluating multiple model metrics, we determined that the ANN model performed the best and selected it as the final risk prediction model for PHLF. Additionally, the predictive performance of the ANN model far exceeded that of commonly used clinical scores. To our knowledge, this is the first study that provides preoperative and postoperative dynamic risk prediction models for clinically significant PHLF, combining ML algorithms with multidimensional clinical features.

First, one advantage of our study is that the risk prediction model we constructed can predict PHLF accurately and efficiently. Through the comparison of multiple ML algorithms using various evaluation metrics, we found that the ANN model showed the best overall prediction performance and outperformed clinical models such as the MELD, FIB-4, ALBI, and APRI. The ANN algorithm is capable of effectively handling complex nonlinear relationships between variables by simulating the structure of biological neural networks. Notably, the ANN model, which has been shown to be more effective than traditional discriminant analysis, has become an alternative or even a new standard for the prediction of disease risk[39–41]. Recently, Lu *et al.*[15] developed an ANN-based model to predict the risk of PHLF in patients with hepatocellular carcinoma, and this model demonstrated good predictive performance. These findings are consistent with our study and further support our conclusions.

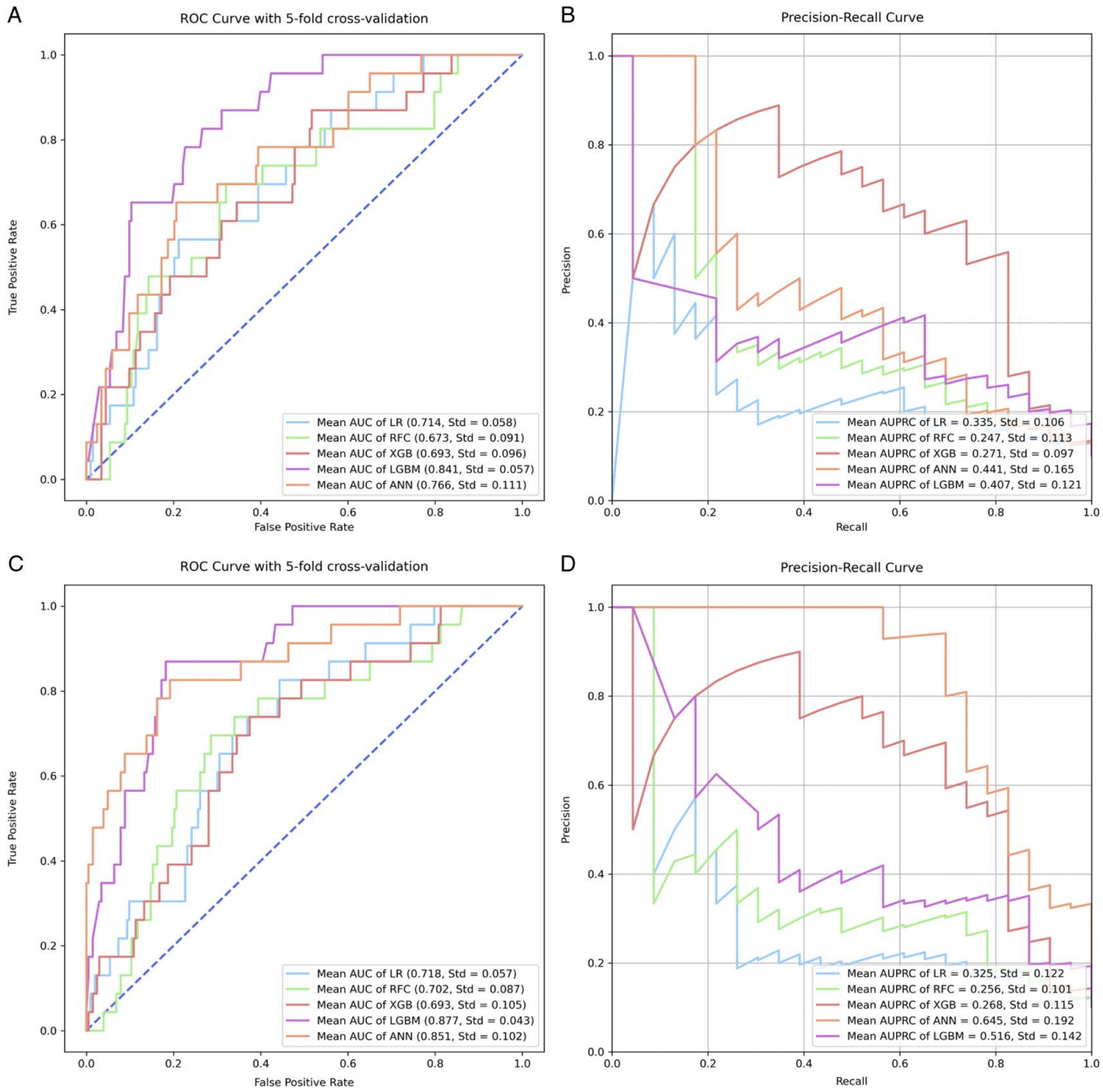

Figure 2. ROC curves and PRCs of the machine learning models for predicting PHLF. ROC curve with fivefold cross-validation (A, C) and PRC curve (B, D) of the preoperative and postoperative models for the whole cohort. PHLF, posthepatectomy liver failure.

Second, our prediction model can achieve multitime point risk stratification. Our risk prediction models for PHLF showed comparable predictive performance between the preoperative and postoperative periods. Considering the life-threatening potential and severity of PHLF, it is necessary to develop an accurate preoperative model to identify patients at high risk at an early stage. Unfortunately, most ML-based studies have focused exclusively on developing postoperative models to predict PHLF. Remarkably, only one model built by Mai *et al.*[14] was available for use in the preoperative period, but it was still unclear whether the model had satisfactory performance in predicting clinically significant PHLF postoperatively. Therefore, our prediction model, which is capable of multitime point risk stratification, can potentially be used in clinical practice to facilitate optimized clinical decision-making and early personalized interventions.

In addition, our risk prediction models are simple and easy to use. In this study, we constructed a preoperative model based on three variables, namely, the preoperative CREA, TBIL, and Child–Pugh score, while the postoperative model further

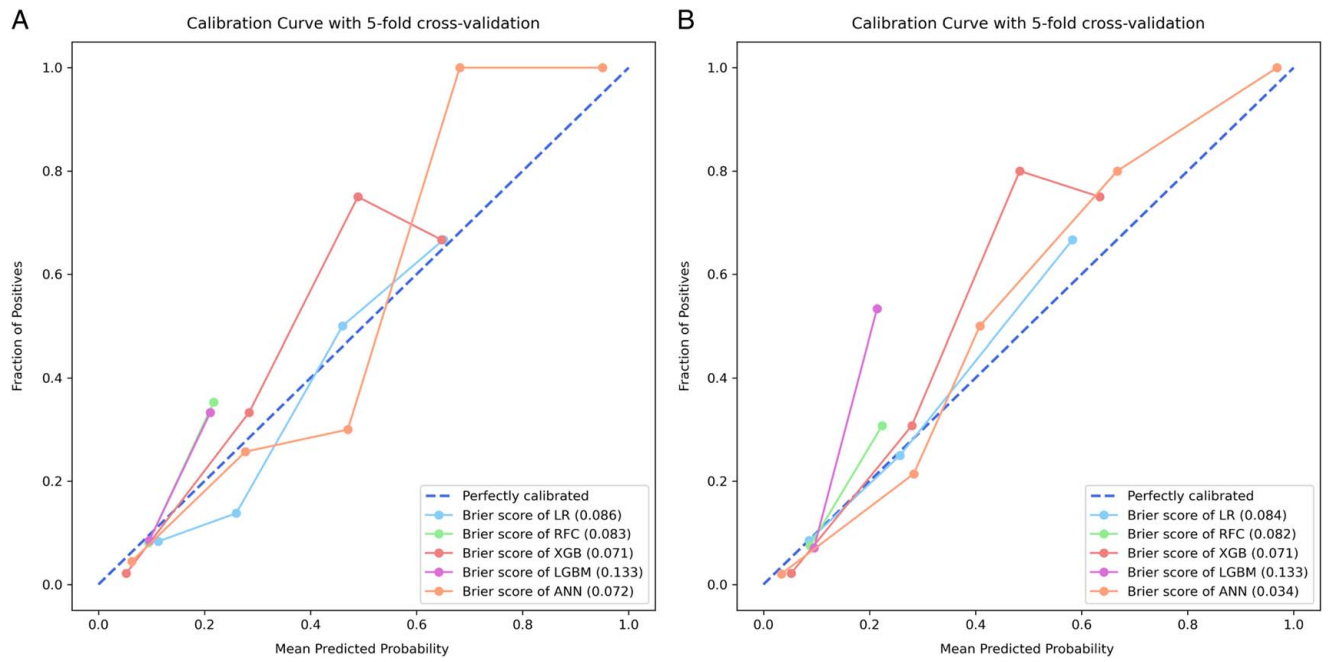

Figure 3. Calibration curve for PHLF. Calibration curves (A, B) of the preoperative and postoperative PHLF models. PHLF, posthepatectomy liver failure.

included the extent of resection. Despite the relatively small number of variables, their predictive effects are comparable to those of other predictive models[14–16]. Hence, the variables included in our model are simple, easy to obtain, and do not rely on expensive or difficult-to-obtain data such as radiomic features. More importantly, to enhance the interaction between the model and the user, we provided a graphical user interface for preoperative and postoperative ANN models on a web page. By accessing the web page online, clinicians are capable of achieving dynamic and instant assessment of the preoperative and postoperative PHLF risk for a specific patient, respectively. The web-based tool we developed can also be used on portable devices

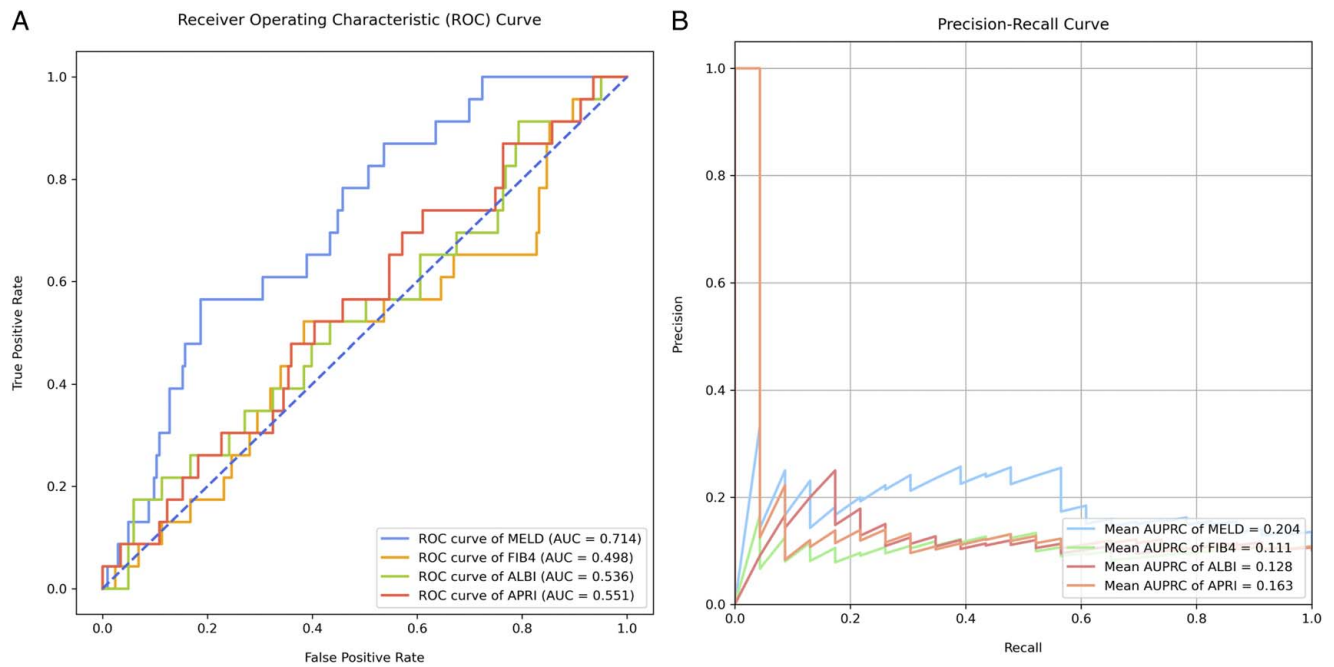

Figure 4. ROC and AUPRC curves of the four clinical scores for predicting PHLF. ROC curve for the four scores (A), AUPRC curve for the four scores (B). PHLF, posthepatectomy liver failure.

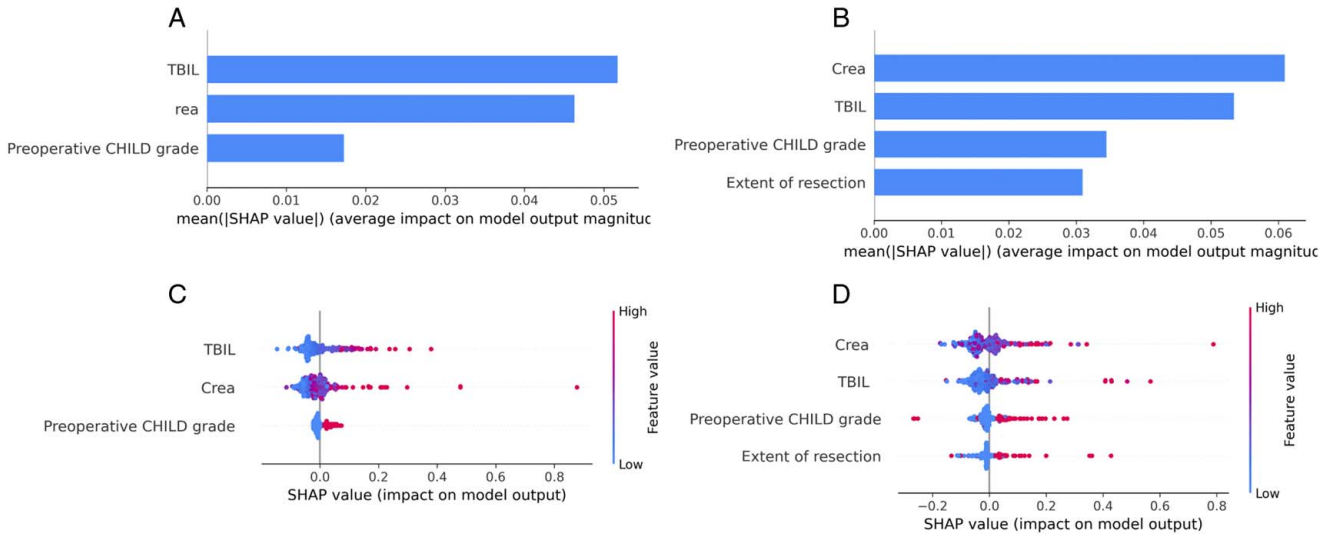

Figure 5. SHAP summary plots of the interpretations of the significant predictors contributing to the preoperative and postoperative ANN models. Bar chart of the average absolute SHAP value for each significant predictor of the preoperative and postoperative ANN models (A, B). Dot chart of each significant predictor contributing to the preoperative and postoperative ANN models (C, D). SHAP, SHapley Additive explanation; ANN, artificial neural network.

such as mobile phones and tablets, or even be integrated into electronic medical record systems to fulfill automatic valuation at any location and time. Therefore, our predictive models can not only reduce user workload and improve prediction efficiency but also reduce prediction costs and increase patient adherence. Even for hospitals in rural and remote areas, our models can also be effectively applied without specialized equipment and personnel. In summary, the models we developed are suitable for routine use in clinical practice.

Our study is the first to address the importance of the management of preoperative CREA, TBIL, and Child–Pugh grade to reduce PHLF risk. In this study, the SHAP algorithm showed that the preoperative CREA, TBIL, and preoperative Child–Pugh grade were significant in both the preoperative and postoperative models. Although previous studies have reported correlations between these factors and PHLF, they have not emphasized the importance of their preoperative management[42]. Our study indicated that the preoperative management of these three factors is crucial for predicting PHLF. Therefore, physicians should focus on these factors and provide relevant preventive or interventional treatments during the preoperative period, especially for patients with high preoperative CREA levels, TBIL levels, or high Child–Pugh grades.

Our study has several limitations. First, this was a retrospective study conducted at a single center, which may introduce selection bias. Second, the independent datasets utilized for external validation were too small to further test the extrapolation and generalizability of our constructed models. Therefore, larger, prospective, multicenter studies are needed to validate our ML model in the future.

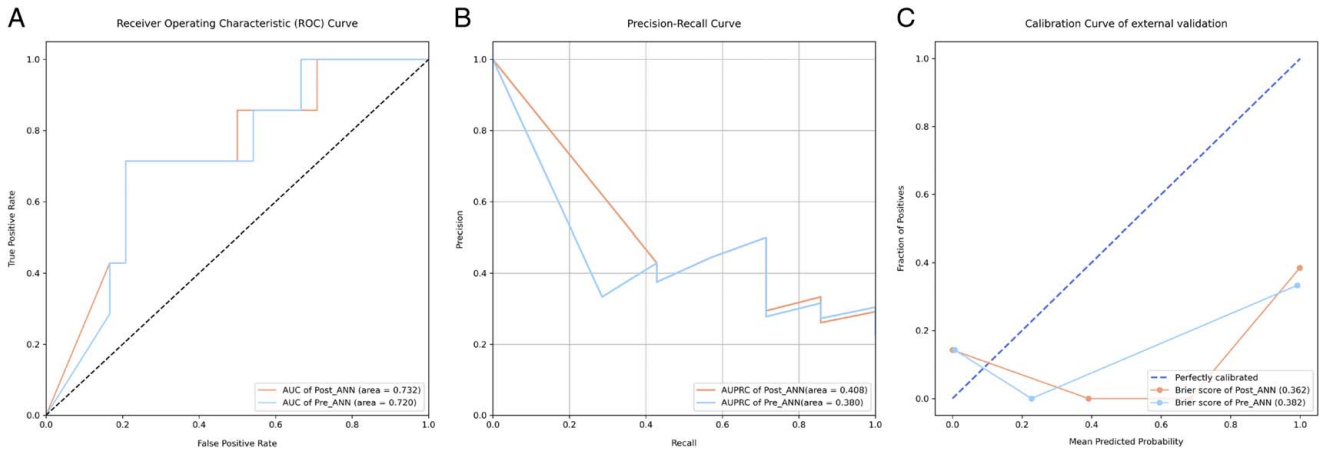

Figure 6. ROC curves, PRC curves and calibration curves of the pre- and postoperative ANN models for predicting PHLF in the temporal external validation cohort. The ROC curves (A) of the preoperative ANN model (blue line) and the postoperative ANN model (orange line). The PRC curves (B) of the preoperative ANN model (blue line) and the postoperative ANN model (orange line). The calibration curves plot (C) of the preoperative ANN model (blue line) and the postoperative ANN model (orange line).

Conclusion

Using ML algorithms and a few simple and easily accessible variables, we successfully developed online interpretable and clinically applicable preoperative and postoperative ANN models that can accurately predict the risk of PHLF. Our models show potential utility for identifying patients at high risk of clinically significant PHLF before and after surgery, which will help clinicians perform better personalized clinical decision-making and determine early individualized interventions for patients undergoing hepatectomy.

Ethical approval

The First Affiliated Hospital of University of South China Medical Ethics (No. 2023ll0223001).

Consent

The requirement for informed consent was waived by the ethics committee because the study was retrospective.

Source of funding

This work was supported by grants from the National Natural Science Foundation of China (82173899), Natural Science Foundation of Hunan Province (2022JJ70119), the Clinical Medicine Technological Innovation Leading Project of Hunan Province (2020SK51818), and Jiangsu Pharmaceutical Association (H202108, A2021024, Q202202, JY202207).

Author contribution

Y.J.: writing – original draft, writing – review and editing, formal analysis, and methodology; W.L.: writing – original draft and writing – review and editing; Y.W.: data curation and writing – original draft; Q.W., Z.X., Z.L., and H.L.: data curation; J.Z.: conceptualization, supervision, and writing – review and editing; Z.Z. and X.D.: conceptualization, supervision, writing – review and editing, and resource.

Conflicts of interest disclosure

All authors have no conflicts of interest to declare.

Research registration unique identifying number (UIN)

1. Registry used: http://www.chictr.org.cn/index.aspx.
2. Unique identifying number or registration ID: ChiCTR-2400083151.
3. Hyperlink to your specific registration: https://www.chictr.org.cn/showproj.html?proj=222899.

Guarantor

Xiaoming Dai, Zhu Zhu, and Jianjun Zou.

Data availability statement

Our research team could provide original data under reasonable request and with permission from our corresponding authors.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Acknowledgement

Assistance with the study: none.

References

[1] Lafaro K, Buettner S, Maqsood H, *et al*. Defining post hepatectomy liver insufficiency: where do we stand? J Gastrointest Surg 2015;19:2079–92.
[2] Xiang ZQ, Zhu FF, Zhao SQ, *et al*. Laparoscopic versus open repeat hepatectomy for recurrent hepatocellular carcinoma: a systematic review and meta-analysis of propensity score-matched cohort studies. Int J Surg 2023;109:963–71.
[3] Wang Q, Li HJ, Dai XM, *et al*. Laparoscopic versus open liver resection for hepatocellular carcinoma in elderly patients: systematic review and meta-analysis of propensity-score matched studies. Int J Surg 2022;105:106821.
[4] Melloul E, Hübner M, Scott M, *et al*. Guidelines for perioperative care for liver surgery: enhanced recovery after surgery (ERAS) society recommendations. World J Surg 2016;40:2425–40.
[5] Schreckenbach T, Liese J, Bechstein WO, *et al*. Posthepatectomy liver failure. Dig Surg 2012;29:79–85.
[6] Rahbari NN, Garden OJ, Padbury R, *et al*. Post-hepatectomy haemorrhage: a definition and grading by the International Study Group of Liver Surgery (ISGLS). HPB 2011;13:528–35.
[7] Paugam-Burtz C, Janny S, Delefosse D, *et al*. Prospective validation of the 'fifty-fifty' criteria as an early and accurate predictor of death after liver resection in intensive care unit patients. Ann Surg 2009;249:124–8.
[8] Kauffmann R, Fong Y. Post-hepatectomy liver failure. Hepatobiliary Surg Nutr 2014;3:238–46.
[9] Morandi A, Risaliti M, Montori M, *et al*. Predicting post-hepatectomy liver failure in HCC patients: a review of liver function assessment based on laboratory tests scores. Medicina (Kaunas) 2023;59:1099.
[10] Durand F, Valla D. Assessment of the prognosis of cirrhosis: Child-Pugh versus MELD. J Hepatol 2005;42(Suppl(1)):S100–7.
[11] Johnson PJ, Berhane S, Kagebayashi C, *et al*. Assessment of liver function in patients with hepatocellular carcinoma: a new evidence-based approach-the ALBI grade. J Clin Oncol 2015;33:550–8.
[12] Vallet-Pichard A, Mallet V, Nalpas B, *et al*. FIB-4: an inexpensive and accurate marker of fibrosis in HCV infection. comparison with liver biopsy and fibrotest. Hepatology (Baltimore, Md) 2007;46:32–6.
[13] Wang YY, Zhong JH, Su ZY, *et al*. Albumin-bilirubin versus Child-Pugh score as a predictor of outcome after liver resection for hepatocellular carcinoma. Br J Surg 2016;103:725–34.
[14] Mai RY, Lu HZ, Bai T, *et al*. Artificial neural network model for preoperative prediction of severe liver failure after hemihepatectomy in patients with hepatocellular carcinoma. Surgery 2020;168:643–52.
[15] Lu HZ, Mai RY, Wang XB, *et al*. Developmental artificial neural network model to evaluate the preoperative safe limit of future liver remnant volume for HCC combined with clinically significant portal hypertension. Future Oncol 2022;18:2683–94.
[16] Wang J, Zheng T, Liao Y, *et al*. Machine learning prediction model for post- hepatectomy liver failure in hepatocellular carcinoma: a multicenter study. Front Oncol 2022;12:986867.
[17] Xiang F, Liang X, Yang L, *et al*. CT radiomics nomogram for the preoperative prediction of severe post-hepatectomy liver failure in patients with huge ($\geq$ 10 cm) hepatocellular carcinoma. World J Surg Oncol 2021;19:344.
[18] Wang J, Zhang Z, Shang D, *et al*. A novel nomogram for prediction of post-hepatectomy liver failure in patients with resectable hepatocellular carcinoma: a multicenter study. J Hepatocell Carcinoma 2022;9:901–12.

[19] Lei Z, Cheng N, Si A, *et al*. A novel nomogram for predicting postoperative liver failure after major hepatectomy for hepatocellular carcinoma. Front Oncol 2022;12:817895.
[20] Xu B, Li XL, Ye F, *et al*. Development and validation of a nomogram based on perioperative factors to predict post-hepatectomy liver failure. J Clin Transl Hepatol 2021;9:291–300.
[21] Dhir M, Samson KK, Yepuri N, *et al*. Preoperative nomogram to predict posthepatectomy liver failure. J Surg Oncol 2021;123:1750–6.
[22] Li B, Qin Y, Qiu Z, *et al*. A cohort study of hepatectomy-related complications and prediction model for postoperative liver failure after major liver resection in 1,441 patients without obstructive jaundice. Ann Transl Med 2021;9:305.
[23] Shen YN, Tang TY, Yao WY, *et al*. A nomogram for prediction of posthepatectomy liver failure in patients with hepatocellular carcinoma: a retrospective study. Medicine 2019;98:e18490.
[24] Kawaguchi T, Tokushige K, Hyogo H, *et al*. A data mining-based prognostic algorithm for NAFLD-related hepatoma patients: a nationwide study by the japan study group of NAFLD. Sci Rep 2018; 8:10434.
[25] Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30.
[26] Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ (Clinical research ed) 2015;350: g7594.
[27] Mathew G, Agha R, Albrecht J, *et al*. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. Int J Surg 2021;96:106165.
[28] Freeman RB Jr. Model for end-stage liver disease (MELD) for liver allocation: a 5-year score card. Hepatology (Baltimore, Md) 2008;47:1052–7.
[29] Sterling RK, Lissen E, Clumeck N, *et al*. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology (Baltimore, Md) 2006;43:1317–25.
[30] Shiha G, Ibrahim A, Helmy A, *et al*. Asian-pacific association for the study of the liver (APASL) consensus guidelines on invasive and non-invasive assessment of hepatic fibrosis: a 2016 update. Hepatol Int 2017; 11:1–30.
[31] Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. BMJ (Clinical research ed) 2020;368:m441.
[32] Rahbari NN, Garden OJ, Padbury R, *et al*. Posthepatectomy liver failure: a definition and grading by the International Study Group of Liver Surgery (ISGLS). Surgery 2011;149:713–24.
[33] Vasquez MM, Hu C, Roe DJ, *et al*. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. BMC Med Res Methodol 2016;16:154.
[34] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. 2010:40-79.
[35] Berrar D. Cross-Validation. 2019:542-545.
[36] Emmert-Streib F, Dehmer M. Evaluation of regression models: model assessment, model selection and generalization error. Mach Learn Knowledge Extract 2019;1:521–51.
[37] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:6785–95.
[38] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel, Switzerland) 2020;23:18.
[39] Hu X, Cammann H, Meyer HA, *et al*. Artificial neural networks and prostate cancer--tools for diagnosis and management. Nat Rev Urol 2013;10:174–82.
[40] Dal Moro F, Abate A, Lanckriet GR, *et al*. A novel approach for accurate prediction of spontaneous passage of ureteral stones: support vector machines. Kidney Int 2006;69:157–60.
[41] Snow PB, Kerr DJ, Brandt JM, *et al*. Neural network and regression predictions of 5-year survival after colon carcinoma treatment. Cancer 2001;91(8 Suppl):1673–8.
[42] Sparrelid E, Olthof PB, Dasari BVM, *et al*. Current evidence on posthepatectomy liver failure: comprehensive review. BJS open 2022;6:zrac142.