pISSN: 2508-5778 • eISSN: 2508-5859
Ann Hepatobiliary Pancreat Surg 2025;29:269-278
https://doi.org/10.14701/ahbps.25-046

AHBPS
Annals of Hepato-Biliary-Pancreatic Surgery

Original Article

A comparative study of machine learning models predicting post-hepatectomy liver failure: Enhancing risk estimation in over 25,000 National Surgical Quality Improvement Program patients

Gautham Nair[1], Ali Hadi[1], Kartik Gupta[1], Edward Tran[1], Geerthan Srikantharajah[2], Evelyn Waugh[1,3], Ephraim Tang[1,3,4], Anton Skaro[1,3], Juan Glinka[1,3]

[1]Schulich School of Medicine, University of Western Ontario, London, ON, Canada,
[2]Toronto Metropolitan University, Toronto, ON, Canada,
[3]Department of Surgery, Schulich School of Medicine, London, ON, Canada,
[4]Department of Oncology Schulich School of Medicine, London, ON, Canada

Backgrounds/Aims: Post-hepatectomy liver failure (PHLF) is a significant complication with an incidence rate between 8% and 12%. Machine learning (ML) can analyze large datasets to uncover patterns not apparent through traditional methods, enhancing PHLF prediction and potentially mitigate complications.
Methods: Using the National Surgical Quality Improvement Program (NSQIP) database, patients who underwent hepatectomy were randomized into training and testing sets. ML algorithms, including LightGBM, Random Forest, XGBoost, and Deep Neural Networks, were evaluated against logistic regression. Performance metrics included receiver operating characteristic area under the curve (ROC AUC) and Brier score loss. Shapley Additive exPlanations was used to identify individual variable relevance.
Results: 28,192 patients from 2013 to 2021 who underwent hepatectomy were included; PHLF occurred in 1,305 patients (4.6%). Preoperative and intraoperative factors most contributed to PHLF. Preoperative factors were international normalized ratio > 1.0, sodium < 139 mEq/L, albumin < 3.9 g/dL, American Society of Anesthesiologists score > 2, total bilirubin > 0.65 mg/dL. Intraoperative risks include transfusion requirements, trisectionectomy, operative time > 266.5 minutes, open surgical approach. The LightGBM model performed best with an ROC AUC of 0.8349 and a Brier Score loss of 0.0834.
Conclusions: While topical, the role of ML models in surgical risk stratification is evolving. This paper shows the potential of ML algorithms in identifying important subclinical changes that could affect surgical outcomes. Thresholds explored should not be taken as clinical cutoffs but as a proof of concept of how ML models could provide clinicians more information. Such integration could lead to improved clinical outcomes and efficiency in patient care.

Key Words: Machine learning; Hepatectomy; Liver failure; Risk; National Surgical Quality Improvement Program

Received: March 10, 2025, **Revised:** May 9, 2025,
Accepted: May 23, 2025, **Published online:** July 7, 2025

Corresponding author: Juan Glinka, MD
University Hospital, London Health Sciences Centre, Room C8-005, 339 Windermere Road, London, ON N6A 5A5, Canada
Tel: +1-647-471-5847, Fax: +1-519-633-3858, E-mail: jglinka@uwo.ca
ORCID: https://orcid.org/0000-0003-4373-965X

INTRODUCTION

Liver resection (LR) is a surgical intervention offering a potentially curative treatment for a variety of benign and malignant conditions [1,2]. Despite advancements in surgical techniques and postoperative care that have improved perioperative and oncologic outcomes, LR still carries significant risks, including surgical site infection, bile leak, life-threatening hemorrhage, and death [3,4]. Among these, post-hepatectomy liver failure (PHLF) stands out as a severe but possibly preventable,

life-threatening complication [4].

PHLF is defined as a significant deterioration in liver synthesis, excretion, and detoxification functions, characterized by increased international normalized ratio (INR) and bilirubin levels, occurring after 5 days post-liver surgery [5]. Its incidence is estimated at approximately 8% to 12% [1,2,5], making it a leading cause of mortality following major LR, with mortality rates attributed to PHLF ranging from 60% to 100% [3,5,6]. Approximately 25% of PHLF-related deaths occur within the first-month post-surgery [4].

Identifying patients at higher risk of PHLF is critical to improving postoperative outcomes and reducing mortality rates associated with LR. Traditional risk assessment tools such as the Child-Pugh classification and MELD (Model for End-Stage Liver Disease) score have limitations in accurately predicting PHLF, particularly in capturing the complex interplay of risk factors [7]. Factors such as cholestasis, preoperative chemotherapy, and underlying liver dysfunction are well-established risks, yet there may be additional nuanced factors contributing to PHLF.

Machine learning (ML) techniques offer a promising approach to enhancing predictive accuracy by identifying novel risk factors and complex relationships between risk factors within large datasets [8,9]. Leveraging the comprehensive data of the National Surgical Quality Improvement Program (NSQIP), our study aims to perform a comparative methodological exploration of ML techniques to estimate the likelihood of PHLF post-hepatectomy based on the NSQIP population. Our primary goal is to illustrate how different algorithms detect signals within this large-scale data, contributing to the understanding of ML's evolving role in surgical risk stratification. By exploring various ML algorithms such as LightGBM, logistic regression, Random Forest, XGBoost, and Deep Neural Networks, we seek to uncover key predictors and compare model performance in this specific context.

Using this data and analysis, we developed a ML-based risk calculator tailored to estimate PHLF risk, thus facilitating personalized surgical management and optimizing patient outcomes. By enhancing our understanding of PHLF through advanced analytics, we aim to contribute to the ongoing efforts to improve the safety and efficacy of LR procedures.

MATERIALS AND METHODS

Dataset

This study is a secondary analysis of prospectively collected data from the NSQIP database, focusing on patients who underwent hepatectomy between 2013 and 2021. The NSQIP database is a nationally validated, risk-adjusted, outcomes-based program to measure and improve the quality of surgical care and contains data from over 130 institutions. We included patients who underwent elective hepatectomy procedures for any indication or approach, including but not limited to colorectal liver metastasis, neuroendocrine tumors, cholangiocarcinoma, and hepatocellular carcinoma. ML techniques were applied to the NSQIP dataset to develop models for assessing surgical outcomes and complications post-hepatectomy to better understand patient risk profiles.

This study was approved by the Lawson Research Institute and Schulich School of Medicine's research ethics board ReDA approval number 14002.

Data preprocessing

All data processing and statistical analyses were performed using Python (version 3.10) and the scikit-learn library (version 1.6.1) [10]. Scikit-learn was utilized for ML modeling, feature selection, and evaluation of model performance.

Two types of categorical variables were encountered in the NSQIP dataset: nominal variables that lack any inherent order and ordinal variables that possess a specific order. One-Hot Encoding was used to convert nominal variables into binary outcomes, where each category becomes a separate binary feature. Ordinal Encoding was used to transform ordinal variable features into a single column of integers, preserving the order of the categories.

For missing quantitative data, we imputed values based on the median of the respective feature.

For missing qualitative data, we transformed the missing data in features with more than 55% into an additional 'Unknown' category. Features with less than 55% of data missing were handled using a random imputation function. This method replaces every missing value with a randomly selected non-missing entry from the same column.

Feature selection

The Boruta algorithm was used to identify significant features for predicting postoperative complications [11]. This was accomplished by comparing the importance of each feature with randomized versions known as shadow features. Features that outperformed their shadow counterparts were retained for further analysis. By applying this strategy, the dimensionality of the dataset was reduced, including only the most impertinent columns. This facilitated efficient model training and easier interpretability by focusing on the most relevant predictors.

The features were standardized using the StandardScaler from the sklearn.preprocessing library to ensure fair comparisons between features with different units. This process ensures that the features have a mean of 0 and a variance of 1.

Modeling

The NSQIP database has issues with class imbalance as some of the outcomes being predicted were uncommon within the dataset. Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors (SMOTEENN) was used in order to create a balanced dataset. This technique over-samples the minority class and undersamples the majority class, in order

to balance the classes. This step is necessary to ensure that the model has sufficient data for all classes and is not biased toward the majority class.

The data was divided into an 80% training set and a 20% test set. Five different ML algorithms were deployed to estimate the risk of potential postoperative complications. The algorithms used were LightGBM, XGBoost, Random Forest, logistic regression, and a Neural Network (Deep Neural Network).

Data dictionary

An in-depth data dictionary for the NSQIP database can be referred to here, https://www.facs.org/media/1nrdyqmr/nsqip_puf_userguide_2022.pdf , and for features specific to hepatectomy, it can be referred to here, https://www.facs.org/media/ancplyqs/pt_nsqip_puf_userguide_2022.pdf.

Metrics/performance

The performance of each model was then evaluated on the test set using metrics such as receiver operating characteristic area under the curve (ROC AUC), area under precision-recall curve (AUPRC), and Brier score.

ROC is a plot wherein the x-axis represents the false positive rate and the y-axis represents the true positive rate. In the extremes, an ROC AUC of 1 implies that the model has perfect discriminatory power between positive and negative examples, a value of 0.5 suggests that the model has the equivalent power as random guessing, and finally 0 suggests it is completely unable to discriminate between positive and negative examples and is worse than random guessing.

The precision-recall curve (PRC) plots precision against recall/sensitivity. AUPRC measures the model's performance, similar to ROC AUC. An AUPRC value of 1 indicates perfect discriminatory power, 0.5 suggests equivalence to random guessing, and 0 suggests an inability to predict positive examples. AUPRC is useful for determining the model's ability to predict the outcome of interest (PHLF) better than random chance, which for PHLF is 0.048. Although the AUPRC value appears smaller than the ROC AUC, it should be compared against the chance rate of the outcome variable in the dataset, while ROC AUC has a default chance rate of 0.5.

The third evaluation metric, Brier score loss, is a metric used to evaluate the accuracy of a predictive machine learning model. It outputs values between 0 and 1, and values closer to 0 are indicative of a stronger model, where the predicted probability of an event occurring more closely matches the actual outcome. In essence, the Brier score quantifies the average squared difference between predicted probabilities and actual outcomes, providing a measure of both calibration and discrimination. It should be noted that a lower Brier score loss reflects a better-performing model as lower values imply less deviation between model prediction and true value.

Each model is cross-validated 5 times, allowing for the models to train and evaluate on 5 different combinations of training and testing data. Overall cross-validation trials, means, standard deviations, and 95% confidence intervals (95% CIs) of all performance metrics are calculated for model comparison.

Feature importance using SHAP values

Using Shapley Additive exPlanations (SHAP) on the classification models, we identified the most pivotal features in predicting hepatectomy-related complications within the database [12]. SHAP values, based on cooperative game theory, provide a method to fairly attribute the contribution of each feature to an individual prediction. They are calculated by considering all possible combinations of features to determine how much each individual factor contributes to the prediction, providing a clear explanation of the influence and importance of each factor on the model's outcome. These insights are instrumental in understanding the impact of patient risk factors on postoperative complications.

A bar graph will be created with each row displaying the influence of a specific feature on the model's predictions. Features with higher absolute SHAP values have more impact on model prediction than those with lower absolute SHAP values such as number of benign lesions. Features are then ordered by the sum of SHAP values, ranging from the most influential feature at the top to the least influential at the bottom.

Beeswarm plots are then derived to determine the directionality of influence for each feature towards predicting PHLF. On the Beeswarm plot, red dots have higher specific feature value and blue dots have lower feature value. By analyzing where these feature values lie on the x axis of SHAP values allows one to identify how that feature is correlated with PHLF prediction in the model.

The combination of the bar graph and the Beeswarm plot will aid in understanding the most important features that the ML model utilizes to predict PHLF.

Threshold values

We used the strongest performing model's predicted probability of the outcome of interest occurring to calculate cutoffs of feature values for certain percentage risk thresholds. A risk threshold (5%) was established to stratify patients into high and low-risk categories for PHLF, based on what is typically considered low risk for an elective operation. This cutoff was chosen because it approximates the contemporary incidence of clinically significant PHLF (ISGLS [International Study Group of Liver Surgery] grades B/C) after major hepatectomy in large NSQIP series. For example, Vitello et al. [13] analyzed 6,274 elective major hepatectomies in the 2014–2020 NSQIP cohort and reported grade B/C PHLF in 5.3% of patients, which carried significantly higher mortality than grade A PHLF or no PHLF (25.4% versus 1.1% and 1.2%, respectively). Conversely, when all hepatectomy types are pooled, the rate of grade B/C PHLF falls to approximately 2.8% [14]. Setting the risk threshold at 5% therefore identifies patients whose predicted

probability is roughly twice the background risk for clinically significant PHLF in the broader hepatectomy population and aligns with the point at which our institutional hepatobiliary team might consider intensifying perioperative assessment or management strategies (e.g., measures to increase future liver remnant [FLR], parenchyma-sparing techniques, and flow modulation considerations). For each threshold, the dataset was re-evaluated and descriptive statistics were calculated for the high-risk group. For clinical cutoff values, values within a margin of 0.5% risk around the cutoff was used to calculate average values of all features of patients in that risk grouping.

RESULTS

Dataset

The study population had a median age of 61 years, with 14,109 males (50.04%) and 14,083 females (49.96%). PHLF occurred in 1,305 patients (4.6%). Primary hepatobiliary cancer was diagnosed in 8,769 patients (31.09%) and secondary metastatic cancer was present in 13,282 patients (47.13%). Table 1 displays complete demographic details.

The study population consisted of 28,192 patients with a median age of 61 years. Gender distribution was nearly equal, with 50.04% male and 49.96% female. The median body mass index (BMI) was 27.60. ASA (American Society of Anesthesiologists) classification included 1.3% in Class 0, 23.1% in Class 1, 68.0% in Class 2, 7.4% in Class 3, 0.1% in Class 4, and 0.2% in Class 5. PHLF was observed in 4.6% of patients, and ascites in 0.6%. Smoking was reported by 14.6% of patients, diabetes by 17.55%, chronic obstructive pulmonary disease by 3.55%, and hypertension requiring medication by 45.67%. Preoperative laboratory values were as follows: median sodium level of 139 (interquartile range 3), median platelet count of 223, median prothrombin time (PT) of 29.2 seconds, median INR of 1, median albumin level of 4.1 g/dL, median aspartate aminotransferase (AST) level of 26 U/L, and median alkaline phosphatase (ALP) level of 90 U/L. Primary hepatobiliary cancer was present in 31.09% of patients, and metastatic cancer to the liver in 47.13%. Tumor staging revealed 78.45% as M0, 19.45% as M1, and 0.49% as M2. The median number of tumors was 2. Liver texture assessments showed 57.39% normal, 27.41% fatty liver, 4.48% fibrosis, and 4.10% congested liver. Neoadjuvant chemotherapy was administered to 30.77% of patients. Surgical procedures included formal right lobectomy in 15.02%, formal left lobectomy in 8.86%, and trisegmentectomy in 7.93%. A planned open approach was used in 72.48% of cases, with the Pringle maneuver performed in 25.85%. Hepaticojejunostomy was done in 6.37% of patients.

Model performance

LightGBM achieved an ROC AUC of 0.8349 (95% CI: 0.8272, 0.8427), Random Forest achieved an ROC AUC of 0.8363 (95% CI: 0.8236, 0.8491), and logistic regression achieved an ROC

Table 1. Descriptive statistics of the features used for the machine learning models of all the patients with hepatectomy from the NSQIP database

Variable	Data
Age	61 [18]
Sex	Male: 14,109 (50.04) Female: 14,083 (49.96)
BMI	27.60 [7.76]
ASA class	0: 361 (1.3) 1: 6,524 (23.1) 2: 19,160 (68.0) 3: 2,082 (7.4) 4: 17 (0.1) 5: 47 (0.2)
Post-hepatectomy liver failure	1305 (4.6)
Ascites	No: 28,033 (99.4) Yes: 159 (0.6)
Smoker	No: 24,083 (85.4) Yes: 4,109 (14.6)
Diabetes	No: 23,246 (82.45) Yes: 4,946 (17.55)
Chronic obstructive pulmonary disease	No: 27,191 (96.45) Yes: 1,000 (3.55)
Hypertension (with medication)	No: 15,311 (54.33) Yes: 12,880 (45.67)
Preoperative sodium	139 [3]
Preoperative platelet count	223 [99]
Preoperative prothrombin Time	29.2 [1.6]
Preoperative international normalized ratio	1 [0.1]
Preoperative albumin (g/dL)	4.1 [0.5]
Preoperative AST (U/L)	26 [15]
Preoperative ALP (U/L)	90 [47]
Primary hepatobiliary cancer	No: 19,432 (68.91) Yes: 8,769 (31.09)
Metastatic cancer to liver	No: 14,910 (52.87) Yes: 13,282 (47.13)
Tumor M stage	M0: 22,480 (78.45) M1: 5,571 (19.45) M2: 140 (0.49)
Number of tumors	9 [8]
Liver texture	Normal: 7,727 (57.39) Fatty: 3,689 (27.41) Fibrosis: 603 (4.48) Congested: 552 (4.10)
Neoadjuvant chemotherapy	No: 19,517 (69.23) Yes: 8,675 (30.77)
Formal right lobectomy	No: 23,961 (84.98) Yes: 4,231 (15.02)
Formal left lobectomy	No: 25,692 (91.14) Yes: 2,500 (8.86)
Trisegmentectomy	No: 25,956 (92.07) Yes: 2,236 (7.93)
Planned open	Yes: 20,434 (72.48) No: 7,758 (27.52)
Pringle maneuver	No: 20,906 (74.15) Yes: 7,285 (25.85)
Hepaticojejunostomy	No: 26,395 (93.63) Yes: 1,796 (6.37)

Values are presented as median [interquartile range] or number (%).
BMI, body mass index; ASA, American Society of Anesthesiologists; AST, aspartate aminotransferase; ALP, alkaline phosphatase.

Table 2. Mean performance metrics across all the cross-validation runs for each model with confidence intervals (CIs)

Model	ROC AUC (95% CI)	AUPRC (95% CI)	Brier score loss (95% CI)
LightGBM	0.8349 (0.8272, 0.8427)	0.1136 (0.1057, 0.1216)	0.0834 (0.0806, 0.0862)
XGBoost	0.8178 (0.8114, 0.8242)	0.1077 (0.0953, 0.1200)	0.0838 (0.0811, 0.0864)
Random forest	0.8363 (0.8236, 0.8491)	0.1182 (0.1079, 0.1285)	0.0915 (0.0855, 0.0975)
Logistic regression	0.8357 (0.8283, 0.8430)	0.1037 (0.0953, 0.1122)	0.2943 (0.2896, 0.2990)
Neural Network	0.8222 (0.8130, 0.8314)	0.1028 (0.0874, 0.1181)	0.2663 (0.1650, 0.3676)

ROC, receiver operating characteristic; AUC, area under the curve; AUPRC, area under the precision-recall curve.

AUC of 0.8357 (95% CI: 0.8283, 0.8430). Detailed performance metrics for all ML models tested with cross-validation are described in Table 2 and Fig. 1.

LightGBM achieved an ROC AUC of 0.8349, Random Forest achieved an ROC AUC of 0.8363, and logistic regression achieved an ROC AUC of 0.8357. Random Forest had an AUPRC of 0.1182, LightGBM had an AUPRC of 0.1136, and XGBoost had an AUPRC of 0.1077. Logistic regression and Neural Network had AUPRC values of 0.1037 and 0.1028, respectively. The Brier score loss was 0.0834 for LightGBM, 0.0838 for XGBoost, 0.0915 for Random Forest, 0.2943 for logistic regression, and 0.2663 for Neural Network.

SHAP values were calculated to identify the most influential features of our best-performing model, LightGBM. According to the SHAP values, the top 10 most influential features for estimating PHLF risk in our model were: the use of intraoperative or postoperative transfusions for bleeding, use of total right lobectomy, preoperative INR, preoperative sodium, operative time, planned open surgical approach, use of trisegmentectomy, preoperative albumin, ASA class, and preoperative bilirubin. Of these, the use of intra or postoperative transfusions for bleeding, use of total right lobectomy, preoperative INR, operative time, open surgical approach (planned), use of trisegmentectomy, ASA class, and preoperative bilirubin were all positively correlated with PHLF. On the other hand, preoperative sodium and preoperative albumin were inversely

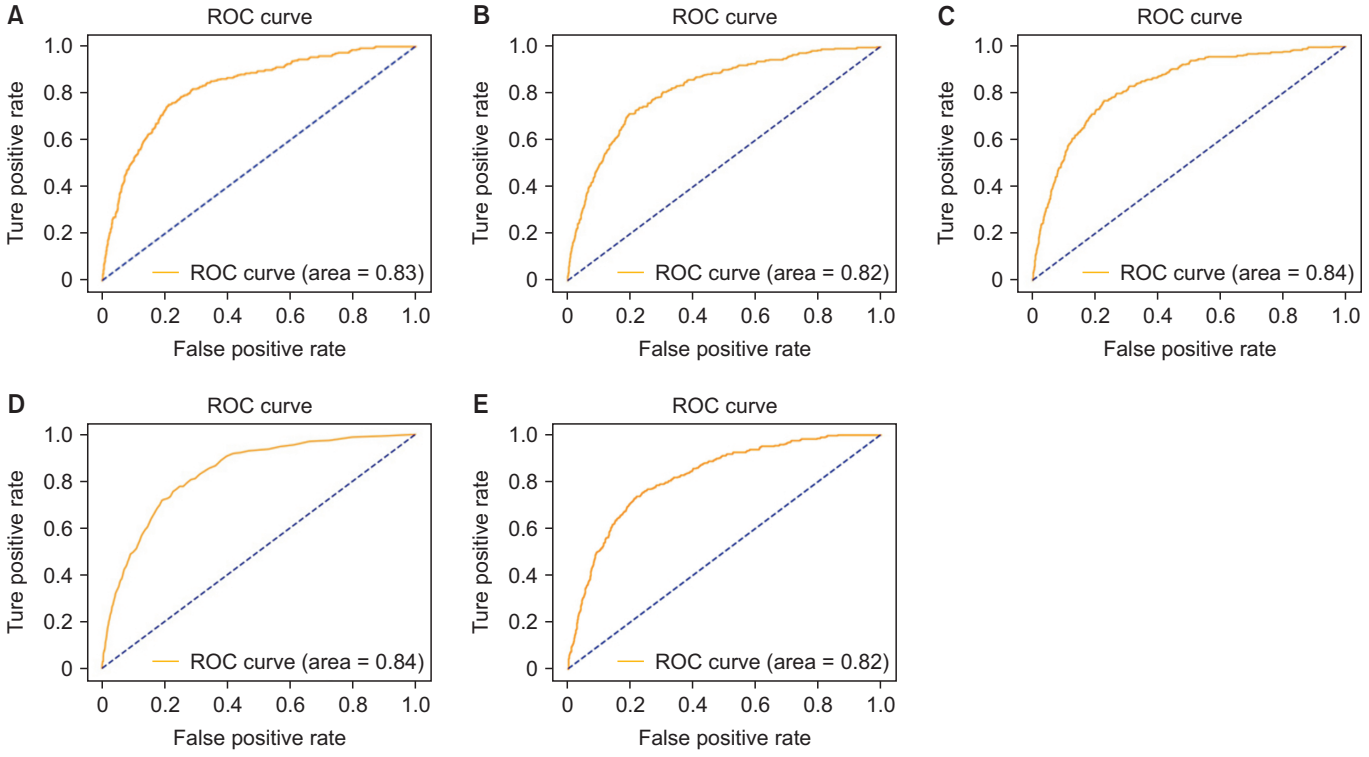

Fig. 1. Representative ROC AUC graphs for each of machine learning models. (A) LightGBM. (B) XGBoost. (C) Logistic regression. (D) Random forest. (E) Neural Network. LightGBM achieved an ROC AUC of 0.8349, Random Forest achieved an ROC AUC of 0.8363, and logistic regression achieved an ROC AUC of 0.8357. An ROC AUC of 1 indicates perfect discriminatory power between positive and negative outcomes, 0.5 indicates performance equivalent to random guessing, and 0 indicates it is worse than random guessing. AUC, area under the curve; ROC, receiver operating characteristic.

correlated with PHLF in the model. Detailed SHAP values and the directionality of the most influential features for PHLF risk are depicted in Fig. 2.

The threshold values of the most influential features when the ML model estimates a 5% PHLF risk are detailed in Table 3. It should be noted that binary values are absolute and do not have thresholds. They simply reflect the percentage chance that a patient with a 5% PHLF risk will have that feature and do not imply directionality.

Within the top 20 preoperative and intraoperative risk factors identified by the models, both clinical and biochemical factors contributed to the likelihood of developing PHLF. Demographic factors for increased PHLF risk included male sex and ASA class of 2 or higher. Biochemical risk factors included both liver function tests and liver enzymes. Increased risk of PHLF was associated with INR levels above 1.04, sodium levels below 139 mEq/L, albumin levels below 3.94 mg/dL, bilirubin levels above 0.65 mg/dL, platelet counts over 223×10^9/L, PT over 29.2 seconds, and AST levels over 35.67 Units/L. Liver-related factors contributing to PHLF risk included having a tumor M stage over 2, undergoing neoadjuvant therapy, more than 5 malignant lesions, more than 3 benign lesions, and abnormal liver texture. Intraoperative risk factors included transfusion requirements, performing a formal right lobectomy or trisegmentectomy, operating times exceeding 266.5 minutes, open surgical approach, use of the Pringle maneuver, and hepaticojejunostomy reconstruction.

The selected features are based on the SHAP values and graphs shown in Fig. 2. Preoperative risk factors include an INR greater than 1.0405, sodium levels below 139.1796 mEq/L, albumin levels under 3.9497 g/dL, ASA class scores over 1.9419, bilirubin levels higher than 0.65 mg/dL, being male, AST levels above 35.7 units/L, partial thromboplastin time (PTT) over 29.2 seconds, metastasis (M) stage scores exceeding 1.52, platelet counts below 223×10^9/L, use of neoadjuvant therapy, having more than five tumors, abnormal liver texture, and having more than three benign lesions. Intraoperative risk factors that increase the risk of PHLF by 5% include the use of intra or postoperative transfusions for bleeding, performing a total right lobectomy, operative times exceeding 266.51 minutes, a planned open surgical approach, trisegmentectomy, use of the Pringle maneuver, and hepaticojejunostomy reconstruction.

DISCUSSION

This study utilized ML models applied to the large-scale NSQIP database to predict PHLF. Our primary aim was a methodological comparison of different ML algorithms in their ability to identify risk patterns within this extensive, albeit registry-based, dataset. While ML is increasingly explored in surgery, its application requires careful interpretation, particularly when using administrative databases like NSQIP, which offer breadth but may lack clinical depth. The findings presented, including specific risk factor cutoffs, should be viewed primarily as illustrations of how ML can detect subtle, incremental risk variations in a large population, rather than as

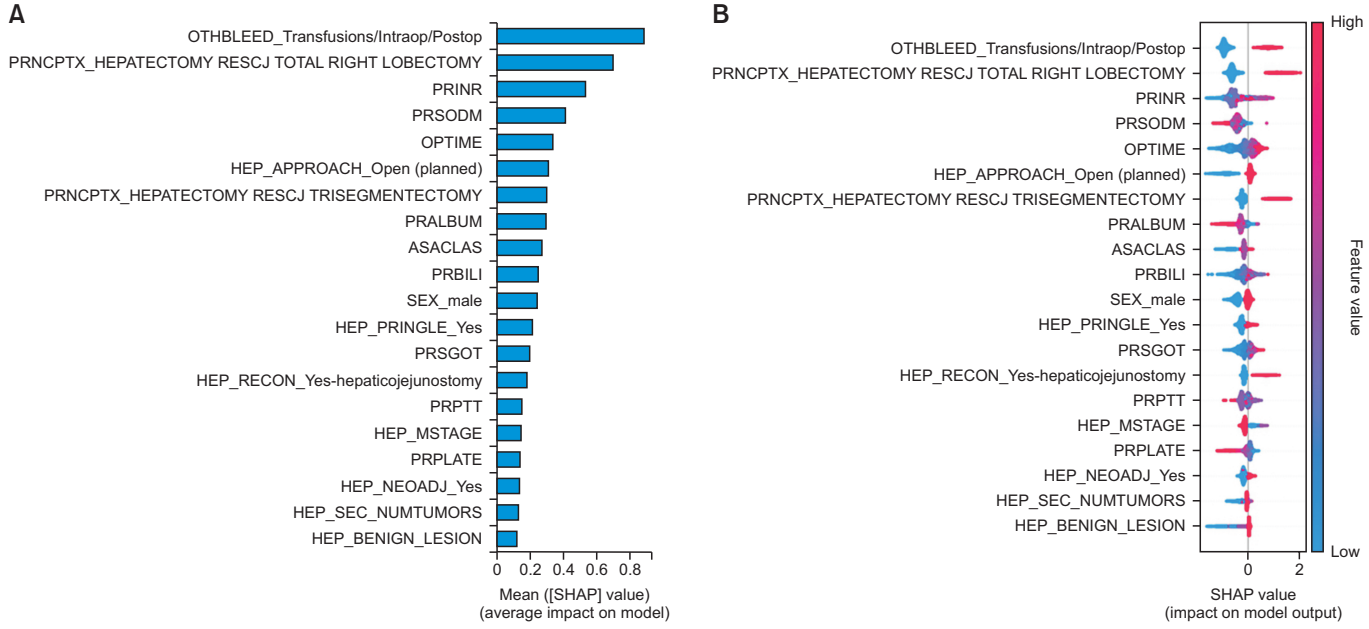

Fig. 2. Top 20 most relevant features and their respective impact on LightGBM prediction of PHLF. (A) SHAP bar plot detailing the top features and their mean SHAP value/impact on the model's prediction of PHLF. (B) SHAP Beeswarm plot showing how the value of the most impactful features positively or negatively affects the model likelihood of predicting a PHLF event. PHLF, post-hepatectomy liver failure; SHAP, Shapley Additive exPlanations.

Table 3. Preoperative and intraoperative risk factors that increase the risk of PHLF by 5%

Feature name	Increases PHLF risk by 5%
Preoperative risk factor	
Preoperative INR	> 1.0405
Preoperative sodium (mEq/L)	< 139.1796
Preoperative albumin (g/dL)	< 3.9497
ASA Class	> 1.9419
Preoperative total bilirubin (mg/dL)	> 0.65
Sex	Male
Preoperative AST (units/L)	> 35.7
Preoperative PTT (sec)	> 29.2
Tumor M stage	> 1.52
Preoperative platelets (x10^9/L)	< 223
Use of neoadjuvant therapy	Yes
Number of tumors	> 5
Normal liver texture	Abnormal
Number of benign lesions	> 3
Intraoperative risk factor	
Use of intra or postoperative transfusions for bleeding	Yes
Use of total right lobectomy	Yes
Operative time (min)	> 266.51
Open surgical approach (planned)	Yes
Use of trisegmentectomy	Yes
Use of Pringle maneuver	Yes
Use of hepaticojejunostomy reconstruction	Yes

immediately actionable clinical guidelines.

PHLF is a severe and potentially life-threatening complication that results in prolonged hospitalization, increased morbidity, and mortality [6]. Artificial intelligence models using data from large databases like NSQIP can enhance the prediction of PHLF by identifying complex and unapparent risk factors not identified by traditional statistical methods, thus providing personalized risk assessments [8,9].

Our study identified several critical preoperative, liver-specific, and intraoperative factors associated with an increased risk of PHLF. Among the preoperative factors, male sex was associated with higher rates of PHLF, consistent with literature suggesting that testosterone may exert an immunosuppressive effect, thereby increasing susceptibility to liver failure [15].

Other preoperative biochemical factors related to impaired liver function were associated with PHLF. These included low sodium levels (< 139 mEq/L), an INR >1.04, prolonged PT (> 29.2 seconds), albumin (< 3.94 mg/dL), high bilirubin (0.65 mg/dL), and elevated AST (35.67 units/L). While these parameters are often used to calculate the MELD score in patients with chronic liver disease [16,17], our study indicates that impaired synthetic function that would affect the MELD score prior to hepatectomy is not necessary for developing PHLF. By establishing precise cutoff points using ML, we found that even subtle biochemical abnormalities can significantly increase the risk of PHLF. This highlights the utility of our model in detecting nuanced risk factors that may not be apparent through traditional scoring systems.

The association between decreased platelet count and PHLF was expected, as low platelet counts are traditionally associated with a certain degree of portal hypertension related to impaired liver function [18]. Typically, patients with cirrhosis and low platelet counts would not qualify for LR, but the cutoff points for this are in general arbitrary. The findings of this study highlight the multifactorial nature of PHLF risk. Among preoperative factors, we identify tumor characteristics such as high M stage and tumor multiplicity as important estimators of PHLF risk. These findings are consistent with existing literature, underscoring that high tumor burden and extensive LRs predispose patients to an increased risk of PHLF [19-23]. We identified that neoadjuvant therapy plays a significant role in increasing the risk of PHLF. The link between neoadjuvant therapy and increased PHLF risk is well understood, involving mechanisms such as hepatic parenchymal injury from receiving oxaliplatin [24] or an increased risk of steatohepatitis due to irinotecan [25].

For intraoperative factors, we identified that extensive resections (e.g., right lobectomy, trisegmentectomy), use of the Pringle maneuver, prolonged operative time, and need for blood transfusion as being correlated with PHLF risk. High tumor burden often necessitates extensive liver resection, which can result in insufficient quality and function of the remnant, leading to PHLF [26-28]. Additionally, extensive resection can elevate portal vein pressure, compromising arterial inflow to the liver and hindering regeneration [29]. This results in "small-for-size" syndrome, which may result in liver failure [29]. Intraoperative blood transfusions are well known to significantly increase the risk of PHLF [30,31]. The amount of blood loss and subsequent blood transfusion during surgery can significantly alter systemic hemodynamics, leading to hypoperfusion of vital organs, and suppressed immune responses; these physiological challenges exacerbate the risk of liver dysfunction, thereby increasing the likelihood of PHLF [32]. Similarly, prolonged operative time subjects the liver to tissue hypoxia and metabolic stress. While the use of the Pringle maneuver is essential for controlling bleeding by restricting inflow to the liver, it also results in ischemia and reperfusion injury [20]. This ischemia-reperfusion cycle can trigger inflammatory responses, oxidative stress, and cellular damage, contributing to liver dysfunction and the development of PHLF [33].

In this study, the LightGBM model was the best in class in its ability to identify true positives and avoid false positives based on the combined metrics of ROC AUC, Brier score, and AUPRC. This model achieved a strong ROC AUC of 0.8349, indicating its accuracy in distinguishing between patients who will and will not develop PHLF. The LightGBM model's low Brier score loss of 0.0834 demonstrates well-calibrated predic-

tions. Additionally, its AUPRC of 0.1136, significantly higher than the observed PHLF rate of 0.048, confirms its effectiveness in accurately predicting positive cases. LightGBM's ability to handle large datasets efficiently, along with its innovative leaf-wise growth approach and built-in regularization, prevents overfitting and enhances predictive accuracy, making it a robust choice for predicting PHLF.

Previous risk calculators often utilized logistic regression with fewer variables and smaller datasets [14,20-25,27-28,34-37]. In contrast, our model, which employs LightGBM, utilized a comprehensive NSQIP dataset from 2013 to 2021, encompassing 28,192 patients. Unlike the ACS risk calculator, which lacks procedure-specific complications [9], our model integrates biochemical markers, providing a more tailored and accurate risk assessment for PHLF. Our model was used to create a calculator that can give more nuanced risk estimates based on a patient's profile and comparing them to the NSQIP population's features and outcomes.

Our study has several limitations inherent to registry-based research using large datasets like NSQIP. While the large sample size enhances statistical power for identifying population trends, it comes with trade-offs. NSQIP lacks granular clinical detail; for instance, it does not differentiate between ISGLS grades of PHLF. Our definition relied on postoperative day 5 laboratory values (bilirubin and INR) as per the ISGLS criteria available in NSQIP, without grading. This means our outcome likely includes grade A PHLF, which often represents transient laboratory abnormalities without significant clinical consequence, potentially diluting the clinical impact of the findings. Furthermore, the observed PHLF incidence of 4.6% in our cohort is lower than some reports focusing on all ISGLS grades but aligns closely with rates reported for clinically significant (Grade B/C) PHLF or grade A PHLF in specific NSQIP analyses of major hepatectomies (e.g., Vitello et al. [13] reported 5.3% Grade B/C and 4.3% Grade A PHLF in major hepatectomies; Liu et al. [14] reported 2.8% clinically significant PHLF across all hepatectomies). The binary nature of NSQIP outcome reporting might also lead to under-recognition of milder cases, potentially skewing our captured rate towards more severe instances or specific definitions captured by the database fields. Consequently, the 4.6% rate likely reflects a combination of these factors.

Another key limitation is the potential divergence between statistical significance, more easily achieved in large datasets, and true clinical relevance. For example, cutoff values identified by our models (e.g., bilirubin > 0.65 mg/dL, sodium < 139 mEq/L) may represent statistically valid points of inflection in risk within this large dataset but are not proposed as new clinical diagnostic thresholds. They primarily demonstrate the ML models' sensitivity to subtle deviations. Reliance on the NSQIP database also excludes genetic information and detailed imaging (e.g., specific findings on MRI/CT scans beyond basic variables), which could refine risk stratification. Additionally, the study does not address real-time clinical implementation challenges, such as handling incomplete or inconsistent data and adapting to new patient information in dynamic clinical settings.

Since our models are trained on aggregate data, the probability it produces is calibrated to the average patient profile in the derivation cohort. In practice, this means the reported risk is a population-level estimate: it indicates how frequently an outcome is expected to occur among patients who share similar covariates, not a deterministic forecast for a single individual. For any one patient, unmeasured factors (e.g., nuanced operative findings, center-specific practices, and genomic variation) can shift true risk above or below the model's point estimate. Consequently, the output should be interpreted as a baseline probability that informs but does not replace clinical judgment.

Our ML models use population-level data, so the risk estimate is based on the characteristics of the entire NSQIP population with similar risk factors. This means that the calculated risk represents the population's risk, not an individual patient's risk. Thus, risk calculators can only provide an estimate, not a precise individual prediction.

It is important to note that our model has been internally validated using the NSQIP dataset, which limits its generalizability. While internal validation confirms consistency within this population, external validation with independent datasets is needed to ensure broader reliability and applicability.

In conclusion, ML models and their utility in medicine and other fields have been topical areas of research. Their role in surgical risk stratification is still evolving and this paper explored further possible utility from these tools. While not intended for immediate clinical adoption without further validation, the risk stratification insights provided by the LightGBM model could potentially inform future perioperative management considerations. These cutoffs are a proof of concept for the potential of these models to detect incremental and potentially subclinical changes to further inform clinician decision making. For instance, a high estimated risk might prompt closer evaluation or consideration of established strategies aimed at mitigating PHLF, such as optimizing the future liver remnant through portal vein embolization or considering staged resections in select complex cases, contingent on prospective validation of the model's utility in specific clinical scenarios. This highlights a potential pathway towards improving patient safety and optimizing resource use. Future research should focus on externally validating and prospectively evaluating these models, refining their predictive capabilities, and exploring interventions to mitigate identified risk factors, thereby reducing the incidence of PHLF and improving the overall prognosis for patients undergoing hepatectomy.

FUNDING

This research was supported by the Big Data/ICES Node

Competition Grant. The grant was awarded by the Department of Surgery at the University of Western Ontario. The funding source had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ORCID

Gautham Nair, https://orcid.org/0009-0005-7615-0827
Ali Hadi, https://orcid.org/0000-0002-5992-0835
Kartik Gupta, https://orcid.org/0000-0002-1310-9895
Edward Tran, https://orcid.org/0000-0002-6215-1053
Geerthan Srikantharajah, https://orcid.org/0009-0004-6576-1182
Evelyn Waugh, https://orcid.org/0000-0002-1720-7281
Ephraim Tang, https://orcid.org/0000-0002-2392-3380
Anton Skaro, https://orcid.org/0000-0002-9215-647X
Juan Glinka, https://orcid.org/0000-0003-4373-965X

AUTHOR CONTRIBUTIONS

Conceptualization: GN, AH, KG, E Tran, E Tang, AS, JG. Data curation: GN, AH, KG, E Tran, GS. Formal analysis: GN, AH, KG, E Tran. Investigation: GN, AH, KG, E Tran. Methodology: GN, AH, KG, E Tran, JG. Project administration: GN, AH, KG, E Tran, JG. Software: GN, AH, KG, E Tran, GS. Visualization: GN, AH, KG, E Tran. Funding acquisition: JG. Resources: JG. Supervision: JG. Writing - original draft: GN, AH, KG, E Tran. Writing - review and editing: GN, AH, KG, E Tran, EW, JG.

REFERENCES

1. Dimitroulis D, Tsaparas P, Valsami S, Mantas D, Spartalis E, Markakis C, et al. Indications, limitations and maneuvers to enable extended hepatectomy: current trends. World J Gastroenterol 2014;20:7887-7893.
2. Jin S, Fu Q, Wuyun G, Wuyun T. Management of post-hepatectomy complications. World J Gastroenterol 2013;19:7983-7991.
3. Jarnagin WR, Gonen M, Fong Y, DeMatteo RP, Ben-Porat L, Little S, et al. Improvement in perioperative outcome after hepatic resection: analysis of 1,803 consecutive cases over the past decade. Ann Surg 2002;236:397-406; discussion 406-397.
4. Ishii M, Mizuguchi T, Harada K, Ota S, Meguro M, Ueki T, et al. Comprehensive review of post-liver resection surgical complications and a new universal classification and grading system. World J Hepatol 2014;6:745-751.
5. Rahbari NN, Garden OJ, Padbury R, Brooke-Smith M, Crawford M, Adam R, et al. Posthepatectomy liver failure: a definition and grading by the International Study Group of Liver Surgery (ISGLS). Surgery 2011;149:713-724.
6. Søreide JA, Deshpande R. Post hepatectomy liver failure (PHLF) - recent advances in prevention and clinical management. Eur J Surg Oncol 2021;47:216-224.
7. Durand F, Valla D. Assessment of the prognosis of cirrhosis: Child-Pugh versus MELD. J Hepatol 2005;42 Suppl:S100-S107.
8. Rogers MP, Janjua H, DeSantis AJ, Grimsley E, Pietrobon R, Kuo PC. Machine learning refinement of the NSQIP risk calculator: who survives the "Hail Mary" case? J Am Coll Surg 2022;234:652-659.
9. Liu Y, Ko CY, Hall BL, Cohen ME. American College of Surgeons NSQIP risk calculator accuracy using a machine learning algorithm compared with regression. J Am Coll Surg 2023;236(5):1024-1030.
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830.
11. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36:1-13.
12. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., eds. Advances in neural information processing systems 30. Curran Associates, Inc., 2017:4765-4774.
13. Vitello DJ, Shah D, Ko B, Brajcich BC, Peters XD, Merkow RP, et al. Establishing the clinical relevance of grade A post-hepatectomy liver failure. J Surg Oncol 2024;129:745-753.
14. Liu JY, Ellis RJ, Hu QL, Cohen ME, Hoyt DB, Yang AD, et al. Post hepatectomy liver failure risk calculator for preoperative and early postoperative period following major hepatectomy. Ann Surg Oncol 2020;27:2868-2876.
15. Yokoyama Y, Schwacha MG, Samy TS, Bland KI, Chaudry IH. Gender dimorphism in immune responses following trauma and hemorrhage. Immunol Res 2002;26:63-76.
16. Schroeder RA, Marroquin CE, Bute BP, Khuri S, Henderson WG, Kuo PC. Predictive indices of morbidity and mortality after liver resection. Ann Surg 2006;243:373-379.
17. Chin KM, Allen JC, Teo JY, Kam JH, Tan EK, Koh Y, et al. Predictors of post-hepatectomy liver failure in patients undergoing extensive liver resections for hepatocellular carcinoma. Ann Hepatobiliary Pancreat Surg 2018;22:185-196.
18. Meyer J, Balaphas A, Combescure C, Morel P, Gonelle-Gispert C, Bühler L. Systematic review and meta-analysis of thrombocytopenia as a predictor of post-hepatectomy liver failure. HPB (Oxford) 2019;21:1419-1426.
19. Gulhar R, Ashraf MA, Jialal I. Physiology, acute phase reactants. In: StatPearls [Internet]. StatPearls Publishing, 2025 [Updated 2023 Apr 24]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519570/
20. Garcea G, Maddern GJ. Liver failure after major hepatic resection. J Hepatobiliary Pancreat Surg 2009;16:145-155.
21. Xiao Y, Yuan Q, Wang H, Zhang X, Yang Y. Predicting post-hepatectomy liver failure in patients with hepatocellular carcinoma: a novel radiomics-based nomogram and scoring system. Eur Radiol. 2021;31:3298-3307.

22. Shoup M, Gonen M, D'Angelica M, Jarnagin WR, DeMatteo RP, Schwartz LH, et al. Volumetric analysis predicts hepatic dysfunction in patients undergoing major liver resection. J Gastrointest Surg 2003;7:325-330.
23. Gazzaniga GM, Cappato S, Belli FE, Bagarolo C, Filauro M. Assessment of hepatic reserve for the indication of hepatic resection: how I do it. J Hepatobiliary Pancreat Surg 2005;12:27-30.
24. Vauthey JN, Pawlik TM, Ribero D, Wu TT, Zorzi D, Hoff PM, et al. Chemotherapy regimen predicts steatohepatitis and an increase in 90-day mortality after surgery for hepatic colorectal metastases. J Clin Oncol 2006;24:2065-2072.
25. Fong Y, Bentrem DJ. CASH (Chemotherapy-Associated Steatohepatitis) costs. Ann Surg 2006;243:8-9.
26. Guglielmi A, Ruzzenente A, Conci S, Valdegamberi A, Iacono C. How much remnant is enough in liver resection? Dig Surg 2012;29:6-17.
27. Bruckmann NM, Kirchner J, Grueneisen J, Li Y, McCutcheon A, Aigner C, et al. Correlation of the apparent diffusion coefficient (ADC) and standardized uptake values (SUV) with overall survival in patients with primary non-small cell lung cancer (NSCLC) using ^{18}F-FDG PET/MRI. Eur J Radiol 2021;134:109422.
28. Shirata C, Hasegawa K, Kokudo N, Makuuchi M, Izumi N. Predictors of post-hepatectomy liver failure in patients with hepatocellular carcinoma. Liver Cancer 2019;8:99-109.
29. Allard MA, Adam R, Bucur PO, Termos S, Cunha AS, Bismuth H, et al. Posthepatectomy portal vein pressure predicts liver failure and mortality after major liver resection on noncirrhotic liver. Ann Surg 2013;258:822-829; discussion 829-830.
30. Fu J, Chen Q, Yu Y, You W, Ding Z, Gao Y, et al. Impact of portal hypertension on short- and long-term outcomes after liver resection for intrahepatic cholangiocarcinoma: a propensity score matching analysis. Cancer Med 2021;10:6985-6997.
31. Kriengkrai W, Somjaivong B, Titapun A, Wonggom P. Predictive factors for post-hepatectomy liver failure in patients with cholangiocarcinoma. Asian Pac J Cancer Prev 2023;24:575-580.
32. Yugawa K, Maeda T, Nagata S, Shiraishi J, Sakai A, Yamaguchi S, et al. Impact of aspartate aminotransferase-to-platelet ratio index based score to assess posthepatectomy liver failure in patients with hepatocellular carcninoma. World J Surg Oncol 2022;20:248.
33. Wei X, Zheng W, Yang Z, Liu H, Tang T, Li X, et al. Effect of the intermittent Pringle maneuver on liver damage after hepatectomy: a retrospective cohort study. World J Surg Oncol 2019;17:142.
34. Doi S, Yasuda S, Hokuto D, Kamitani N, Matsuo Y, Sakata T, et al. Impact of the prolonged intermittent Pringle maneuver on post-hepatectomy liver failure: comparison of open and laparoscopic approaches. World J Surg 2023;47:3328-3337.
35. Wang JJ, Feng J, Gomes C, Calthorpe L, Ashraf Ganjouei A, Romero-Hernandez F, et al.; International Post-Hepatectomy Liver Failure Study Group. Development and validation of prediction models and risk calculators for posthepatectomy liver failure and postoperative complications using a diverse international cohort of major hepatectomies. Ann Surg 2023;278:976-984.
36. Dasari BVM, Hodson J, Roberts KJ, Sutcliffe RP, Marudanayagam R, Mirza DF, et al. Developing and validating a pre-operative risk score to predict post-hepatectomy liver failure. HPB (Oxford) 2019;21:539-546.
37. Chin KM, Koh YX, Syn N, Teo JY, Goh BKP, Cheow PC, et al. Early prediction of post-hepatectomy liver failure in patients undergoing major hepatectomy using a PHLF prognostic nomogram. World J Surg 2020;44:4197-4206.