Original Study

Annals of Surgery Open

OPEN

Development and Validation of a Machine Learning Prediction Model for Textbook Outcome in Liver Surgery: Results From a Multicenter, International Cohort

Jane Wang, MD,[*] Amir Ashraf Ganjouei, MD, MPH,[*] Taizo Hibi, MD, PhD,[†] Nuria Lluis, MD, PhD,[‡] Camilla Gomes, MD,[*] Fernanda Romero-Hernandez, MD,[*] Han Yin, BA,[*] Lucia Calthorpe, MD,[*] Yukiyasu Okamura, MD, PhD,[§] Yuta Abe, MD, PhD,[‖] Shogo Tanaka, MD, PhD,[¶] Minoru Tanabe, MD, PhD,[#] Zeniche Morise, MD, PhD,[**] Horacio Asbun, MD, PhD,[‡] David Geller, MD,[††] Mohammed Abu Hilal, MD, PhD,[‡‡] Mohamed Adam, MD,[*] Adnan Alseidi, MD, EdM,[*] and International Hepatectomy Study Group

Objective: This study aimed to (1) develop a machine learning (ML) model that predicts the textbook outcome in liver surgery (TOLS) using preoperative variables and (2) validate the TOLS criteria by determining whether TOLS is associated with long-term survival after hepatectomy.

Background: Textbook outcome is a composite measure that combines several favorable outcomes into a single metric and represents the optimal postoperative course. Recently, an expert panel of surgeons proposed a Delphi consensus-based definition of TOLS.

Methods: Adult patients who underwent hepatectomies were identified from a multicenter, international cohort (2010–2022). After data preprocessing and train-test splitting (80:20), 4 models for predicting TOLS were trained and tested. Following model optimization, the performance of the models was evaluated using receiver operating characteristic curves, and a web-based calculator was developed. In addition, a multivariable Cox proportional hazards analysis was conducted to determine the association between TOLS and overall survival (OS).

Results: A total of 2059 patients were included, with 62.8% meeting the criteria for TOLS. The XGBoost model, which had the best performance with an area under the curve of 0.73, was chosen for the web-based calculator. The most predictive variables for having TOLS were a minimally invasive approach, fewer lesions, lower Charlson Comorbidity Index, lower preoperative creatinine levels, and smaller lesions. In the multivariable analysis, having TOLS was associated with improved OS (hazard ratio = 0.82, *P* = 0.015).

Conclusions: Our ML model can predict TOLS with acceptable discrimination. We validated the TOLS criteria by demonstrating a significant association with improved OS, thus supporting their use in informing patient care.

Keywords: hepatectomy, liver surgery, machine learning, prediction model, textbook outcome

INTRODUCTION

Textbook outcome is a novel composite measure that incorporates multiple favorable outcomes into a single metric and reflects the ideal postoperative course.[1] It has a few advantages over the more traditional, individual outcome measures such as morbidity, mortality, and hospital length of stay. For one, textbook outcome can better capture the multidimensionality of a patient's perioperative course and may be a better marker of the quality of surgical care.[2–5] Thus, it can be useful for assessing and comparing outcomes across various hospitals for a particular procedure.[2] In addition, single outcome measures may have a low event rate, limiting their ability to serve as a quality metric.[6] Furthermore, patients may find textbook outcome to be a more digestible summary measure than needing to synthesize

From the [*]Department of Surgery, University of California, San Francisco, San Francisco, CA; [†]Department of Pediatric Surgery and Transplantation, Kumamoto University Graduate School of Medical Sciences, Kumamoto, Japan; [‡]Hepato-Biliary and Pancreas Surgery, Miami Cancer Institute, Miami, FL; [§]Division of Hepato-Biliary-Pancreatic Surgery, Shizuoka Cancer Center, Shizuoka, Japan; [‖]Department of Surgery, Keio University School of Medicine, Tokyo, Japan; [¶]Department of Hepato-Biliary-Pancreatic Surgery, Osaka City University Graduate School of Medicine, Osaka, Japan; [#]Department of Hepatobiliary and Pancreatic Surgery, Graduate School of Medicine, Tokyo Medical and Dental University, Tokyo, Japan; [**]Department of Surgery, Okazaki Medical Center, Fujita Health University School of Medicine, Okazaki, Japan; [††]Department of Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA; and [‡‡]Department of General Surgery, Poliambulanza Foundation Hospital, Brescia, Italy.

J.W. and A.A.G. are cofirst authors.

J.W. was supported by the UCSF Noyce Initiative for Digital Transformation in Computational Biology & Health, Computational Innovator Postdoctoral Fellowship Award.

Disclosure: The authors declare that they have nothing to disclose.

The datasets generated during and/or analyzed during the current study are not publicly available, but are available from the corresponding author on reasonable request.

SDC Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

Reprints: Adnan Alseidi, MD, EdM, Department of Surgery, University of California, San Francisco, 550 16th Street, San Francisco, CA 94143. E-mail: adnan.alseidi@ucsf.edu.

Annals of Surgery Open (2025) 1:e539

Received: 8 May 2024; Accepted 13 December 2024

Published online 10 February 2025

DOI: 10.1097/AS9.0000000000000539

the relevance of multiple, individual outcomes.[7] Notably, textbook outcome has been described for multiple general surgery procedures, including colorectal, hepatobiliary, and esophageal surgeries.[1,2,6,8,9]

Recently, an international, Delphi consensus-based definition of textbook outcome in liver surgery (TOLS) was proposed by an expert panel of 44 liver surgeons across 22 countries.[7] TOLS was defined by the following 7 criteria: the absence of intraoperative complications, postoperative bile leakage, postoperative liver failure, 90-day major complications, 90-day readmission, 90-day mortality, and the presence of a negative (R0) resection margin. TOLS was defined as meeting all 7 criteria. While the Delphi consensus technique is well described and widely utilized in the surgical literature,[10–14] it is ultimately based on expert opinion. If the developed criteria are to be used to inform patient care or quality improvement initiatives, it is important to validate them using real-world patient data.

In this setting, our study group wanted to explore the proposed TOLS criteria using an international cohort of hepatectomy patients. Specifically, our study aims are as follows: (1) develop a machine learning model to predict TOLS using only preoperative variables; (2) identify which variables are most predictive of TOLS; and (3) validate the TOLS criteria by determining whether TOLS impacts long-term survival after hepatectomy.

METHODS

Cohort Selection and Data Collection

Patients were identified from a multicenter, international cohort that included 6 centers in Japan and 3 in the United States. The list of all participating centers is available in Supplemental Table 1, see http://links.lww.com/AOSO/A447. Each institution collected the variables of interest from their own prospectively maintained database. Each center gathered additional variables via manual chart review of the electronic medical record of their respective institutions. This study received Institutional Review Board approval (# 22-38059).

We included adult patients age 18 and older who underwent liver surgery at a participating institution from 2010 to 2022. All indications (eg, both benign and malignant), surgical approaches (eg, open, pure laparoscopic, robotic, and hand-assisted), and types of resections (eg, anatomic and nonanatomic; major and minor) were included.

Demographic and Clinical Variables

Demographic variables such as patient age, sex, and body mass index (BMI) were extracted. Clinical variables including American Society of Anesthesiologists (ASA) classification, Charlson Comorbidity Index (CCI), parameters reflecting baseline liver function (portal hypertension, esophageal varices, liver impairment, clinical cirrhosis, and ascites), model for end-stage liver disease (MELD) score, prior history of abdominal surgery, prior history of liver resection, indication for surgery (benign, hepatocellular carcinoma [HCC], colorectal liver metastases [CRLM], cholangiocarcinoma, and other malignancy), number of lesions, maximum lesion size, preoperative lab values (hemoglobin, platelet count, international normalized ratio, creatinine, and total bilirubin), and intended operative approach were extracted. Intraoperative and postoperative variables were also collected and included type of resection, number of segments resected, estimated blood loss, transfusion of blood products, Pringle maneuver, concurrent surgery, synchronous ablation, vascular reconstruction, intraoperative incidents, margin status (for malignant indications), and postoperative complications, readmission, and mortality.

Definition of Textbook Outcome in Liver Surgery

TOLS was defined[7] as not experiencing intraoperative incidents (grade 2 or higher as defined by the Oslo Classification[15]), postoperative bile leakage (grade B or C as defined by the International Study Group of Liver Surgery criteria[16]), postoperative liver failure (grade B or C as defined by the International Study Group of Liver Surgery criteria[17]), 90-day major postoperative complications (Clavien–Dindo grade 3 or higher), 90-day readmission due to surgery-related major complications, 90-day or inhospital mortality, and having an R0 resection margin (for malignant indications).

Statistical Analysis

Demographic and clinical characteristics were analyzed using the chi-square test for categorical variables and the Wilcoxon rank sum test for continuous variables. The Kaplan–Meier method was used to evaluate overall survival (OS) for patients with and without TOLS, with differences assessed using the log-rank test. The reference group for these analyses was the non-TOLS cohort. To evaluate survival trends over time, we included survival data up to 80 months, which reflects the 90th percentile of the cohort's follow-up data. A multivariable Cox proportional hazards model, which adjusted for age, sex, indication for surgery, CCI, MELD score, number of lesions, and type of resection, was performed to determine the association between TOLS and OS. All analyses were 2-sided, and a *P*-value of 0.05 was considered statistically significant.

Machine Learning Pipeline

Preoperative variables including age, sex, BMI, ASA class, CCI, portal hypertension, esophageal varices, liver impairment, clinical cirrhosis, ascites, MELD score, prior history of abdominal surgery, prior history of liver resection, indication for surgery, number of lesions, maximum lesion size, lab values, and operative approach were included as potential predictors of TOLS. Categorical variables with more than 1 level underwent one-hot encoding. Median imputation was used for the following variables, which had 1% to 13% missing data: BMI, number of lesions, maximum lesion size, hemoglobin, platelet count, international normalized ratio, creatinine, and total bilirubin.

After preprocessing, the data were randomly divided into training and test sets using the 'train_test_split' function from the scikit-learn library. This function was employed with an 80:20 ratio to split the data, and a random state was set to ensure the replicability of the results. Four prediction models were then trained and tested: logistic regression, neural network, random forest, and Extreme Gradient Boosting (XGBoost). The Grid Search algorithm from the scikit-learn library was employed to fine-tune the machine learning models. This algorithm systematically worked through multiple combinations of hyperparameters to determine which combination led to the most accurate predictive model. Key hyperparameters adjusted during the fine-tuning included the learning rate and the number of estimators. Recursive feature elimination, coupled with fivefold cross-validation, was also used to assess the performance of the models on the training set using varying numbers of input variables. Each model's ability to predict TOLS was then evaluated using the area under the receiver operating characteristic curve (AUC), and calibration was assessed using calibration plots. The performance of each model was evaluated on the test set and the best-performing model was identified. SHapley Additive exPlanations (SHAP) were then analyzed to gain insight into the most important variables for predicting TOLS. Finally, the best-performing model was used to develop a web-based calculator. A detailed description of the machine-learning pipeline is presented in Figure 1.

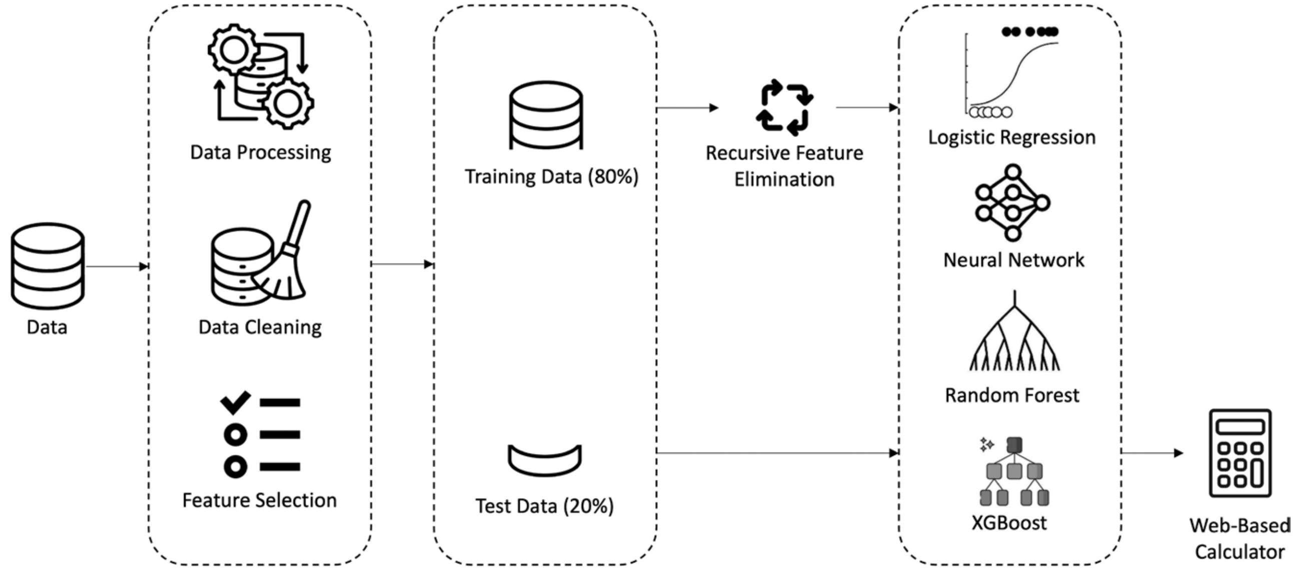

FIGURE 1. Diagram visualizing the machine learning pipeline for predicting textbook outcome in liver surgery.

Data extraction and descriptive analyses were conducted using R v4.3.0 (the R Foundation for Statistical Computing, Vienna, Austria). The machine learning pipeline was created using Python v3.11 (Python Software Foundation, Wilmington, DE).

RESULTS

Characteristics of the Cohort

A total of 2059 patients who underwent liver surgery between 2010 and 2022 were included. The median age was 63 years (interquartile range: 53–71), and the majority of patients were male (60%). The median CCI was 5 (interquartile range: 3–7) and 88% of patients had an ASA class of at least 2. The majority of patients (70.2%) did not have baseline liver impairment, and the primary indication for surgery was HCC (33%) followed by CRLM (25%). Complete demographic and preoperative clinical characteristics are detailed in Table 1.

A total of 1293 (62.8%) patients met criteria for TOLS. These patients were younger (median, 62 vs 65 years), had lower CCI (median, 4 vs 5), and less frequently had baseline liver impairment (27% vs 32.6%) compared with those without TOLS (Table 1). Furthermore, patients who did not meet the criteria for TOLS experienced higher estimated blood loss (median, 400 vs 200 mL) and were more likely to undergo conversion to open (14.7% vs 5.4%) if the initial approach was minimally invasive surgery (MIS). Further data on intraoperative characteristics are presented in Supplemental Table 2, see http://links.lww.com/AOSO/A447. Notably, the most common reasons for not having TOLS were the presence of a positive surgical margin (20%) and the presence of 90-day major postoperative complications (16%) (Table 2).

Predicting TOLS

Recursive feature elimination showed that the performance of the logistic regression model improved as the number of features increased. Specifically, the maximum AUC was achieved with 32 variables. In contrast, the machine learning models reached their highest AUC with fewer features (random forest: 17 variables; XGBoost: 15 variables) (Supplemental Fig. 1, see http://links.lww.com/AOSO/A447). After fine-tuning, the random forest and XGBoost algorithms had the best performances on the testing set, with each achieving an AUC of 0.73 (95% confidence interval: 0.68–0.78). In contrast, the logistic regression and neural network models had the lowest performances with an AUC of 0.71 (logistic regression, 95% confidence interval: 0.66–0.75; neural network, 0.66–0.76) (Fig. 2). Furthermore, the models exhibited good calibration in the testing set (Supplemental Fig. 2, see http://links.lww.com/AOSO/A447).

Given its superior performance, the XGBoost model was used to evaluate the significance of the input variables. The model identified the operative approach as the most important preoperative variable in predicting TOLS, followed by the number of lesions, CCI, creatinine, and lesion size (Fig. 3). The SHAP values analysis indicated that the MIS approach (SHAP values: 0.16–0.46), a lower number of lesions (SHAP values: 0.01–0.22), lower CCI scores (SHAP values: 0.01–0.72), lower creatinine levels (SHAP values: 0.01–0.33), and smaller lesion sizes (SHAP values: 0.01–0.38) were associated with a higher likelihood of TOLS. Finally, the XGBoost model was used as the algorithm for the web-based calculator, which estimates the probability of TOLS (https://tolsprediction.streamlit.app).

Long-Term Impact of TOLS

Patients who met the criteria for TOLS had an increased median OS compared with those who did not (90 vs 54 months, respectively, P < 0.001). This association remained significant even after excluding patients who died within 90 days after surgery (90 vs 58 months, P < 0.001) (Fig. 4A). On multivariable analysis, the factors that were independently associated with OS included TOLS (hazard ratio [HR] = 0.82, P = 0.015), lower age (HR = 1.02, P < 0.001), indication for surgery (benign: HR = 0.06, P < 0.001; CRLM: HR = 0.59, P < 0.001; other malignancies: HR = 0.62, P = 0.003; reference: HCC), lower MELD score (HR = 1.04, P = 0.026), fewer number of lesions (HR = 1.06, P < 0.001), and type of resection (extended right hepatectomy: HR = 1.64, P = 0.011; reference: left hepatectomy) (Table 3). Subset analyses using the HCC and CRLM cohorts also showed that patients with TOLS had an increased median OS compared with those without TOLS (Figs. 4B, C). This association remained significant in the multivariable Cox regression analysis (HR for HCC cohort: 0.8, P = 0.046; HR for CRLM cohort: 0.66, P = 0.017) (Supplemental Table 3, see http://links.lww.com/AOSO/A447).

TABLE 1.

Demographic and Preoperative Clinical Characteristics

	Overall	TOLS		
Characteristic	**N = 2059**	**No, N = 766**	**Yes, N = 1293**	***P***
Age (years), median (IQR)	63 (53–71)	65 (56–73)	62 (51–70)	**<0.001**
Sex (male), n (%)	1228 (60.0)	499 (65.0)	729 (56.0)	**<0.001**
BMI (kg/m^2), median (IQR)	24.5 (21.9–28.1)	24.7 (21.9–28.4)	24.5 (21.9–28.0)	0.7
ASA class, n (%)				
1	253 (12.0)	46 (6.0)	207 (16.0)	**<0.001**
2	999 (49.0)	403 (53.0)	596 (46.0)	
3	758 (37.0)	291 (38.0)	467 (36.0)	
4	48 (2.0)	25 (3.0)	23 (2.0)	
CCI, median (IQR)	5 (3–7)	5 (3–8)	4 (2–7)	**<0.001**
MELD, median (IQR)	7 (6–8)	7 (6–8)	7 (6–7)	**<0.001**
Preoperative portal HTN, n (%)	94 (4.6)	43 (5.6)	51 (3.9)	0.079
Preoperative varices, n (%)	88 (4.3)	46 (6.0)	42 (3.2)	**0.003**
Liver impairment, n (%)				
No	1455 (70.7)	516 (67.4)	939 (72.6)	**0.02**
Chemotherapy	37 (1.8)	15 (2.0)	22 (1.7)	
Alcohol	97 (4.7)	39 (5.1)	58 (4.5)	
Hepatitis B	128 (6.2)	53 (6.9)	75 (5.8)	
Hepatitis C	251 (12.2)	107 (13.9)	144 (11.1)	
NAFLD	91 (4.4)	36 (4.7)	55 (4.3)	
Cirrhosis, n (%)				
No	1250 (61.0)	402 (52.5)	848 (65.6)	**<0.001**
Child A	788 (38.0)	348 (45.4)	440 (34.0)	
Child B	21 (1.0)	16 (2.1)	5 (0.4)	
Ascites, n (%)	23 (1.1)	11 (1.4)	12 (0.9)	0.3
Previous abdominal surgery, n (%)	869 (42.0)	335 (44.0)	534 (41.0)	0.3
Previous liver resection, n (%)	185 (9.0)	75 (9.8)	110 (8.5)	0.3
Indication for surgery, n (%)				
Benign	406 (19.7)	56 (7.3)	350 (27.0)	**<0.001**
HCC	671 (32.6)	292 (38.1)	79 (29.3)	
CCC	192 (9.3)	106 (13.8)	86 (6.7)	
CRLM	514 (25.0)	186 (24.3)	328 (25.4)	
Other malignancies	276 (13.4)	126 (16.5)	150 (11.6)	
Number of lesions, median (IQR)	1 (1–2)	1 (1–2)	1 (1–2)	**<0.001**
Lesion size (mm), median (IQR)	31 (18–56)	35 (21–70)	28 (15–50)	**<0.001**
Preoperative hemoglobin (g/dL), median (IQR)	13.3 (12.1–14.4)	13.10 (11.7–14.3)	13.40 (12.3–14.4)	**<0.001**
Preoperative platelets (×10^9/L), median (IQR)	208 (160–261)	207 (154–254)	209 (163–264)	0.06
Preoperative INR, median (IQR)	1.0 (1.0–1.1)	1.0 (1.0–1.1)	1.0 (1.0–1.1)	**<0.001**
Preoperative creatinine (mg/dL), median (IQR)	0.8 (0.7–1.0)	0.8 (0.7–1.0)	0.8 (0.7–0.9)	**0.007**
Preoperative t-bili (mg/dL), median (IQR)	0.6 (0.5–0.8)	0.7 (0.5–0.9)	0.6 (0.5–0.8)	**0.003**
Operative approach, n (%)				
Open	1274 (61.9)	549 (71.7)	725 (56.1)	**<0.001**
Hand-assisted	136 (6.6)	38 (5.0)	98 (7.6)	
Laparoscopic	581 (28.2)	169 (22.0)	412 (31.8)	
Robotic	68 (3.3)	10 (1.3)	58 (4.5)	

P-values less than 0.05 are considered statistically significant and are presented in boldface.

HTN indicates hypertension; INR, international normalized ratio; IQR, interquartile range; NAFLD, nonalcoholic fatty liver disease; t-bili, total bilirubin.

TABLE 2.

Percentage of Patients Who Did Not Meet the TOLS Criteria

	N = 2059
Criteria	**n (%)**
Intraoperative incidents	149 (7.2)
Postoperative bile leakage	155 (7.5)
Postoperative liver failure	71 (3.4)
90-day major complications	321 (16.0)
90-day readmission	131 (6.4)
90-day or inhospital mortality	23 (1.1)
Positive resection margin	418 (20.0)

DISCUSSION

Using a large, international cohort of hepatectomy patients, we developed a machine learning model to predict TOLS with acceptable discrimination (AUC: 0.73).[18] Our model identified operative approach, number of liver lesions, CCI, creatinine, and lesion size as the most important preoperative variables for predicting TOLS. Importantly, we also demonstrated that TOLS was independently associated with improved OS on multivariable analysis.

Textbook outcome is an increasingly popular composite measure that reflects the ideal postoperative course and has been described in multiple general surgery procedures.[1,2,6,8,9] Görgec et al[7] recently defined TOLS based on the expert opinions of 44 surgeons from 22 countries and 3 international societies. The authors noted several potential advantages of using their proposed TOLS criteria over individual outcome metrics. For example, TOLS may be a more digestible summary measure for patients than needing to understand the significance of multiple, separate variables. In addition, TOLS can be used for quality improvement initiatives, as hospitals can compare their rate of TOLS to that of other centers for various procedures. In this setting, the next logical step was to create a tool that could provide individualized predictions for TOLS and also validate these Delphi consensus-based criteria using real-world patient data.

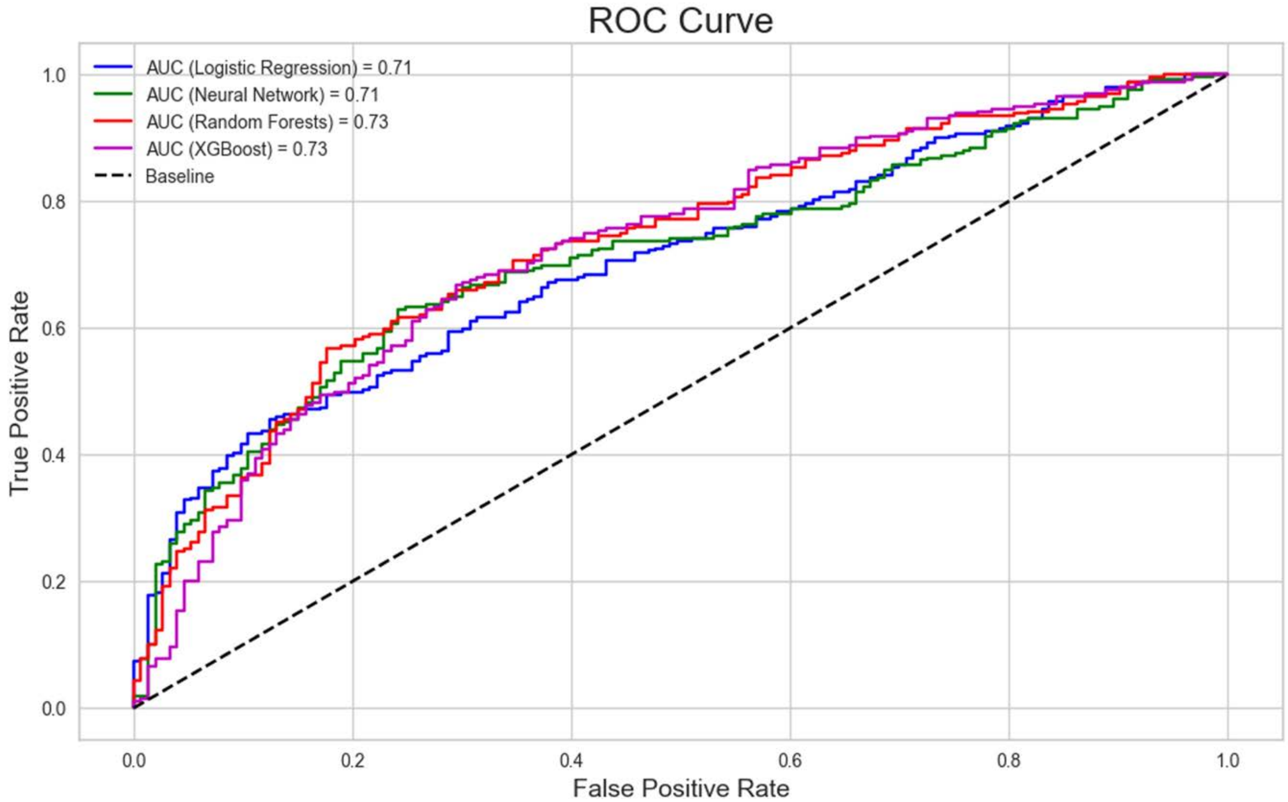

FIGURE 2. Receiver operating characteristic (ROC) curves for the logistic regression, neural network, random forest, and XGBoost models. ROC curves were plotted for the logistic regression, neural network, random forest, and XGBoost models, with AUC of 0.71, 0.71, 0.73, and 0.73, respectively.

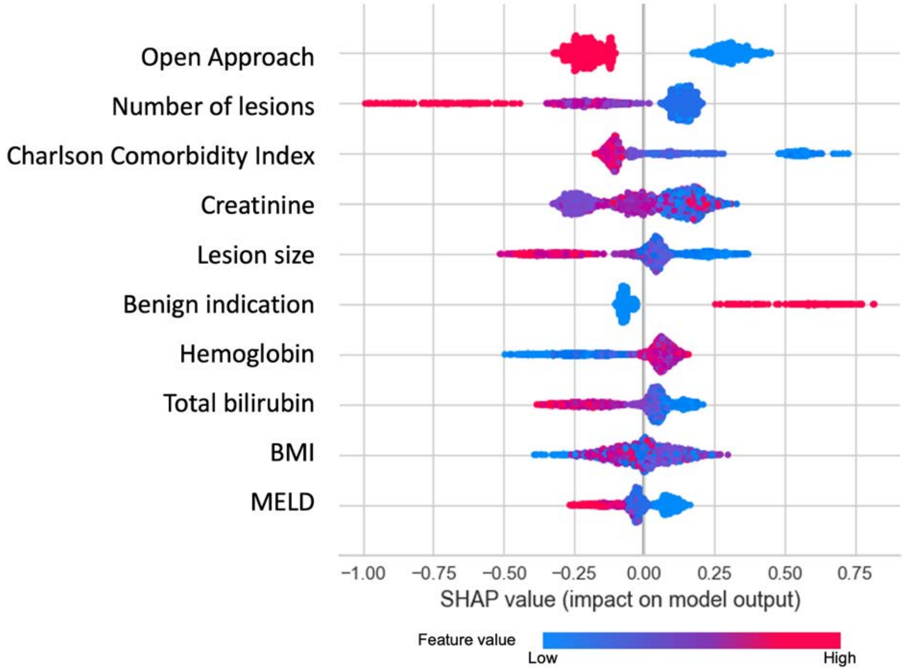

FIGURE 3. SHAP values of the XGBoost model. As seen on the *x*-axis, positive SHAP values suggest that a variable predicts a higher likelihood of TOLS, whereas negative SHAP values indicate that a variable predicts a lower likelihood of TOLS. In the SHAP plot, red represents a high value of the variable (eg, an open approach or a higher number of lesions), while blue represents a low value of the variable (eg, a nonopen (MIS) approach or a lower number of lesions). The *y*-axis reflects the overall importance of each variable in predicting TOLS in descending order. For example, if one examines the variable "number of lesions," the blue (lower number of lesions) has a positive SHAP value, indicating a higher likelihood of TOLS. Based on its location on the *y*-axis, the number of lesions is the second most important variable in predicting TOLS.

While models that incorporate intraoperative variables may have improved performance, they have limited clinical use as they cannot reliably inform care in the preoperative setting. Importantly, our machine learning model predicted TOLS with acceptable discrimination using only preoperative variables. Using the resulting web-based calculator, surgeons can

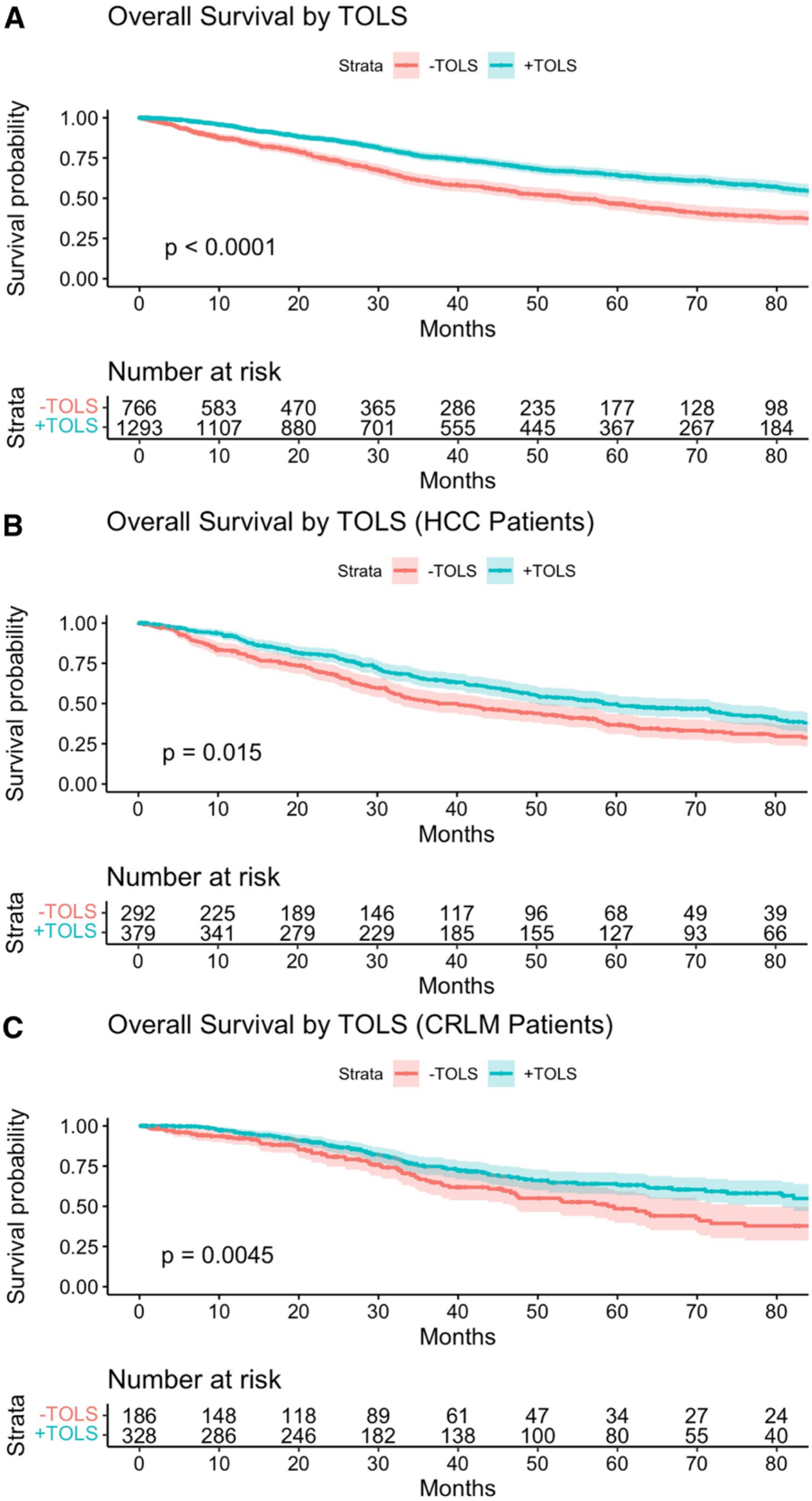

FIGURE 4. Impact of TOLS on overall survival. (A) This Kaplan–Meier survival plot compares the probability of overall survival between patients with and without TOLS. The TOLS group had a significantly higher survival probability over time (*P* < 0.0001 derived from the log-rank test). (B) HCC patients with TOLS have better overall survival than those without TOLS (*P* = 0.015). (C) CRLM patients with TOLS have better overall survival than those without TOLS (*P* = 0.0045).

counsel patients on their individualized chances of TOLS during informed consent. In addition, while patients who need surgery for malignant indications may have less flexibility, those who are considering surgery for benign indications may reconsider operative treatment if their chances of having TOLS are very slim. By using machine learning methods, we were able to achieve a slightly better model performance than the more traditional logistic regression-based model. More importantly,

TABLE 3.

Multivariable Cox Proportional Hazards Model of Overall Survival

Characteristic	HR (95% CI)	*P*
TOLS	0.82 (0.70–0.96)	**0.014**
Age	1.02 (1.01–1.03)	**<0.001**
Sex		
Female	—	—
Male	0.94 (0.80–1.11)	0.5
Indication for surgery		
HCC	—	—
Benign	0.11 (0.05–0.25)	**<0.001**
CCC	0.77 (0.58–1.02)	0.07
CRLM	0.61 (0.46–0.82)	**<0.001**
Other malignancies	0.7 (0.51–0.95)	**0.024**
CCI	1.03 (0.98–1.09)	0.2
MELD	1.05 (1.02–1.08)	**0.002**
Number of lesions	1.06 (1.03–1.09)	**<0.001**
Type of resection		
Left	—	—
Left lateral	1.3 (0.94–1.78)	0.11
Bisegmentectomy	0.98 (0.67–1.45)	>0.9
Central	1.16 (0.56–2.40)	0.7
Extended left	0.79 (0.54–1.15)	0.2
Extended right	1.74 (1.21–2.51)	**0.003**
Left medial	1.5 (0.88–2.56)	0.14
Right	1.22 (0.93–1.61)	0.2
Right anterior	1.21 (0.80–1.83)	0.4
Right posterior	1.12 (0.78–1.62)	0.5
Other	0.99 (0.76–1.30)	>0.9

P-values less than 0.05 are considered statistically significant and are presented in boldface.
CI indicates confidence interval; CCC, cholangiocarcinoma.

the machine learning models also required significantly fewer features. For example, the XGBoost model, which was ultimately selected to develop the calculator, needed 15 variables to achieve optimal model performance (AUC: 0.73), as compared with 32 variables for the logistic regression model (AUC: 0.71). This improves the usability of the web-based calculator without compromising performance.

Notably, using an MIS approach was found to be the most significant predictor of TOLS. This is certainly not surprising, as MIS hepato-pancreato-biliary surgery has been shown to yield superior perioperative outcomes compared with open surgery. For example, 1 study reported that patients undergoing MIS hepato-pancreato-biliary surgery had lower rates of postoperative morbidity and mortality compared with those who underwent open surgery.[19] Furthermore, using an MIS approach has even been shown to yield equivalent or improved oncologic outcomes.[20,21] Having fewer and smaller liver lesions was found to be the second and fifth most important factors in predicting TOLS, respectively. This makes intuitive sense, as patients with more and/or larger lesions may need more extensive resections to remove the underlying pathology, which would predispose them to such complications as posthepatectomy liver failure, an important determinant of mortality after major hepatectomy.[17] Furthermore, larger and bulkier tumors may be more challenging to resect, increasing the chances of having an R1 resection margin or an intraoperative complication. Finally, having a lower CCI score and a lower preoperative creatinine level were found to be the third and fourth most important factors in predicting TOLS, respectively. These variables reflect the underlying comorbidities of a patient, and thus it makes sense that patients who are healthier before surgery have better outcomes.

Naturally, one may conclude that clinicians should attempt to optimize these variables to increase the chances of having TOLS. For example, a surgeon may consider switching their approach from open to MIS if the initial TOLS prediction is poor. However, this raises a particularly important point regarding risk prediction models in general. Specifically, while risk calculators may serve as clinical adjuncts, their predictions reflect learned associations from the data and do not necessarily represent causal relationships between a specific variable and the outcome of interest. Thus, surgeons should not rely on such tools in isolation to inform clinical decision-making. Furthermore, while optimizing such variables as BMI may be reasonable in the appropriate clinical context, surgeons should not significantly alter their perioperative planning to “appease” the calculator. Rather, understanding the most important features for predicting TOLS may provide some insight into how the model’s prediction was derived.

Perhaps the most noteworthy finding of this study was that TOLS was independently associated with increased OS on multivariable analysis. There are a few possible explanations for this observation. For one, patients who have significant postoperative complications – particularly those who undergo surgery for malignant indications – have previously been shown to experience worse long-term outcomes.[22,23] This may be due to further immunosuppression caused by the severe physiological stress associated with major complications or from a delay in return to intended oncologic therapy. In addition, patients with positive margins ultimately failed to undergo an oncologically complete resection and thus likely experience decreased survival as well.[24] Importantly, this finding has 2 significant implications. First, it highlights the clinical relevance of the proposed TOLS criteria. Similar studies that use Delphi consensus methodologies to define textbook outcome should consider validating their proposed criteria using real-world patient data when possible. This ensures that these expert-based definitions are supported by evidence, which is essential if they are to be used to counsel patients, aid in clinical decision-making, and guide quality improvement initiatives. Second, although we validated only 1 study’s textbook outcome criteria, our findings suggest that the collective experience and expertise of a diverse group of surgeons can serve as a powerful tool. Future studies should similarly represent a wide range of countries and societies in their expert panels.

Some limitations of this study must be acknowledged. First, selection bias is likely present given the retrospective nature of this study. In addition, the cohort had significant heterogeneity in regard to patient and disease characteristics; however, not only does this increase the generalizability of our study findings, but this was intentionally done to reflect the broadness of the original TOLS criteria, which was not specific to any particular patient population, indication for surgery, operative approach, or type of resection. Furthermore, one may argue that operative approach is a technically an intraoperative variable; however, this is determined in the preoperative setting, and the rate of conversion in our cohort was relatively low. Finally, all 9 centers included in this study were either academic centers or had their own affiliated research institution; thus, the findings may not be applicable to all centers as other practice models were not represented in the cohort.

CONCLUSIONS

Our machine learning model can predict TOLS with acceptable discrimination. We also demonstrated that patients with TOLS had improved OS compared with those without TOLS; this held true on subset analysis. The next steps will include both external validation of the model as well as validation of other definitions of textbook outcome, particularly those that were also derived via expert consensus.

ACKNOWLEDGMENTS

Statistical analysis was performed by A.A.G. and J.W., under the supervision and guidance of M.A. and A.A.

The authors of the International Hepatectomy Study Group are as follows: Alison Baskin, Department of Surgery, University of California, San Francisco, San Francisco, CA; Annie Wong-On-Wing, Department of Surgery, University of California, San Francisco, San Francisco, CA; Annie Yang, Department of Surgery, University of California, San Francisco, San Francisco, CA; Devesh Sharma, Department of Surgery, University of California, San Francisco, San Francisco, CA; Taisuke Imamura, Division of Hepato-Biliary-Pancreatic Surgery, Shizuoka Cancer Center, Sunto-Nagaizumi, Shizuoka, Japan; Masanori Nakamura, Division of Hepato-Biliary-Pancreatic Surgery, Shizuoka Cancer Center, Sunto-Nagaizumi, Shizuoka, Japan; Yuya Miura, Division of Hepato-Biliary-Pancreatic Surgery, Shizuoka Cancer Center, Sunto-Nagaizumi, Shizuoka, Japan; Koki Hayashi, Department of Surgery, Keio University School of Medicine, Tokyo, Japan; Masatsugu Ishii, Department of Pediatric Surgery and Transplantation, Kumamoto University Graduate School of Medical Sciences, Kumamoto, Japan; Keita Shimata, Department of Pediatric Surgery and Transplantation, Kumamoto University Graduate School of Medical Sciences, Kumamoto, Japan; Kazuya Hirukawa, Department of Pediatric Surgery and Transplantation, Kumamoto University Graduate School of Medical Sciences, Kumamoto, Japan; Hiroki Ueda, Department of Hepatobiliary and Pancreatic Surgery, Graduate School of Medicine, Tokyo Medical and Dental University, Tokyo, Japan; June S. Peng, Department of Surgery, University of California, San Francisco, San Francisco, CA; Lucas Thornblade, Department of Surgery, University of California, San Francisco, San Francisco, CA; Kenzo Hirose, Department of Surgery, University of California, San Francisco, San Francisco, CA; Kimberly Kirkwood, Department of Surgery, University of California, San Francisco, San Francisco, CA; Eric Nakakura, Department of Surgery, University of California, San Francisco, San Francisco, CA; Carlos Corvera, Department of Surgery, University of California, San Francisco, San Francisco, CA; and Ajay Maker, Department of Surgery, University of California, San Francisco, San Francisco, CA.

REFERENCES

1. Kolfschoten NE, Kievit J, Gooiker GA, et al. Focusing on desired outcomes of care after colon cancer resections; hospital variations in “textbook outcome.”. *Eur J Surg Oncol*. 2013;39:156–163.
2. Merath K, Chen Q, Bagante F, et al. Textbook outcomes among Medicare patients undergoing hepatopancreatic surgery. *Ann Surg*. 2020;271:1116–1123.
3. Dimick JB, Staiger DO, Baser O, et al. Composite measures for predicting surgical mortality in the hospital. *Health Aff (Millwood)*. 2009;28:1189–1198.
4. Dimick JB, Birkmeyer NJ, Finks JF, et al. Composite measures for profiling hospitals on bariatric surgery performance. *JAMA Surg*. 2014;149:10–16.
5. Dimick JB, Staiger DO, Osborne NH, et al. Composite measures for rating hospital quality with major surgery. *Health Serv Res*. 2012;47:1861–1879.
6. Görgec B, Benedetti Cacciaguerra A, Lanari J, et al. Assessment of textbook outcome in laparoscopic and open liver surgery. *JAMA Surg*. 2021;156:e212064.
7. Görgec B, Benedetti Cacciaguerra A, Pawlik TM, et al. An international expert Delphi consensus on defining textbook outcome in liver surgery (TOLS). *Ann Surg*. 2023;277:821–828.
8. Kalff MC, van Berge Henegouwen MI, Gisbertz SS. Textbook outcome for esophageal cancer surgery: an international consensus-based update of a quality measure. *Dis Esophagus*. 2021;34:doab011.
9. Lim C, Llado L, Salloum C, et al. Textbook outcome following liver transplantation. *World J Surg*. 2021;45:3414–3423.
10. Ruiz-Tovar J, Boermeester MA, Bordeianou L, et al; Colorectal Delphi Facilitating Group. Delphi consensus on intraoperative technical/surgical aspects to prevent surgical site infection after colorectal surgery. *J Am Coll Surg*. 2022;234:1–11.
11. Omar I, Miller K, Madhok B, et al. The first international Delphi consensus statement on laparoscopic gastrointestinal surgery. *Int J Surg*. 2022;104:106766.
12. Korrel M, Lof S, Alseidi AA, et al; International Consortium on Minimally Invasive Pancreatic Surgery (I-MIPS). Framework for training in minimally invasive pancreatic surgery: an international Delphi consensus study. *J Am Coll Surg*. 2022;235:383–390.
13. Müller PC, Kapp JR, Vetter D, et al. Fit-for-discharge criteria after esophagectomy: an international expert Delphi consensus. *Dis Esophagus*. 2021;34:doaa101.
14. Grove TN, Kontovounisios C, Montgomery A, et al; AWR Europe Collaborative. Perioperative optimization in complex abdominal wall hernias: Delphi consensus statement. *BJS Open*. 2021;5:zrab082.
15. Kazaryan AM, Røsok BI, Edwin B. Morbidity assessment in surgery: refinement proposal based on a concept of perioperative adverse events. *ISRN Surg*. 2013;2013:625093.
16. Koch M, Garden OJ, Padbury R, et al. Bile leakage after hepatobiliary and pancreatic surgery: a definition and grading of severity by the International Study Group of Liver Surgery. *Surgery*. 2011;149:680–688.
17. 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;149:713–724.
18. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Wiley; 2013.
19. Ejaz A, Sachs T, He J, et al. A comparison of open and minimally invasive surgery for hepatic and pancreatic resections using the Nationwide Inpatient Sample. *Surgery*. 2014;156:538–547.
20. Mirnezami R, Mirnezami AH, Chandrakumaran K, et al. Short- and long-term outcomes after laparoscopic and open hepatic resection: systematic review and meta-analysis. *HPB (Oxford)*. 2011;13:295–308.
21. Topal H, Aerts R, Laenen A, et al. Survival after minimally invasive vs open surgery for pancreatic adenocarcinoma. *JAMA Netw Open*. 2022;5:e2248147.
22. Kong J, Li G, Chai J, et al. Impact of postoperative complications on long-term survival after resection of hepatocellular carcinoma: a systematic review and meta-analysis. *Ann Surg Oncol*. 2021;28:8221–8233.
23. Dorcaratto D, Mazzinari G, Fernandez M, et al. Impact of postoperative complications on survival and recurrence after resection of colorectal liver metastases: systematic review and meta-analysis. *Ann Surg*. 2019;270:1018–1027.
24. Goh BKP, Chow PKH, Teo JY, et al. Number of nodules, Child-Pugh status, margin positivity, and microvascular invasion, but not tumor size, are prognostic factors of survival after liver resection for multifocal hepatocellular carcinoma. *J Gastrointest Surg*. 2014;18:1477–1485.