ANNALS OF SURGERY OPEN
OPEN

Call to Improve the Quality of Prediction Tools for Intrahepatic Cholangiocarcinoma Resection: A Critical Appraisal, Systematic Review, and External Validation Study

Woo Jin Choi, MD,*† Richard Walker, MD,*† Luckshi Rajendran, MD, MEd,*
Owen Jones, BSc,‡ Annie Gravely,‡ Marina Englesakis, BA, MLIS,§ Steven Gallinger, MD, MSc,*‡
Gideon Hirschfield, MD, PhD,†||¶ Bettina Hansen, PhD,†¶# and Gonzalo Sapisochin, MD, PhD, MSc*†‡

Objective: To conduct a systematic review, critical appraisal, and external validation of survival prediction tools for patients undergoing intrahepatic cholangiocarcinoma (iCCA) resection.

Summary background data: Despite the development of several survival prediction tools in recent years for patients undergoing iCCA resections, there is a lack of critical appraisal and external validation of these models.

Methods: We conducted a systematic review and critical appraisal of survival and recurrence prediction models for patients undergoing curative-intent iCCA resections. Studies were evaluated based on their model design, risk of bias, reporting, performance, and validation results. We identified the best model and externally validated it using our institution's data.

Results: This review included a total of 31 studies, consisting of 26 studies with original prediction tools and 5 studies that only conducted external validations. Among the 26, 54% of the studies conducted internal validations, 46% conducted external validations, and only 1 study scored a low risk of bias. Harrell's C-statistics ranged from 0.67 to 0.76 for internal validation and from 0.64 to 0.75 for external validation. Only 81% of the studies reported model calibration. Our external validation of the best model (Intrahepatic Cholangiocarcinoma [ICC]-Metroticket) estimated Harrell's and Uno's C-statistics of 0.67 (95% CI: 0.56–0.77) and Uno's time-dependent area under the receiver operating characteristic curve (AUC) of 0.71 (95% CI: 0.53–0.88), with a Brier score of 0.20 (95% CI: 0.15–0.26) and good calibration plots.

Conclusions: Many prediction models have been published in recent years, but their quality remains poor, and minimal methodological quality improvement has been observed. The ICC-Metroticket was selected as the best model (Uno's time-dependent AUC of 0.71) for 5-year overall survival prediction in patients undergoing curative-intent iCCA resection.

Keywords: cholangiocarcinoma, prediction, resection, tool, validation

INTRODUCTION

Many outcome prediction tools have been developed in the context of intrahepatic cholangiocarcinoma (iCCA) resections, mainly to predict 5-year overall survival (OS) and recurrence-free survival (RFS).[1] Prediction tools can be presented in the form of nomograms, scores, or online calculators that are designed to assist clinicians in making individual risk calculations for their patients.[2] Having robustly designed, well-validated (both internally and externally) prediction tools for survival and recurrence after iCCA resection are needed to identify high-risk patients and recommend different and/or augmented treatment and disease surveillance strategies.[3–5]

Büttner et al.[1] conducted the most recent systematic review to assess the performance of survival prediction models for resected iCCAs, with their final study search dated July 18, 2019. However, their review only included models that underwent

*From the *Department of Surgery, University of Toronto, Toronto, Ontario, Canada; †Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; ‡University Health Network, HPB Surgical Oncology, Toronto, Ontario, Canada; §Library and Information Services, University Health Network, Toronto, Canada; ||Department of Medicine, University of Toronto, Toronto, Ontario, Canada; ¶Toronto Centre for Liver Disease, Toronto General Hospital, University Health Network, Toronto, Canada; #Department of Epidemiology & Biostatistics, Erasmus MC, Rotterdam, the Netherlands.*

This work has not previously or concurrently been submitted for publication.

The protocol for this study was registered with PROSPERO (CRD42022384741). The reporting for this study adheres to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement and the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist. The Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines were also used to report outcomes.

G.S. discloses consultancy for Astra-Zeneca, Roche, Novartis, Evidera, Integra and HepaRegeniX. G.S. has received financial compensation for talks for Roche, Astra-Zeneca, Chiesi, and Integra. G.S. has received a grant from Roche. The other authors declare that they have nothing to disclose.

W.J.C. has been supported by the Canadian Institutes of Health Research (CIHR) [FRN: 181365, 2022] for his PhD studies.

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: Gonzalo Sapisochin, MD, PhD, MSc, University of Toronto, Staff Surgeon, HBP & Multi Organ Transplant Program, Division of General Surgery. University Health Network, 585 University Avenue, 9-MaRS-9047B, Toronto, M5G 2N2, ON, Canada. Email: Gonzalo.sapisochin@uhn.ca.

Annals of Surgery Open (2023) 3:e328

Received: 19 July 2023; Accepted 24 July 2023

Published online 1 September 2023

DOI: 10.1097/AS9.0000000000000328

external validation.[1] Since then, several contemporary prediction tools have been published on this topic, but it is uncertain how extensively these models have been critically appraised or validated.[6–9] Reporting critical appraisal and external validation results will help assess model accuracy, generalizability to other populations, and potential clinical utility.[10]

The primary aim of this systematic review was to summarize the evidence and critically appraise prognostic models for survival and recurrence in patients who underwent curative-intent iCCA resection. The secondary aim was to externally validate and assess the performance of the selected best models.

METHODS

Protocol and Reporting

The protocol for this study was registered with PROSPERO (CRD42022384741). The reporting for this study adheres to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement and the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies checklist.[11,12] The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines were also used to report outcomes.[13]

Eligibility Criteria

All randomized/quasi-randomized trials and cohort studies were included. Review articles, meta-analyses, case series, and cross-sectional studies were excluded. The population included were adults, 18 years of age or older, who underwent curative-intent iCCA resection. We excluded those diagnosed with extrahepatic cholangiocarcinoma, hilar or perihilar cholangiocarcinoma (or Klatskin tumor), gallbladder cancer, hepatocellular carcinoma (HCC), or mixed-type HCC-cholangiocarcinoma. For the interventions, we included studies that developed, validated (internal or external), or updated a prognostic model based on a statistical method (excluding consensus statements) and produced a readily-usable quantitative clinical tool, such as scoring systems, nomograms, or online calculators designed for individual patient risk calculation. However, we excluded tools designed for the diagnosis or screening of iCCA, those analyzing only 1 or 2 prognostic/risk factors, those based on previously established cancer staging systems (ie, American Joint Committee on Cancer system), and those incorporating prognostic factors that are not readily obtainable in the clinical setting (ie, protein or DNA expression).

Information Sources and Search Strategy

Our academic hospital information specialist (M.E.) developed the search strategies, in conjunction with the authors (Supplemental Table 5, http://links.lww.com/AOSO/A246). Key search terms were determined from a scoping search of the literature and consultation with experts in the field. Preliminary searches were conducted, and full-text literature was mined for potential keywords and appropriate controlled vocabulary terms (such as Medical Subject Headings for MEDLINE and EMTREE descriptors for Embase). The Yale MeSH Analyzer was used to facilitate the MeSH and text word analysis.[14] In addition, citation searching (backward and forward) of target citations was conducted to glean additional potential terms.

The databases MEDLINE, MEDLINE In-Process/ePubs, Embase, Cochrane Central Register of Controlled Trials, and the Cochrane Database of Systematic Reviews, all via the Ovid platform, were searched from inception, but the results were limited to 2010 Jan 1, 2010 to Aug 16, 2022. The 2010 cut-off was selected based on the publication year of the ABC-02 trial, which has changed practice for iCCA treatment.[15] The search component blocks used were: "cholangiocarcinoma" and "intrahepatic" and "surgery" and "prediction tools", and "survival". All components included controlled vocabulary and text word terms. The searches were limited to humans, adults, and conference materials were removed when possible. No language limits were applied. No gray literature was searched. Study authors were planned to be contacted only if clarification was needed. The outcome eligibility was to include a dimension of time (survival analysis) estimating at least 3-year RFS or OS after curative-intent iCCA resection.

Selection Process

PRISMA 2020 flow diagram is presented in **Figure 1**. Article abstracts identified in the search were independently screened by 5 authors (W.C., R.W., L.R., O.J., and A.G.). These same 5 authors then assessed the full-text articles. Reviewer disagreements were resolved by consensus and involvement with a senior reviewer (G.S.) as needed. Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia) was used for screening and full-text selections.

Data Collection Process

Included studies had baseline characteristics and outcome data extracted in duplicates using a standardized template designed based on the TRIPOD reporting guidelines.[13] The critical appraisal of the model focused on assessing its design, performance, in terms of discrimination and calibration, and the internal and external validation results.

Data Items

The primary outcome measure for this study was the model's discrimination performance, which was evaluated through apparent performance or internal validation. The secondary outcome measures included the model's discrimination performance through external validation and calibration, as well as the 3- and 5-year OS and RFS of the overall cohort, reported with 95% confidence intervals (CI), where available. Additional measured outcomes in this study included the predictor selection method and the comparison of predictor distribution between the internal and external validation cohorts (if applicable). The study authors, publication years, population, design, model predictor characteristics, and reporting/handling of missing data were collected (Supplemental Table 1, http://links.lww.com/AOSO/A246). TRIPOD analysis types were collected and categorized as the following: Type 1a (model development only), Type 1b (development and validation using resampling), Type 2a (random split-sample development and validation), Type 2b (nonrandom split-sample development and validation), Type 3 (development and validation using separate data), and Type 4 (validation only).[13]

Study Risk of Bias Assessment

The prediction Model Risk of Bias Assessment Tool (PROBAST) was used to assess the studies in duplicates.[2] Disagreements between reviewers were resolved by discussion and consensus. Each study obtained an additional overall low *versus* moderate *versus* high-risk rating depending on individual component gradings. If at least 1 component received a moderate or high risk of bias, the overall rating was reflected with the same highest risk of bias grading.

Synthesis and Selection of Best Model

Nonquantitative data were described in table formats and qualitatively synthesized in the discussion section. External validation

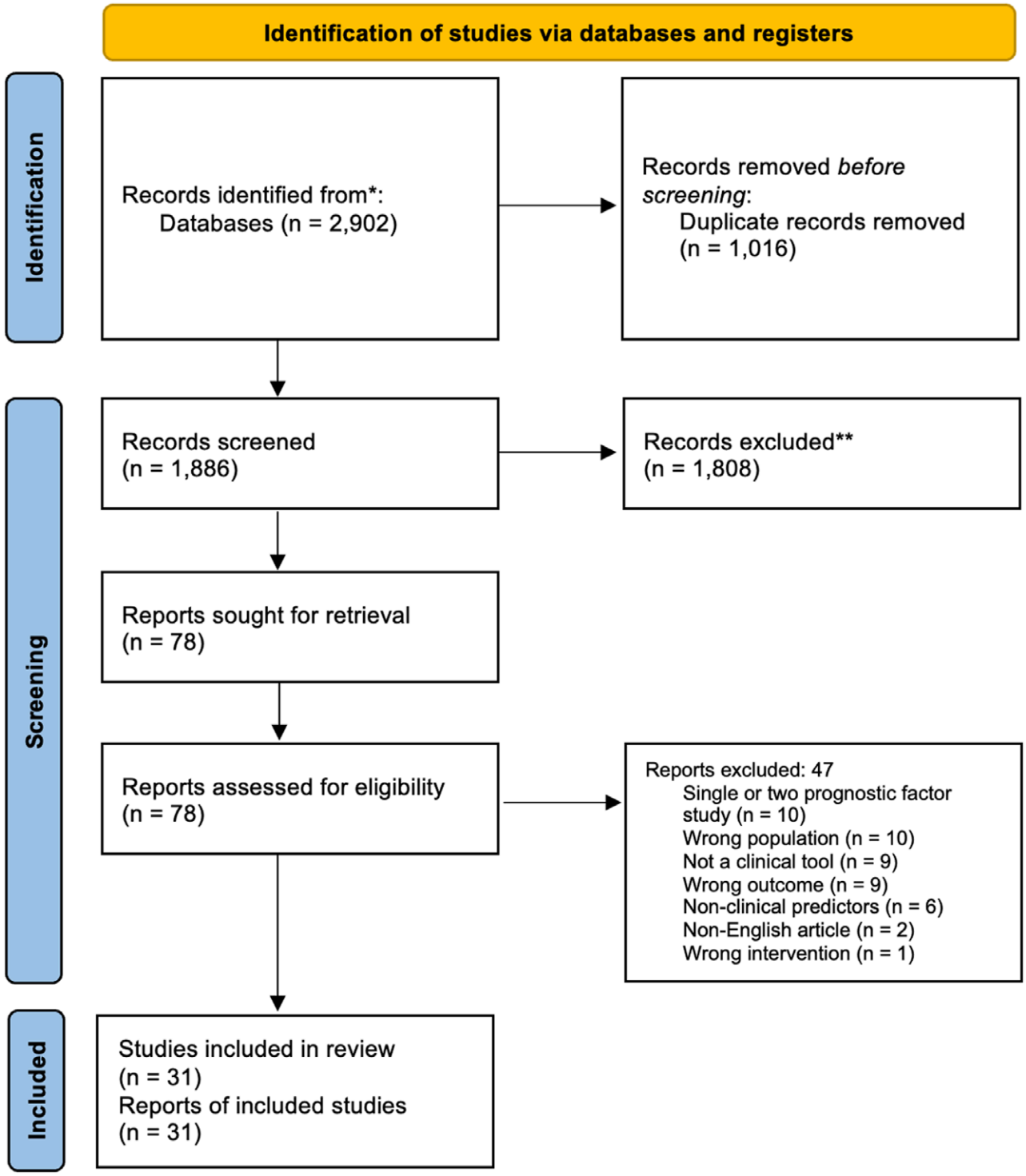

FIGURE 1. PRISMA 2020 flow diagram.

was conducted for the selected best model (TRIPOD analysis type 4). The best model was selected in the ranked order of: (1) lowest risk of bias in the tool development process (2) best performance in discrimination (3) use of comprehensive and clinically relevant predictor variables, and (4) tool that is readily available for use by clinicians. We used our institution's database to externally validate the selected best model. For missing information on the prediction model, study authors were contacted to retrieve the baseline survival and the model equation. The baseline characteristics and relevant survival analysis metrics of the original cohort were compared to those of our validation cohort. As a supplementary analysis, the external validation was repeated for the best model selected from among the models that solely utilized preoperative predictors.

External Validation Cohort

A retrospective cohort dataset from the Toronto General Hospital, University Health Network was used for external validation. All patients 18 years old or older who underwent a curative-intent operation for histologically documented iCCA between October 2005 and October 2017 were included.

Patients histologically diagnosed with extrahepatic cholangiocarcinoma, HCC, or mixed-type HCC-cholangiocarcinoma were excluded. The University Health Network's Research Ethics Board approved this study protocol (#18-5233).

External Validation Methods

Measures of performance, including discrimination, calibration, and overall performance, were estimated. $S_0(t)$ or baseline risk, as well as the model equation, including beta coefficients, were obtained from the original study. Three different discrimination measures were used in evaluating the predictive accuracy of the model, in both fixed time and time range approach: (1) Harrell's C-statistic (2) Uno's C-statistic for time range discrimination measures, and (3) Uno's time-dependent area under the receiver operating characteristic curve (AUC) calculated as a fixed-time discrimination measure at 5 years.[16] For Harrel's C-statistic, we measured the degree of concordance by assessing the probability that a randomly selected patient who experiences an event at a given time has a higher predicted risk of experiencing the event compared with a patient who experiences the event at a later time.[16] Uno's C-statistic additionally applied time-dependent weighting that better adjusts for censoring.[16] For Uno's time-dependent AUC, we measured the discrimination between patients who experienced an event and those who did not, assessing whether the predicted probabilities of experiencing the event were higher for patients who actually experienced the event over 5 years, compared with patients who remained event-free at 5 years.[16]

Calibration at a fixed-time point (ie, 5 years) was measured using the mean (level 1), weak (level 2), and moderate (level 3) calibrations in the order of increasingly robust checks described in a previously proposed framework.[16,17] Mean calibration at 5 years was estimated using the observed *versus* expected ratio (OE ratio). A complementary of the Kaplan–Meier curve was used to estimate the observed survival fraction, and the average predicted risk was used to estimate the expected.[16] Mean calibration was reported as a ratio with 95% CI, and a ratio closer to 1 indicates better calibration. Weak calibration at 5 years was estimated by measuring the calibration slope and fitting a Cox proportional hazards model with the prognostic index from the original model as the only covariate in the validation dataset with censoring at 5 years.[16] Weak calibration was reported as slope (or regression coefficient of the prognostic index) with 95% CI, and a slope closer to 1 indicates better calibration. Moderate calibration at 5 years was estimated using a flexible calibration curve (instead of linear), using the predicted risk from a Cox proportional hazards model against the predicted risk from the developed model to detect miscalibration, which could not be detected in mean and weak calibrations.[16] The moderate calibration was presented as a calibration graph, in which a 45-degree line indicates perfect calibration.

To assess the overall performance of the model, we measured the Brier score by estimating the mean squared difference between the observed event indicators and the predicted risks of experiencing the event at 5 years. The Brier score was reported with 95% CI, and a lower score closer to zero indicates better performance.

Missing Data

All variables with missing values greater than 5% were plotted to determine if they were missing completely at random. If so, several methods for handling missing data were implemented, and the resulting datasets' external validation performance measures were compared. The following methods were used to handle missing values: (1) complete-case analysis, (2) multiple imputations, (3) replacement of missing values with the median value of the missing variable, and (4) replacement of missing values with the weighted mean of the missing variable. The 'mice' package was used for multiple imputations. For the inverse-probability weighting (IPW) approach, the probability of the variable missing was estimated using logistic regression, the inverse of these probabilities was used to calculate weights, which then the calculated weighted means were used to replace missing values.

RESULTS

Study Selection

Our initial search strategy identified a total of 2902 studies, of which 1016 were duplicates. After the initial title and abstract screening, 1808 abstracts were excluded for not meeting our inclusion criteria. We retrieved a total of 78 full-text articles. Of the 78 full-text articles screened, 47 were excluded with reasons presented in the PRISMA flow diagram (**Fig. 1**). A total of 31 studies were included in this study.

Study Characteristics

The study characteristics of all the included studies are presented in Supplemental Table 1, http://links.lww.com/AOSO/A246 **and Table 1**.[6–9,18–44] Twenty-six studies[6–9,18–31,33–38,42,44] developed their own original prediction models (Supplemental Table 1, http://links.lww.com/AOSO/A246), and 5 studies[32,39–41,43] conducted external validations on previously published models (**Table 1**). Eighteen studies (69%) published nomograms,[8,18,19,21–30,33,36,37,42,44] 7 studies (27%) published scoring systems,[6,9,20,31,34,35,38] and 1 study (4%) published an online calculator.[7] All 26 original models were developed using a retrospective cohort, and 9 studies (35%) used single-center data to develop the models. The population used to develop the models ranged from 1990 to 2019, and the model derivation sample size ranged from 83 to 1323. Model derivation cohorts were from the following: 15 studies (58%) from China, 7 studies (27%) from international collaborations, 2 studies (8%) from the USA, 1 study (4%) from Japan, and 1 study (4%) from Taiwan. The following were the distribution of the TRIPOD analysis types: 6 Type 1a (model development only), 7 Type 1b (development and validation using resampling), 1 Type 2a (random split-sample development and validation), 2 Type 2b (nonrandom split-sample development and validation), 10 Type 3 (development and validation using separate data), and 5 Type 4 (validation only). Lymph node metastases, adjuvant chemotherapy, and R1 resection rates ranged from 10.5% to 45.3%, 19.2% to 44.9%, and 3.9% to 39.5%, respectively. Most frequently used predictors ($n > 1$) are presented in **Figure 2**. Among the preoperative predictors, albumin, tumor size on imaging, and neutrophil-to-lymphocyte ratio (NLR) were the 3 predictors most frequently utilized. Among the postoperative predictors, lymph node status, tumor size on pathology, and tumor number on pathology were the 3 predictors most frequently utilized.

Risk of Bias in Studies

The risk of bias assessments using the PROBAST tool is presented in **Table 2.** A total of 26 studies that produced an original tool were rated using the PROBAST tool. One study (4%) was rated with a low overall risk of bias.[7] Two studies (8%) were rated as unclear risk of bias, and the remaining 23 studies (88%) were rated as high risk of bias.[6,34]

Model Performance and Evaluation

Model performance and evaluation of original models are presented in Supplemental Table 2, http://links.lww.com/AOSO/

TABLE 1.
Characteristics of Studies That Have Conducted External Validation Only

First/Last Author, Year (Country)	TRIPOD Analysis Type	Original Model for Validation	Source of Data	Population	Exclusion	Study Period	Validation Sample Size	# of Events	Adjuvant Chemotherapy	R1 (%)	Median Follow-up (Month)	Overall OS (%, 3/5 yr)	# of Participants with Any Missing Value	Handling of Missing Data	External Validation Results	Comparison of the Distribution of Predictors
Brustia/Scatton, 2020, France	4	PRS (preoperative risk score, Sasaski et al.)	R/11 centers, International	Curative-intent iCCA resection	R2 resection, missing data to calculate PRS	2001–2018	355	174	21.4	17.5	41.7 (IQR 32.8–50.6)	40.0/20.0	Reported	None	C-index, 0.61 (95% CI 0.56–0.67)	Yes
Buettner/Pawlik,2017, USA	4	Wang nomogram, AJCC 7th, LCSGJ, SHPBSJ, Okabayashi, Nathan, Hyder nomogram	R/12 centers, International	curative-intent iCCA resection	Liver Transplant	1990–2016	1054	…	…	13.1	37.7	51.5/39.3	Reported	Multiple imputations	C-index, Wang 0.668/ Nathan 0.639, LCSGJ 0.63, SHSJ 0.61, Okabayashi 0.61, Hyder 0.60	Yes
Doussot/Jarnagin, 2015, USA	4	Wang nomogram, Hyder nomogram, AJCC 7th, Fudan score	R/Single-cen-ter, USA	Curative-intent iCCA resection	Distant metastasis, postop death within 90days, mixed-type tumor	1993–2013	188	98	27.1	19.6	42.5 (range 5–192)	59.0/45.0	Reported	Complete-case	C-index, Wang 0.72, Hyder 0.66, AJCC 0.63, Fudan 0.55	Yes
Hahn/Koechkner, 2020, Germany	4	MEGNA, AJCC 8th	R/Single-cen-ter, Germany	Curative-intent iCCA resection	Missing imaging and lost to f/u	1997–2018	417	166	…	14.8	21.7	…	…	Complete-case	C-index, 0.58, Brier scores 0.193	Yes
Schnitzbauer/Ruckert, 2020, Germany	4	MEGNA	R/10 centers, Germany	Curative-intent iCCA resection	metastatic disease, exploratory surgeries	2004–2013	488	…	…	25.0	18 (range 0–204)	…	Reported	Excluded if 25% or more missing	KM stratification based on scores	Yes

"…" indicates not reported; AJCC, American Joint Committee on Cancer; ALP, alkaline phosphatase; AFP, alpha-fetoprotein; bin, binary variable; CEA, carcinoembryonic antigen; CT, computed tomography; DFS, disease-free survival; f/u, follow-up; iCCA, intrahepatic cholangiocarcinoma; IQR, interquartile range; KM, Kaplan–Meier, LCSGJ, Liver Cancer Study Group of Japan; OS, overall survival; R1, positive margin resection; R2, macroscopic residual tumor; SHPBSJ, Society of Hepatobiliary Surgery Japan; yr, year.

A246. Among the 26 studies that have developed an original model, 8 studies used backward selection, 17 studies (65%) used forward selection, and 1 study (4%) used a priori selection to choose predictors for building the model. The selected prognostic factors varied significantly across the studies. Eight studies (31%) designed their tools using preoperative predictors only. Of the 26 studies that developed original tools, 24 studies (92%) used Cox regression analyses, and the remaining 2 studies (8%) used either a combined Cox and logistic ranking system or a machine-learning method for model development. Median follow-up time range from 12.2 to 50.3 months. The 5-year OS and RFS of included patient samples across the studies ranged from 6.6 to 42.7 months and 5.8 to 59.0 months, respectively. One study (Ma *et al.*) used the date of diagnosis as time-zero (anchor point in survival analysis), 19 studies (73%) used the date of surgery, and 6 studies (23%) did not report their time-zero definitions. Only 6 studies (23%) reported their absolute number of outcome events.[6,22,34,36,38,44]

For measurement of model discrimination, Harrell's C-statistic was most widely used across 20 studies (77%), followed by Kaplan–Meier methods in 3 studies (11.5%), AUC in 2 studies (8%), and discriminatory capacity index (R^2) in 1 study (4%). Of the 26 studies that have developed an original model, only 14 studies (54%) conducted internal validation, in which 9 studies (35%) used bootstrapping and 5 studies (19%) used random splitting methods. Among the studies that have conducted internal validations, Harrell's C-statistics ranged from 0.67 to 0.76. Of the 26 studies that have developed an original model, only 12 studies (46%) conducted external validation as part of the original study, in which 3 studies (11%) used temporal validation (data from the same institution but different time periods) and 9 studies (35%) used data from an external institution(s). Among the studies that have conducted external validations, the Harrell's C-statistics ranged from 0.64 to 0.75. Out of the 26 studies, only 21 studies (81%) reported the calibration of their developed model, in which 17 studies (65%) reported calibration curves, 1 study (4%) reported Brier score, 1 study (4%) reported χ^2, 1 study (4%) reported Kaplan–Meier curves, and 1 study (4%) used R^2 value.

Results From the Studies That Have Conducted External Validation Only

Five studies[32,39–41,43] only conducted external validations on previously published models (**Table 1**).

In total, the Wang nomogram[29] was externally validated by 2 studies,[40,41] the MEGNA score[34] was externally validated by 2 studies,[32,39] and the Hyder nomogram[37] was externally validated by 2 studies.[40,41] The preoperative risk score (PRS)[6] and the Fudan score[35] were externally validated by 1 study.[43] Among these external validation studies, the Harrell's C-statistics ranged from 0.58 to 0.72.

External Validation of the Selected Best Model

Out of the 26 original studies that have developed a prediction model, the study from Sahara et al.[7] (from here on, "ICC-Metroticket") was selected for designing and producing the best model/prediction tool. Intrahepatic Cholangiocarcinoma (ICC)-Metroticket (1) scored low risk of bias in every domain relating to model design, (2) reported Harrell's C-statistics of 0.70 on apparent performance, and 0.67 on internal validation through 5000 times bootstrapping, (3) utilized a comprehensive and clinically relevant list of 11 predictor variables (CA19-9, tumor size, tumor number, tumor grade, resection margin status [R0 vs. R1], N stage [N0, N1, or Nx], type of hepatectomy [major, defined as resecting 3 or more Couinaud segments vs. minor], cirrhosis status, major vascular invasion, minor vascular invasion, and receipt of adjuvant chemotherapy), and (4) published

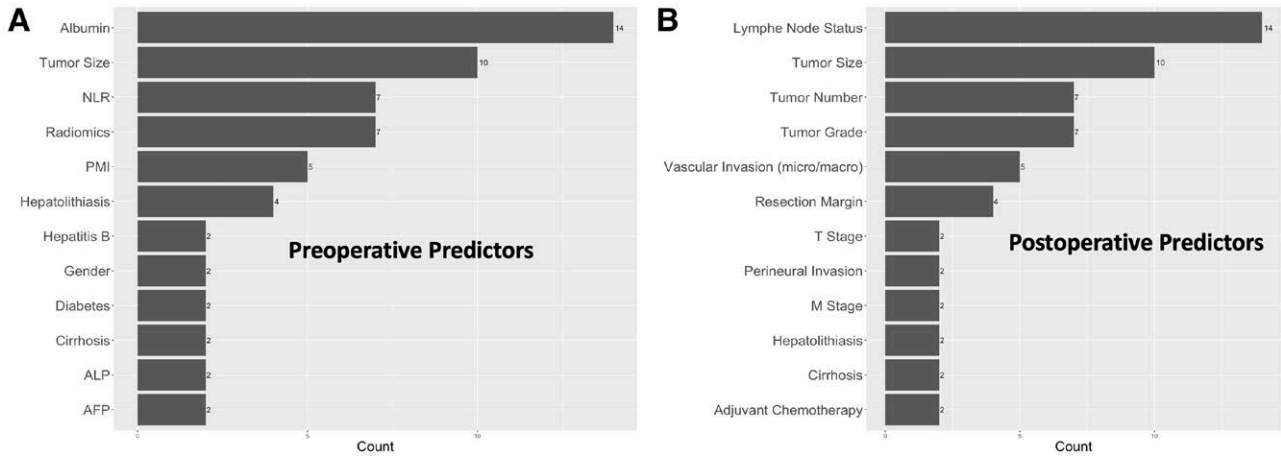

FIGURE 2. Most frequently used predictors (n > 1) divided into preoperative and postoperative variables.

TABLE 2.

Risk of Bias Assessment Using the Prediction Model Risk of Bias Assessment (PROBAST) tool

	Risk of Bias					Applicability			
First/Last Author, Year (Country)	Population	Predictors	Outcome	Analysis	Overall	Population	Predictors	Outcome	Overall
Bagante/Pawlik, 2018, USA	+	+	+	–	–	+	+	+	+
Cai/Wu, 2021, China	+	+	–	–	–	+	+	+	+
Deng/Chen, 2021, China	+	+	+	–	–	+	+	+	+
Deng/Chen, 2021, China	+	+	+	–	–	+	+	+	+
He/Lin, 2021, China	+	+	+	–	–	+	+	+	+
Hyder/Pawlik, 2013, USA	+	+	+	–	–	+	+	+	+
Hyder/Pawlik, 2014, USA	+	+	+	–	–	+	+	+	+
Jeong/Chen, 2021, China	+	+	+	–	–	+	+	+	+
Jeong/Xia, 2017, China	+	+	+	–	–	+	+	+	+
Jiang/Chen, 2010, China	+	+	+	–	–	+	+	+	+
Li/Chen, 2021, China	+	+	+	–	–	+	+	+	+
Li/Jiang, 2021, China	+	+	?	–	–	+	+	+	+
Li/Tang, 2021, China	+	+	–	–	–	+	+	+	+
Ma/Liang, 2019, China	+	+	+	–	–	+	+	+	+
Raoof/Singh, 2017, USA	+	+	+	?	?	+	+	+	+
Sahara/Pawlik, 2019, USA	+	+	+	+	+	+	+	+	+
Sasaki/Pawlik, 2018, USA	+	+	+	?	?	+	+	+	+
Sotiropoulos/Sgourakis, 2010, Germany	+	+	+	–	–	+	+	+	+
Sui/Fujiwara, 2021, Japan	+	+	+	–	–	+	+	+	+
Tang/Ma, 2021, China	+	+	+	–	–	+	+	+	+
Tsilimigras/Pawlik, 2020, USA	+	+	+	–	–	+	+	+	+
Wang/Shen. 2013, China	+	+	+	–	–	+	+	+	+
Wu/Hu, 2019, China	+	+	+	–	–	–	+	+	-
Yeh/Chen, 2016, Taiwan	+	+	+	–	–	+	+	+	+
Yu/Chen, 2021, China	+	+	+	–	–	+	+	+	+
Zhao/Zhang, 2021, China	+	?	?	–	–	+	+	+	+

"+" indicates low risk of bias (ROB)/low concern regarding applicability; "–", high ROB/high concern regarding applicability; and "?" indicates unclear ROB/unclear concern regarding applicability.

as an online calculator readily available for use in clinical practice.[7] The first author of this study has provided us with both the baseline survival, $S_0(t)$ or baseline risk and the model equation including beta coefficients (0.239 * [cirrhosis yes: 1, no: 0] + 0.963 * [CA19-9 >200: 1, <=200: 0] –0.094 * [type of resection major: 1, minor:0] +0.631 * [number summed to size >7: 1, <=7: 0] +[Nx: 0.519, N1: 1.054, N0: 0] + 0.542 * [margin R1: 1, R0: 0] + 0.554 * [poor to undifferentiated: 1, well to mod: 0] + 0.385 * [major vascular invasion yes: 1, no: 0] + 0.058 * [minor vascular invasion yes: 1, no: 0] –0.431 * [adjuvant chemotherapy yes: 1, no: 0]). The comparison of the characteristics between the original study cohort and our validation cohort is presented in Supplemental Table 3, http://links.lww.com/AOSO/A246.

Results of the external validation are presented in Table 3, Figure 3, and Supplemental Figure 1, http://links.lww.com/AOSO/A246. In the validation dataset, 70 (58%) CA19-9 and 2 (1%) tumor differentiation/grade were missing. The CA19-9 was missing completely at random, and the external validation was conducted in several methods of handling missing data: (1) complete-case analysis (n = 50), (2) multiple imputations (n = 121), (3) replacement of missing values with the median value of the CA19-9 from the original cohort (n = 119), and (4) IPW approach, replacing the missing values with the weighted mean of the missing variable (n = 119). The validation data used in the complete-case form performed best, with Harrell's and Uno's C-statistics of 0.67 (95% CI:0.56–0.77), Uno's time-dependent AUC of 0.71 (95% CI: 0.53–0.88), and Brier score of 0.20 (95% CI:0.15–0.26). The moderate calibration plots are presented in **Figure 3**. Following the complete-case analysis, the IPW approach performed best with Harrell's and Uno's C-statistics of 0.65 (95% CI:0.58–0.71) and 0.65 (0.58–0.72), respectively, Uno's time-dependent AUC of

TABLE 3.

External Validation of the ICC Metroticket Model by Sahara *et al*

Performance Measures	External Validation (Complete-case)	External Validation (Multiple Imputations)	External Validation (Median Value ca19-9)	External Validation (IPW for ca19-9)	Apparent Model Performance	Internal Validation
n=	50	121	119	119	643	643
Discrimination						
Time range:						
Harrell's C-statistic (95% CI)	0.67 (0.56–0.77)	0.63 (0.57–0.70)	0.65 (0.58–0.72)	0.65 (0.58–0.71)	0.70	0.67 (5000 bootstrapping)
Uno's C-statistic (95% CI)	0.67 (0.56–0.77)	0.64 (0.57–0.70)	0.66 (0.59–0.72)	0.65 (0.59–0.72)	…	…
Fixed time at 5 years:						
Uno's time-dependent AUC (95% CI)	0.71 (0.53–0.88)	0.70 (0.59–0.82)	0.70 (0.59–0.81)	0.70 (0.59–0.82)	…	…
Calibration						
Fixed time at 5 years:						
Mean calibration (O/E, 95% CI)	1.19 (0.83–1.69)	1.05 (0.83–1.34)	0.90 (0.71–1.16)	1.07 (0.84–1.37)	…	…
Weak calibration (Slope, 95% CI)	0.72 (0.22–1.22)	0.60 (0.26–0.94)	0.86 (0.47–1.25)	0.70 (0.35–1.05)	…	…
Overall performance						
Fixed time at 5 years:						
Brier score (95%CI)	0.20 (0.15–0.26)	0.22 (0.18–0.26)	0.22 (0.18–0.27)	0.22 (0.18–0.26)	…	…

AUC indicates area under the curve; CI, confidence intervals; IPW, inverse-probability weighting; O/E, Observed number of events/Expected number of events.

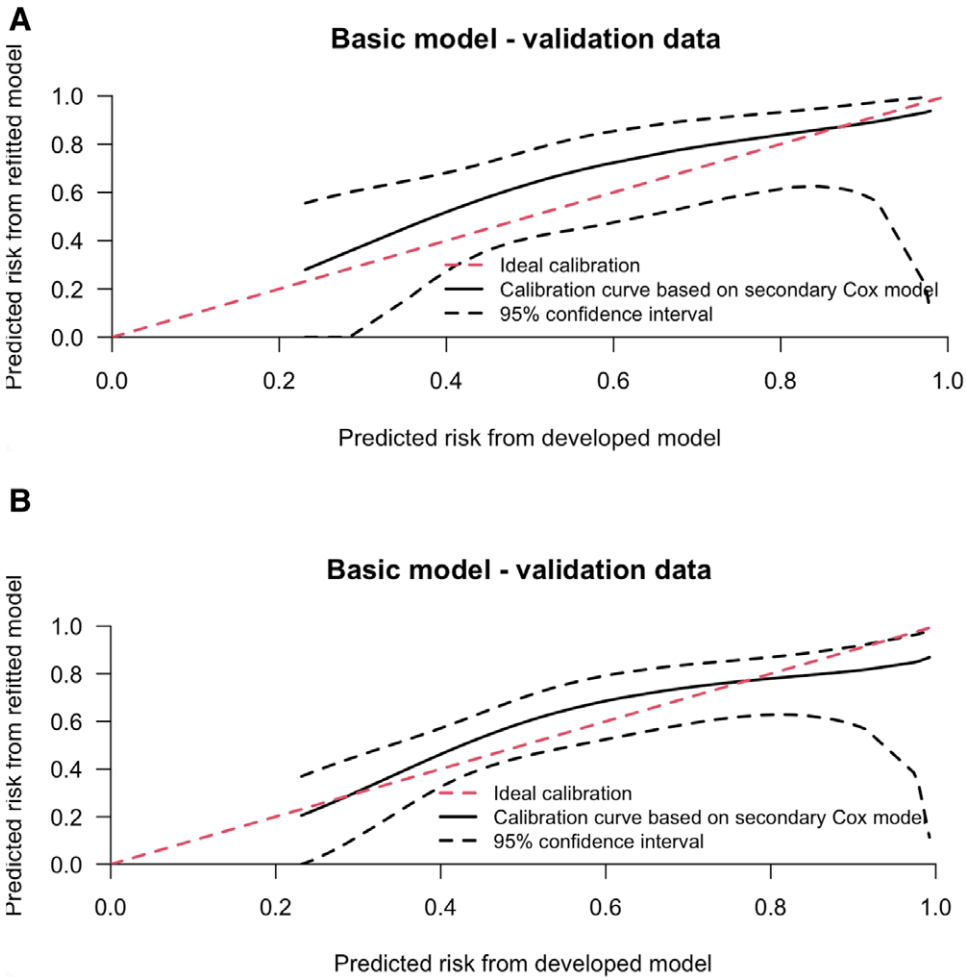

FIGURE 3. Moderate calibration plots at 5 years after handling missing Ca19-9 with: (a) Complete-case analysis (b) inverse-probability weighting (IPW) approach.

0.70 (95% CI:0.59–0.82), and Brier score of 0.22 (95% CI:0.18–0.26). The supplementary results of externally validating the best pro-operative model are presented in Supplemental Table 4, http://links.lww.com/AOSO/A246 and Supplemental Figure 2 http://

links.lww.com/AOSO/A246. Among the models that solely utilized preoperative predictors, the study from Sasaki et al.[6] (from here on, "preoperative risk score") was selected based on the same criteria above. Preoperative predictors used in this score were albumin, NLR, CA19-9 and tumor size (9 + [−2.79 × Alb] + [0.50 × NLR] + [2.81 × logN CA19-9] + [1.12 × Tumor size]).[6] The validation dataset was missing preoperative albumin values and therefore the median serum albumin level (4.2 g/dL) reported by Sasaki et al.[6] was used instead. Complete-case analysis (n = 51) was conducted for external validation, and Harrell's and Uno's C-statistics were 0.65 (95% CI:0.56–0.76) and 0.67 (0.57–0.75), respectively, and Uno's time-dependent AUC of 0.73 (95% CI:0.57–0.89). The Kaplan–Meier curve, stratified into score categories following the methods outlined in the Sasaki et al.[6] article, demonstrated a significant log-rank value (P = 0.002, Supplemental Figure 2, http://links.lww.com/AOSO/A246).

DISCUSSION

In this study, we systematically reviewed and critically appraised 31 studies consisting of 26 original survival prediction tools for resected iCCAs. Additionally, we conducted an external validation of the selected best model to evaluate the tool's generalizability. Out of the 26 studies that developed an original model, only 14 studies, (54%) conducted internal validation, with Harrell's C-statistics ranging from 0.67 to 0.76. Only 12 studies (46%) conducted external validation as part of the original study, with Harrell's C-statistics ranging from 0.64 to 0.75. Furthermore, only 21 studies (81%) reported model calibration. We selected ICC-Metroticket by Sahara et al.[7] as the overall best model, and external validation using our institution's data estimated Harrell's and Uno's C-statistics of 0.67 (95% CI:0.56–0.77), Uno's time-dependent AUC of 0.71 (95% CI:0.53–0.88) at 5 years, Brier score of 0.20 (95% CI:0.15–0.26), and good moderate calibration plots. We also externally validated the best preoperative model by Sasaki et al.[6] and estimated Harrell's and Uno's C-statistics of 0.65 (95% CI:0.56–0.76) and 0.67 (0.57–0.75), respectively, and Uno's time-dependent AUC of 0.73 (95% CI:0.57–0.89).

The most recent systematic review evaluating the performance of survival prediction models for resected iCCAs was conducted by Büttner et al.[1] with their last study search dated July 18, 2019. Our systematic review provides the most up-to-date evidence, with a search date of August 16, 2022. In contrast to Büttner et al.[1] who only included externally validated models, we included all original prediction tools regardless of their validation status, and studies published from January 1, 2010, to account for the practice-changing ABC-02 trial publication.[15] Despite Büttner et al.'s[1] report of poor methodological quality in most models published before July 2019 and their recommendations to follow the TRIPOD guidelines for prediction model reporting, there was little improvement observed among studies published between July 2019 and August 2022.[1,13] We identified 16 prediction models published since 2019, and all but one of these models were found to have moderate to high risk of bias, primarily due to insufficient reporting on handling missing data, the use of P value based forward selection for predictor selection, and a lack of reporting on key study characteristics such as median follow-up or time-zero definitions. Furthermore, none of these studies reported their $S_0(t)$ or baseline risk function, which is an essential component when validating a Cox prognostic model in an external dataset.[16,45] To improve the quality of prediction tools, key strategies include utilizing comprehensive multicenter databases, employing proper methodologies (with the assistance of a biostatistician) for variable selection, handling missing data, model parameter optimization, discrimination and calibration measurement, and validation, as well as standardizing reporting practices and implementing stricter peer-review processes.

Büttner et al.[1] identified the Wang nomogram model as the best model based on its pooled C-index of 0.70 after external validation, as well as its "good" calibration and model quality.[1,29] However, using the PROBAST tool, we rated the Wang model as high risk of bias, as the study did not report the number of missing values or how they were handled.[2,29] Additionally, it was unclear whether complete-case analysis was conducted or not, leaving it susceptible to selection bias.[29] In our assessment of the updated list of models, we found the ICC-Metroticket by Sahara et al.[7] to be a better model, with a lower risk of bias (*a priori* model design and reporting complete-case analysis), a user-friendly online calculator, and a more comprehensive list of predictors compared with the Wang nomogram.[7,29] Predictors not included in the Wang nomogram but in the ICC-Metroticket included tumor size, tumor grade, margin status, type of hepatectomy, cirrhosis, and adjuvant chemotherapy status.[7,29] However, it is worth noting that both of these prediction models rely primarily on postoperative factors, specifically information from the final surgical pathology report, which could explain their relatively higher model performances compared with those with more preoperative predictor variables. As we anticipate an important role for preoperative prediction models with the evolving treatment strategies for iCCA, we also externally validated the Preoperative Risk Score by Sasaki et al.[6,46]

To further validate our findings from the critical appraisal, we conducted an external validation on the ICC-Metroticket and confirmed its sustained performance. Despite our validation dataset having a longer follow-up (median follow-up, 36 vs. 21 months), worse OS (5-year OS, 40 vs. 43%), a higher major hepatectomy rate (87% vs. 50%), and more contemporary patient samples (2005–2017 vs. 1990–2016), the ICC-Metroticket maintained an Uno's time-dependent AUC of 0.71 (95% CI: 0.53–0.88) at 5 years and demonstrated good moderate calibration plot. Therefore, we suggest the ICC-Metroticket model to be further validated in a larger global dataset and undergo an update using a more contemporary dataset to reflect the rapidly evolving practice in the field of iCCA.[10] There are several reasons why we were unable to identify a model with better performance. First, the heterogeneous nature of iCCA poses challenges for accurate prediction. Additionally, variations in study populations and methodologies can introduce bias, overfitting, or generalizability issues. Ongoing research that focuses on novel biomarkers, genomic data, and advanced machine-learning techniques holds the potential to produce better models in the future.[47,48]

This study has several limitations. First, all studies reporting original prediction tool development used retrospective datasets that may have been affected by missing data and potential biases. Second, although we conducted a risk of bias assessment based on the PROBAST tool, there may have been some subjectivity in this evaluation. To address this, we conducted the assessment in duplicate. Third, the external validation dataset was small, and one of the predictors, CA19-9, had a significant number of missing values. To overcome this limitation, we compared several statistical approaches for replacing missing values and measured how they correlated with the model's performances using the external dataset. Finally, we were unable to perform a meta-analysis as pooling the C-statistics violates the Rubin's rule.

CONCLUSIONS

In this systematic review and external validation study, we summarized the evidence and critically appraised the survival and recurrence prediction tools for patients who underwent curative-intent iCCA resection. We externally validated the ICC-Metroticket model using our local institution data and showed that it has good overall performance. Future initiatives should focus not only on educating those interested in developing prediction models but also on evaluating the potential

clinical utility of existing models, some of which may benefit from updating and further validation.

Acknoeledgments

W.J.C.: Conception, literature search, data collection, statistical analysis, article write-up; R.W.: Data collection, statistical analysis, article write-up; L.R.: Data collection and article write-up; O.J.: Data collection and article write-up; A.G.: Data collection and article write-up; M.E.: systematic literature search and article write-up; S.G.: Conception, article write-up; G.H.: Conception, article write-up; BH: Conception, statistical analysis, manuscript write-up; GS: Conception, statistical analysis, manuscript write-up.

REFERENCES

1. Büttner S, Galjart B, Beumer BR, et al. Quality and performance of validated prognostic models for survival after resection of intrahepatic cholangiocarcinoma: a systematic review and meta-analysis. *HPB (Oxford)*. 2021;23:25–36.
2. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51–58.
3. Wang C, Pang S, Si-Ma H, et al. Specific risk factors contributing to early and late recurrences of intrahepatic cholangiocarcinoma after curative resection. *World J Surg Oncol*. 2019;17:2.
4. Zhang XF, Beal EW, Bagante F, et al. Early versus late recurrence of intrahepatic cholangiocarcinoma after resection with curative intent. *Br J Surg*. 2018;105:848–856.
5. Tsilimigras DI, Sahara K, Wu L, et al. Very early recurrence after liver resection for intrahepatic cholangiocarcinoma: considering alternative treatment approaches. *JAMA Surg*. 2020;155:823–831.
6. Sasaki K, Margonis GA, Andreatos N, et al. Preoperative risk score and prediction of long-term outcomes after hepatectomy for intrahepatic cholangiocarcinoma. *J Am Coll Surg*. 2018;226:393–403.
7. Sahara K, Tsilimigras DI, Mehta R, et al. A novel online prognostic tool to predict long-term survival after liver resection for intrahepatic cholangiocarcinoma: the "metro-ticket" paradigm. *J Surg Oncol*. 2019;120:223–230.
8. Sui K, Okabayashi T, Umeda Y, et al. Prognostic utility of the glasgow prognostic score for the long-term outcomes after liver resection for intrahepatic cholangiocarcinoma: a multi-institutional study. *World J Surg*. 2021;45:279–290.
9. Tsilimigras DI, Mehta R, Aldrighetti L, et al; International Intrahepatic Cholangiocarcinoma Study Group. Development and validation of a laboratory risk score (LabScore) to predict outcomes after resection for intrahepatic cholangiocarcinoma. *J Am Coll Surg*. 2020;230:381–391.e2.
10. Tan YL, Saffari SE, Tan NCK. A framework for evaluating predictive models. *J Clin Epidemiol*. 2022;150:188–190.
11. Moher D, Liberati A, Tetzlaff J, et al; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6:e1000097.
12. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11:e1001744.
13. 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*. 2015;350:g7594–g7594.
14. Grossetta Nardini HK, Wang L. *The Yale MeSH Analyzer [Internet]*. Cushing/Whitney Medical Library.
15. Valle J, Wasan H, Palmer DH, et al; ABC-02 Trial Investigators. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *N Engl J Med*. 2010;362:1273–1281.
16. McLernon DJ, Giardiello D, Van Calster B, et al; topic groups 6 and 8 of the STRATOS Initiative. Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for Cox proportional hazards models. *Ann Intern Med*. 2023;176:105–114.
17. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–176.
18. Wu Y, Ren F, Chai Y, et al. Prognostic value of inflammation-based indexes for intrahepatic cholangiocarcinoma following curative resection. *Oncol Lett*. 2019;17:165–174.
19. Ma K, Dong B, Wang L, et al. Nomograms for predicting overall survival and cancer-specific survival in patients with surgically resected intrahepatic cholangiocarcinoma. *Cancer Manag Res*. 2019;11:6907–6929.
20. Li Z, Yuan L, Zhang C, et al. A novel prognostic scoring system of intrahepatic cholangiocarcinoma with machine learning basing on real-world data. *Front Oncol*. 2020;10:576901.
21. Li Q, Chen C, Zhang J, et al. Prediction efficacy of prognostic nutritional index and albumin-bilirubin grade in patients with intrahepatic cholangiocarcinoma after radical resection: a multi-institutional analysis of 535 patients. *Front Oncol*. 2021;11:769696.
22. Jeong S, Luo G, Gao Q, et al. A combined Cox and logistic model provides accurate predictive performance in estimation of time-dependent probabilities for recurrence of intrahepatic cholangiocarcinoma after resection. *Hepatobiliary Surg Nutr*. 2021;10:464–475.
23. He C, Zhao C, Zhang Y, et al. An inflammation-index signature predicts prognosis of patients with intrahepatic cholangiocarcinoma after curative resection. *J Inflamm Res*. 2021;14:1859–1872.
24. Deng LM, Wang Y, Yang JH, et al. Diffuse reduction of spleen density is a novel prognostic marker for intrahepatic cholangiocarcinoma after curative resection. *World J Gastrointest Oncol*. 2021;13:929–942.
25. Cai Y, Zhang B, Li J, et al. A novel nomogram based on hepatic and coagulation function for evaluating outcomes of intrahepatic cholangiocarcinoma after curative hepatectomy: a multi-center study of 653 patients. *Front Oncol*. 2021;11:711061.
26. Zhao J, Chen Y, Wang J, et al. Preoperative risk grade predicts the long-term prognosis of intrahepatic cholangiocarcinoma: a retrospective cohort analysis. *BMC Surg*. 2021;21:113.
27. Yu H, Wang M, Wang Y, et al. The prognostic value of sarcopenia combined with preoperative fibrinogen-albumin ratio in patients with intrahepatic cholangiocarcinoma after surgery: A multicenter, prospective study. *Cancer Med*. 2021;10:4768–4780.
28. Yeh CN, Wang SY, Chen YY, et al. A prognostic nomogram for overall survival of patients after hepatectomy for intrahepatic cholangiocarcinoma. *Anticancer Res*. 2016;36:4249–4258.
29. Wang Y, Li J, Xia Y, et al. Prognostic nomogram for intrahepatic cholangiocarcinoma after partial hepatectomy. *J Clin Oncol*. 2013;31:1188–1195.
30. Tang Y, Zhang T, Zhou X, et al. The preoperative prognostic value of the radiomics nomogram based on CT combined with machine learning in patients with intrahepatic cholangiocarcinoma. *World J Surg Oncol*. 2021;19:45.
31. Sotiropoulos GC, Miyazaki M, Konstadoulakis MM, et al. Multicentric evaluation of a clinical and prognostic scoring system predictive of survival after resection of intrahepatic cholangiocarcinomas. *Liver Int*. 2010;30:996–1002.
32. Schnitzbauer AA, Eberhard J, Bartsch F, et al. The MEGNA score and preoperative anemia are major prognostic factors after resection in the German intrahepatic cholangiocarcinoma cohort. *Ann Surg Oncol*. 2020;27:1147–1155.
33. Li MD, Lu XZ, Liu JF, et al. Preoperative survival prediction in intrahepatic cholangiocarcinoma using an ultrasound-based radiographic-radiomics signature. *J Ultrasound Med*. 2022;41:1483–1495.
34. Raoof M, Dumitra S, Ituarte PHG, et al. Development and validation of a prognostic score for intrahepatic cholangiocarcinoma. *JAMA Surg*. 2017;152:e170117.
35. Jiang W, Zeng ZC, Tang ZY, et al. A prognostic scoring system based on clinical features of intrahepatic cholangiocarcinoma: the Fudan score. *Ann Oncol*. 2011;22:1644–1652.
36. Jeong S, Cheng Q, Huang L, et al. Risk stratification system to predict recurrence of intrahepatic cholangiocarcinoma after hepatic resection. *BMC Cancer*. 2017;17:464.
37. Hyder O, Marques H, Pulitano C, et al. A nomogram to predict long-term survival after resection for intrahepatic cholangiocarcinoma: an Eastern and Western experience. *JAMA Surg*. 2014;149:432–438.
38. Hyder O, Hatzaras I, Sotiropoulos GC, et al. Recurrence after operative management of intrahepatic cholangiocarcinoma. *Surgery*. 2013;153:811–818.
39. Hahn F, Muller L, Mahringer-Kunz A, et al. Risk prediction in intrahepatic cholangiocarcinoma: direct comparison of the MEGNA score and the 8th edition of the UICC/AJCC Cancer staging system. *PLoS One*. 2020;15:e0228501.
40. Doussot A, Groot-Koerkamp B, Wiggers JK, et al. Outcomes after resection of intrahepatic cholangiocarcinoma: external validation and comparison of prognostic models. *J Am Coll Surg*. 2015;221:452–461.
41. Buettner S, Galjart B, van Vugt JLA, et al. Performance of prognostic scores and staging systems in predicting long-term survival outcomes after surgery for intrahepatic cholangiocarcinoma. *J Surg Oncol*. 2017;116:1085–1095.

42. Deng L, Wang Y, Zhao J, et al. The prognostic value of sarcopenia combined with hepatolithiasis in intrahepatic cholangiocarcinoma patients after surgery: a prospective cohort study. *Eur J Surg Oncol*. 2021;47:603–612.
43. Brustia R, Langella S, Kawai T, et al; Study GROUP. Preoperative risk score for prediction of long-term outcomes after hepatectomy for intrahepatic cholangiocarcinoma: report of a collaborative, international-based, external validation study. *Eur J Surg Oncol*. 2020;46:560–571.
44. Bagante F, Merath K, Squires MH, et al. The limitations of standard clinicopathologic features to accurately risk-stratify prognosis after resection of intrahepatic cholangiocarcinoma. *J Gastrointest Surg*. 2018;22:477–485.
45. Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:1–5.
46. Akateh C, Ejaz AM, Pawlik TM, et al. Neoadjuvant treatment strategies for intrahepatic cholangiocarcinoma. *World J Hepatol*. 2020;12:693–708.
47. Banales JM, Marin JJG, Lamarca A, et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. *Nat Rev Gastroenterol Hepatol*. 2020;17:557–588.
48. Lamarca A, Barriuso J, McNamara MG, et al. Molecular targeted therapies: ready for "prime time" in biliary tract cancer. *J Hepatol*. 2020;73:170–185.