

Review

Diagnostic Performance of Computed Tomography–Based Artificial Intelligence for Early Recurrence of Cholangiocarcinoma: Systematic Review and Meta-Analysis

Jie Chen¹, MM; Jianxin Xi², MM; Tianyu Chen³, MM; Lulu Yang¹, MM; Kaijia Liu¹, MM; Xiaobo Ding¹, MD

¹Department of Radiology, The First Hospital of Jilin University, Jilin, China

²General Surgery Center, Department of Hepatobiliary and Pancreatic Surgery, The First Hospital of Jilin University, Changchun, China

³Department of Urology, The First Hospital of Jilin University, Changchun, China

Corresponding Author:

Xiaobo Ding, MD

Department of Radiology

The First Hospital of Jilin University

71 Xinmin Street, Chaoyang District, Changchun City

Jilin, 130000

China

Phone: 86 15804300156

Email: Dingxiaobo_2008@126.com

Abstract

Background: Despite artificial intelligence (AI) models demonstrating high predictive accuracy for early cholangiocarcinoma recurrence, their clinical application faces challenges, such as reproducibility, generalizability, hidden biases, and uncertain performance across diverse datasets and populations, raising concerns about their practical applicability.

Objective: This meta-analysis aims to systematically assess the diagnostic performance of AI models using computed tomography (CT) imaging to predict early recurrence of cholangiocarcinoma.

Methods: A systematic search was conducted in PubMed, Embase, and Web of Science for studies published up to May 2025. Studies were selected based on the Participants, Index test, Target condition, Reference standard, Outcomes, and Setting (PITROS) framework. Participants included patients diagnosed with cholangiocarcinoma (including intrahepatic and extrahepatic locations). The index test was AI techniques applied to CT imaging for early recurrence prediction (defined as within 1 year), while the target condition was early recurrence of cholangiocarcinoma (positive group: recurrence; negative group: no recurrence). The reference standard was pathological diagnosis or imaging follow-up confirming recurrence. Outcomes included sensitivity, specificity, diagnostic odds ratio (DOR), and area under the receiver operating characteristic curve (AUC), assessed in both internal and external validation cohorts. The setting comprised retrospective or prospective studies using hospital datasets. Methodological quality was assessed using an optimized version of the revised Quality Assessment of Diagnostic Accuracy Studies-2 tool. Heterogeneity was assessed using the I^2 statistic. Pooled sensitivity, specificity, DOR, and AUC were calculated using a bivariate random-effects model.

Results: A total of 9 studies with 30 datasets involving 1537 patients were included. In internal validation cohorts, CT-based AI models showed a pooled sensitivity of 0.87 (95% CI 0.81-0.92), specificity of 0.85 (95% CI 0.79-0.89), DOR of 37.71 (95% CI 18.35-77.51), and AUC of 0.93 (95% CI 0.90-0.94). In external validation cohorts, pooled sensitivity was 0.87 (95% CI 0.81-0.91), specificity was 0.82 (95% CI 0.77-0.86), DOR was 30.81 (95% CI 18.79-50.52), and AUC was 0.85 (95% CI 0.82-0.88). The AUC was significantly lower in external validation cohorts compared to internal validation cohorts ($P<.001$).

Conclusions: Our results show that CT-based AI models predict early cholangiocarcinoma recurrence with high performance in internal validation sets and moderate performance in external validation sets. However, the high heterogeneity observed may impact the robustness of these results. Future research should focus on prospective studies and establishing standardized gold standards to further validate the clinical applicability and generalizability of AI models.

(*J Med Internet Res* 2025;27:e78306) doi: [10.2196/78306](https://doi.org/10.2196/78306)

KEYWORDS

artificial intelligence; computed tomography; cholangiocarcinoma; early recurrence; meta-analysis

Introduction

Among primary hepatic malignancies, cholangiocarcinoma is the second most lethal tumor, following hepatocellular carcinoma in biological aggressiveness [1]. Cholangiocarcinoma can be categorized into 2 primary subtypes—intrahepatic cholangiocarcinoma (IHC) and extrahepatic cholangiocarcinoma (EHC)—based on the specific anatomical location of biliary tract involvement [2]. Globally, the incidence of cholangiocarcinoma has been steadily increasing. Radical surgical resection remains the only recognized curative treatment; however, the long-term prognosis remains poor, with 5-year survival rates limited to 20%-35% [3]. More distressingly, the early recurrence rate (defined as recurrence within 1 year) following radical resection reaches as high as 40%-65% [4]. Li et al [5] identified early recurrence as an adverse prognostic factor after curative resection of cholangiocarcinoma. Accordingly, accurately predicting early recurrence in patients with cholangiocarcinoma after surgery has become an important clinical focus. Precisely identifying high-risk populations for early recurrence is of critical importance for guiding treatment decisions, developing individualized therapeutic strategies, and ultimately improving the prognosis of patients with cholangiocarcinoma [6,7].

The conventional diagnostic approaches for detecting cholangiocarcinoma recurrence include serum tumor biomarkers (notably CA19-9), imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasonography, and pathological biopsy [8,9]. The diagnostic utility of serum biomarkers is compromised by their limited sensitivity and specificity, with performance substantially impacted by variables, including bilirubin concentrations and inflammatory states [10,11]. While traditional imaging modalities offer noninvasive insights into the tumor microenvironment, they are fundamentally constrained by inherent methodological limitations [12]. These approaches mainly rely on morphological features and basic quantitative measurements, which are often subjective and susceptible to interobserver variability [13]. As a result, they may not fully reflect the complex biological features of the tumor beyond what is visually apparent, reducing diagnostic accuracy. Pathological biopsy, on the other hand, has several challenges. It is invasive and prone to sampling bias, and it does not fully reflect the complex heterogeneity of the tumor. Moreover, its main use is for postsurgical assessment, which greatly limits its value for preoperative risk evaluation and stratification [14,15].

Recent studies have demonstrated significant predictive performance of artificial intelligence (AI) models, with some

reports showing area under the receiver operating characteristic curve (AUC) values approaching 0.99 for predicting early recurrence in cholangiocarcinoma [16,17]. However, the clinical application of these technologies remains controversial. The field faces critical challenges, including model reproducibility, generalizability, and maintaining consistent performance across different patient populations and imaging protocols. There is also uncertainty about model performance on independent datasets, largely due to hidden biases and the lack of extensive external validation [14,18]. These limitations restrict the clinical translation of AI-based predictive tools and hinder their adoption across diverse medical settings [19].

Therefore, this systematic review and meta-analysis aims to provide a comprehensive, critical evaluation of the diagnostic performance of CT-based AI technologies in predicting early recurrence of cholangiocarcinoma.

Methods

We rigorously implemented the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies (PRISMA-DTA) protocol to ensure comprehensive methodology in our systematic review and meta-analysis [20]. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist is provided in [Multimedia Appendix 1](#).

Search Strategy

We comprehensively searched multiple databases (PubMed, Embase, and Web of Science) with an initial cutoff on March 25, 2025, and a supplementary search in May 2025 to ensure inclusion of the latest research. The search strategy incorporated 3 keyword groups: the first group comprised AI-related terms (such as “artificial intelligence,” “radiomic,” and “deep learning”), the second group focused on target-related terms (including “early recurrence” and “tumor recurrence”), and the third group included disease-related terms (such as “cholangiocarcinoma”). We used a combined approach of free-text terms and Medical Subject Headings, with no initial restrictions on language or publication year. Detailed search strategy information is provided in [Multimedia Appendix 2](#). Additionally, we manually reviewed the reference lists of included studies to identify further relevant literature.

Inclusion and Exclusion Criteria

Studies were included based on the Participants, Index test, Target condition, Reference standard, Outcomes, and Setting (PITROS) framework. For comprehensive details, please consult [Textbox 1](#).

Textbox 1. Summary of inclusion criteria using the Participants, Index test, Target condition, Reference standard, Outcomes, and Setting framework.

Participants
<ul style="list-style-type: none">Patients diagnosed with cholangiocarcinoma, covering both intrahepatic and extrahepatic anatomical locations, were included in this study.
Index test
<ul style="list-style-type: none">Artificial intelligence techniques were applied to analyze computed tomography imaging to predict early recurrence of cholangiocarcinoma (defined as within 1 year).
Target condition
<ul style="list-style-type: none">The target condition was early recurrence of cholangiocarcinoma, with the positive group defined as patients who developed early recurrence and the negative group as those without early recurrence.
Reference standard
<ul style="list-style-type: none">Pathological biopsy or clinical imaging follow-up was used as the reference standard to confirm recurrence status.
Outcomes
<ul style="list-style-type: none">The primary outcomes were sensitivity, specificity, diagnostic odds ratio, and area under the receiver operating characteristic curve, evaluated in both internal and external validation cohorts.
Setting
<ul style="list-style-type: none">Studies with retrospective or prospective designs were considered, utilizing data from local hospitals.

Additionally, we systematically excluded studies with obviously irrelevant titles and abstracts, as well as inappropriate literature types, including reviews, conference abstracts, case reports, meta-analyses, and letters to editors. Furthermore, non-English studies were excluded due to inaccessibility. Moreover, articles that did not explicitly specify CT-based methodologies and those from which true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values could not be obtained were also excluded.

Quality Assessment

We used an improved version of the revised Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) assessment protocol to systematically evaluate the methodological characteristics of the studies [21]. QUADAS-2 is specifically designed to assess bias risk and applicability in diagnostic studies, with its earlier versions having been widely applied across various research contexts. However, we identified certain irrelevant criteria that were not applicable in our context and incorporated the Prediction Model Risk of Bias Assessment Tool (PROBAST) to improve and adapt the risk of bias assessment standards for predictive models [22]. The modified QUADAS-2 tool encompassed 4 primary domains: participants, index test (ie, the applied AI algorithm), reference standard, and analysis. Detailed scoring criteria are presented in [Multimedia Appendix 3](#). Moreover, for certainty rating, the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) tool was used to assess the evidence level for each standard [23], with scoring details provided in [Multimedia Appendix 4](#). To ensure the objectivity and accuracy of the assessment process, 2 independent reviewers (JC and JX) comprehensively evaluated the bias risk of included studies. During the review process, any disagreements between reviewers were resolved through in-depth discussion and analysis.

Data Extraction

The 2 reviewers (JC and JX) independently conducted preliminary screenings of the remaining articles' titles and abstracts to determine their potential eligibility. In cases of disagreement, a third reviewer (KL) would serve as a supervisor to assist in reaching a consensus. The extracted data included patient and study-level information (first author's name, publication year, patient origin country, type of cholangiocarcinoma, reference standard, lesion-based or patient-based analysis, training set, and total number of patients in internal and external validation sets). Imaging technology-level information included CT type, AI method, AI algorithm, data segmentation method, and TP, TN, FP, and FN values in training, internal, and external validation sets. TP referred to the outcome where the AI model judged as positive based on CT and confirmed as early recurrence of cholangiocarcinoma through pathology or clinical imaging follow-up. TN referred to the AI model judging as negative based on CT and the reference standard confirming the absence of cholangiocarcinoma (ie, no disease present). FP indicated the AI model judged as positive based on CT but confirmed as negative by reference standard (eg, misdiagnosed as a disease but actually a benign lesion or normal variation). FN represented cases where the AI model judged as negative based on CT but confirmed as positive by reference standard (ie, missed disease cases). For studies included in the systematic review but lacking data available for meta-analysis, we sent emails to corresponding authors to obtain the necessary information.

Given the limited availability of diagnostic contingency tables in the majority of studies, we implemented 2 primary methodological approaches to construct 4-cell diagnostic matrices. First, we retroactively calculated TP, FP, TN, and FN using sensitivity, specificity, and the number of positive cases under the reference standard and total patient numbers. Second,

through receiver operating characteristic curve analysis, we used GetData Graph Digitizer software (S Fedorov) to replot points, extracting the optimal sensitivity and specificity based on the best Youden index, and then retroactively calculated TP, FP, TN, and FN in combination with the number of positive cases under the reference standard and total patient numbers. In studies providing multiple nonoverlapping patient datasets, we assumed these contingency tables to be independent and therefore extracted them in full. For studies presenting AI models with multiple algorithms, we also performed a comprehensive extraction to enhance the feasibility of comparing algorithms between different approaches.

Outcome Measures

The primary outcome measures included sensitivity, specificity, diagnostic odds ratio (DOR), and AUC from both internal and external validation sets. Sensitivity (also known as recall or TP rate) measures the ability of the AI model to accurately identify true cases, calculated by the formula $TP/(TP+FN)$. Specificity (also known as TN rate) reflects the model's capability to correctly identify healthy cases, calculated by the formula $TN/(TN+FP)$. AUC, a comprehensive indicator for evaluating the model's ability to distinguish between positive and negative cases. DOR is a comprehensive diagnostic performance metric that combines sensitivity and specificity. This metric quantifies the diagnostic test's discriminative capability by comparing the probability of a positive result in populations with disease versus without disease.

Statistical Analysis

Using a bivariate random-effects statistical framework, we comprehensively analyzed the diagnostic capabilities of AI applied to CT imaging for early cholangiocarcinoma recurrence detection. We separately summarized diagnostic results for internal and external validation sets, calculating pooled sensitivity, specificity, and DOR values. Forest plots showed combined sensitivity and specificity, and a summary receiver

operating characteristic curve illustrated the 95% confidence and prediction ranges, with pooled AUC values calculated.

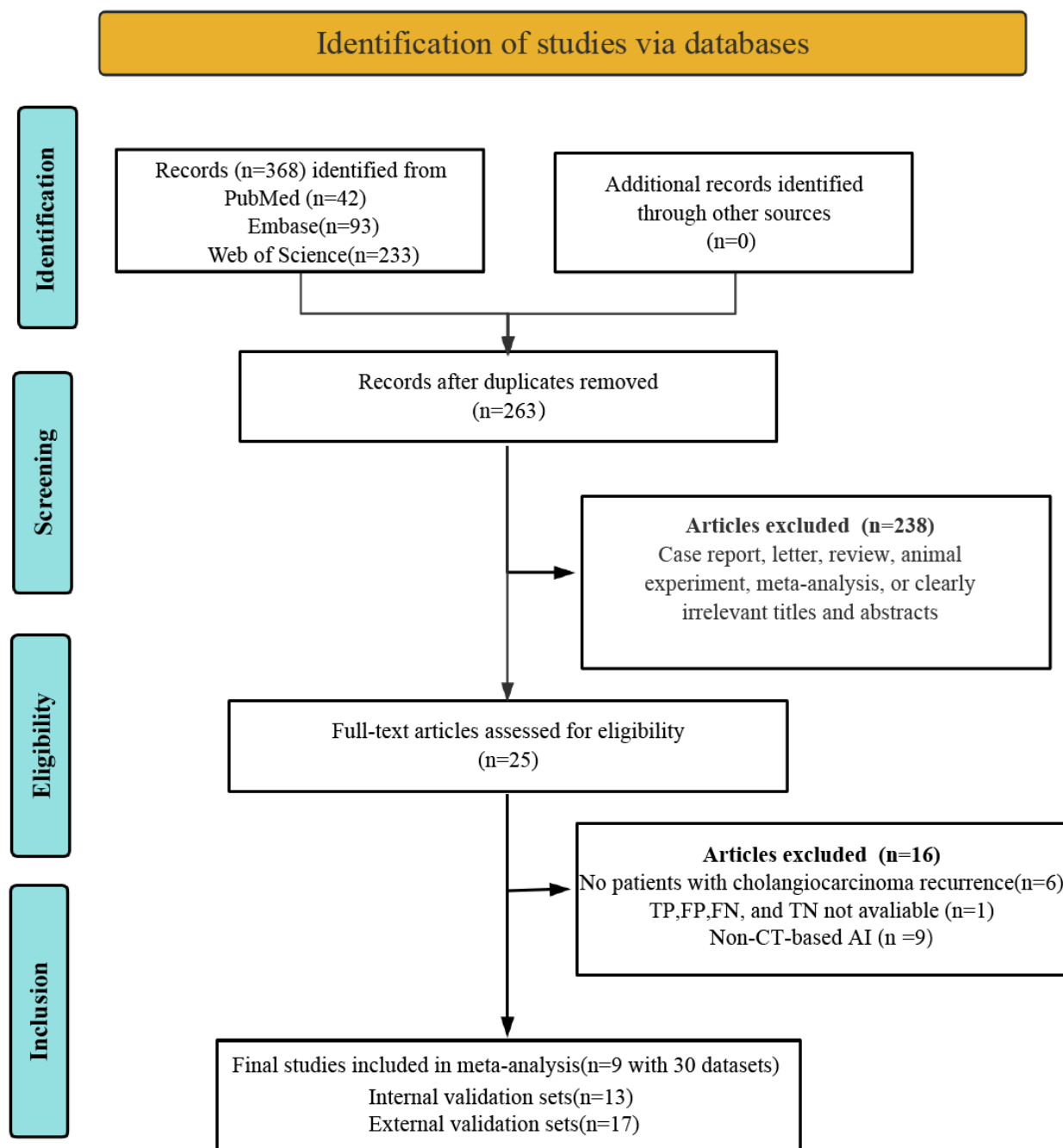
To assess interstudy heterogeneity, we used the Higgins I^2 statistic, where I^2 values of 25%, 50%, and 75% represent low, moderate, and high heterogeneity, respectively. For substantial heterogeneity ($I^2 > 50\%$), we conducted subgroup analysis and meta-regression to explore potential sources of heterogeneity. Included subgroup analysis and meta-regression variables comprised type of cholangiocarcinoma, analysis type, reference standard, type of CT, AI method, AI model, and data splitting method. Additionally, we used bubble plots to assess the variation of AI model DOR values over time. We also used Fagan plots to evaluate the clinical application impact of AI models. Furthermore, we used Deeks' funnel plot asymmetry test, assessing publication bias through effective sample size and DOR logarithmic regression. Statistical analyses were performed with Midas in Stata (version 15.1; StataCorp LLC), and risk of bias was assessed using RevMan (version 5.4; Cochrane). All tests were 2-sided, with $P < .05$ considered significant.

Results

Study Selection

A total of 3 initial database searches discovered 368 potentially relevant literatures. After removing 105 duplicate articles, 263 articles entered the preliminary screening phase. During the initial screening, 238 articles were excluded due to irrelevant titles and abstracts and inappropriate literature types, leaving 25 articles for full-text review. After detailed examination, we excluded 1 study with insufficient or incomplete diagnostic data (including TPs, FPs, FNs, and TNs), 9 studies not based on CT-based AI models, and 6 studies that failed to evaluate cholangiocarcinoma recurrence in patients. Ultimately, 9 studies met the inclusion criteria and were incorporated into the meta-analysis [17,24-31]. The literature screening process followed the standardized PRISMA protocol, which is detailed in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection for the diagnostic performance of computed tomography (CT)-based artificial intelligence (AI) in early cholangiocarcinoma recurrence. FN: false negative; FP: false positive; TN: true negative; TP: true positive.



Study Description and Quality Assessment

A total of 9 eligible studies were identified, with internal validation sets comprising 8 studies and 13 datasets [17,25-31], encompassing 1234 patients. External validation involved 5 studies with 17 datasets [24,26,27,29,30], totaling 303 patients. These included studies were published between 2021 and 2023. A total of 78% (7/9) of the studies were conducted in Asia (Japan=1 and China=6), with the remaining 22% (2/9) in North America (United States=2). All 9 studies were retrospective. Overall, 5 studies used pathology and clinical imaging follow-up as the reference standard [24,26,27,29,31], while 4 studies used

clinical imaging follow-up alone as the reference standard [17,25,28,30]. Additionally, 8 studies focused on IHC [17,24-28,30,31], while 1 study addressed EHC [29]. Furthermore, 8 studies conducted patient-based analysis [24-31], and 1 study performed image-based analysis [17]. Regarding data splitting methods, 4 studies used random splitting [28-31], 4 studies used cross-validation [17,25-27], and 1 study divided data based on 2 independent hospitals [24]. In terms of AI approaches, 8 studies used machine learning [24-31] and 1 study used deep learning [17]. For internal validation sets, the most frequently used modeling methods were random forest (2/13, 15%) and logistic regression (2/13, 15%). For external validation

sets, the most common modeling methods included light gradient boosting machine (LightGBM; 3/17, 18%), logistic regression (2/17, 12%), random forest (2/17, 12%), neural network (2/17, 12%), Bayesian classifier (2/17, 12%), support vector machine (2/17, 12%), and extreme gradient boosting (XGBoost; 2/17, 12%). Study, patient, and technical characteristics are summarized in [Table 1](#) and [Multimedia Appendix 5](#).

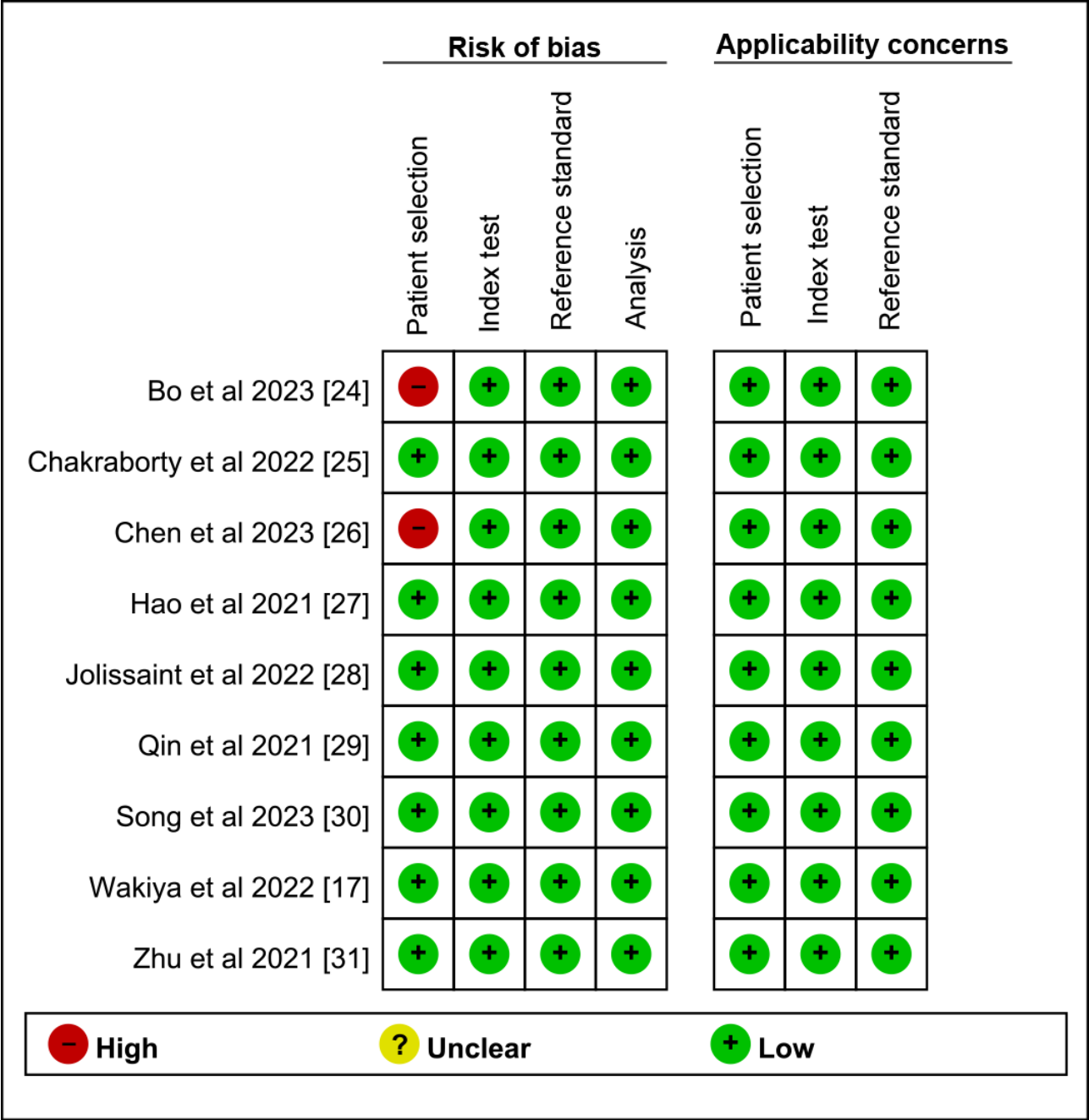
Table 1. Study and patient characteristics of the included studies.

Reference	Country	Study design	Type of CCA ^a	Reference standard	Analysis	Patients, lesions, or images per set, n			Early recurrence (patients, lesions, or images), n		
						Training	Internal validation	External validation	Training	Internal validation	External validation
Hao et al (2021) [27]	China	Retro ^b	IHC ^c	Pathology and clinical imaging follow-up	PB ^d	124	124	53	N/A ^e	66	35
Song et al (2023) [30]	China	Retro	IHC	Clinical imaging follow-up	PB	140	36	135	75	19	71
Wakiya et al (2022) [17]	Japan	Retro	IHC	Clinical imaging follow-up	IB ^f	71,081	71,081	N/A	N/A	45,316	N/A
Jolissaint et al (2022) [28]	America	Retro	IHC	Clinical imaging follow-up	PB	97	41	N/A	28	11	N/A
Bo et al (2023) [24]	China	Retro	IHC	Pathology and clinical imaging follow-up	PB	90	N/A	37	49	22	N/A
Qin et al (2021) [29]	China	Retro	EHC ^g	Pathology and clinical imaging follow-up	PB	167	70	37	84	46	30
Chen et al (2023) [26]	China	Retro	IHC	Pathology and clinical imaging	PB	95	95	41	N/A	31	13
Zhu et al (2021) [31]	China	Retro	IHC	Pathology and clinical imaging	PB	92	33	N/A	24	11	N/A
Chakraborty et al (2022) [25]	America	Retro	IHC	Clinical imaging follow-up	PB	139	139	N/A	N/A	39	N/A

^aCCA: cholangiocarcinoma.
^bRetro: retrospective.
^cIHC: intrahepatic cholangiocarcinoma.
^dPB: patient-based.
^eN/A: not available.
^fIB: image-based.
^gEHC: extrahepatic cholangiocarcinoma.

Bias risk was assessed using the revised QUADAS-2 tool, as shown in [Figure 2](#). In patient selection, 2 studies were rated as “high risk.” Bo et al [24] inappropriately excluded patients with Child-Pugh scores >7, and Chen et al [26] excluded patients with performance status scores >2 or Child-Pugh scores >7. Overall, high-risk items were few, with low-risk items predominantly represented, indicating that the included studies were acceptable in overall quality ([Multimedia Appendix 3](#)). According to GRADE standards, the evidence quality assessment for outcome indicators ranged from very low to low, suggesting weak certainty of evidence ([Multimedia Appendix 4](#)).

Figure 2. Risk of bias and applicability concerns of the included studies using the revised Quality Assessment of Diagnostic Performance Studies-2 [17,24-31].



Diagnostic Performance of Different AI Algorithms in Internal and External Validation Sets

Emerging computational intelligence technologies, particularly machine learning and deep learning, have demonstrated remarkable progress through enhanced algorithmic designs and increased data accessibility. In internal validation sets, the DOR remained relatively stable from 2021 to 2023, with some algorithms (residual network 50 [ResNet50] and XGBoost) achieving higher DOR values, while others maintained lower

levels (Figure 3A). In external validation sets, DOR values showed a slow increase during the same period, with some algorithms (random forest and LightGBM) achieving higher DOR values, suggesting potential improvements or refinements in AI algorithms (Figure 3B).

In internal validation sets, ResNet50 simultaneously achieved the highest sensitivity (0.98) and specificity (0.94). In external validation sets, the neural network demonstrated the highest sensitivity (0.94), while the support vector machine showed the highest specificity (0.88), as detailed in Table 2.

Figure 3. Bubble plot showing diagnostic odds ratios of various artificial intelligence algorithms for cholangiocarcinoma recurrence prediction across publication years. (A) Bubble plot of diagnostic odds ratios for the internal validation cohort. (B) Bubble plot of diagnostic odds ratios for the external validation cohort. AdaBoost: adaptive boosting; Bayes: Bayesian classifier; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; LR: logistic regression; MRMR-GBM: minimum redundancy maximum relevance-gradient boosting machine; NN: neural network; ResNet50: residual network 50; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

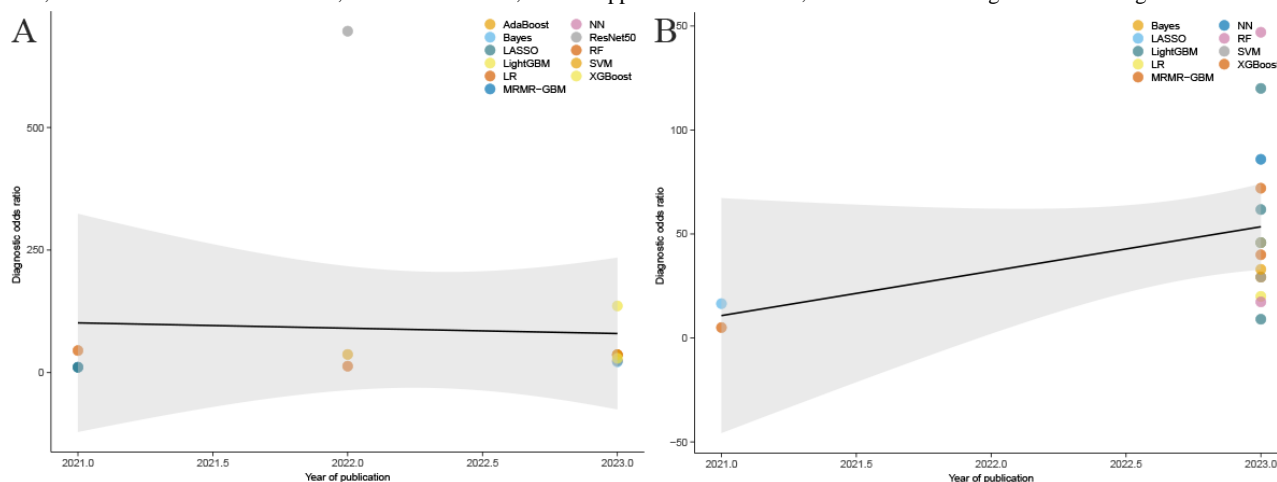


Table 2. Subgroup analysis based on different artificial intelligence (AI) algorithms.

AI algorithms	Interval validation			External validation		
	Studies, n (%)	Sensitivity (95% CI)	Specificity (95% CI)	Studies, n (%)	Sensitivity (95% CI)	Specificity (95% CI)
LR ^a	2 (15.4)	0.79 (0.64-0.88)	0.88 (0.80-0.94)	2 (11.8)	0.89 (0.73-0.96)	0.81 (0.67-0.90)
RF ^b	2 (15.4)	0.88 (0.72-0.93)	0.78 (0.68-0.85)	2 (11.8)	0.83 (0.67-0.92)	0.86 (0.72-0.94)
MRMR-GBM ^c	1 (7.7)	0.88 (0.78-0.95)	0.60 (0.47-0.73)	1 (5.9)	0.71 (0.54-0.85)	0.67 (0.41-0.87)
LightGBM ^d	1 (7.7)	0.89 (0.67-0.99)	0.94 (0.71-1.00)	3 (17.6)	0.92 (0.85-0.96)	0.76 (0.65-0.84)
ResNet50 ^e	1 (7.7)	0.98 (0.98-0.98)	0.94 (0.94-0.94)	N/A ^f	N/A	N/A
LASSO ^g	1 (7.7)	0.74 (0.59-0.86)	0.79 (0.58-0.93)	1 (5.9)	0.73 (0.54-0.88)	0.86 (0.42-1.00)
NN ^h	1 (7.7)	0.84 (0.66-0.95)	0.88 (0.77-0.94)	2 (11.8)	0.94 (0.80-0.99)	0.81 (0.67-0.90)
Bayes ⁱ	1 (7.7)	0.87 (0.70-0.96)	0.77 (0.64-0.86)	2 (11.8)	0.83 (0.67-0.92)	0.86 (0.72-0.94)
SVM ^j	1 (7.7)	0.84 (0.66-0.95)	0.88 (0.77-0.94)	2 (11.8)	0.83 (0.67-0.92)	0.88 (0.75-0.95)
XGBoost ^k	1 (7.7)	0.81 (0.63-0.93)	0.88 (0.77-0.94)	2 (11.8)	0.91 (0.77-0.97)	0.84 (0.70-0.92)
AdaBoost ^l	1 (7.7)	0.82 (0.66-0.92)	0.89 (0.81-0.94)	N/A	N/A	N/A

^aLR: logistic regression.

^bRF: random forest.

^cMRMR-GBM: minimum redundancy maximum relevance-gradient boosting machine.

^dLightGBM: light gradient boosting machine.

^eResNet50: residual network 50.

^fN/A: not available.

^gLASSO: least absolute shrinkage and selection operator.

^hNN: neural network.

ⁱBayes: Bayesian classifier.

^jSVM: support vector machine.

^kXGBoost: extreme gradient boosting.

^lAdaBoost: adaptive boosting.

Diagnostic Performance of CT-Based AI Models for Early Recurrence of Cholangiocarcinoma in Internal Validation Sets

In the internal validation sets, the CT-based AI model demonstrated a sensitivity of 0.87 (95% CI 0.81-0.92; very low certainty), specificity of 0.85 (95% CI 0.79-0.89; very low

certainty), and a DOR of 37.71 (95% CI 18.35-77.51; very low certainty), as shown in [Figures 4 and 5A](#). Additionally, the AUC of the model was 0.93 (95% CI 0.90-0.94; [Figure 6A](#)). Based on the predefined 20% test probability, the Fagan nomogram revealed a positive likelihood ratio of 59% and a negative likelihood ratio of 4% ([Multimedia Appendix 6](#)).

Figure 4. Forest plots of the diagnostic odds ratio of computed tomography–based artificial intelligence on the (A) internal validation set and (B) external validation set for diagnosing early recurrence of cholangiocarcinoma [17,24-31]. Bayes: Bayesian classifier; LR: logistic regression; NN: neural network; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

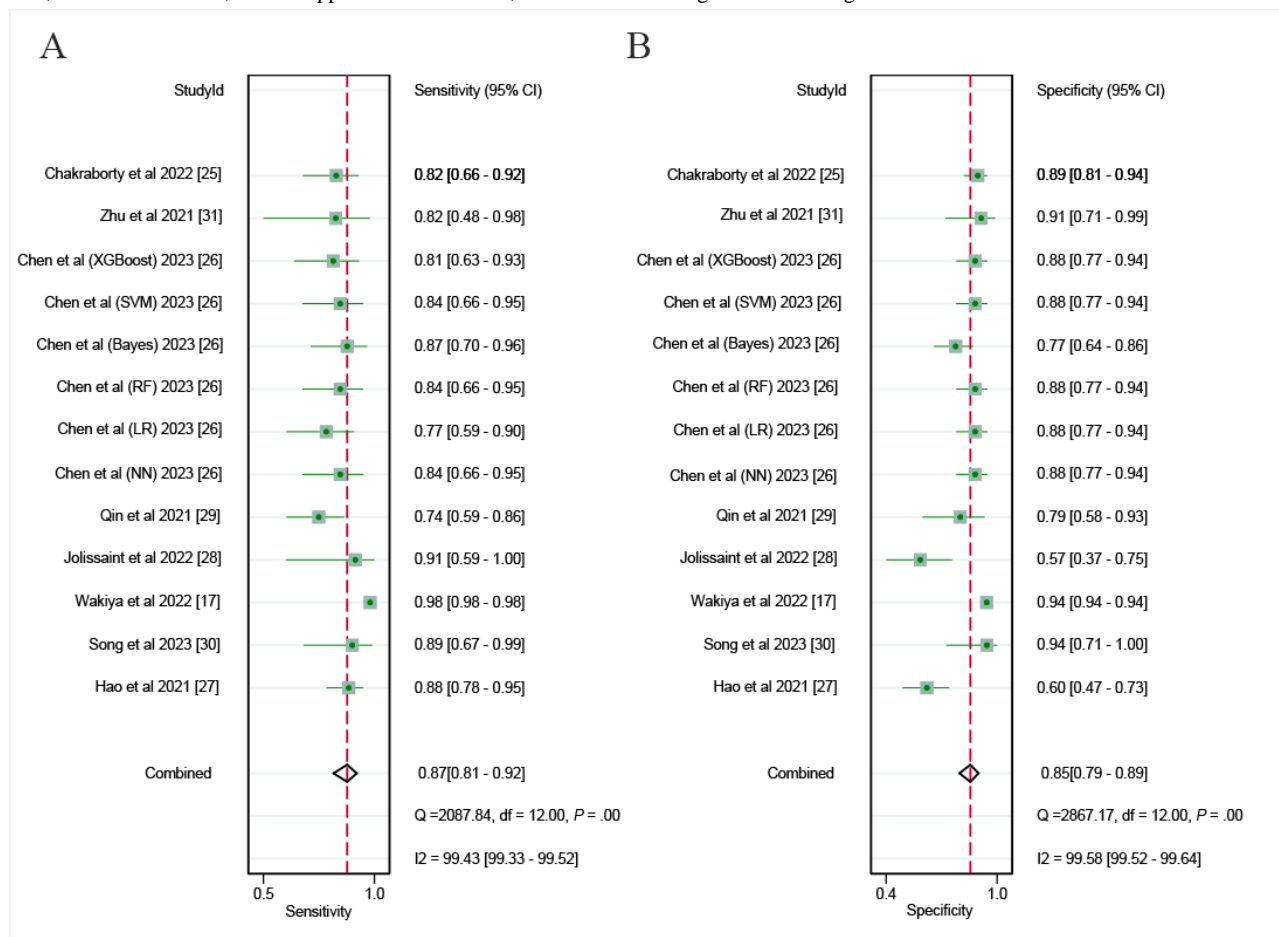


Figure 5. Forest plots of the diagnostic odds ratio of computed tomography–based artificial intelligence on the (A) internal validation set and (B) external validation set for diagnosing early recurrence of cholangiocarcinoma. CI [17,24-31]. Bayes: Bayesian classifier; LR: logistic regression; NN: neural network; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

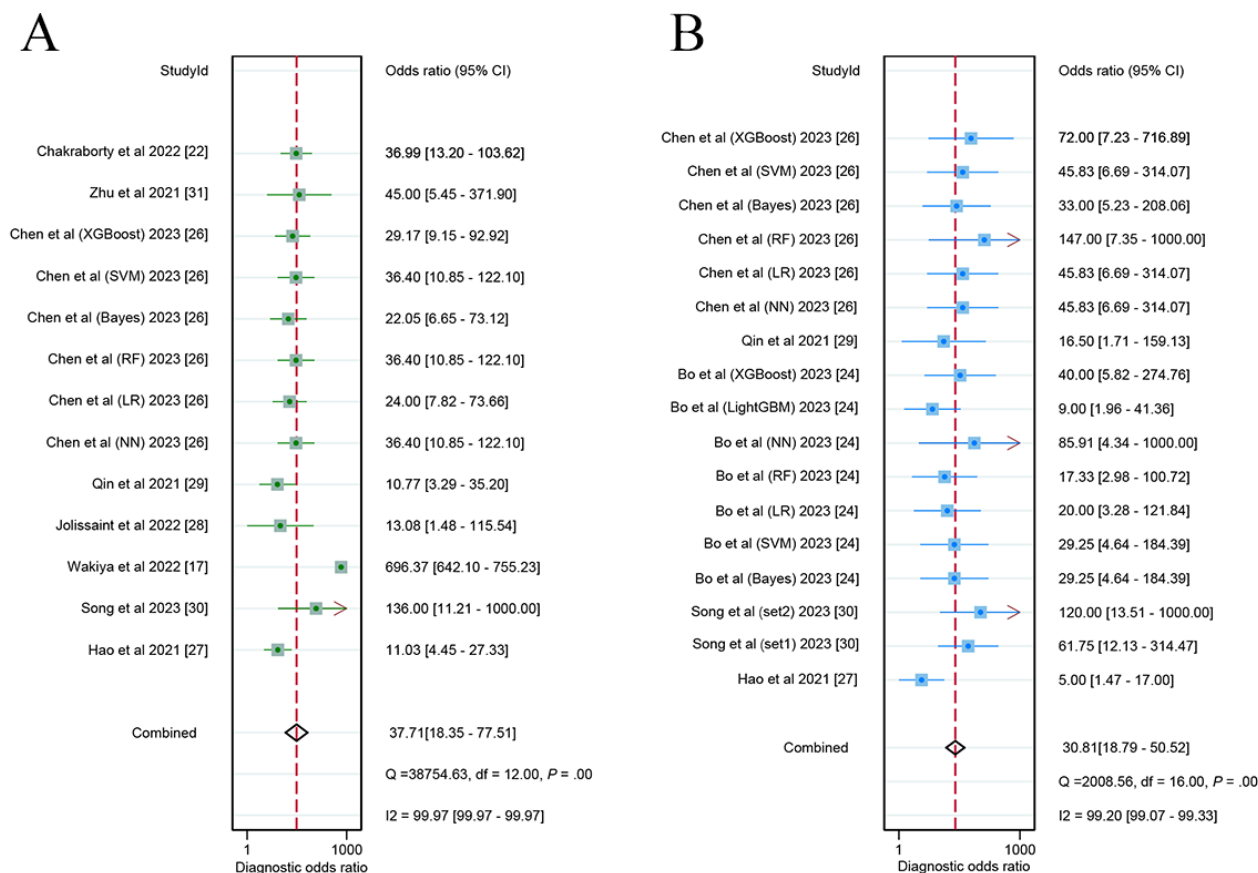
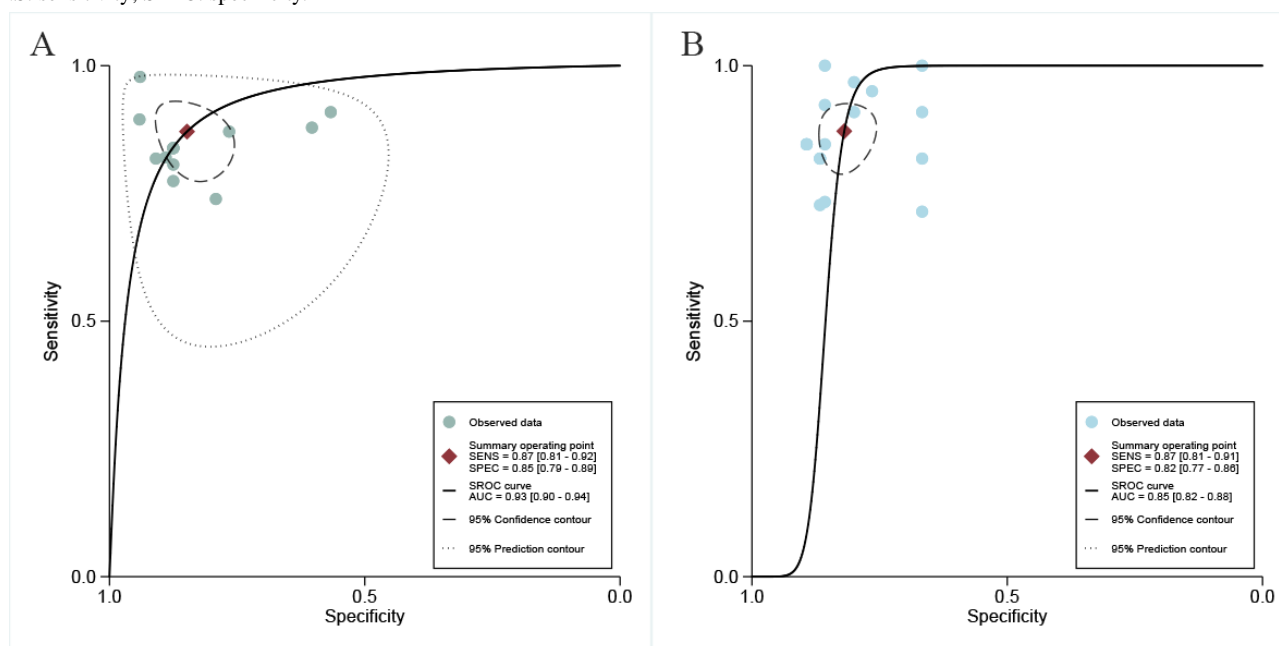


Figure 6. Summary receiver operating characteristic (SROC) curves of computed tomography–based artificial intelligence on the (A) internal validation set and (B) external validation set for predicting early recurrence of cholangiocarcinoma. AUC: area under the receiver operating characteristic curve; SENS: sensitivity; SPEC: specificity.



Meta-Regression and Subgroup Analysis in Internal Validation Sets

Meta-regression revealed high heterogeneity in sensitivity ($P=99.43\%$) and specificity ($P=99.58\%$) for internal validation sets. The meta-regression analysis results showed no identifiable

sources of potential heterogeneity (Table 2). Subgroup analysis demonstrated no statistically significant differences in sensitivity and specificity of CT-based AI models across various categories, including type of cholangiocarcinoma, analysis, reference standard, AI model, AI method, and data splitting method (all $P>.05$; Table 3).

Table 3. Subgroup analysis and meta-regression analysis of the diagnostic performance of computed tomography (CT)–based artificial intelligence (AI) for early recurrence of cholangiocarcinoma within internal validation cohorts.

Subgroup	Studies, n (%)	Sensitivity (95% CI)	Meta-regression, <i>P</i> value	Specificity (95% CI)	Meta-regression, <i>P</i> value
Type of cholangiocarcinoma			.13		.49
IHC ^a	12 (92.3)	0.88 (0.83-0.93)		0.85 (0.80-0.90)	
PHC ^b	1 (7.7)	0.74 (0.46-1.00)		0.80 (0.54-1.00)	
Analysis			.91		.22
Lesion-based	1 (7.7)	0.83 (0.60-1.00)		0.89 (0.76-1.00)	
Patient-based	12 (92.3)	0.87 (0.82-0.93)		0.84 (0.79-0.90)	
Reference standard			.57		.57
Pathology and clinical imaging follow-up	9 (69.2)	0.83 (0.77-0.90)		0.84 (0.77-0.91)	
Clinical imaging follow-up	4 (30.8)	0.94 (0.90-0.99)		0.87 (0.79-0.95)	
Type of CT			.47		.87
CECT ^c	12 (92.3)	0.87 (0.82-0.93)		0.84 (0.79-0.90)	
Plain CT	1 (7.7)	0.83 (0.60-1.00)		0.89 (0.76-1.00)	
AI model			.35		.56
Radiomic model	2 (15.4)	0.86 (0.72-1.00)		0.78 (0.62-0.94)	
Radiomic and clinical model	11 (84.6)	0.87 (0.81-0.93)		0.86 (0.81-0.91)	
AI method			.95		.16
Deep learning	1 (7.7)	0.92 (0.73-1.00)		0.57 (0.26-0.87)	
Machine learning	12 (92.3)	0.87 (0.82-0.93)		0.86 (0.82-0.90)	
Data splitting method			.32		.22
Random split	4 (30.8)	0.84 (0.71-0.96)		0.81 (0.68-0.94)	
K-fold cross-validation	9 (69.2)	0.88 (0.83-0.94)		0.86 (0.80-0.91)	

^aIHC: intrahepatic cholangiocarcinoma.
^bPHC: perihilar cholangiocarcinoma.
^cCECT: contrast-enhanced computed tomography.

Diagnostic Performance of CT-Based AI Models for Early Cholangiocarcinoma Recurrence in External Validation Sets

In the validation sets conducted externally, the AI model based on CT exhibited a sensitivity of 0.87 (95% CI 0.81-0.91; low certainty) and a specificity of 0.82 (95% CI 0.77-0.86; very low certainty; Multimedia Appendix 7). Additionally, it showed a DOR of 30.81 (95% CI 18.79-50.52; very low certainty; Figure 5B) and an AUC measuring 0.85 (95% CI 0.82-0.88; Figure 6B). When applying a 20% pretest probability, the Fagan nomogram indicated a positive likelihood ratio of 55% along with a negative likelihood ratio of 4% (Multimedia Appendix 6).

No statistically significant variances were found between the internal and external validation sets regarding sensitivity (z score=0.00; $P>.99$), specificity (z score=0.87; $P=.38$), and DOR (z score=0.00; $P=.08$). However, the AUC for the internal validation set was significantly greater than that of the external validation set (z score=4.35; $P<.001$).

Meta-Regression and Subgroup Analysis in External Validation Sets

Meta-regression revealed high heterogeneity in sensitivity ($P=42.07\%$), and although specificity ($P=0\%$) did not show high heterogeneity, we attempted to identify potential sources of heterogeneity. Meta-regression analysis results indicated that different reference standards (specificity, $P<.001$) and type of



cholangiocarcinoma (sensitivity, $P=.05$) might be potential sources of heterogeneity. In the cholangiocarcinoma-type subgroups, the sensitivity was 0.88 (95% CI 0.83-0.92) for IHC and 0.74 (95% CI 0.50-0.98) for EHC, with IHC showing significantly higher sensitivity compared to EHC ($P=.05$). In the recurrence reference standard subgroups, the specificity of pathology combined with clinical imaging follow-up was 0.83 (95% CI 0.78-0.87), which was significantly higher than that of clinical imaging follow-up alone (0.78, 95% CI 0.68-0.88; $P<.001$; [Multimedia Appendix 8](#)).

Sensitivity Analysis and Bivariate Box Plot

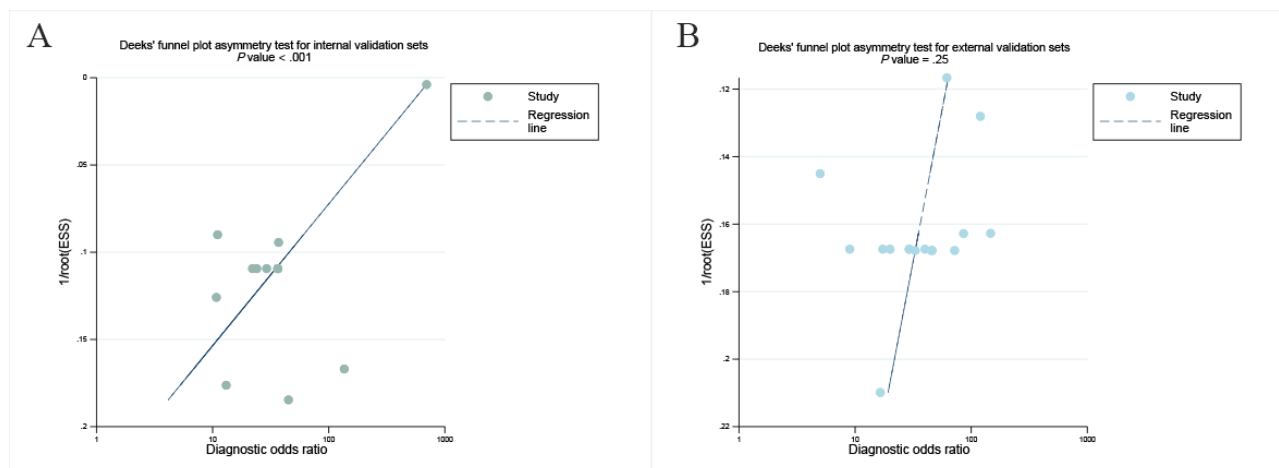
For the internal validation sets, after excluding low-quality studies, the AI model's sensitivity was 0.90 (95% CI 0.80-0.95), specificity was 0.84 (95% CI 0.71-0.91), DOR was 43.43 (95% CI 12.31-153.31), and AUC was 0.93 (95% CI 0.91-0.95). For the external validation sets, after excluding low-quality studies, the AI model's sensitivity was 0.88 (95% CI 0.70-0.96), specificity was 0.76 (95% CI 0.44-0.84), DOR was 21.79 (95% CI 5.74-82.68), and AUC was 0.80 (95% CI 0.77-0.84; [Multimedia Appendix 9](#)).

For the internal validation sets, the bivariate box plot suggested that Wakiya et al [17], Song et al [30], and Zhu et al [31] may represent potential contributors to statistical heterogeneity. After excluding the studies identified in the bivariate box plot, the AI model's sensitivity was 0.83 (95% CI 0.78-0.87), specificity was 0.82 (95% CI 0.75-0.88), DOR was 22.28 (95% CI 14.43-34.40), and AUC was 0.87 (95% CI 0.84-0.90). For the external validation sets, the bivariate box plot suggested that Chen et al [26] and Hao et al [27] might be potential sources of heterogeneity ([Multimedia Appendix 10](#)). After excluding the studies identified in the bivariate box plot, the AI model's sensitivity was 0.88 (95% CI 0.82-0.93), specificity was 0.80 (95% CI 0.74-0.85), DOR was 30.48 (95% CI 17.27-53.79), and AUC was 0.87 (95% CI 0.84-0.90).

Publication Bias

Using Deeks' funnel plot methodology, we identified substantial publication bias in the internal validation cohorts ($P<.001$), whereas for the external validation sets, there was no evidence of small-study effects ($P=.25$; [Figures 7A-7B](#)).

Figure 7. Deek's funnel plot was used to evaluate the publication bias of computed tomography-based artificial intelligence. $P<.05$ was considered significant.



Discussion

Principal Findings

Our study demonstrated the exceptional performance of CT-based AI models in diagnosing early recurrence of cholangiocarcinoma. In the internal validation sets, the model exhibited high diagnostic performance (sensitivity=0.87, specificity=0.85, DOR=37.71, and AUC=0.93). However, a moderate decline in performance was observed during external validation (sensitivity=0.87, specificity=0.82, DOR=30.81, and AUC=0.85), with a statistically significant reduction in AUC ($P<.001$). This finding suggests that despite the model's excellent performance in internal datasets, its generalizability across different patient populations and imaging conditions may be limited. The outstanding performance of the CT-based AI model primarily stems from its powerful feature extraction capabilities, enabling it to capture high-dimensional radiomics features that reflect tumor heterogeneity and microenvironmental characteristics [32]. These features are particularly crucial for highly invasive and heterogeneous cholangiocarcinoma, which

are often challenging to detect through conventional imaging techniques. By using advanced algorithms, such as machine learning and radiomics, the model can effectively identify complex patterns within imaging data, thereby enabling precise prediction of early recurrence risk in cholangiocarcinoma [29,33,34]. However, the performance decline observed in external validation highlighted several potential limitations. First, heterogeneity in CT acquisition techniques across different institutions (such as scanner models, contrast agent protocols, and reconstruction parameters) may compromise the model's stability [35]. Second, variability in patient demographics and pathological characteristics may further contribute to decreased performance in external cohorts. These findings underscore the importance of conducting large-scale, multicenter external validation studies to ensure the robustness and generalizability of AI models across different clinical environments [36].

Our subgroup analysis results based on different cholangiocarcinoma subtypes revealed that in the external validation sets, the sensitivity of IHC was significantly higher than that of EHC ($P=.05$). A similar trend was observed in the

internal validation sets, although the difference was not statistically significant. This difference may be attributed to the distinct biological characteristics and imaging features of IHC and EHC. IHC typically demonstrates pronounced tumor heterogeneity and microenvironmental characteristics, which are more readily captured by AI algorithms on CT imaging, thereby enhancing diagnostic sensitivity. In contrast, the anatomical complexity of EHC and its less distinctive imaging characteristics may limit the diagnostic performance of AI models [37,38]. It is important to note that the EHC dataset was limited (only 1 group of data), which may constrain the model's learning capacity for EHC. Future multicenter studies are needed to validate these findings.

Comparison to Previous Work

In 2023, Yang et al [39] conducted a meta-analysis evaluating the diagnostic performance of machine learning models for early recurrence of IHC. The analysis included 5 studies comprising a total of 1247 patients, using paired and network meta-analysis to compare the diagnostic accuracy of machine learning models with traditional clinical models. The results demonstrated that machine learning models exhibited a pooled diagnostic sensitivity of 0.92, a specificity of 0.79, and an AUC of 0.81, revealing superior diagnostic value compared to traditional clinical models. In contrast, our AI model demonstrated higher diagnostic performance for predicting early recurrence of cholangiocarcinoma in the internal validation set, with an AUC value of 0.93. This may be attributed to the larger sample size included in our study, as well as the use of more advanced deep learning algorithms. However, it is important to note that, in contrast to the previous meta-analysis, our study focused solely on CT-based AI models and used a more specific data source for model training.

In 2025, Xu et al [40] conducted a meta-analysis evaluating the application of machine learning based on radiomics in IHC. Their results showed that AI diagnostic models integrating radiomics and clinical features achieved a sensitivity of 0.85 and specificity of 0.77. Our study showed that the CT-based AI model achieved a sensitivity of 0.87 and a specificity of 0.85. Compared to previous research, our diagnostic performance was superior, which may be related to differences in patient populations and AI models included. Xu et al [40] focused primarily on IHC cases, whereas our study included both IHC and EHC. Additionally, our analysis exclusively evaluated CT-based AI models. Compared with previous meta-analyses, our unique strength lies in merging different specific algorithms separately and stratifying the study population into internal and external validation sets to assess the generalizability of the AI models, thus providing more rigorous and comprehensive diagnostic evidence.

Heterogeneity

The high heterogeneity of included studies might affect the overall sensitivity and specificity of AI models in internal and external validation sets. For internal validation sets, we sought to identify sources of heterogeneity through meta-regression and box plot analysis. While meta-regression did not identify any significant factors, box plots indicated that studies by Wakiya et al [17], Zhu et al [31], and Song et al [30] might be

primary sources of heterogeneity. In the external validation sets, meta-regression indicated that cholangiocarcinoma types and reference standards were potential sources of heterogeneity, while box plots also identified studies by Chen et al [26] and Hao et al [27] as significant contributors to heterogeneity. Nevertheless, high heterogeneity may still result from multiple interacting factors, including patient age and status, tumor stage, geographical region, preprocessing methods, classification algorithms and validation approaches, sample size, and AI model characteristics, such as feature selection, hyperparameter optimization, and modeling algorithms [41-44]. The complex interplay of these factors may cause differences between studies, highlighting the need for future research to standardize variables and reduce heterogeneity for more accurate and reliable results. To address the challenges posed by heterogeneity, future studies should consider implementing harmonization techniques, which standardize imaging parameters and processes used across different institutions. Robust data augmentation methods can enhance the diversity of training datasets, enabling models to generalize effectively across diverse populations. Additionally, domain adaptation strategies can facilitate the transfer of learned features from source datasets to new, unseen populations, thereby enhancing model robustness and generalizability [45].

Future Directions

Our results demonstrate that CT-based AI models achieve high diagnostic performance in internal validation sets and moderate performance in external validation sets. AI can perform preliminary reading, enabling clinicians to process cases more quickly, improve turnaround time, expand the accessibility of specialized reports, and ultimately alleviate pressure on the health care system. Implementing CT-based AI models in primary health care systems, such as general practice, may lead to early disease detection and timely intervention. However, it is worth emphasizing that these models should not be viewed as independent standards or decision-making tools, but rather as useful resources in emergency situations (when expert consultation is unavailable) or for residents and clinicians lacking expertise in detecting this disease. In addition, none of the studies included in our research reported the comparative performance of AI-assisted human interpretation against that of AI alone, making it challenging to assess the specific added value of AI tools for human readers. Additionally, the studies did not detail how AI models perform relative to radiologists, which may limit the practical interpretability of the results. Future research could benefit from exploring comparative analyses that examine the interplay between AI and radiologists. Moreover, while our study primarily focused on AI models based on CT imaging, it is important to acknowledge the inherent limitations of relying solely on this modality. CT imaging does not account for the comprehensive insights that can be gained from MRI and ultrasound imaging techniques. Each of these imaging modalities has unique strengths; for example, MRI excels in soft tissue contrast and provides functional information, while ultrasound offers real-time imaging capabilities and is more accessible in certain clinical settings. Future AI model evaluations should consider integrating cross-modal imaging data, associating observations

with patients' clinical backgrounds, and effectively communicating synthesized insights through reports [46,47].

Additionally, our study primarily focused on traditional machine learning algorithms, with 8 of the 9 included studies using methods like logistic regression, random forests, and support vector machines. Only 1 study used a deep learning approach (ResNet50). Deep learning models typically use image data as their primary input, allowing them to automatically learn hierarchical features from raw data. In contrast, traditional machine learning models rely on parametric values and features that are often manually crafted from the data [48]. Although deep learning excels at handling complex, high-dimensional data, its "black box" nature may result in a lack of interpretability and transparency in clinical settings, which could affect its practical acceptability and the reproducibility of model decisions [49]. Therefore, future research should incorporate a greater number of deep learning models to comprehensively assess their effectiveness and applicability relative to traditional algorithms. At the same time, the importance of explainability in the clinical deployment of AI models cannot be overstated. Future studies should incorporate interpretable frameworks, such as Gradient-Weighted Class Activation Mapping and Shapley Additive Explanations, and attention maps, to ensure that AI-generated predictions can be understood by clinicians [50]. This not only fosters trust in AI systems among health care providers but also facilitates their broader adoption in real-world clinical environments [51].

Acknowledgments

This study was funded by the Jilin Province Science and Technology Development Plan Project (grant 20250203042SF). During the preparation of this work, the authors used Sider (Sider Inc) in order to improve readability and language quality. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

The original findings of this study are encompassed within this paper. For additional inquiries, please contact the corresponding authors.

Authors' Contributions

JC and JX conceived and designed the study. JC, JX, TC, LY, KL, and XD extracted and analyzed the data, while JC wrote the first version of the manuscript. All authors contributed to the manuscript and approved the final version for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA checklist 2020.

[\[DOCX File , 22 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Search strategy in PubMed, Embase, and Web of Science.

[\[DOCX File , 14 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Revised Quality Assessment of Diagnostic Accuracy Studies-2 tool for the included studies.

[\[DOCX File , 18 KB-Multimedia Appendix 3\]](#)

Limitations

Our study has several limitations. First, all included studies used retrospective designs, which may introduce potential bias. Well-designed prospective studies are needed to further validate our meta-analysis findings. Second, not all studies defined the gold standard as histopathological examination, which could influence diagnostic performance. However, our subgroup analysis showed no statistically significant differences in sensitivity or specificity among different gold standards, suggesting that the choice of gold standard may have a limited impact. Third, most of the included studies (7/9, 78%) were conducted in Asian regions, which may limit the generalizability of our conclusions.

Conclusions

Our meta-analysis systematically evaluated the diagnostic performance of CT-based AI models for predicting early recurrence of cholangiocarcinoma. The results indicate that while these models demonstrate high diagnostic accuracy in internal validation cohorts, their performance is only moderate in external validation cohorts. Significant heterogeneity between studies may affect the robustness of these findings. Future research should prioritize prospective study designs and the adoption of standardized reference standards to validate these findings and improve the clinical utility and generalizability of AI models.

Multimedia Appendix 4

Grading of Recommendations, Assessment, Development, and Evaluations scoring assessments in all of the pooled outcomes.

[\[DOCX File , 15 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Technical aspects of included studies.

[\[DOCX File , 23 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Fagan's nomogram for the computed tomography-based artificial intelligence model in predicting early recurrence of cholangiocarcinoma. (A) Internal validation sets. (B) External validation sets.

[\[PNG File , 72 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Forest plots of the combined sensitivity and specificity of computed tomography-based artificial intelligence in patients with early cholangiocarcinoma recurrence. Squares denote the sensitivity and specificity for each study, while horizontal bars indicate the 95% CI.

[\[PNG File , 158 KB-Multimedia Appendix 7\]](#)

Multimedia Appendix 8

Subgroup analysis and meta-regression analysis of the diagnostic performance of computed tomography-based artificial intelligence for early recurrence of cholangiocarcinoma within external validation sets.

[\[DOCX File , 15 KB-Multimedia Appendix 8\]](#)

Multimedia Appendix 9

Sensitivity analysis.

[\[DOCX File , 18 KB-Multimedia Appendix 9\]](#)

Multimedia Appendix 10

Bivariate boxplot for the computed tomography-based artificial intelligence model in predicting early recurrence of cholangiocarcinoma. (A) Internal validation sets. (B) External validation sets.

[\[PNG File , 53 KB-Multimedia Appendix 10\]](#)

References

1. Ilyas SI, Gores GJ. Pathogenesis, diagnosis, and management of cholangiocarcinoma. *Gastroenterology*. 2013;145(6):1215-1229. [\[FREE Full text\]](#) [doi: [10.1053/j.gastro.2013.10.013](https://doi.org/10.1053/j.gastro.2013.10.013)] [Medline: [24140396](#)]
2. Al Mahjoub A, Bouvier V, Menahem B, Bazille C, Fohlen A, Alves A, et al. Epidemiology of intrahepatic, perihilar, and distal cholangiocarcinoma in the French population. *Eur J Gastroenterol Hepatol*. 2019;31(6):678-684. [doi: [10.1097/MEG.0000000000001337](https://doi.org/10.1097/MEG.0000000000001337)] [Medline: [30633038](#)]
3. Hyder O, Hatzaras I, Sotiropoulos GC, Paul A, Alexandrescu S, Marques H, et al. Recurrence after operative management of intrahepatic cholangiocarcinoma. *Surgery*. 2013;153(6):811-818. [\[FREE Full text\]](#) [doi: [10.1016/j.surg.2012.12.005](https://doi.org/10.1016/j.surg.2012.12.005)] [Medline: [23499016](#)]
4. Zhang X, Beal EW, Bagante F, Chakedis J, Weiss M, Popescu I, et al. Early versus late recurrence of intrahepatic cholangiocarcinoma after resection with curative intent. *Br J Surg*. 2018;105(7):848-856. [doi: [10.1002/bjs.10676](https://doi.org/10.1002/bjs.10676)] [Medline: [29193010](#)]
5. Li Q, Zhang J, Chen C, Song T, Qiu Y, Mao X, et al. A nomogram model to predict early recurrence of patients with intrahepatic cholangiocarcinoma for adjuvant chemotherapy guidance: a multi-institutional analysis. *Front Oncol*. 2022;12:896764. [\[FREE Full text\]](#) [doi: [10.3389/fonc.2022.896764](https://doi.org/10.3389/fonc.2022.896764)] [Medline: [35814440](#)]
6. Wang X, Liu L, Liu Z, Wang J, Dai H, Ou X, et al. Machine learning model to predict early recurrence in patients with perihilar cholangiocarcinoma planned treatment with curative resection: a multicenter study. *J Gastrointest Surg*. 2024;28(12):2039-2047. [\[FREE Full text\]](#) [doi: [10.1016/j.gassur.2024.09.027](https://doi.org/10.1016/j.gassur.2024.09.027)] [Medline: [39368645](#)]
7. Gupta P, Basu S, Arora C. Applications of artificial intelligence in biliary tract cancers. *Indian J Gastroenterol*. 2024;43(4):717-728. [doi: [10.1007/s12664-024-01518-0](https://doi.org/10.1007/s12664-024-01518-0)] [Medline: [38427281](#)]
8. Mar WA, Chan HK, Trivedi SB, Berggruen SM. Imaging of intrahepatic cholangiocarcinoma. *Semin Ultrasound CT MR*. 2021;42(4):366-380. [doi: [10.1053/j.sult.2021.04.001](https://doi.org/10.1053/j.sult.2021.04.001)] [Medline: [34130849](#)]

9. Moris D, Palta M, Kim C, Allen PJ, Morse MA, Lidsky ME. Advances in the treatment of intrahepatic cholangiocarcinoma: an overview of the current and future therapeutic landscape for clinicians. *CA Cancer J Clin*. 2023;73(2):198-222. [FREE Full text] [doi: [10.3322/caac.21759](https://doi.org/10.3322/caac.21759)] [Medline: [36260350](https://pubmed.ncbi.nlm.nih.gov/36260350/)]
10. Loosen SH, Roderburg C, Kauertz KL, Koch A, Vucur M, Schneider AT, et al. CEA but not CA19-9 is an independent prognostic factor in patients undergoing resection of cholangiocarcinoma. *Sci Rep*. 2017;7(1):16975. [FREE Full text] [doi: [10.1038/s41598-017-17175-7](https://doi.org/10.1038/s41598-017-17175-7)] [Medline: [29208940](https://pubmed.ncbi.nlm.nih.gov/29208940/)]
11. Izquierdo-Sanchez L, Lamarca A, La Casta A, Buettner S, Utpatel K, Klumpen HJ, Krupa, et al. Cholangiocarcinoma landscape in Europe: diagnostic, prognostic and therapeutic insights from the ENSCCA Registry. *J Hepatol*. 2022;76(5):1109-1121. [FREE Full text] [doi: [10.1016/j.jhep.2021.12.010](https://doi.org/10.1016/j.jhep.2021.12.010)] [Medline: [35167909](https://pubmed.ncbi.nlm.nih.gov/35167909/)]
12. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446. [FREE Full text] [doi: [10.1016/j.ejca.2011.11.036](https://doi.org/10.1016/j.ejca.2011.11.036)] [Medline: [22257792](https://pubmed.ncbi.nlm.nih.gov/22257792/)]
13. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpst WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*. 2013;52(7):1391-1397. [FREE Full text] [doi: [10.3109/0284186X.2013.812798](https://doi.org/10.3109/0284186X.2013.812798)] [Medline: [24047337](https://pubmed.ncbi.nlm.nih.gov/24047337/)]
14. Chu H, Liu Z, Liang W, Zhou Q, Zhang Y, Lei K, et al. Radiomics using CT images for preoperative prediction of futile resection in intrahepatic cholangiocarcinoma. *Eur Radiol*. 2021;31(4):2368-2376. [doi: [10.1007/s00330-020-07250-5](https://doi.org/10.1007/s00330-020-07250-5)] [Medline: [33033863](https://pubmed.ncbi.nlm.nih.gov/33033863/)]
15. Haghbin H, Aziz M. Artificial intelligence and cholangiocarcinoma: updates and prospects. *World J Clin Oncol*. 2022;13(2):125-134. [FREE Full text] [doi: [10.5306/wjco.v13.i2.125](https://doi.org/10.5306/wjco.v13.i2.125)] [Medline: [35316928](https://pubmed.ncbi.nlm.nih.gov/35316928/)]
16. Xu H, Li X, Jia M, Ma Q, Zhang Y, Liu F, et al. AI-derived blood biomarkers for ovarian cancer diagnosis: systematic review and meta-analysis. *J Med Internet Res*. 2025;27:e67922. [FREE Full text] [doi: [10.2196/67922](https://doi.org/10.2196/67922)] [Medline: [40126546](https://pubmed.ncbi.nlm.nih.gov/40126546/)]
17. Wakiya T, Ishido K, Kimura N, Nagase H, Kanda T, Ichiyama S, et al. CT-based deep learning enables early postoperative recurrence prediction for intrahepatic cholangiocarcinoma. *Sci Rep*. 2022;12(1):8428. [FREE Full text] [doi: [10.1038/s41598-022-12604-8](https://doi.org/10.1038/s41598-022-12604-8)] [Medline: [35590089](https://pubmed.ncbi.nlm.nih.gov/35590089/)]
18. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;9(5):1303-1322. [FREE Full text] [doi: [10.7150/thno.30309](https://doi.org/10.7150/thno.30309)] [Medline: [30867832](https://pubmed.ncbi.nlm.nih.gov/30867832/)]
19. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *Eur Radiol*. 2020;30(1):523-536. [doi: [10.1007/s00330-019-06360-z](https://doi.org/10.1007/s00330-019-06360-z)] [Medline: [31350588](https://pubmed.ncbi.nlm.nih.gov/31350588/)]
20. Salameh J, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred Reporting Items for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ*. 2020;370:m2632. [FREE Full text] [doi: [10.1136/bmj.m2632](https://doi.org/10.1136/bmj.m2632)] [Medline: [32816740](https://pubmed.ncbi.nlm.nih.gov/32816740/)]
21. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. [FREE Full text] [doi: [10.7326/0003-4819-155-8-201110180-00009](https://doi.org/10.7326/0003-4819-155-8-201110180-00009)] [Medline: [22007046](https://pubmed.ncbi.nlm.nih.gov/22007046/)]
22. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. [FREE Full text] [doi: [10.7326/M18-1376](https://doi.org/10.7326/M18-1376)] [Medline: [30596875](https://pubmed.ncbi.nlm.nih.gov/30596875/)]
23. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJPM, Hyde C, Brozek J, et al. Applying grading of recommendations assessment, development and evaluation (GRADE) to diagnostic tests was challenging but doable. *J Clin Epidemiol*. 2014;67(7):760-768. [FREE Full text] [doi: [10.1016/j.jclinepi.2014.01.006](https://doi.org/10.1016/j.jclinepi.2014.01.006)] [Medline: [24725643](https://pubmed.ncbi.nlm.nih.gov/24725643/)]
24. Bo Z, Chen B, Yang Y, Yao F, Mao Y, Yao J, et al. Machine learning radiomics to predict the early recurrence of intrahepatic cholangiocarcinoma after curative resection: a multicentre cohort study. *Eur J Nucl Med Mol Imaging*. 2023;50(8):2501-2513. [doi: [10.1007/s00259-023-06184-6](https://doi.org/10.1007/s00259-023-06184-6)] [Medline: [36922449](https://pubmed.ncbi.nlm.nih.gov/36922449/)]
25. Chakraborty J, Jolissaint J, Wang T, Soares K, Gonen M, Pak L. CT radiomics to predict early hepatic recurrence after resection for intrahepatic cholangiocarcinoma. In: *Medical Imaging 2022: Computer-Aided Diagnosis*. 2022. Presented at: SPIE Medical Imaging; 2022 April 4:120333C; San Diego, California. [doi: [10.1117/12.2612889](https://doi.org/10.1117/12.2612889)]
26. Chen B, Mao Y, Li J, Zhao Z, Chen Q, Yu Y, et al. Predicting very early recurrence in intrahepatic cholangiocarcinoma after curative hepatectomy using machine learning radiomics based on CECT: a multi-institutional study. *Comput Biol Med*. 2023;167:107612. [doi: [10.1016/j.combiomed.2023.107612](https://doi.org/10.1016/j.combiomed.2023.107612)] [Medline: [37939408](https://pubmed.ncbi.nlm.nih.gov/37939408/)]
27. Hao X, Liu B, Hu X, Wei J, Han Y, Liu X, et al. A radiomics-based approach for predicting early recurrence in intrahepatic cholangiocarcinoma after surgical resection: a multicenter study. *Annu Int Conf IEEE Eng Med Biol Soc*. 2021;2021:3659-3662. [doi: [10.1109/EMBC46164.2021.9630029](https://doi.org/10.1109/EMBC46164.2021.9630029)] [Medline: [34892030](https://pubmed.ncbi.nlm.nih.gov/34892030/)]
28. Jolissaint JS, Wang T, Soares KC, Chou JF, Gonen M, Pak LM, et al. Machine learning radiomics can predict early liver recurrence after resection of intrahepatic cholangiocarcinoma. *HPB (Oxford)*. 2022;24(8):1341-1350. [FREE Full text] [doi: [10.1016/j.hpb.2022.02.004](https://doi.org/10.1016/j.hpb.2022.02.004)] [Medline: [35283010](https://pubmed.ncbi.nlm.nih.gov/35283010/)]

29. Qin H, Hu X, Zhang J, Dai H, He Y, Zhao Z, et al. Machine-learning radiomics to predict early recurrence in perihilar cholangiocarcinoma after curative resection. *Liver Int.* 2021;41(4):837-850. [doi: [10.1111/liv.14763](https://doi.org/10.1111/liv.14763)] [Medline: [33306240](https://pubmed.ncbi.nlm.nih.gov/33306240/)]
30. Song Y, Zhou G, Zhou Y, Xu Y, Zhang J, Zhang K, et al. Artificial intelligence CT radiomics to predict early recurrence of intrahepatic cholangiocarcinoma: a multicenter study. *Hepatol Int.* 2023;17(4):1016-1027. [doi: [10.1007/s12072-023-10487-z](https://doi.org/10.1007/s12072-023-10487-z)] [Medline: [36821045](https://pubmed.ncbi.nlm.nih.gov/36821045/)]
31. Zhu Y, Mao Y, Chen J, Qiu Y, Guan Y, Wang Z, et al. Radiomics-based model for predicting early recurrence of intrahepatic mass-forming cholangiocarcinoma after curative tumor resection. *Sci Rep.* 2021;11(1):18347. [FREE Full text] [doi: [10.1038/s41598-021-97796-1](https://doi.org/10.1038/s41598-021-97796-1)] [Medline: [34526604](https://pubmed.ncbi.nlm.nih.gov/34526604/)]
32. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762. [doi: [10.1038/nrclinonc.2017.141](https://doi.org/10.1038/nrclinonc.2017.141)] [Medline: [28975929](https://pubmed.ncbi.nlm.nih.gov/28975929/)]
33. Zhao J, Zhang W, Zhu Y, Zheng H, Xu L, Zhang J, et al. Development and validation of noninvasive MRI-based signature for preoperative prediction of early recurrence in perihilar cholangiocarcinoma. *J Magn Reson Imaging.* 2022;55(3):787-802. [doi: [10.1002/jmri.27846](https://doi.org/10.1002/jmri.27846)] [Medline: [34296802](https://pubmed.ncbi.nlm.nih.gov/34296802/)]
34. Chen P, Yang Z, Zhang H, Huang G, Li Q, Ning P, et al. Personalized intrahepatic cholangiocarcinoma prognosis prediction using radiomics: application and development trend. *Front Oncol.* 2023;13:1133867. [FREE Full text] [doi: [10.3389/fonc.2023.1133867](https://doi.org/10.3389/fonc.2023.1133867)] [Medline: [37035147](https://pubmed.ncbi.nlm.nih.gov/37035147/)]
35. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020;295(2):328-338. [FREE Full text] [doi: [10.1148/radiol.2020191145](https://doi.org/10.1148/radiol.2020191145)] [Medline: [32154773](https://pubmed.ncbi.nlm.nih.gov/32154773/)]
36. Kann BH, Hosny A, Aerts HJ. Artificial intelligence for clinical oncology. *Cancer Cell.* 2021;39(7):916-927. [FREE Full text] [doi: [10.1016/j.ccell.2021.04.002](https://doi.org/10.1016/j.ccell.2021.04.002)] [Medline: [33930310](https://pubmed.ncbi.nlm.nih.gov/33930310/)]
37. Cholangiocarcinoma Working Group. Italian clinical practice guidelines on cholangiocarcinoma-part I: classification, diagnosis and staging. *Dig Liver Dis.* 2020;52(11):1282-1293. [doi: [10.1016/j.dld.2020.06.045](https://doi.org/10.1016/j.dld.2020.06.045)] [Medline: [32893173](https://pubmed.ncbi.nlm.nih.gov/32893173/)]
38. Bertuccio P, Malvezzi M, Carioli G, Hashim D, Boffetta P, El-Serag HB, et al. Global trends in mortality from intrahepatic and extrahepatic cholangiocarcinoma. *J Hepatol.* 2019;71(1):104-114. [FREE Full text] [doi: [10.1016/j.jhep.2019.03.013](https://doi.org/10.1016/j.jhep.2019.03.013)] [Medline: [30910538](https://pubmed.ncbi.nlm.nih.gov/30910538/)]
39. Yang C, Xu J, Wang S, Wang Y, Zhang Y, Piao C. Machine learning to predict the early recurrence of intrahepatic cholangiocarcinoma: a systematic review and meta-analysis. *Oncol Lett.* 2024;28(2):385. [doi: [10.3892/ol.2024.14518](https://doi.org/10.3892/ol.2024.14518)] [Medline: [38966582](https://pubmed.ncbi.nlm.nih.gov/38966582/)]
40. Xu L, Chen Z, Zhu D, Wang Y. The application status of radiomics-based machine learning in intrahepatic cholangiocarcinoma: systematic review and meta-analysis. *J Med Internet Res.* 2025;27:e69906. [FREE Full text] [doi: [10.2196/69906](https://doi.org/10.2196/69906)] [Medline: [40323647](https://pubmed.ncbi.nlm.nih.gov/40323647/)]
41. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol.* 2015;50(11):757-765. [FREE Full text] [doi: [10.1097/RLI.0000000000000180](https://doi.org/10.1097/RLI.0000000000000180)] [Medline: [26115366](https://pubmed.ncbi.nlm.nih.gov/26115366/)]
42. Jensen LJ, Kim D, Elgeti T, Steffen IG, Hamm B, Nagel SN. Stability of radiomic features across different region of interest sizes: a CT and MR phantom study. *Tomography.* 2021;7(2):238-252. [FREE Full text] [doi: [10.3390/tomography7020022](https://doi.org/10.3390/tomography7020022)] [Medline: [34201012](https://pubmed.ncbi.nlm.nih.gov/34201012/)]
43. Kakino R, Nakamura M, Mitsuyoshi T, Shintani T, Hirashima H, Matsuo Y, et al. Comparison of radiomic features in diagnostic CT images with and without contrast enhancement in the delayed phase for NSCLC patients. *Phys Med.* 2020;69:176-182. [doi: [10.1016/j.ejmp.2019.12.019](https://doi.org/10.1016/j.ejmp.2019.12.019)] [Medline: [31918370](https://pubmed.ncbi.nlm.nih.gov/31918370/)]
44. Hau H, Devantier M, Jahn N, Sucher E, Rademacher S, Seehofer D, et al. Impact of body mass index on tumor recurrence in patients undergoing liver resection for perihilar cholangiocarcinoma (pCCA). *Cancers (Basel).* 2021;13(19):4772. [FREE Full text] [doi: [10.3390/cancers13194772](https://doi.org/10.3390/cancers13194772)] [Medline: [34638257](https://pubmed.ncbi.nlm.nih.gov/34638257/)]
45. Su J, Yu X, Wang X, Wang Z, Chao G. Enhanced transfer learning with data augmentation. *Eng Appl Artif Intell.* 2024;129:107602. [doi: [10.1016/j.engappai.2023.107602](https://doi.org/10.1016/j.engappai.2023.107602)]
46. Park HJ, Park B, Park SY, Choi SH, Rhee H, Park JH, et al. Preoperative prediction of postsurgical outcomes in mass-forming intrahepatic cholangiocarcinoma based on clinical, radiologic, and radiomics features. *Eur Radiol.* 2021;31(11):8638-8648. [doi: [10.1007/s00330-021-07926-6](https://doi.org/10.1007/s00330-021-07926-6)] [Medline: [33890153](https://pubmed.ncbi.nlm.nih.gov/33890153/)]
47. Langlotz CP. The future of AI and informatics in radiology: 10 predictions. *Radiology.* 2023;309(1):e231114. [doi: [10.1148/radiol.231114](https://doi.org/10.1148/radiol.231114)] [Medline: [37874234](https://pubmed.ncbi.nlm.nih.gov/37874234/)]
48. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Mark.* 2021;31(3):685-695. [doi: [10.1007/s12525-021-00475-2](https://doi.org/10.1007/s12525-021-00475-2)]
49. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: an overview for clinical practitioners--Saliency-based XAI approaches. *Eur J Radiol.* 2023;162:110787. [doi: [10.1016/j.ejrad.2023.110787](https://doi.org/10.1016/j.ejrad.2023.110787)] [Medline: [37001254](https://pubmed.ncbi.nlm.nih.gov/37001254/)]
50. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med Image Anal.* 2023;84:102684. [doi: [10.1016/j.media.2022.102684](https://doi.org/10.1016/j.media.2022.102684)] [Medline: [36516555](https://pubmed.ncbi.nlm.nih.gov/36516555/)]

51. van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal.* 2022;79:102470. [FREE Full text] [doi: [10.1016/j.media.2022.102470](https://doi.org/10.1016/j.media.2022.102470)] [Medline: [35576821](https://pubmed.ncbi.nlm.nih.gov/35576821/)]

Abbreviations

AI: artificial intelligence

AUC: area under the receiver operating characteristic curve

CT: computed tomography

DOR: diagnostic odds ratio

EHC: extrahepatic cholangiocarcinoma

FN: false negative

FP: false positive

GRADE: Grading of Recommendations, Assessment, Development, and Evaluations

IHC: intrahepatic cholangiocarcinoma

LightGBM: light gradient boosting machine

MRI: magnetic resonance imaging

PITROS: Participants, Index test, Target condition, Reference standard, Outcomes, and Setting

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies

PROBAST: Prediction Model Risk of Bias Assessment Tool

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2

ResNet50: residual network 50

TN: true negative

TP: true positive

XGBoost: extreme gradient boosting

Edited by T Leung, A Coristine; submitted 30.May.2025; peer-reviewed by J Jagtap, P Shingru; comments to author 22.Jul.2025; revised version received 08.Aug.2025; accepted 30.Aug.2025; published 18.Sep.2025

Please cite as:

Chen J, Xi J, Chen T, Yang L, Liu K, Ding X

Diagnostic Performance of Computed Tomography-Based Artificial Intelligence for Early Recurrence of Cholangiocarcinoma: Systematic Review and Meta-Analysis

J Med Internet Res 2025;27:e78306

URL: <https://www.jmir.org/2025/1/e78306>

doi: [10.2196/78306](https://doi.org/10.2196/78306)

PMID:

©Jie Chen, Jianxin Xi, Tianyu Chen, Lulu Yang, Kaijia Liu, Xiaobo Ding. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 18.Sep.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.