

JOURNAL OF TAZEEZ FOR PUBLIC HEALTH

AN OFFICIAL JOURNAL OF SAUDI HEALTH PROMOTION AND EDUCATION ASSOCIATION

Systematic review

Performance and external validation of prediction models for gangrenous cholecystitis in emergency surgical patients: a systematic review

Mazi Mohammed Alanazi¹, Osamah Bin Bakheet², Sulaf Ahmed Alyami³, Rahaf Hussain Ali Alwadai⁴, Sarah Habib Alhindi⁵, Lana Hasan Abdullah Alamri⁶

1. Saudi and Jordanian Board Emergency Medicine, Emergency Department, Head of Emergency Research Unit, First Health Cluster, Riyadh, Saudi Arabia. ORCID: 0009-0003-5836-8086.
2. Emergency Consultant, Head of the Emergency Medicine Department, First Health Cluster, Riyadh, Saudi Arabia. Email: Dr-osamah@live.com.
3. Medical Intern, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia. Email: suealyami@gmail.com. ORCID: 0009-0003-1199-6376.
4. General Practitioner, College of Medicine, Najran University, Najran, Saudi Arabia. Email: Rahaf.hu01@gmail.com.
5. Medical Intern, College of Medicine, Fakeeh College for Medical Sciences, Jeddah, Saudi Arabia. Email: Saronalhindi@gmail.com.
6. Medical Intern, College of Medicine, Umm Al-Qura University, Makkah, Saudi Arabia. Email: lanahasan1423@gmail.com.

Abstract

Background: Gangrenous cholecystitis complicates emergency surgical assessment because routine clinical findings overlap with nongangrenous inflammation. Preoperative prediction offers a basis for risk stratification and operative planning. This review evaluated model performance and the strength of external validation. **Methods:** Two authors searched PubMed, Scopus, Web of Science, and the Cochrane Library through Jan 2026, screened studies, and extracted data. Original multivariable development and validation studies using operative or pathological gangrene outcomes were synthesized narratively. Discrimination, threshold accuracy, calibration, and validation design were examined separately. **Results:** eleven original articles described clinical equations,

Published: 8 Oct, 2026

<https://doi.org/10.65759/h5kgtv41>

Copyright © 2025 The Author(s). Published by Lizzy B. This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited).

bedside scores, imaging approaches, and machine learning. Independent evaluations of established scores yielded areas under the receiver operating characteristic curve of 0.72-0.78. A six-factor clinoradiological score reported an area of 0.83, although cohort-specific validation reporting required interpretation. A random forest achieved 0.818 in another-centre validation, while a fusion computed tomography model achieved 0.879 and 0.887 in two external cohorts. Same-centre temporal testing was distinguished from geographical validation. **Conclusions:** External validation supports several approaches, without establishing a universally transferable tool or improved clinical outcomes. Representative multicentre evaluation, transparent calibration, and prospective assessment of model-guided decisions remain priorities for emergency surgical implementation.

Keywords: Gangrenous cholecystitis; acute cholecystitis; prediction models; external validation; emergency surgery; machine learning; systematic review.

Introduction

Acute cholecystitis requires systematic clinical assessment of local inflammation, systemic inflammatory responses, and imaging abnormalities, with severity grading informing the urgency and complexity of surgical management. Gangrenous cholecystitis (GC) represents destructive gallbladder inflammation, and the Tokyo Guidelines classify this finding in the marked local inflammatory conditions defining Grade II disease (1). Early laparoscopic cholecystectomy is central to management of operable patients with acute calculous cholecystitis, supported by the World Society of Emergency Surgery recommendations (2). The Tokyo management algorithm integrates disease severity, comorbidity, physiological fitness, surgical expertise, and institutional resources when selecting early surgery, drainage, or delayed intervention (3).

The clinical importance of timely treatment is supported by randomized evidence comparing immediate and delayed surgical strategies in acute cholecystitis (4). In the ACDC trial, immediate cholecystectomy produced lower morbidity than initial antibiotic treatment followed by delayed surgery, with shorter hospitalization and lower costs (4). A Cochrane review found shorter total hospital stay with early surgery, without statistically significant differences in conversion or serious complications, while identifying substantial trial limitations (5). These studies establish the importance of efficient surgical pathways, although

their treatment comparisons do not validate a diagnostic model for gallbladder gangrene or establish the benefit of score-directed operating priorities (4,5).

Earlier observational investigations identified patient characteristics associated with gangrenous disease, providing a clinical foundation for subsequent prediction research. Merriam et al. identified older age, male sex, cardiovascular disease, and pronounced leukocytosis as risk factors in patients undergoing cholecystectomy for acute cholecystitis (6). Fagan et al. identified diabetes and white blood cell count as independent predictors in a retrospective surgical cohort (7). The differing predictor sets indicate that clinical associations depend on the population examined, while distinguishing association from reliable individual risk estimation remains essential when translating these findings into practical emergency surgical assessment (6,7).

Within this clinical context, preoperative risk assessment addresses the recognition of destructive local disease before its operative or pathological confirmation (1,6). Routine demographic, laboratory, and imaging findings provide the information underpinning that assessment, although isolated risk factors do not establish the performance of a combined prediction tool (6,7). The present review evaluates published multivariable approaches for identifying GC, emphasizing discrimination, threshold

performance, calibration, and validation in patients assessed for surgical treatment.

Methods

This systematic review followed the PRISMA 2020 reporting framework. Two authors independently searched PubMed, Scopus, Web of Science, and the Cochrane Library from database inception through Jan 2026. Searches combined controlled vocabulary and free-text terms for cholecystitis and gangrene with prediction, model, score, nomogram, risk stratification, discrimination, validation, and machine learning. Database-specific field tags and syntax were applied. Reference lists of relevant reports and reviews were examined to identify additional studies.

Original studies were eligible when they evaluated preoperative multivariable prediction of GC in surgical patients, including model development, validation, or evaluation of a combined diagnostic rule for gangrene. Histopathological confirmation or defined operative gangrene constituted acceptable reference standards. Predictor-only reports provided contextual evidence. Mixed surgical cohorts and imaging studies containing nongangrenous or healthy controls were retained, with applicability to emergency presentations assessed separately. We exclude case reports, reviews, postoperative prediction tools, isolated biomarker analyses, and studies predicting composite severe cholecystitis without separate gangrene information. English-language reports containing sufficient information for extraction were retained for synthesis.

Two authors screened titles and abstracts, examined full texts, and extracted data using a standardized form. Disagreements were resolved through discussion and reassessment of the source report. Extracted items comprised publication details, setting, recruitment design, sample size, outcome definition, predictor availability, model

structure, missing-data handling, and evaluation procedures. Performance extraction covered area under the receiver operating characteristic curve, sensitivity, specificity, predictive values, calibration, and decision-analytic measures. Development estimates, internal resampling, same-centre temporal evaluation, geographical validation, and independent testing of previously published models were distinguished. Cohorts and alternative models within the same publication were recorded separately.

Two authors examined methodological limitations in participants and data sources, predictors, outcomes, and analysis. Particular attention addressed selected surgical populations, healthy-control inclusion, incomplete predictor data, outcome heterogeneity, optimism, and separation of model development from evaluation. Findings were synthesized narratively and tabulated according to model family and validation design. Meta-analysis was not undertaken because the reports evaluated different models, thresholds, populations, and performance measures, with insufficient comparable independent evaluations of individual models. Accuracy, discrimination, and clinical utility were interpreted separately. The review used published aggregate information and involved no collection of identifiable patient data.

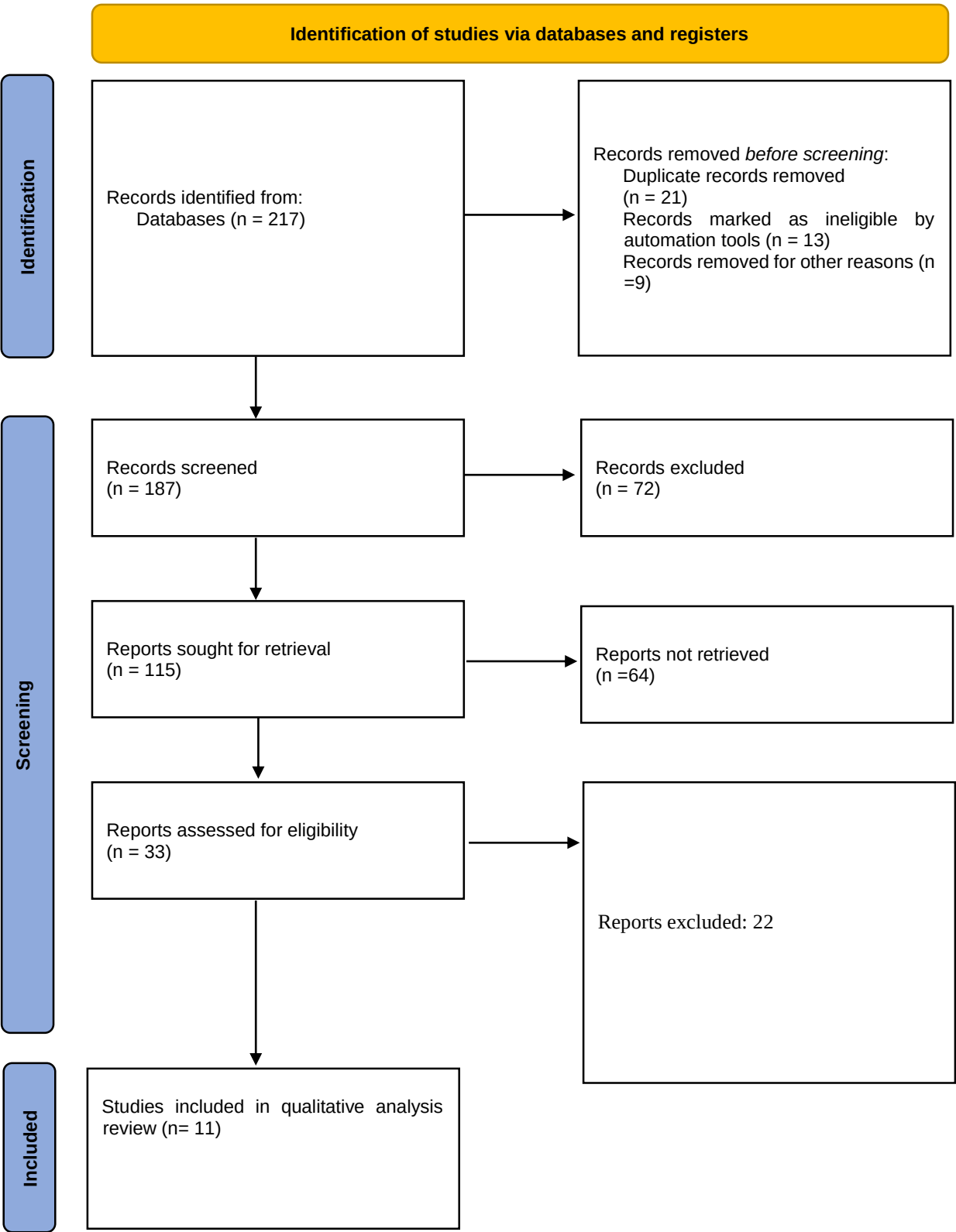


Figure 1: PRISMA flow chart

Results

We included 11 articles in our qualitative analysis, Bennett et al. evaluated CT findings in 75 patients, including 23 with gangrenous disease, and reported 86.7% accuracy for a multivariable imaging model. This model achieved 85.7% sensitivity and 87.5% specificity, whereas the overall observer diagnosis of gangrene had 29.3% sensitivity (8). Nguyen et al. prospectively tested a previously derived equation in 98 patients, including 18 with GC. At a probability threshold of 0.724, sensitivity was 83%, specificity 93%, positive predictive value 71%, and negative predictive value 96% (9).

Yacoub et al. developed a five-variable score in 245 patients and evaluated its risk categories in 313 subsequently admitted patients. The score incorporated age, sex, heart rate, leukocytosis, and ultrasound wall thickness, with derivation-cohort gangrene frequencies of 13% and 87% in the low-risk and high-risk categories, respectively. The later cohort represented same-centre temporal evaluation rather than geographical validation (10). Falor and colleagues identified male sex, elevated white blood cell count, tachycardia, and hyponatraemia through multivariable analysis of 391 urgent cholecystectomies, including 89 gangrenous cases, without reporting model discrimination in the abstract (11).

Wu et al. provided independent evaluation of the Yacoub score in a broader cholecystectomy cohort containing 351 gangrenous cases among 5,243 eligible patients. Complete predictor information was available for 478 patients, and the original score achieved an AUC of 0.74. Their locally

modified four-variable score achieved an AUC of 0.77 (12).

Chang et al. studied 308 patients, including 28 gangrenous cases, and identified distension, mural striation, and decreased mural enhancement as independent CT predictors (13). Reader-specific specificity was 93.9% and 89.6%, although the comparison population included chronic cholecystitis and healthy gallbladders (13). Mangukiya et al. constructed two equations in 437 patients, combining diabetes, leukocytosis, wall thickening, and pericholecystic fluid (14). Their standard model reported 82% sensitivity and 82.2% specificity within the original single-centre retrospective development population, without independent external validation data (14). Bouassida et al. developed a six-factor clinikoradiological score in 587 patients, reporting an AUC of 0.83, sensitivity of 81.4%, and specificity of 70% (15).

Ma et al. developed machine learning models in 1,006 patients, with 109 reserved for same-centre testing, and reported an XGBoost AUC of 0.944 (16). Guo et al. analyzed 1,228 patients, with external cohorts of 143 and 164 patients, obtaining fusion CT-model AUCs of 0.879 and 0.887 (17). Hu et al. trained models in 744 patients and evaluated a random forest in 300 patients from another centre, reporting an external AUC of 0.818 (18). Calibration and decision curves were reported in the latter study, while comparable quantitative calibration information remained inconsistent across the model reports, supporting separate interpretation of discrimination, validation design, and clinical usefulness.

Table 1. Original reports characteristics

Study	Cohort size	Approach and principal predictors
Bennett, 2002	75; 23 GC	CT regression: irregular wall, enhancement, thickness, distension (8)

Nguyen, 2004	98; 18 GC	Equation: diabetes, WBC, liver enzymes, pericholecystic fluid (9)
Yacoub, 2010	245 development; 313 temporal	Score: age, sex, heart rate, WBC, ultrasound wall thickness (10)
Falor, 2012	391; 89 GC	Contextual predictor analysis: sex, WBC, heart rate, sodium (11)
Wu, 2014	5,243 overall; 478 complete cases	Existing-score validation and four-variable score development (12)
Chang, 2016	308; 28 GC	CT analysis: distension, mural striation, decreased enhancement (13)
Mangukiya, 2019	437; 151 GC	Two equations: diabetes, WBC, wall thickness, pericholecystic fluid (14)
Bouassida, 2021	587 development; prospective cohort size NR [†]	Score: ASA class, temperature, symptom duration, WBC, sex, fluid (15)
Ma, 2025	897 development; 109 temporal	XGBoost: inflammatory/coagulation markers, gallbladder dimensions, electrolyte abnormalities (16)
Guo, 2025	921 training; 143/164 external	Self-supervised fusion of plain and contrast-enhanced CT (17)
Hu, 2025	744 development/internal testing; 300 external	Random forest: clinical, inflammatory, and imaging variables (18)

Table 2. Performance and validation evidence

Study	Evaluation design	Reported performance	Interpretation
Bennett	Development	Accuracy 86.7%; sensitivity/specificity 85.7%/87.5%	Fitted model; selected controls (8)
Nguyen	Prospective equation testing	Sensitivity/specificity 83%/93%; NPV 96%	AUC and geographical transportability NR (9)
Yacoub	Development plus temporal cohort	Derivation risk-category frequencies 13%-87%	Later same-hospital testing (10)
Falor	Contextual predictor analysis	Discrimination NR	Associations; no validated bedside score (11)
Wu	Independent existing-score testing	Yacoub AUC 0.74; revised score 0.77	Revised-score estimate from development data (12)
Chang	CT reader evaluation	Specificity 93.9%/89.6%	Reader estimates; mixed controls (13)
Mangukiya	Development	Standard/abbreviated AUC 0.908/0.861	No independent evaluation; equation-reporting inconsistencies (14)
Bouassida	Development; prospective validation reported [†]	New score AUC 0.83; Yacoub/Wu 0.78/0.72	New-score cohort-specific evaluation unclear [†] (15)

Ma	Same-centre temporal testing	AUC 0.944; 95% CI 0.888-0.984	Temporal generalization; geographical testing absent (16)
Guo	Two external centres	AUC 0.879/0.887 versus training 0.965	Cross-centre imaging evaluation (17)
Hu	Another-centre validation	AUC 0.818 versus internal testing 0.875	Calibration curves and decision curves reported (18)

Discussion

We identify reproducible discrimination in clinical and imaging approaches, with limited evidence of transportability and uneven reporting of absolute risk accuracy. The repeated independent evaluation of established scores provides a stronger basis for judging generalizability than the development performance of newly fitted alternatives. The variation between geographical testing and same-centre temporal evaluation remains central to interpreting the newer machine learning results, because these designs examine different changes in patient populations and clinical practice. The evidence supports prioritizing validation design and performance transparency when comparing tools for emergency surgical risk assessment.

Recurring inflammatory and structural predictors clearly differ between populations. Leukocytosis, tachycardia, gallbladder wall abnormalities, and pericholecystic changes reflect the combinations of systemic inflammation and local tissue injury represented in the included studies. Sex was strongly weighted in the original Yacoub score, while its adjusted association disappeared in the Wu cohort, demonstrating instability in at least one established predictor. PROBAST+AI distinguishes development quality from evaluation bias and examines participants, predictors, outcomes, and analysis, providing an appropriate framework for interpreting such differences without treating every significant association as a dependable clinical prediction (19).

Discrimination alone does not establish whether a predicted probability reflects an individual patient's risk, because ranking and calibration answer different questions (20). This distinction is relevant when clinicians interpret score categories as numerical probabilities and transfer them between institutions with different gangrene prevalence (12,20). The calibration reporting in Hu's study advances evaluation beyond a single AUC, although

external calibration intercepts and slopes offer more informative transportability assessment than a development goodness-of-fit test alone (18,20). Decision curve analysis evaluates net benefit in clinically relevant thresholds, connecting prediction errors with the consequences of acting on a result (21). Its interpretation depends on defining the surgical decision under consideration rather than assuming that identifying gangrene automatically mandates a particular intervention (3,21).

The observed advantage of advanced imaging and machine learning requires cautious comparison because the studies used different cohorts, predictor sets, reference standards, and evaluation procedures (8,13,16-18). Healthy controls and chronic cholecystitis in imaging datasets introduce applicability concerns when the intended population consists exclusively of acute emergency presentations (8,13). Yacoub's grouping of three xanthogranulomatous cases with gangrenous disease adds outcome heterogeneity, while complete-case availability restricted the Wu model analysis substantially (10,12). Mangukiya's report contains inconsistencies between coefficient signs, coded predictors, and illustrative probability calculations, limiting direct equation reuse (14). External validation guidance emphasizes applying a specified model to an appropriate new dataset and evaluating discrimination, calibration, and clinical usefulness, supporting replication across representative emergency surgical services (22).

Clinical integration requires distinguishing diagnosis of existing gangrene from prediction of future deterioration, since the included models predominantly classified disease confirmed during surgery or pathological assessment. Model-assisted prioritization complements established management pathways without establishing a demonstrated reduction in bile duct injury, complications, or mortality (2,3,12,18). TRIPOD+AI reporting supports reproducible evaluation through clear descriptions of development, data handling,

model specification, performance, and intended use across regression and machine learning approaches (23).

Conclusion

Prediction models for gangrenous cholecystitis combine accessible clinical findings with laboratory and imaging information to support individualized preoperative risk stratification. Established scores show moderate discrimination during independent testing, while newer machine learning approaches report stronger performance in selected external cohorts. Differences in patient selection, outcome definitions, missing data, and validation design limit comparisons between models. Current data supports further evaluation within emergency surgical pathways rather than universal adoption of a single tool.

Abbreviations

ASA: American Society of Anesthesiologists; AUC: area under the receiver operating characteristic curve; CI: confidence interval; CT: computed tomography; GC: gangrenous cholecystitis; NPV: negative predictive value; NR: not reported; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROBAST+AI: Prediction model Risk Of Bias ASsessment Tool plus Artificial Intelligence; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence; WBC: white blood cell count; XGBoost: extreme gradient boosting.

References

1. Yokoe M, Hata J, Takada T, Strasberg SM, Asbun HJ, Wakabayashi G, et al. Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). *J Hepatobiliary Pancreat Sci.* 2018;25(1):41-54. doi:10.1002/jhbp.515.
2. Pisano M, Allievi N, Gurusamy K, Borzellino G, Cimbanassi S, Boerna D, et al. 2020 World Society of Emergency Surgery updated guidelines for the diagnosis and treatment of acute calculus cholecystitis. *World J Emerg Surg.* 2020;15:61. doi:10.1186/s13017-020-00336-x.
3. Okamoto K, Suzuki K, Takada T, Strasberg SM, Asbun HJ, Endo I, et al. Tokyo Guidelines 2018: flowchart for the management of acute cholecystitis. *J Hepatobiliary Pancreat Sci.* 2018;25(1):55-72. doi:10.1002/jhbp.516.
4. Gutt CN, Encke J, Köninger J, Harnoss JC, Weigand K, Kipfmüller K, et al. Acute cholecystitis: early versus delayed cholecystectomy, a multicenter randomized trial (ACDC study, NCT00447304). *Ann Surg.* 2013;258(3):385-93. doi:10.1097/SLA.0b013e3182a1599b.
5. Gurusamy KS, Davidson C, Gluud C, Davidson BR. Early versus delayed laparoscopic cholecystectomy for people with acute cholecystitis. *Cochrane Database Syst Rev.* 2013;(6). doi:10.1002/14651858.CD005440.pub3.
6. Merriam LT, Kanaan SA, Dawes LG, Angelos P, Prystowsky JB, Rege RV, et al. Gangrenous cholecystitis: analysis of risk factors and experience with laparoscopic cholecystectomy. *Surgery.* 1999;126(4):680-6.
7. Fagan SP, Awad SS, Rahwan K, Hira K, Aoki N, Itani KMF, et al. Prognostic factors for the development of gangrenous cholecystitis. *Am J Surg.* 2003;186(5):481-5. doi:10.1016/j.amjsurg.2003.08.001.
8. Bennett GL, Rusinek H, Lisi V, Israel GM, Krinsky GA, Slywotzky CM, et al. CT findings in acute gangrenous cholecystitis. *AJR Am J Roentgenol.* 2002;178(2):275-81. doi:10.2214/ajr.178.2.1780275.
9. Nguyen L, Fagan SP, Lee TC, Aoki N, Itani KMF, Berger DH, et al. Use of a predictive

- equation for diagnosis of acute gangrenous cholecystitis. *Am J Surg*. 2004;188(5):463-6. doi:10.1016/j.amjsurg.2004.07.013.
10. Yacoub WN, Petrosyan M, Sehgal I, Ma Y, Chandrasoma P, Mason RJ. Prediction of patients with acute cholecystitis requiring emergent cholecystectomy: a simple score. *Gastroenterol Res Pract*. 2010;2010:901739. doi:10.1155/2010/901739.
 11. Falor AE, Zobel M, Kaji A, Neville A, De Virgilio C. Admission variables predictive of gangrenous cholecystitis. *Am Surg*. 2012;78(10):1075-8. doi:10.1177/000313481207801014.
 12. Wu B, Buddensick TJ, Ferdosi H, Narducci DM, Sautter A, Setiawan L, et al. Predicting gangrenous cholecystitis. *HPB (Oxford)*. 2014;16(9):801-6. doi:10.1111/hpb.12226.
 13. Chang WC, Sun Y, Wu EH, Kim SY, Wang ZJ, Huang GS, et al. CT findings for detecting the presence of gangrenous ischemia in cholecystitis. *AJR Am J Roentgenol*. 2016;207(2):302-9. doi:10.2214/AJR.15.15658.
 14. Mangukiya DO, Bhatt KS, Desai PN, Arora PV, Patel KB, Parekh KK, et al. Predictive models of gangrenous cholecystitis in laparoscopic cholecystectomy treated patients. *J Clin Diagn Res*. 2019;13(9)-3. doi:10.7860/JCDR/2019/42231.13137.
 15. Bouassida M, Madhioub M, Kallel Y, Zribi S, Slama H, Mighri MM, et al. Acute gangrenous cholecystitis: proposal of a score and comparison with previous published scores. *J Gastrointest Surg*. 2021;25(6):1479-86. doi:10.1007/s11605-020-04707-2.
 16. Ma Y, Luo M, Guan G, Liu X, Cui X, Luo F. An explainable predictive machine learning model of gangrenous cholecystitis based on clinical data: a retrospective single center study. *World J Emerg Surg*. 2025;20:1. doi:10.1186/s13017-024-00571-6.
 17. Guo Q, Li Y, Huang Y, Xie S, Li S, Zheng J, et al. Self-supervised learning model integrates plain and contrast-enhanced CT for preoperatively identifying gangrenous cholecystitis: a multicenter retrospective cohort study. *Int J Surg*. 2025;111(11):7622-32. doi:10.1097/JS9.0000000000003063.
 18. Hu Y, Chen Y, Zhao H. Development and validation of an explainable machine learning model for gangrenous cholecystitis prediction: a multicenter retrospective study. *J Inflamm Res*. 2025;18:17747-58. doi:10.2147/JIR.S552926.
 19. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388. doi:10.1136/bmj-2024-082505.
 20. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
 21. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-74. doi:10.1177/0272989X06295361.
 22. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384. doi:10.1136/bmj-2023-074820.
 23. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385. doi:10.1136/bmj-2023-078378.