CANADIAN JOURNAL OF UROLOGY

Tech Science Press

Doi:10.32604/cju.2026.077411
REVIEW

Can AI and predictive models accurately predict stone-free status? a systematic review and meta-analysis

Yahya Ghazwani,[1,2,3] Mohammad Alghafees,[1,2,3*] Mishari Alshasha,[1,2,3] Fahad Brayan,[1,2,3] Abdulrahman Alsayyari,[1,2,3] Ali Alyami[1,2,3]

[1]College of Medicine, King Saud bin Abdulaziz University for Health Sciences (KSAU-HS), Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia

[2]Department of Surgery, Division of Urology, King Abdulaziz Medical City, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia

[3]King Abdullah International Medical Research Centre (KAIMRC), Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia

GHAZWANI Y, ALGHAFEES M, ALSHASHA M, BRAYAN F, ALSAYYARI A, ALYAMI A. Can AI and predictive models accurately predict stone-free status? a systematic review and meta-analysis. *Can J Urol* 2026;33(2):291–308

Objectives: *The emergence of artificial intelligence (AI) and predictive modeling offers prospects for clinical, anatomical, and imaging factor combination, like radiomics, to help with stone-free status (SFS) estimation and peroperative decision-making. The goal of this study was, therefore, to define the present performance range, determine sources of heterogeneity, and determine methodological practices permitting reliable implementation by varied circumstances.*

Methods: *We searched six bibliographic databases through 19 September 2025. Studies deriving or validating AI/predictive models for SFS after ureteroscopy were eligible. Independent dual screening, duplicate data extraction, and risk-of-bias consideration using QUADAS-AI were conducted.*

Results: *Five retrospective cohorts were included. Modeling approaches encompassed multivariable logistic regression, regularized/radiomics pipelines, gradient boosting, and ensembles. SFS definitions ranged from <2 mm residual (day-1 to 3 months) to ≤5 mm residual (1 month), determined by plain radiography, ultrasound, and/or CT. The pooled ratio-scale effect for stone size per 1 mm increase was 1.26 (95% CI 0.91–1.76; $\tau^2 \approx 0.055$; $Q = 18.52$; $I^2 = 94.6\%$; prediction interval 0.03–49.45). Hydronephrosis (moderate–severe vs. mild/none) showed a pooled RR 2.72 (95% CI 0.96–7.72; $\tau^2 \approx 0.821$; $Q = 65.40$; $I^2 = 96.9\%$; prediction interval 0.03–249.87). As continuous contrasts, stone size was larger in the non-stone-free group (SMD 1.36, 95% CI 0.85–1.86; $\tau^2 \approx 0.096$; $I^2 = 72.9\%$; prediction interval −3.77 to 6.48), and HU was higher (SMD 0.64, 95% CI 0.39–0.90; $\tau^2 \approx 0$; $Q = 0.73$; $I^2 = 0\%$; prediction interval −0.99 to 2.27).*

Conclusions: *Across studies evaluating AI and predictive models for ureteroscopy, discrimination was generally acceptable to excellent, and performance appeared highest in models integrating radiomics with anatomic/clinical descriptors. However, the degree of between-study heterogeneity (population mix, outcome definitions, imaging protocols, thresholds, and follow-up windows) was sufficiently large that pooled quantitative estimates should be considered clinically uninterpretable.*

Key Words: ureteroscopy, urolithiasis, artificial intelligence, radiomics, machine learning, stone-free status

Received date 09 December 2025
Accepted for publication 28 February 2026
Published online 15 April 2026

*Corresponding Author: Mohammad Alghafees. Email: alghafees687@gmail.com

Introduction

Urolithiasis remains a common and recurring disease significantly affecting healthcare utilization and quality of life globally.[1] Ureteroscopy has become a first-line approach for management of ureteral and selected renal stones based on high clearances, progressing instrument functionality, and an acceptable profile for morbidity.[2] In this regard, the stone-free status (SFS)—commonly defined by an imaging criterion for remaining fragments and assessed at fixed intervals—serves as a key endpoint offering information regarding subsequent procedures necessary, complications risk, related cost, and vigor of follow-up needed.[3] As a consequence, accurate preoperative SFS predictions are clinically and operationally significant: patient education and per-selection of equipment and planning for operative procedures can all be influenced by them as well as vigor of follow-up after surgery.[3]

Historically, prognosis has depended upon expert-crafted scores and classical regression models, covering only a limited set of quantifiable variables easily available, such as stone size, location/topography, multiplicity, hydronephrosis, and Hounsfield units.[4–6] Though these methods are interpretable and straightforward to program, they are prone to fail to fully benefit from information embedded in cutting-edge cross-sectional imaging, disregard nonlinear interactions and relations, and exhibit performance variability across institutions with differing patient sets and imaging modalities.[5,6] More recent breakthroughs involving radiomics and machine learning (ML) permit an elaborated representational framework with automatic extraction of quantifiable attributes from images and application of versatile function approximators (e.g., regularized logistic regression, gradient boosting, and neural networks) to transform clinical and imaging inputs to SFS outputs.[7,8] These methods can enhance discrimination and calibration but bring novel challenges regarding data quality, harmonization, and transparent reporting.[7,8]

Another perennial challenge to evidence syntheses in this field is heterogeneity of SFS definitions and timing windows, with differences in residual-fragment thresholds (e.g., ≤2–5 mm) and imaging techniques (plain radiography, ultrasound, CT), and early vs. delayed verification of clearance.[3,9,10] These design choices alter numbers of events and can significantly affect apparent model performance and effect sizes, particularly for location- and burden-dependent covariates.[9,10] In addition, there have been predominately single-center and retrospective development studies, to date augmenting risks of spectrum bias and overly favorable internal validation with limited transportability without recalibration.[8,9] External validation and update methods for models and application of AI-specific reporting templates (e.g., TRIPOD-AI and CONSORT-AI), have been recommended as strategies to overcome these risks and achieve safe and reliable translation to practice.[9,11]

In this context, we systematically accumulated and quantitatively summarized studies employing artificial intelligence or predictive modeling techniques to characterize SFS following ureteroscopy. Our focus was upon frequently observed clinical and imaging predictors, measures of discriminative ability, effect sizes, and measures of generalizability. The goal of this study was to define the present performance range, determine sources of heterogeneity, and determine methodological practices permitting reliable implementation by varied circumstances.

Materials and Methods

Review design

We constructed the review with a PECOS statement and applied selection, extraction, and synthesis per PRISMA 2020 reporting recommendations.[12] The Population included patients with urolithiasis who undergo ureteroscopic procedures (semirigid ureteroscopy, flexible ureterorenoscopy/retrograde intrarenal surgery), without regard to laterality of stones or burden, across any care environment. The Exposure/Intervention comprised artificial intelligence/predictive modeling methods—including logistical or other statistical models, machine-learning (e.g., tree-based ensembles, gradient boosting, support vector machines), deep learning, and radiomics pipelines—constructed from pre-operative and/or peri-operative clinical and imaging elements to foretell stone-free status. The Comparator (e.g., clinician judgment, rule-based score(s), or other models) was encouraged but optional. The Outcomes emphasized stone-free status (SFS) by source studies' definition (e.g., residual fragment thresholds ≤2–5 mm with modality-specific follow-up) and performance by model (discrimination, calibration, diagnostic classification metrics) and/or effect sizes relating inputs to SFS. The Study design was limited a priori to retrospective observational cohorts (single- or multi-center), covering development sets, internal validation sets,

and external validation sets evaluating ureteroscopy-only sets. The workflow of the review, eligibility, double independent screening and reasoned exclusions, data extraction doublely, risk-of-bias grading, and synthesis decisions were implemented and documented per PRISMA; a PRISMA flow diagram summarized all records through recognition, checking for non-screenable reasons and eligibility and inclusion. The searches extended to inception through 19 September 2025 with reproducible plans stored before initial screening.

Inclusion and exclusion criteria

We selected studies that: (i) recruited patients who undergo ureteroscopy (semirigid URS, flexible URS/fURSL, RIRS) for urolithiasis; (ii) employed AI/ML or predictive models (inclusive of traditional multivariable statistical models when specifically constructed for prediction) to predict SFS; (iii) gave SFS definitions and/or measures of quantitative performance of models (e.g., AUC/ROC, accuracy, sensitivity, specificity, PPV/NPV, calibration summaries, effect measures like OR/RR) or left extractable 2×2 or continuous data allowing re-extraction and re-estimation; (iv) had a retrospective study design; and (v) were full-text primary research papers published in English. We excluded: (i) prospective study designs (inclusive of randomized studies/perspective studies), editorials, reviews, protocols, conference papers with missing complete data, case reports/series; (ii) those mainly managed with therapy other than ureteroscopy (e.g., ESWL, PCNL) unless there was a separately analyzed ureteroscopy-only subgroup with satisfactory data; (iii) those with neither an AI/predictive element nor a quantitatively expressed relevant SFS outcome; (iv) replicated data sets (latest/most comprehensive report was retained); and (v) non-English-language papers when transfer was unacceptable within study timeframes.

Search strategy for database

We identified database-specific strategies by employing controlled vocabulary and free-text synonyms for condition (urolithiasis; kidney/ureteral stones), procedure (ureteroscopy/fURSL/RIRS/laser lithotripsy), methodology (artificial intelligence, machine learning, deep learning, radiomics, prediction, nomogram), and outcome (stone-free, residual fragments, SFR). The process was further elaborated by employing Boolean operators and adjacency/field restrictions unique to each platform. Our search criteria only encompassed human subjects and English language where possible filters existed, with an additional term for retrospective design if possible. The searches comprised various databases: MEDLINE (PubMed), Embase (Ovid), Scopus (Elsevier), Web of Science Core Collection (Clarivate), Cochrane Library (Wiley), and IEEE Xplore (Table 1). Implied strategies were piloted once, internally peer-checked once, and run to September 2025. The reference lists of studies selected and relevant reviews were scoured to identify further records.

Data extraction protocol and data items

We conducted dual, independent data extraction using piloted template with consensus adjudication by third reviewer. For each eligible study we extracted: bibliographic information (first author, year, journal), study characteristics (-country/region, center(s), time period, retrospective design subtype, sample number, inclusion/exclusion criteria), patient and stone measures (age, sex, BMI if known; stone size/volume, number, location/topography, lower-pole anatomy proxies; hydronephrosis grade; Hounsfield units; stent/nephrostomy; urine culture), imaging information (CT protocol, segmentation/feature extraction, radiomics pipeline, selection of features), intervention/model information (model family—e.g., logistic regression, gradient boosting, random forest, SVM, neural network; hyperparameter tuning, cross-validation specification, training/validation/test splits; internal and external validation), index test information (pre-specification of thresholding, class imbalance management, missing data approach, calibration techniques, decision-curve analysis), outcome definitions (SFS threshold, modality, timing), and performance measures (AUC/ROC; accuracy, sensitivity, specificity, PPV, NPV; calibration intercept/slope; Brier score when reported). Where available, we extracted effect estimates for key predictors (e.g., OR/RR with 95% CI for size, location, hydronephrosis, HU) and raw numerators/denominators for dichotomous contrasts and means/SDs for continuous contrasts, allowing standardized re-estimation. For situations with multiple models/timepoints being reported we used a predefined hierarchy favoring SFS threshold closest to $\leq$2 mm at earliest routine imaging evaluation and most parsimonious clinically deployable model; alternate models/timepoints were recorded for sensitivity/narrative synthesis.

Bias evaluation protocol

Risk of bias and applicability were evaluated using QUADAS-AI,[13] with independent assessments by two

TABLE 1. **Search strings utilised across the databases (Abbreviations: MeSH = Medical Subject Headings; tiab = title/abstract; exp = explode (includes narrower terms); ti,ab = title, abstract; kw = keyword; TITLE-ABS-KEY = title/abstract/keywords (Scopus field); TS = topic (Web of Science); NEAR/3 = within 3 words; W/3 = within 3 words; NEXT = next to/adjacent; RIRS = retrograde intrarenal surgery; fURS = flexible ureteroscopy; SFR = stone-free rate; PCNL = percutaneous nephrolithotomy; ESWL = extracorporeal shock wave lithotripsy; IEEE = Institute of Electrical and Electronics Engineers; Ovid = Ovid platform)**

Database (platform)	Final search string (copied verbatim)
MEDLINE (PubMed)	("Urolithiasis"[MeSH] OR urolithiasis[tiab] OR nephrolithiasis[tiab] OR "Urinary Calculi"[MeSH] OR (kidney[tiab] OR renal[tiab] OR ureter*[tiab]) AND (stone*[tiab] OR calculi[tiab])) AND ("Ureteroscopy"[MeSH] OR ureteroscop*[tiab] OR "retrograde intrarenal surgery"[tiab] OR RIRS[tiab] OR fURS[tiab] OR "laser lithotripsy"[tiab]) AND ("Artificial Intelligence"[MeSH] OR "Machine Learning"[MeSH] OR "Deep Learning"[tiab] OR radiomic*[tiab] OR "predict*"[tiab] OR nomogram*[tiab] OR "logistic regression"[tiab] OR "gradient boosting"[tiab] OR random forest[tiab] OR xgboost[tiab] OR lightgbm[tiab] OR "neural network*"[tiab]) AND (("stone free"[tiab] OR stone-free[tiab] OR "SFR"[tiab] OR "residual fragment*"[tiab] OR clearance[tiab])) AND (retrospective[tiab] OR "Retrospective Studies"[Publication Type]) NOT (pcnl[tiab] OR "percutaneous nephrolithotomy"[tiab] OR eswl[tiab] OR "shock wave lithotripsy"[tiab])
Embase (Ovid)	('urolithiasis'/exp OR urolithiasis:ti,ab OR nephrolithiasis:ti,ab OR 'urinary calculus'/exp) AND ('ureteroscopy'/exp OR ureteroscop*:ti,ab OR 'retrograde intrarenal surgery':ti,ab OR RIRS:ti,ab OR fURS:ti,ab OR 'laser lithotripsy':ti,ab) AND ('artificial intelligence'/exp OR 'machine learning'/exp OR 'deep learning'/exp OR radiomic*:ti,ab OR predict*:ti,ab OR nomogram*:ti,ab OR 'logistic regression':ti,ab OR 'gradient boosting':ti,ab OR 'random forest':ti,ab OR xgboost:ti,ab OR lightgbm:ti,ab OR 'neural network*':ti,ab) AND ('stone free':ti,ab OR 'stone-free':ti,ab OR SFR:ti,ab OR 'residual fragment*':ti,ab OR clearance:ti,ab) AND ([english]/lim AND [human]/lim) AND retrospective:ti,ab NOT ('percutaneous nephrolithotomy'/exp OR 'extracorporeal shock wave lithotripsy'/exp)
Scopus (Elsevier)	TITLE-ABS-KEY((urolithiasis OR nephrolithiasis OR "urinary calculi" OR ("kidney" OR "renal" OR "ureter*") W/3 (stone* OR calculi)) AND (ureteroscop* OR "retrograde intrarenal surgery" OR RIRS OR fURS OR "laser lithotripsy") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR radiomic* OR predict* OR nomogram* OR "logistic regression" OR "gradient boosting" OR "random forest" OR xgboost OR lightgbm OR "neural network*") AND ("stone free" OR stone-free OR SFR OR "residual fragment*" OR clearance) AND retrospective) AND NOT (pcnl OR "percutaneous nephrolithotomy" OR eswl OR "shock wave lithotripsy") AND (LIMIT-TO(LANGUAGE, "English"))
Web of science core collection	TS=((urolithiasis OR nephrolithiasis OR "urinary calculi" OR ((kidney OR renal OR ureter*) NEAR/3 (stone* OR calculi))) AND (ureteroscop* OR "retrograde intrarenal surgery" OR RIRS OR fURS OR "laser lithotripsy") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR radiomic* OR predict* OR nomogram* OR "logistic regression" OR "gradient boosting" OR "random forest" OR xgboost OR lightgbm OR "neural network*") AND ("stone free" OR stone-free OR SFR OR "residual fragment*" OR clearance) AND retrospective) NOT TS=("percutaneous nephrolithotomy" OR PCNL OR "extracorporeal shock wave lithotripsy" OR ESWL)
Cochrane library (Wiley)	ti,ab,kw:(urolithiasis OR nephrolithiasis OR "urinary calculi" OR (kidney OR renal OR ureter* NEXT stone* OR calculi)) AND ti,ab,kw:(ureteroscop* OR "retrograde intrarenal surgery" OR RIRS OR fURS OR "laser lithotripsy") AND ti,ab,kw:("artificial intelligence" OR "machine learning" OR "deep learning" OR radiomic* OR predict* OR nomogram* OR "logistic regression" OR "gradient boosting" OR "random forest" OR xgboost OR lightgbm OR "neural network*") AND ti,ab,kw:("stone free" OR stone-free OR SFR OR "residual fragment*" OR clearance) AND ti,ab,kw:retrospective NOT ti,ab,kw:(PCNL OR "percutaneous nephrolithotomy" OR ESWL OR "shock wave lithotripsy")
IEEE xplore	(("All Metadata": urolithiasis OR "urinary calculi" OR nephrolithiasis) AND ("All Metadata": ureteroscop* OR "retrograde intrarenal surgery" OR RIRS OR fURS OR "laser lithotripsy") AND ("All Metadata": "artificial intelligence" OR "machine learning" OR "deep learning" OR radiomic* OR "neural network" OR "gradient boosting" OR "random forest") AND ("All Metadata": "stone free" OR stone-free OR SFR OR "residual fragment*" OR clearance) AND ("All Metadata": retrospective)) NOT ("All Metadata": PCNL OR "percutaneous nephrolithotomy" OR ESWL OR "shock wave lithotripsy")

reviewers and consensus resolution. Domain assessments were conducted for Patient Selection, Index Test/Analysis, Reference Standard, and Flow and Timing, as well as considerations specific to AI, including data provenance, leakage protection, modeling transparency, and transportability. "Dataset governance/generalizability" was defined by the availability and sufficiency of: (i) data provenance and cohort construction description, (ii) safe leakage and proper resampling of training/validation/test splits, (iii) handling of missing data and class imbalance, (iv) reproducible feature extraction (including radiomics pipeline description when applicable), and (v) transportability support, which included external or temporal validation and/or calibration reporting and recalibration. Studies were considered to have at least "some concerns" regarding this domain if they were retrospective, single-center studies without external validation, failed to report calibration, or chose thresholds a posteriori, as these factors limit generalizability and the validity of governance assurances.

Meta Analysis Protocol and Statistical Analysis

The quantitative synthesis was conducted utilizing METAANALYSISONLINE software package (Version number 1.0).[14] The experimental cohort was designated as No-stone-free (No-SFR; referred to as "cases"), while the control cohort was characterized as SFR ("controls"). For evaluating ratio-scale effects of stone size per 1-mm increment, we aggregated the log-effects reported in studies employing inverse-variance, random-effects methodology (DerSimonian–Laird) and subsequently presented back-transformed ratios accompanied by 95% confidence intervals (CIs); METAANALYSISONLINE provided τ^2, Cochran's Q (χ^2, degrees of freedom, p), I^2, a Z test for overall effect, and a 95% prediction interval, all of which were illustrated beneath the forest plot. In the case of the dichotomous variable hydronephrosis (moderate–severe vs. mild/none), we entered 2×2 contingency data as events and totals for both cases and controls, estimating risk ratios through the Mantel–Haenszel random-effects model with 95% CIs, similarly reporting τ^2, Q, I^2, Z, and the prediction interval precisely as depicted in the figure. A p value of 0.5 was considered to be statistically significant. For continuous outcomes—specifically, stone size (mm) and stone density (HU)—we combined standardized mean differences (Hedges g) using inverse-variance random-effects, with the directional specification as (No-SFR − SFR), thereby indicating that positive values reflected poorer profiles among failures.

Ethics approval and consent to participate:

This study was a systematic review and meta-analysis of previously published studies. No new data were collected from human participants, no identifiable private information was accessed, and no interventions were performed. Therefore, institutional ethics committee approval and informed consent were not required.

RESULTS

The search retrieved 677 database records from which 51 (7.5%) duplicates were excluded (Figure 1). The 626 remaining records were assessed; none were excluded at title/abstract and all proceeded to retrieval, 32/626 (5.1%) being non-retrievable. 594 reports went to full evaluation; exclusions being case reports 288/594 (48.5%), reviews 164/594 (27.6%), and off-topic 137/594 (23.1%), and 5 eligible retrospective studies remaining.[15–19]

Bias observations

As shown through Figure 2, the overall risk of bias was low for three[16,17,19] and some concerns for two.[15,18] Patient selection indicated some concerns in[15,18] and was low elsewhere.[16,17,19] Index AI processes indicated some concerns in[15,18,19]—largely regarding model specification and explainability—while being low in.[16,17] The reference standard domain indicated some concerns only in;[16] all others were low.[15,17–19] Flow and timing were low in[15,17,19] but indicated some concerns in.[16,18] Dataset governance/generizability was low for[15–17,19] and indicated some concerns in[18] owing to single-center focus and transportability constraints.

The QUADAS-AI summary showed that the overall risk of bias was at least "some concerns" for all included studies, rather than "low risk overall" for most. This assessment was done because the evidence base was comprised entirely of retrospective cohorts, most models were not externally validated, there was incomplete reporting of calibration performance across several studies, and decision thresholds were not prespecified for most, introducing non-trivial concerns under QUADAS-AI, especially in the Index Test/Analysis and Dataset Governance/Generalizability domains. Dataset governance/generalizability was assessed using specific criteria, such as adequate documentation of data

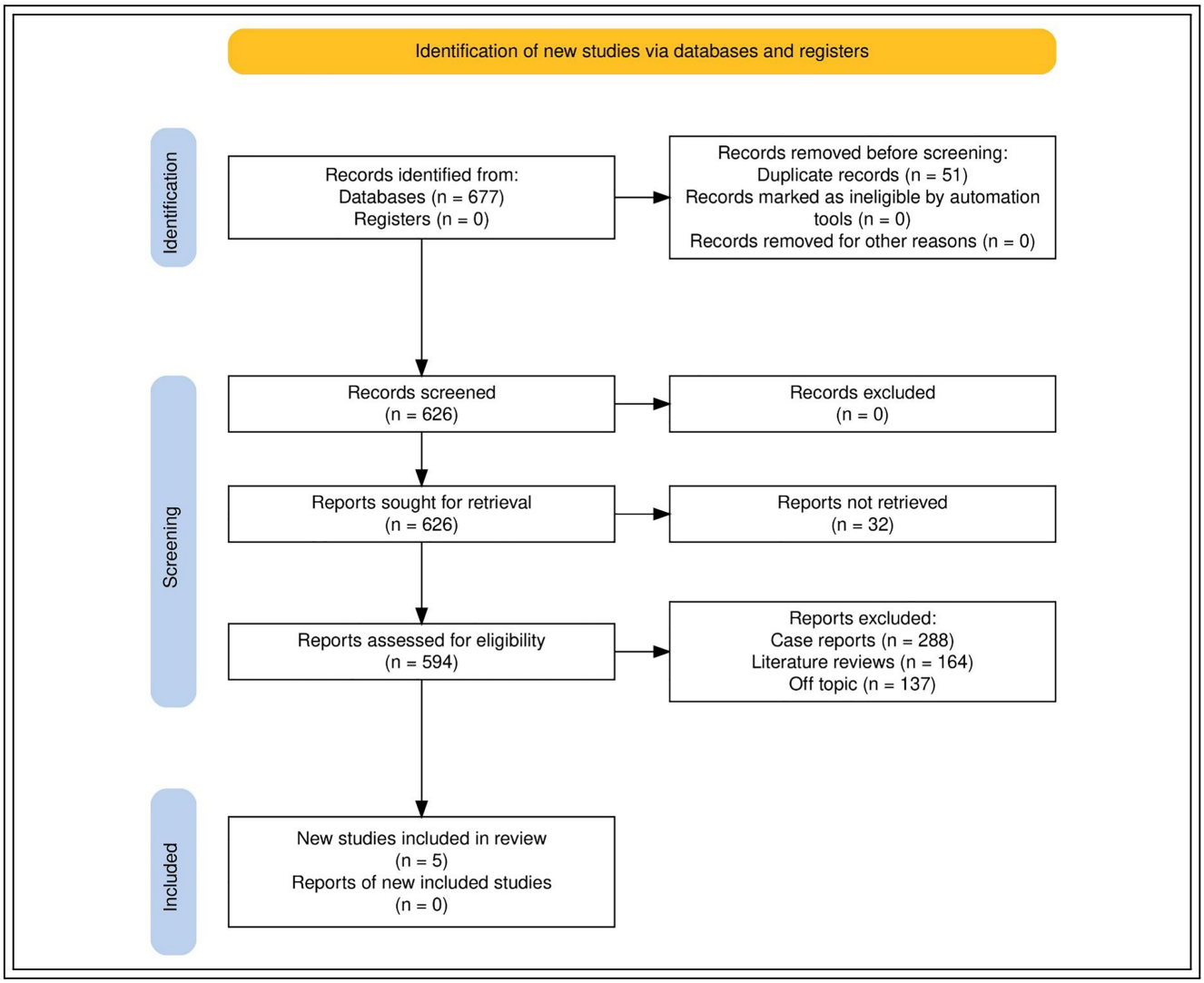

FIGURE 1. Study selection process for the review

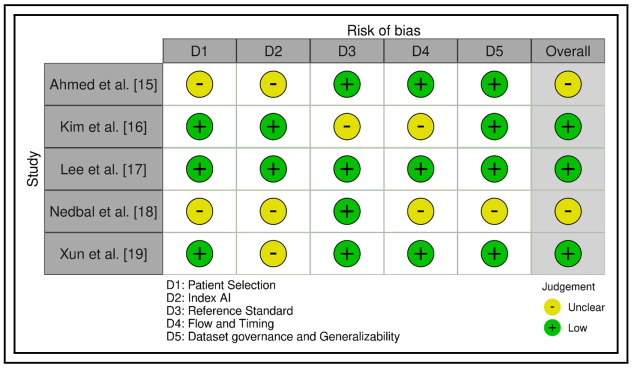

FIGURE 2. Bias levels assessed across the included studies[15–19]

provenance and cohort construction, leakage-safe splitting of training/validation/test datasets (or equivalent sound resampling without information leakage), adequate documentation of missing data and class imbalance handling, reproducible feature extraction (including specification of radiomics pipelines as appropriate), and transportability evidence such as external or temporal validation and/or adequate documentation of calibration performance and recalibration strategies. Since these criteria were not met, especially regarding external validation, calibration reporting, and prespecified thresholding, studies were correctly assigned a rating of "some concerns" regarding dataset governance/generalizability and could not be rated as "low

risk overall" without better assurance of generalizability and model validity.

Demographic variables

The selected studies as presented in Table 2 ranged over 2014,[16] 2020,[19] 2024,[18] and 2025,[15,17] and comprised retrospective cohorts with single-center development/validation or multi-center testing.[15–19] The number of samples ranged from n = 237[16] and n = 266[15] to n = 872,[18] with one two-center data set with internal cross-validation and an external test set.[17]

Intervention framework and approach to modeling

All studies foretold stone-free status at follow-up after ureteroscopy based on clinical/radiologic inputs with or without radiomics with no active comparators.[15–19] Machine learning algorithms comprised logistic regression for manually engineered features,[15–16] LightGBM for tabular ML with SHAP interpretability,[17] CatBoost/ensemble classifiers and a multitask ANN for largesingle-center data,[18] and LASSO-regularized logistic regression combining a radiomics signature with clinical variables.[19]

Characteristics and possible predictors of input

Feature sets consistently captured stone burden/size,[15–19] location/topography (including proximal ureter or lower-pole anatomy proxies),[15–19] hydronephrosis/obstruction,[15–16,19] stone density (HU),[15–17] and multiplicity.[15,18] Additional determinants included tissue rim sign,[16] operator experience thresholds,[19] anatomic channel measures (e.g., pelvic splanchnic angle, renal infundibulopelvic length, infundibular width),[17] microbiology (urine culture) and pre-operative stent/nephrostomy as peri-operative context variables,[18] and high-dimensional NCCT radiomics (604→28 features).[19]

Stone-free definitions and windows for assessment

Definitions of outcomes differed: ≤2 mm residual by endoscopy/fluoroscopy/or imaging at 1 month,[15] <2 mm on day-of/discharge plain radiograph after URSL (POD1),[16] ≤5 mm by CT at 1 month,[17] <2 mm endoscopic with no >2 mm by KUB/US at 3 months,[18] and <2 mm by CT/KUB at 3 months.[19] This diversity in threshold[15–19] and time point[15–19] reflected non-trivial study-to-study variation in classification.

Discrimination (AUC)

Global discrimination was acceptable to excellent with AUC 0.759 (external validation),[17] 0.785,[15] 0.825,[16] and 0.949/0.947 (derivation/validation),[19] while a large-scale ensemble only showed classification performance and did not show AUC.[18] The interval from 0.759[17] to 0.949[19] improved when covariates at the level of operators and when phenotypically constrained cohorts were added.

Classification performance (accuracy, sensitivity, specificity, PPV, NPV)

Accuracy approached 93% using an ensemble approach,[18] 85% using a logistical regression based upon a nomogram,[15] and 77.1% using an externally calibrated gradient-boosted approach.[17] Sensitivity approached 88.9% with external validation[17] and 84.8% with CT-feature LR,[16] with 65% recall at an ensemble with an operating point reflecting a precision-driven approach.[18] Specificity was 81% for a S.T.O.N.E.-based approach[15] and 69.3% for a CT-based LR.[16] PPV was 75% using a score-based approach,[15] 82.8% using a boosted external test,[17] and 87% using a precision at an ensemble approach.[18] NPV was 83% where it was published.[15] Operating points individually favored moderate-to-high rule-in ability with variable performance at excluding with thresholding strategy and class balance-driven differences.[15–18]

Effect sizes and main predictors

Adjusted associations repeatedly showed that a 1-mm increase in stone size correlated with increased odds of non-clearance (OR 1.51)[15] and (OR 1.074);[16] risk was markedly increased when stones lay in the proximal ureter (AOR 15.13);[15] moderate and severe hydronephrosis significantly contributed to risk (AOR 34.23 and AOR 33.75);[15] increased HU per unit added a small but cumulative risk (AOR 1.01);[15] and an itemizing rim sign of ≥2 revealed difficult extraction process (OR 4.635).[16] In radiomics-enhanced modeling, severe hydronephrosis (OR 9.908), favorable stone volume ≤1 cm^3 compared to >1 cm^3 (OR 8.337), operator experience with ≥100 fURS as favorable compared to <100 (OR 11.169), and radiomics score per unit (OR 2.679) became significant variables.[19] The gradient boost modeling unveiled stone burden, HU, pelvic splanchnic angle, and renal infundibulopelvic length as key contributors by feature attribution,[17] whereas logistic coefficients based on a large single-center pipeline highlighted total burden, stenting pre-operatively and positive urine culture as prime areas to prioritize.[18]

TABLE 2. **Study characteristics and outcome table (Abbreviations: AI = artificial intelligence; N/n = sample size; URS = ureteroscopy; semirigid URS = semirigid ureteroscopy; pneumatic URS/URSL = pneumatic ureteroscopic lithotripsy; URSL = ureteroscopic lithotripsy; RIRS = retrograde intrarenal surgery; fURS/fURSL = flexible ureteroscopy/flexible ureteroscopic lithotripsy; S.T.O.N.E. = Stone size, Topography, Obstruction (hydronephrosis), Number of stones, Evaluation of HU; HU = hounsfield units; CT/NCCT = computed tomography/non-contrast CT; US = ultrasound; KUB = kidney–ureter–bladder radiograph; fluoro = fluoroscopy; POD1 = postoperative day 1; SFS/SFR = stone-free status/stone-free rate; AUC = area under the receiver operating characteristic curve; Sens = sensitivity; Spec = specificity; PPV/NPV = positive/negative predictive value; CI = confidence interval; OR = odds ratio; LR = logistic regression; ML = machine learning; LASSO = least absolute shrinkage and selection operator; LightGBM = Light Gradient Boosting Machine; CatBoost = categorical boosting; RF = random forest; ANN = artificial neural network; CV = cross-validation; BMI = body mass index; DM/HTN = diabetes mellitus/hypertension; PSA = pelvic splanchnic angle; RIL = renal infundibulopelvic length; IW = infundibular width; SHAP = shapley additive explanations; NR = not reported)**

Study name	Year	Study type	Intervention (AI/model, inputs, comparator)	Algorithm	Key input variables/features	Stone-free definition (SFS criteria)	AUC	Accuracy	Sens	Spec	PPV	NPV	95% CI/Notable effect sizes	Overall inference drawn
Ahmed et al.[15]	2025	Retrospective cohort (N = 266)	External validation/nomogram of **S.T.O.N.E.** score for semirigid pneumatic URS; no external comparator	Logistic regression	S.T.O.N.E. components: size, topography, obstruction (hydronephrosis), number, HU	No fragments or ≤2 mm on endoscopy/fluoro or imaging (US/X-ray/CT) within 1 month	0.785	85%	72%	81%	75%	83%	HU 1.01 (1.0023–1.0065); proximal 15.13 (1.52–51.13); hydro-moderate 34.23 (8.28–141.45), severe 33.75 (4.55–250.36); size 1.51 (1.30–1.75)	Validated S.T.O.N.E.-based prediction for semirigid URS with solid discrimination; hydronephrosis, proximal location, HU and size were dominant risk factors; clinically deployable nomogram.
Kim et al.[16]	2014	Retrospective (n = 237)	CT-based predictive model for URSL outcome; multivariable LR; no external comparator	Logistic regression	Stone diameter, estimated stone location, tissue rim sign; also VSB, HU, hydronephrosis, perinephric edema, demographics	No stones or fragments <2 mm by URSL + plain radiograph POD1	0.825	NR	84.8%	69.3%	NR	NR	ORs: diameter 1.074 (1.029–1.121); location 1.165 (1.048–1.296); rim sign >2 mm 4.635 (1.463–14.678)	CT-derived LR showed good discrimination for URSL SFR; larger size, proximal location, and rim sign >2 mm predicted failure; lacked external validation and used early imaging endpoint.

(Continued)

TABLE 2. **Continued**

Study name	Year	Study type	Intervention (AI/model, inputs, comparator)	Algorithm	Key input variables/features	Stone-free definition (SFS criteria)	AUC	Accuracy	Sens	Spec	PPV	NPV	95% CI/Notable effect sizes	Overall inference drawn
Lee et al.[17]	2025	Retrospective two-center; internal CV + external validation	ML prediction of SFR after **RIRS** for lower-pole stones	LightGBM (best)	Age, sex, BMI, DM/HTN; stone side/number/multifocality; stone burden; HU; pelvic splanchnic angle (PSA); renal infundibulopelvic length (RIL); infundibular width (IW)	No visible stones or ≤5 mm residuals on CT at 1 month	0.759 (external val.)	77.1%	88.9%	NR	82.8%	NR	SHAP/gain: stone burden, HU, PSA, RIL dominant	Acceptable cross-center generalization; anatomical/stone-burden metrics drive predictions for RIRS; scope for calibration and broader validation.
Nedbal et al.[18]	2024	Retrospective single-surgeon, single-center (n = 872)	16 ML models for **fURSL** outcomes; ensemble (CatBoost+Bagging+RF); multitask ANN	CatBoost (best); ensemble; ANN	Age, sex, urine culture, anatomic variants, stone location, multiplicity, total stone burden, pre-op stent/nephrostomy	Fragments <2 mm endoscopically and no >2 mm on KUB/US at 3 months	NR	93%	65% (recall)	NR	87% (precision)	NR	Logistic model weights: burden 0.34; pre-op stent 0.106; urine culture 0.14; negative weight for location 0.09	High overall accuracy with strong precision but modest recall; stone burden and pre-op factors influential; single-center design limits generalizability.
Xun et al.[19]	2020	Retrospective single-center; dev + internal validation (lower calyx cohort)	Clinical–radiomics nomogram for **fURS** SFR (radiomics + clinical factors)	LASSO + logistic regression	NCCT radiomics (604→28), stone volume (≤1 vs. >1 cm^3), hydronephrosis, operator experience (fURS ≥100)	No stones or fragments <2 mm on CT/KUB at 3 months	0.949 (derivation); 0.947 (internal val.)	NR	NR	NR	NR	NR	ORs: hydro 9.908 (2.224–44.132); volume ≤1 cm^3 8.337 (1.309–53.106); fURS ≥100 11.169 (2.095–59.550); radiomics score 2.679 (1.395–5.146)	Radiomics-augmented model achieved excellent internal performance for fURS SFR; highlights impact of stone volume, hydronephrosis, and surgeon experience; needs external validation.

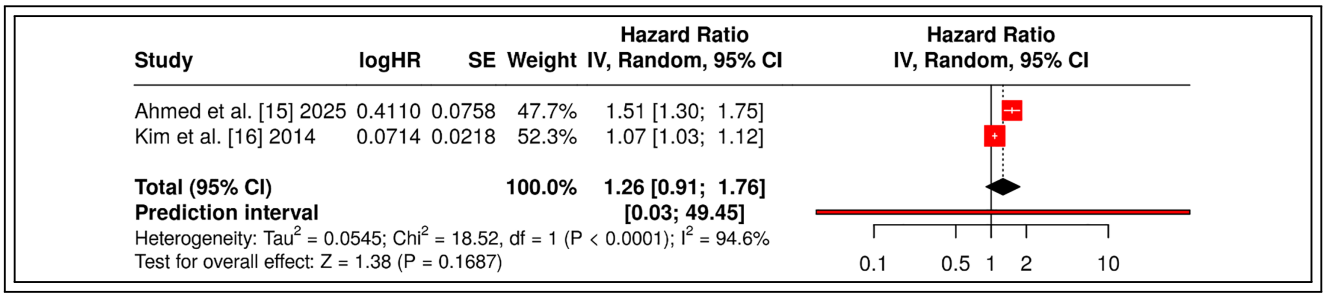

FIGURE 3. Effect estimates (HR/OR/RR) with 95% CI (Abbreviations: **HR** = hazard ratio; **logHR** = logarithm of hazard ratio; **SE** = standard error; **IV** = inverse variance; **CI** = confidence interval; **Tau² (τ^2)** = between-study variance; **Chi² (χ^2)** = Cochran's Q statistic; **df** = degrees of freedom; **I²** = percentage heterogeneity; **Z** = Z statistic for overall effect; **P** = *p*-value)[15,16]

Statistical analysis observations

A doubling of each millimeter increase in stone size (Figure 3) was associated with an increased risk for a non–stone-free result and thus indicated a strong signal present to AI models (RE ratio 1.26, 95% CI 0.91–1.76; $\tau^2 \approx 0.055$; Q = 18.52, $p < 0.0001$; I^2 = 94.6%).[15–16] The effect was large when estimated using the S.T.O.N.E.–based logistic model (1.51, 95% CI 1.30–1.75),[15] while the CT-feature based logistic model signaled a smaller but consistently directional effect (1.07, 95% CI 1.03–1.12).[16] The large prediction interval (0.03–49.45) indicated limitations for transportability for models and emphasized a need for recalibration/domain adaptation when applying size-driven algorithms.

Hydronephrosis provided a high-magnitude, clinically intuitive feature for machine prediction (study-level RRs 2.79, 1.28, 5.84),[15,16,19] but the pooled estimate was imprecise (RR 2.72, 95% CI 0.96–7.72; $\tau^2 \approx 0.821$; Q = 65.40, $p < 0.0001$; I^2 = 96.9%; prediction interval 0.03–249.87) (Figure 4). This dispersion reflected label heterogeneity (e.g., "moderate+severe" vs. "severe only"), indicating that AI pipelines required harmonized obstruction grading and/or probabilistic severity encodings to maintain generalization.

Stone size when modeled continuously revealed a very large standardized separation by No-SFR and SFR (SMD 1.36, 95% CI 0.85–1.86; $\tau^2 \approx 0.096$; I^2 = 72.9%; prediction interval −3.77 to 6.48), suggesting that linear and non-linear learners would equally take advantage of strong margin information, but heterogeneity warned site-specific thresholds would need to be tuned[15–16] (Figure 5).

HU demonstrated moderate, directionally consistent separation (SMD 0.64, 95% CI 0.39–0.90; $\tau^2 \approx 0$; Q = 0.73, $p = 0.394$; I^2 = 0%; prediction interval −0.99 to 2.27) based on the contributing datasets[15–16] (Figure 6). This stability suggested HU was a portable predictor that could enhance robustness when combined with burden/topography features, aligning with feature-importance signals from gradient-boosted and radiomics-augmented models.[17,19]

Across AI/predictive approaches (Figure 7), discrimination and operating performance varied meaningfully. The radiomics-augmented logistic model exhibited the highest discrimination (AUC 94.7%) but lacked paired accuracy/recall reporting, indicating strong separability in internal validation without a fixed threshold. The ensemble pipeline achieved the highest observed accuracy (93.0%) but operated at a lower recall (65.0%), consistent with a precision-oriented decision threshold. The gradient-boosted model prioritized recall (sensitivity 88.9%) with moderate discrimination (AUC 75.9%) and accuracy (77.1%). Score-based logistic regression delivered balanced performance (AUC 78.5%, accuracy 85.0%, sensitivity 72.0%), and CT-feature logistic regression showed high sensitivity (84.8%) at the expense of specificity (heatmap), reflecting a more liberal decision rule.

The cross-metric pattern accommodated two working archetypes: (i) high precision/accuracy with reduced recall (ensemble: 93.0% accuracy, 87.0% precision), and (ii) high recall with decent accuracy (boosting: 88.9% sensitivity, 77.1% accuracy) (Figure 8). The score-based approach yielded acceptable specificity (81.0%) combined with decent PPV (75.0%) and only reported NPV (83.0%), while CT-based LR exchanged specificity (69.3%) for sensitivity (84.8%) but at reduced comparability when combined with score thresholding. NA cells constrained full cross-study comparability but showed through the matrix how thresholding and

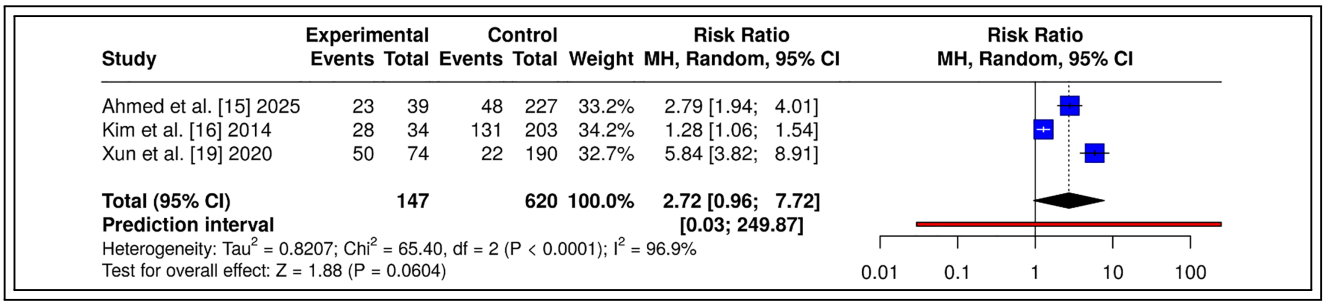

FIGURE 4. Hydronephrosis (moderate–severe vs. mild/none; Notes: Kim et al.[16] grades 2–3 counted as moderate–severe; Ahmed et al.[15] "moderate" + "severe" grouped as exposed; Xun et al.[19] reported "severe" vs. "no/mild"; Abbreviations: **RR** = risk ratio; **MH** = Mantel–Haenszel method; **CI** = confidence interval; **Tau² (τ^2)** = between-study variance; **Chi² (χ^2)** = Cochran's Q statistic; **df** = degrees of freedom; **I²** = percentage heterogeneity; **Z** = Z statistic for overall effect; **P** = *p*-value)[15,16,19]

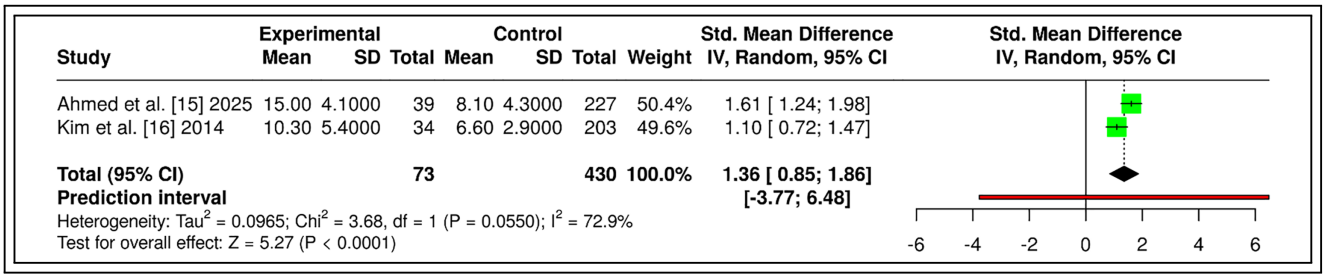

FIGURE 5. Stone size (Abbreviations: **SMD** = standardized mean difference; **SD** = standard deviation; **IV** = inverse variance; **CI** = confidence interval; **Tau² (τ^2)** = between-study variance; **Chi² (χ^2)** = Cochran's Q statistic; **df** = degrees of freedom; **I²** = percentage heterogeneity; **Z** = Z statistic for overall effect; **P** = *p*-value[15,16]

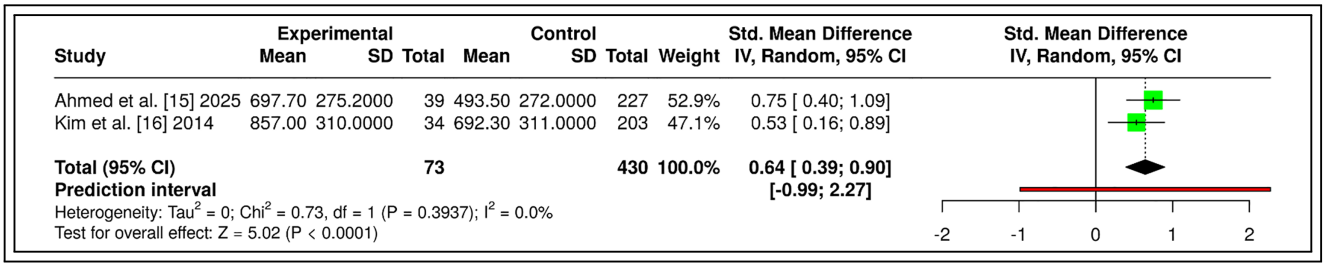

FIGURE 6. Stone density (Abbreviations: **SMD** = standardized mean difference; **SD** = standard deviation; **IV** = inverse variance; **CI** = confidence interval; **Tau² (τ^2)** = between-study variance; **Chi² (χ^2)** = Cochran's Q statistic; **df** = degrees of freedom; **I²** = percentage heterogeneity; **Z** = Z statistic for overall effect; **P** = *p*-value; **HU** = Hounsfield units)[15,16]

model family changed the trade-off in terms of precision and recall—due consideration for AI decision support deployment.

Of inputs to AI models (Figure 9), tissue rim sign ≥2 revealed a large single-study odds ratio with non–stone-free outcome (OR 6.60, 95% CI 2.24–19.41), hydronephrosis transmitted a large but imprecise pooled risk (RR 2.72, 95% CI 0.96–7.72), stone size revealed a large standardized separation (SMD 1.36, 95% CI 0.85–1.86), and Hounsfield units revealed a consistent, moderate signal (SMD 0.64, 95% CI 0.39–0.90). Altogether, these effects suggested that models based on CT-derived burden, obstruction, and composition variables

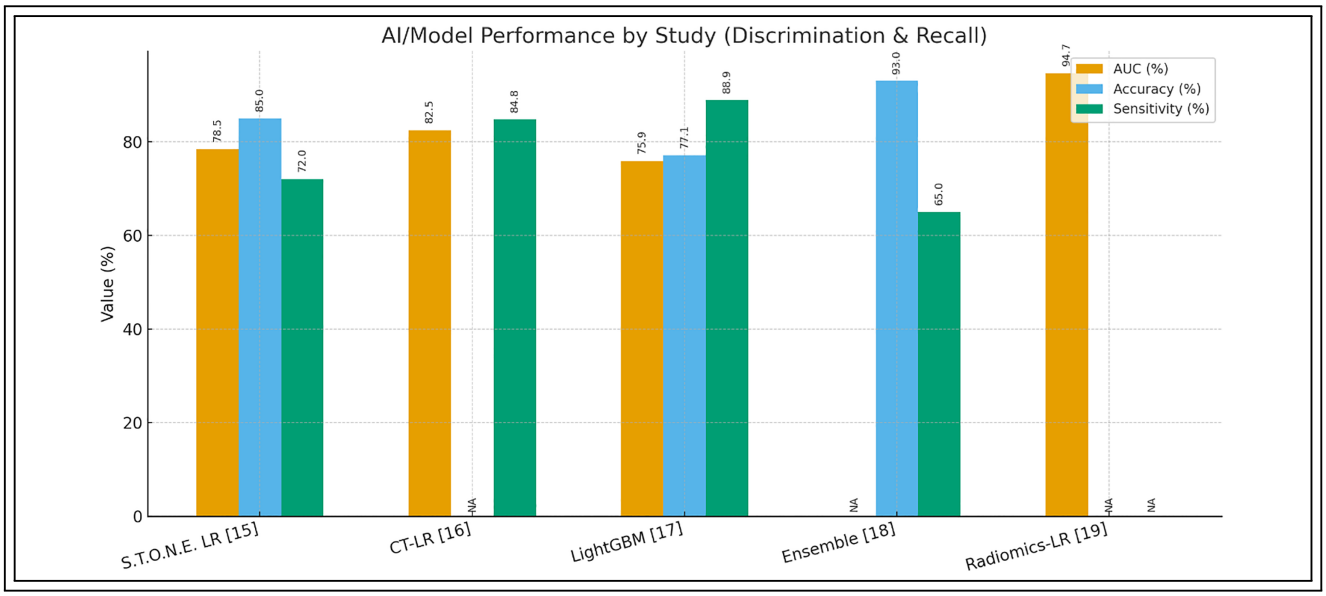

FIGURE 7. AI/Model Performance by Study (Abbreviations: **AI** = artificial intelligence; **AUC** = area under the receiver operating characteristic curve; **LR** = logistic regression; **CT-LR** = computed tomography–based logistic regression; **LightGBM** = Light Gradient Boosting Machine; **NA** = not available/not reported)[15–19]

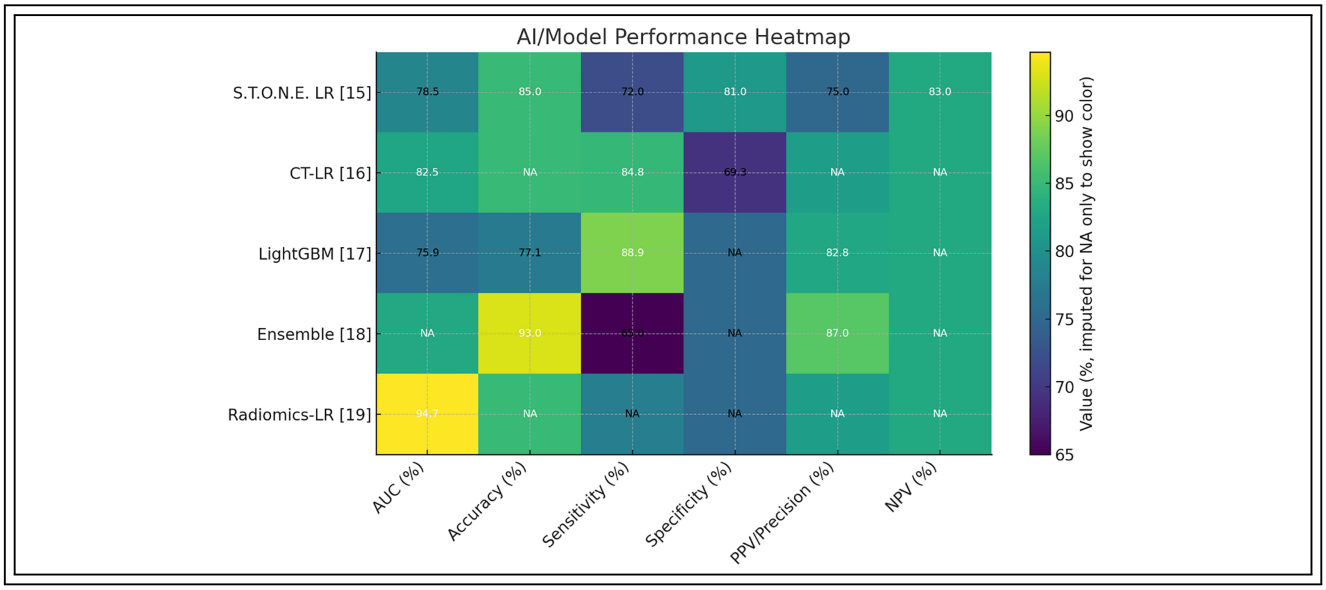

FIGURE 8. AI/Model performance heatmap (Abbreviations: **AI** = artificial intelligence; **AUC** = area under the receiver operating characteristic curve; **LR** = logistic regression; **CT-LR** = computed tomography–based logistic regression; **LightGBM** = Light Gradient Boosting Machine; **PPV** = positive predictive value; **NPV** = negative predictive value; **NA** = not available/not reported)[15–19]

would retain most predictive signal, while variability between studies (wide CIs for hydronephrosis) suggested a requirement for calibration and harmonized definitions when translating AI models to other sites.

Discussion

Endourologic AI began with proof-of-concept classifiers and extended to interacting with imaging and scope navigation to and beyond peri-operative

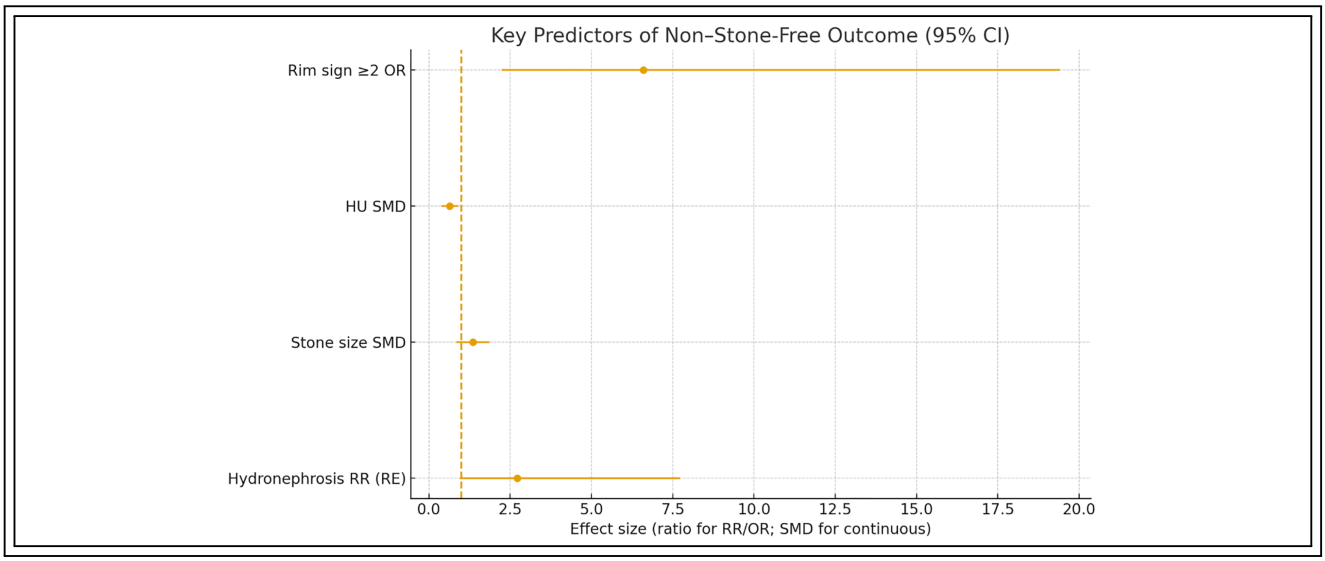

FIGURE 9. Key predictors of non–stone-free outcome (Abbreviations: **OR** = odds ratio; **RR** = risk ratio; **RE** = random-effects model; **SMD** = standardized mean difference; **HU** = Hounsfield units; **CI** = confidence interval)

decision support through the entire continuum of stone care. Cross-pollination from ancillary urologic specialties has been useful: breakthroughs with diagnostic pipelines for upper tract urothelial carcinoma illustrated how multimodal data ingestion, feature harmonization, and prospective validation can be translated to anatomically constrained endoscopy and transferable lessons to ureteroscopic prediction and guidance.[20] Contemporary stone-management roadmaps substantiated technology adoption being grounded in relevance to outcomes, resource stewardship, and reproducibility and positioning AI as an ancillary adjunct to complement tried-and-trusted care pathways and not usurp them.[21–23] Narrative reviews chronicling AI's emergence to prominence in endourology emphasized transparent reporting, external validation, and transfer of calibration as prerequisites to transferrable prediction with heterogeneous patient populations and imaging protocols.[22–25]

This review defined performance parameters for artificial intelligence and predictive models of SFS after ureteroscopy and defined a reproducible set of clinically interpretable predictors—stone size/burden, location/topography, hydronephrosis, and Hounsfield units (HU)—that enabled parsimonious models suitable for near-term clinical application. The findings showed that radiomic features and anatomic attributes improved discrimination capacities compared with those afforded by customary attributes, whereas heterogeneity in endpoint definitions and operating thresholds between studies influenced observed recall and calibration.

Device and workflow advances have widened the substrate upon which AI can work. Iterative refinements to stone-removal procedures produced stable procedural end-points supporting supervised learning and benchmarking by center.[23,26–28] Concurrently, robotic ureteroscopy has reached maturity in studies of feasibility, raising questions regarding how algorithmic support can augment teleoperation, haptics, and micro-manipulation by constrained kinematics.[8,24,29] Visual phenotypes from endoscopic papillary observations and stone-surface recognition defined a semantic lexicon to support intraoperative scene interpretation with linking to phenotypes seen by vision, composition proxies, and clearable likelihood by AI models.[25,30] Motion-aware semantic segmentation for ureteroscopy and lithotripsy demonstrated temporal cues to enhance instrument–tissue separation and formed a basis for context-aware safety interlocks and guidance.[26,31] More extensive experience with robotics for stone disease identified further applications for shared control and autonomy layers amenable to auditing and calibration to clinical concerns like safety margins and dwell time by clinical priority.[27–33]

Preoperative risk stratification has also advanced with deployment of machine-learning nomograms to detect obstructed ureteral stones and drive access

strategy planning, choice of energy, and ancillary planning accordingly. These methods have described the ability to integrate tabulated clinical and imaging descriptors with anatomy-aware variables to limit decision-making pathways prior to intervention.[34] At a research design level, trial design innovations are key: endourology trials have been encouraged to perform prospective cohorts, standardized outcomes and registries to prevent spectrum bias and allow direct comparative evaluations of artificial intelligence vs. recommended risk scoring tools.[35] Concurrent education reforms identified competency-based education and simulation, generating formatted data streams suitable for model building as well as offering performance feedback to surgeons.[36–42] The implementation of single-use flexible ureteroscopes has sparked discussion regarding procurement and viability but has eliminated variability related to optical elements and sensor noise and thereby enhanced consistency of video inputs to computer vision application.[31]

Beyond the surgical suite, conversational robots have evolved scalable models for symptom tracking, adherence support, and complication triage following endourologic surgery; integrating predictive outputs can synchronize thresholds with patient-reported outcomes and minimize attrition during follow-up stages.[32] Robotic flexible ureteroscopy presents a novel challenge to workflow implementation, whereby artificial intelligence can potentially facilitate cooperation between camera and laser systems and automate centering while offering stable views irrespective of fluctuating irrigation conditions.[43] The implementation of intricate reconstructions involving robot-assisted pyeloplasty with retrograde flexible ureteroscopy has illustrated that multi-team cooperation with plural teams and cross-modality navigation are possible targets for AI-augmented schedule planning, guidance, and logistical management of instrumentation.[44]

Augmented intraoperative cognition is developing from a number of vantage points. Mixed-reality overlays for urology defined registration accuracy requirements, latency budgets, and human-factors design, all of which limit the utility of AI-predicted guidance within a changing endoscopic field.[45] Editorial opinions regarding minimally invasive evolution contended that innovation cycles needed to be accompanied by implementation science and stiff post-market surveillance, a dictum that maps directly onto learning systems prone to drift and domain shift.[46–51] Fusion-based visual-electromagnetic localization provided a roadmap to robust monocular tracking, reconstruction, and metrically scaled measurement to overcome the decades-old hurdle of spatial context during flexible ureteroscopy.[37] More generally, robotics and intraoperative navigation reviews all converged upon modularity and standardized interfaces and upon verification of autonomy levels as prerequisites to safe AI augmentation of endoscopy.[38]

Studies involving automated recognition of urinary stones has revealed trade-offs involved with hand-crafted descriptors, deep features, and domain adaptation to inform composition inference. This work suggests computer vision has the ability to synthesize pre-operative imaging with texture cues at intraoperative time-points to improve real-time prediction for clearance.[52–55] In further extension, risk modeling has begun to include adverse outcomes with machine learning instruments constructed for post-RIRS infection, which showed promise in synthesizing pre-, intra-, and early post-operative variables to actionable alerting and antibiotic stewardship pathways.[40] Lastly, automated composite efficiency scores based on simulated ureteroscopic tasks imply a quantitative and scalable approach to measuring technical performance and hence feedback loops whereby operational metrics inform and improve predictive models and an iterative reverse process.[41]

Limitations

The meta-analysis combined retrospective cohorts alone and faced heterogeneous SFS definitions and follow-up points, heightening statistical heterogeneity and restricted comparability. The performance metrics and calibration reporting was scant in some datasets, excluding common pooling (e.g., AUC not necessarily being present). The majority of models were single-center and thus prompted concerns regarding spectrum biases and restricted external generalizability. The restricted number of qualified studies restricted small-study effect and meta-regression evaluation and the broad prediction intervals reflected doubt regarding transportability to other settings.

Clinical recommendations

Given the marked heterogeneity that limits the interpretability of pooled effects, implementation guidance should be treated as provisional and grounded in variables that show consistent qualitative importance across settings rather than relying on pooled quantitative magnitudes. For ureteroscopy, clinical deployment should preferentially begin with parsimonious, interpretable clinical–CT models, using stone size/burden,

location/topography, hydronephrosis, and Hounsfield units as mandatory baseline inputs because these features most consistently carried discriminative signal across studies. Radiomics augmentation should be considered only when acquisition, reconstruction, and feature-extraction pipelines are standardized, version-frozen, and audited, and should be added on top of (not substituted for) core clinical–CT variables.

Model choice and thresholding should remain use-case specific: precision-oriented operating points (higher PPV at lower recall) may fit confirmation-type decisions (e.g., identifying low-likelihood failure before discharge), whereas recall-oriented operating points (higher sensitivity at moderate accuracy) may fit pre-operative risk flagging and resource planning. Threshold selection should be supported by decision-curve analysis, followed by site-level calibration and domain adaptation, explicitly acknowledging that broad prediction intervals may signal transportability limits. Obstruction severity should be harmonized using common grading conventions and time windows and modeled, where feasible, as ordinal/probabilistic inputs rather than binary categories, because hydronephrosis showed large but imprecise contributions under heterogeneous definitions.

To reduce between-study variance and enable future clinically meaningful pooling, centers should standardize stone-free status definitions (e.g., $\leq$2 mm vs. $\leq$5 mm), adopt fixed follow-up windows (e.g., 1 vs. 3 months), and align a primary imaging modality for endpoint determination. Reports should present full operating profiles (AUC plus accuracy, sensitivity, specificity, PPV/NPV) alongside calibration (intercept/slope), Brier score, and confusion matrices at the chosen threshold. Internal validation must be leakage-safe with transparent handling of class imbalance and missingness, and external validation should include pre-specified recalibration (e.g., intercept/scale updating) before performance claims. For adoption, a tiered trajectory remains most defensible: (i) deploy a compact clinical–CT model as the maintainable reference; (ii) permit a radiomics-enhanced tier only under strict protocol control; and (iii) monitor drift post-deployment with scheduled recalibration and threshold retuning. Finally, shared data dictionaries, harmonized obstruction grading, and CT intensity normalization are required to stabilize performance across sites and support methodologically consistent future meta-analyses.

Conclusion

In heterogeneous, multi-site ureteroscopy cohorts with variable definitions of stone-free status and follow-up timing, AI and predictive models frequently demonstrated discrimination in the AUC ~0.8 range, with some studies reporting $\geq$0.9 when radiomics features and proxies of operator experience were incorporated. Nevertheless, the extent of heterogeneity across study populations, endpoints, imaging workflows, and analytic thresholds renders pooled quantitative summaries clinically uninterpretable, and the quantitative synthesis should be viewed as exploratory only. Therefore, the most defensible inference is that predictive parsimony is often achievable: stone size, location/topography, hydronephrosis, and Hounsfield units repeatedly emerged as clinically interpretable predictors supporting preoperative risk stratification. Single-center reports with high accuracy but modest recall likely reflect thresholding and case-mix differences and should not be assumed transportable without external validation, calibration transfer, and explicit domain adaptation. Standardizing stone-free thresholds and follow-up windows is likely to reduce variance and may permit more clinically meaningful pooled estimation in future studies.

Acknowledgement

The author would like to express sincere gratitude to the Department of Surgery at King Abdulaziz Medical City for their continuous support and encouragement throughout the development of this manuscript. The department's commitment to academic excellence, clinical training, and research has provided an outstanding environment that greatly facilitated the completion of this work. Their dedication to advancing surgical knowledge and patient care is deeply appreciated.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Yahya Ghazwani contributed to the study conception and design, literature screening, data extraction, and manuscript drafting. Mohammad Alghafees contributed to study design, supervision of the project, data interpretation, statistical analysis, and critical revision of the manuscript. Mishari Alshasha and Fahad Brayan contributed to data extraction, quality assessment, and drafting of the methods section. Abdulrahman Alsayyari assisted in data analysis, figure preparation, and manuscript editing. Ali Alyami contributed to study supervision, methodological oversight, and final critical review of the manuscript. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials

The data that support the findings of this study are available from the Corresponding Author, MG, upon reasonable request.

Ethics Approval

This study was conducted in accordance with the principles of the Declaration of Helsinki. As this work represents a systematic review and meta-analysis of previously published studies, no new human participants were enrolled, no identifiable personal data were accessed, and no direct patient intervention was performed. Therefore, institutional review board (IRB) approval and informed consent were not required.

Conflicts of Interest

The authors declare no conflicts of interest.

Supplementary Materials

The supplementary material is available online at https://www.techscience.com/doi/10.32604/cju.2026.077411/s1.

References

1. Poulakis V, Dahm P, Witzsch U, de Vries R, Remplik J, Becht E. Prediction of lower pole stone clearance after shock wave lithotripsy using an artificial neural network. *J Urol* 2003;169(4):1250–1256. doi:10.1097/01.ju.0000055624.65386.b9.
2. Yang SW, Hyon YK, Na HS et al. Machine learning prediction of stone-free success in patients with urinary stone after treatment of shock wave lithotripsy. *BMC Urol* 2020;20(1):88. doi:10.1186/s12894-020-00662-x.
3. Zhao H, Li W, Li J, Li L, Wang H, Guo J. Predicting the stone-free status of percutaneous nephrolithotomy with the machine learning system: comparative analysis with guy's stone score and the S.T.O.N.E score system. *Front Mol Biosci* 2022;9:880291. doi:10.3389/fmolb.2022.880291.
4. AlAzab R, Ghammaz O, Ardah N et al. Predicting the stone-free status of percutaneous nephrolithotomy with the machine learning system. *Int J Nephrol Renovasc Dis* 2023;16:197–206. doi:10.2147/IJNRD.S427404.
5. Choo MS, Uhmn S, Kim JK et al. A prediction model using machine learning algorithm for assessing stone-free status after single session shock wave lithotripsy to treat ureteral stones. *J Urol* 2018;200(6):1371–1377. doi:10.1016/j.juro.2018.06.077.
6. Liu ZR, Yu ZJ, Zhou J, Huang JB. Predictive value of the stone-free rate after percutaneous nephrolithotomy based on multiple machine learning models. *Front Med* 2025;12:1559613. doi:10.3389/fmed.2025.1559613.
7. Noble PA, Hamilton BD, Gerber G. Stone decision engine accurately predicts stone removal and treatment complications for shock wave lithotripsy and laser ureterorenoscopy patients. *PLoS One* 2024;19(5):e0301812. doi:10.1371/journal.pone.0301812.
8. Yang B, Veneziano D, Somani BK. Artificial intelligence in the diagnosis, treatment and prevention of urinary stones. *Curr Opin Urol* 2020;30(6):782–787. doi:10.1097/MOU.0000000000000820.
9. Tano ZE, Cumpanas AD, Gorgen ARH, Rojhani A, Altamirano-Villarroel J, Landman J. Surgical artificial intelligence: endourology. *Urol Clin North Am* 2024;51(1):77–89. doi:10.1016/j.ucl.2023.06.004.
10. Alves BM, Belkovsky M, Passerotti CC et al. Use of artificial intelligence for sepsis risk prediction after flexible ureteroscopy: a systematic review. *Rev Col Bras Cir* 2023;50(9):e20233561. doi:10.1590/0100-6991e-20233561-en.
11. Balasubramanian A, Bhambhvani H, Lee J, Shah O. Artificial intelligence and machine learning for stone management. *Urol Clin North Am* 2025;52(3):465–474. doi:10.1016/j.ucl.2025.04.011.
12. Page MJ, Moher D, Bossuyt PM et al. PRISMA, 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021;372:n160. doi:10.1136/bmj.n160.
13. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the quality assessment of diagnostic accuracy studies using AI (QUADAS-AI): protocol for a qualitative study. *JMIR Res Protoc* 2024;13:e58202. doi:10.2196/58202.
14. Fekete JT, Győrffy B. MetaAnalysisOnline.com: web-based tool for the rapid meta-analysis of clinical and epidemiological studies. *J Med Internet Res* 2025;27(4):e64016. doi:10.2196/64016.
15. Ahmed F, Al-Kohlany K, Al-Naggar K, Alnadhari I, Altam AY, Badheeb M. Assessing the predictive accuracy of the S.T.O.N.E. score for stone-free rates in semirigid pneumatic ureteral lithotripsy: implications for validation. *Res Rep Urol* 2025;17:139–152. doi:10.2147/RRU.S515846.

16. Kim JW, Chae JY, Kim JW et al. Computed tomography-based novel prediction model for the stone-free rate of ureteroscopic lithotripsy. *Urolithiasis* 2014;42(1):75–79. doi:10.1007/s00240-013-0609-0.
17. Lee HY, Tung YH, Elises JC, Wang YC, Gauhar V, Cho SY. Machine learning-based prediction of stone-free rate after retrograde intrarenal surgery for lower pole renal stones. *World J Urol* 2025;43(1):433. doi:10.1007/s00345-025-05762-7.
18. Nedbal C, Adithya S, Naik N, Gite S, Juliebø-Jones P, Somani BK. Can machine learning correctly predict outcomes of flexible ureteroscopy with laser lithotripsy for kidney stone disease? results from a large endourology university centre. *Eur Urol Open Sci* 2024;64(Suppl 1):30–37. doi:10.1016/j.euros.2024.05.004.
19. Xun Y, Chen M, Liang P et al. A novel clinical-radiomics model pre-operatively predicted the stone-free rate of flexible ureteroscopy strategy in kidney stone patients. *Front Med* 2020;7:576925. doi:10.3389/fmed.2020.576925.
20. Kostakopoulos N, Argyropoulos V, Bellos T, Katsimperis S, Kostakopoulos A. Artificial intelligence and novel technologies for the diagnosis of upper tract urothelial carcinoma. *Medicina* 2025;61(5):923. doi:10.3390/medicina61050923.
21. Papatsoris A, Alba AB, Galán Llopis JA et al. Management of urinary stones: state of the art and future perspectives by experts in stone disease. *Arch Ital Urol Androl* 2024;96(2):12703. doi:10.4081/aiua.2024.12703.
22. Zeeshan Hameed BM, Shah M, Naik N et al. The ascent of artificial intelligence in endourology: a systematic review over the last 2 decades. *Curr Urol Rep* 2021;22(10):53. doi:10.1007/s11934-021-01069-3.
23. Tzelves L, Geraghty RM, Hughes T, Juliebø-Jones P, Somani BK. Innovations in kidney stone removal. *Res Rep Urol* 2023;15:131–139. doi:10.2147/RRU.S386844.
24. Sinha MM, Gauhar V, Tzelves L et al. Technical aspects and clinical outcomes of robotic ureteroscopy: is it ready for prime-time? *Curr Urol Rep* 2023;24(8):391–400. doi:10.1007/s11934-023-01167-4.
25. Almeras C, Pradere B, Estrade V, Meria P, On behalf of the lithiasis committee of the french urological association. Endoscopic papillary abnormalities and stone recognition (EPSR) during flexible ureteroscopy: a comprehensive review. *J Clin Med* 2021;10(13):2888. doi:10.3390/jcm10132888.
26. Gupta S, Ali S, Goldsmith L, Turney B, Rittscher J. Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy. *Comput Med Imaging Graph* 2022;101(9):102112. doi:10.1016/j.compmedimag.2022.102112.
27. Hasan O, Reed A, Shahait M, Crivellaro S, Dobbs RW. Robotic surgery for stone disease. *Curr Urol Rep* 2023;24(3):127–133. doi:10.1007/s11934-022-01131-8.
28. Qi Y, Yang S, Li J et al. Development and validation of a nomogram to predict impacted ureteral stones via machine learning. *Minerva Urol Nephrol* 2024;76(6):736–747. doi:10.23736/S2724-6051.24.05856-7.
29. Donaldson JF, McClinton S. Evidence and clinical trials in endourology: where are we going. *Curr Opin Urol* 2021;31(2):120–124. doi:10.1097/MOU.0000000000000851.
30. Antoniou V, Gauhar V, Kallidonis P et al. Education and training evolution in urolithiasis: a perspective from European School of Urology. *Asian J Urol* 2023;10(3):281–288. doi:10.1016/j.ajur.2023.01.004.
31. Bahaee J, Plott J, Ghani KR. Single-use flexible ureteroscopes: how to choose and what is around the corner? *Curr Opin Urol* 2021;31(2):87–94. doi:10.1097/MOU.0000000000000852.
32. Geoghegan L, Scarborough A, Wormald JCR et al. Automated conversational agents for post-intervention follow-up: a systematic review. *BJS Open* 2021;5(4):zrab070. doi:10.1093/bjsopen/zrab070.
33. Lee JY, Jeon SH. Robotic flexible ureteroscopy: a new challenge in endourology. *Investig Clin Urol* 2022;63(5):483–485. doi:10.4111/icu.20220256.
34. Krings G, Ayoub E, Campi R et al. Ureteropelvic junction obstruction and renal calculi: simultaneous treatment by robot-assisted laparoscopic pyeloplasty and transcutaneous retrograde flexible ureteroscopy. Technique description and early outcomes. *Prog Urol* 2023;33(5):279–284. doi:10.1016/j.purol.2023.01.006.
35. Reis G, Yilmaz M, Rambach J et al. Mixed reality applications in urology: requirements and future potential. *Ann Med Surg* 2021;66:102394. doi:10.1016/j.amsu.2021.102394.
36. Tzelves L, Juliebø-Jones P, Somani B. Editorial: the evolution of minimally invasive urologic surgery: innovations, challenges, and opportunities. *Front Surg* 2024;11:1525713. doi:10.3389/fsurg.2024.1525713.
37. Fu Z, Jin Z, Zhang C et al. Visual-electromagnetic system: a novel fusion-based monocular localization, reconstruction, and measurement for flexible ureteroscopy. *Int J Med Robot* 2021;17(4):e2274. doi:10.1002/rcs.2274.
38. Schoeb DS, Rassweiler J, Sigle A et al. Robotics and intraoperative navigation. *Urologe A* 2021;60(1):27–38. doi:10.1007/s00120-020-01405-4.
39. El Beze J, Mazeaud C, Daul C et al. Evaluation and understanding of automated urinary stone recognition methods. *BJU Int* 2022;130(6):786–798. doi:10.1111/bju.15767.
40. Castellani D, De Stefano V, Brocca C et al. The infection post flexible UreteroreNoscopy (I-FUN) predictive model based on machine learning: a new clinical tool to assess the risk of sepsis post retrograde intrarenal surgery for kidney stone disease. *World J Urol* 2024;42(1):612. doi:10.1007/s00345-024-05314-5.
41. Valovska MT, Yang J, Chen N, Yu D, Wollin DA. Development of an automated composite ureteroscopic efficiency score through simulated ureteroscopic skills assessment. *J Endourol* 2023;37(8):956–964. doi:10.1089/end.2022.0820.
42. Resorlu B, Unsal A, Gulec H, Oztuna D. A new scoring system for predicting stone-free rate after retrograde intrarenal surgery: the resorlu-unsal stone score. *Urology* 2012;80(3):512–518. doi:10.1016/j.urology.2012.02.072.
43. Sfoungaristos S, Gofrit ON, Mykoniatis I et al. External validation of Resorlu-Unsal stone score as predictor of outcomes after retrograde intrarenal surgery. *Int Urol Nephrol* 2016;48(8):1247–1252. doi:10.1007/s11255-016-1311-2.
44. Xiao Y, Li D, Chen L et al. The R.I.R.S. scoring system: an innovative scoring system for predicting stone-free rate following retrograde intrarenal surgery. *BMC Urol* 2017;17(1):105. doi:10.1186/s12894-017-0297-0.
45. Wang C, Wang S, Wang X, Lu J. External validation of the R.I.R.S. scoring system to predict stone-free rate after retrograde intrarenal surgery. *BMC Urol* 2021;21(1):33. doi:10.1186/s12894-021-00801-y.
46. Jung JW, Lee BK, Park YH et al. Modified seoul national university renal stone complexity score for retrograde intrarenal surgery. *Urolithiasis* 2014;42(4):335–340. doi:10.1007/s00240-014-0650-7.

47. Özman O, Başataç C, Akgül HM et al. External validation of modified seoul national university renal stone complexity score to predict outcome and complications of retrograde intrarenal surgery: a RIRSearch group study. *Minim Invasive Ther Allied Technol* 2022;31(6):917–922. doi:10.1080/13645706.2021.2025112.
48. Elmohamady BN, Farag MM, Sherif HW, El Ghobashy A, Al Hefnawy MA. Predicting stone free rate after retrograde intrarenal surgery using RIRS scoring system versus Resorlu Unsal stone score (RUSS). *Arab J Urol* 2023;22(2):102–108. doi:10.1080/20905998.2023.2252227.
49. Selmi V, Sari S, Oztekin U, Caniklioglu M, Isikay L. External validation and comparison of nephrolithometric scoring systems predicting outcomes of retrograde intrarenal surgery. *J Endourol* 2021;35(6):781–788. doi:10.1089/end.2020.0491.
50. Molina WR, Kim FJ, Spendlove J et al. Score: a new assessment tool to predict stone free rates in ureteroscopy from pre-operative radiological features. *Int Braz J Urol* 2014;40(1):23–29. doi:10.1590/s1677-5538.ibju.2014.01.04.
51. Zhang Y, Xie Z, Wu L et al. A nomogram for predicting the risk of residual stone fragments after ureteroscopy. *Transl Androl Urol* 2023;12(3):364–374. doi:10.21037/tau-22-609.
52. Polat S, Danacioglu YO, Yarimoglu S et al. External validation of the current scoring systems and derivation of a novel scoring system to predict stone free rates after retrograde intrarenal surgery in patients with cumulative stone diameter of 2-4 cm. *Actas Urol Esp* 2023;47(4):211–220. doi:10.1016/j.acuroe.2022.08.015.
53. Tung YH, Li WM, Juan YS et al. New infundibulopelvic angle measurement method can predict stone-free rates following retrograde intrarenal surgery. *Sci Rep* 2024;14(1):9891. doi:10.1038/s41598-024-60248-7.
54. Elises JC, Parikesit D, Zogan M et al. Validating pelvic stone angle as a substitute for infundibulopelvic angle in RIRS outcome prediction. *J Formos Med Assoc* 2025;33(11):1033. doi:10.1016/j.jfma.2025.12.011.
55. Zhang Z, Chen L, Niu K et al. A novel nomogram for predicting stone-free status after retrograde intrarenal surgery in patients with kidney stones. *BMC Urol* 2025;26(1):12. doi:10.1186/s12894-025-02022-z.