Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction models for sepsis-associated acute kidney injury and mortality in adult intensive care: a systematic review and stratified meta-analysis

Jing Yang
Tongde Hospital of Zhejiang Province
Mingfang Qian
Tongde Hospital of Zhejiang Province
Jitao Zou
Tongde Hospital of Zhejiang Province
Haihua Wu
135884526590163.com

Tongde Hospital of Zhejiang Province

Research Article

Keywords: sepsis, acute kidney injury, prediction models, intensive care unit, systematic review, meta-analysis

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11725522/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Prediction models may support risk assessment for sepsis-associated acute kidney injury (SA-AKI), but their clinical applicability remains uncertain. We assessed models for SA-AKI occurrence or identification in adults with sepsis in intensive care units (ICUs), and mortality in patients with established SA-AKI.

Methods

We systematically searched PubMed/MEDLINE, Embase, Web of Science, Cochrane Library, and ClinicalTrials.gov through October 1, 2026, for model development or validation studies with compatible quantitative performance data. Two reviewers independently verified the extracted data. PROBAST + AI was used to assess model development quality and risk of bias in performance evaluation separately. Different models were synthesized in exploratory strata defined by clinical task, prediction horizon, and assessment stage. Areas under the receiver operating characteristic curve (AUROCs) were pooled on the logit scale using restricted maximum likelihood random-effects models. The modified Hartung–Knapp method was used to calculate 95% confidence intervals (CIs). Sensitivity analyses examined prediction timing, risk of bias, possible patient overlap across reports, and sources of precision information.

Results

We included 39 reports. Validation strata covered any AKI within 7 days (5 reports), any AKI within 48 hours (3 reports), and in-hospital mortality (7 reports). Pooled AUROCs were 0.828 (95% CI 0.700–0.908), 0.725 (95% CI 0.631–0.803), and 0.778 (95% CI 0.747–0.806), respectively. Corresponding I^2 values were 89.1%, 70.8%, and 67.4%. Of 116 performance assessment groups, 61 were at high risk of bias and 55 at unclear risk; none was at low risk.

Conclusions

Existing models showed some discrimination for SA-AKI occurrence or identification and mortality risk under the conditions evaluated. However, unclear prediction timing, bias in performance evaluation, possible patient overlap across reports, and limited calibration evidence restrict inferences about early warning and use across settings. Before clinical use, fixed models require external validation with clearly defined prediction timing and independent patient sources, alongside assessment of calibration and clinical net benefit.

Background

Sepsis-associated acute kidney injury (SA-AKI) is a common form of organ dysfunction among adults with sepsis in intensive care units (ICUs). It is associated with mortality and adverse long-term kidney outcomes. The 2023 Acute Disease Quality Initiative (ADQI) consensus defines SA-AKI as acute kidney injury (AKI) occurring within 7 days of sepsis onset [1]. In a retrospective cohort study from two US centers using this definition, 46.6% of 63536 ICU patients with sepsis developed SA-AKI. In-hospital mortality among those with SA-AKI was 25.0% [2].

Clinical identification of AKI relies primarily on changes in serum creatinine and urine output. However, an increase in serum creatinine may lag behind kidney injury, limiting the use of these diagnostic measures alone for risk stratification before AKI develops [3]. In patients with sepsis who have not yet developed AKI, risk assessment may inform closer kidney function monitoring, review of exposure to nephrotoxic medications, and supportive care. In patients with established SA-AKI, mortality risk assessment aims to identify those at higher risk of a poor prognosis and inform subsequent monitoring [1, 3]. These tasks concern different target populations and decision points; performance must therefore be interpreted separately in relation to the prediction time point and outcome observation window.

Electronic health records and critical care databases have enabled the development of multivariable prediction models that integrate comorbidities, physiological measurements, and laboratory data. Models and risk scores based on conventional regression or machine learning continue to emerge, but high reported discrimination alone does not establish their clinical value. Large systematic reviews have documented substantial differences in study populations, outcome definitions, and validation approaches. Study bias and uncertainty about performance across settings continue to limit implementation [4, 5]. The area under the receiver operating characteristic curve (AUROC) primarily measures discrimination: a high AUROC does not establish agreement between predicted risks and observed outcome probabilities. Calibration and clinical utility require separate assessment [6].

Previous reviews have synthesized models for SA-AKI occurrence and machine learning models for mortality. More recent reviews have extended their scope to AKI occurrence and post-AKI mortality across clinical settings [7–9]. Reported discrimination in adults with sepsis in ICUs may not represent prediction of future outcomes at a defined clinical time point. This distinction requires closer examination of the original evaluation conditions. If predictor collection overlaps with AKI ascertainment, performance may primarily reflect concurrent disease identification. Apparent performance in the development data, internal validation, and external validation also provide different types of evidence [10, 11]. Multiple reports may reuse patients from the same public database, so the number of reports may not reflect the amount of independent evidence [8]. Within compatible clinical tasks and prediction horizons, interpreting performance by assessment stage and patient source can clarify its applicability. Calibration and performance at specified thresholds also inform this interpretation.

We therefore conducted a systematic review and exploratory stratified meta-analysis of multivariable models with compatible quantitative performance data. The tasks were SA-AKI occurrence or identification in adults with sepsis in ICUs, and mortality among patients with established SA-AKI. We

synthesized discrimination by clinical target, prediction horizon, and assessment stage. Analyses examined how prediction timing and possible patient overlap across reports affect interpretation. We also summarized reporting of calibration and performance at specified thresholds.

Methods

This systematic review and exploratory stratified meta-analysis of prediction models used the Checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) to structure study information and followed Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA) guidance [12, 13]. PROSPERO (International Prospective Register of Systematic Reviews) identifier: CRD420261299822. Completed reporting checklists are provided in Additional files 2 and 3.

Information sources and search strategy

We systematically searched PubMed/MEDLINE, Embase, Web of Science, and the Cochrane Library on October 1, 2026. The strategies combined controlled vocabulary and free-text terms relating to artificial intelligence and machine learning, sepsis, acute kidney injury, and intensive care. To reduce the possibility of missing eligible reports, we also searched ClinicalTrials.gov and manually examined the reference lists of included reports and relevant reviews. Potentially eligible reports underwent full-text assessment. Complete search strategies and the number of records retrieved from each source are provided in Additional file 1, Search strategies.

Eligibility criteria

Inclusion criteria

(1) Population: Adults with sepsis in the ICU. Studies predicting mortality were required to concern patients with sepsis-associated acute kidney injury (SA-AKI) or provide separately extractable model performance data for an SA-AKI subgroup.

(2) Prediction models: Multivariable clinical prediction models, risk scores, or machine learning models developed using ICU clinical information.

(3) Target outcomes and reference standards: SA-AKI occurrence or diagnosis, or mortality among patients with SA-AKI. AKI was ascertained using Kidney Disease: Improving Global Outcomes (KDIGO), Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease (RIFLE), or Acute Kidney Injury Network (AKIN) criteria, or an explicitly reported and verifiable clinical standard. Mortality was determined from observed survival status.

(4) Performance data: At least one AUROC or corresponding concordance statistic for a binary outcome, with precision information matched to the model evaluation dataset. Precision information could be obtained directly from the report. Alternatively, it could be approximated from actual event and non-event counts in the same evaluation dataset, or from eligible confidence intervals reaching a boundary. Other performance measures and threshold-specific data were extracted as reported.

(5) Study design: Formally published original model development or validation studies, including prospective or retrospective studies, single-center or multicenter studies, and studies using public ICU databases.

(6) Conditions for quantitative synthesis: Each report had to provide at least one eligible model performance record for quantitative synthesis. Each corresponding synthesis stratum had to contain at least two reports, with compatible clinical tasks, prediction horizons, performance measures, and model assessment stages. The requirement for a report to contribute to a quantitative synthesis was introduced during analysis.

Exclusion criteria

(1) Reviews, commentaries, conference abstracts, case reports, animal studies, simulation studies, retracted articles, and preprints that had not been formally published.

(2) Studies without separately extractable model performance data for the target population, or with AKI ascertainment criteria that could not be verified. Mortality prediction studies were also excluded if performance data for patients with SA-AKI could not be extracted separately.

(3) Studies examining associations or reporting receiver operating characteristic curves for a single biomarker without developing a combined prediction model.

(4) Reports providing no model performance records that met the conditions for quantitative synthesis specified above.

Study selection and data extraction

Two reviewers independently screened titles and abstracts and then assessed full texts against the eligibility criteria. Disagreements were resolved through discussion, with a third reviewer assisting if consensus could not be reached. Two reviewers independently verified the extracted data using a standardized form, and discrepancies were resolved through discussion. Extracted study information included authors, publication year, design, country or region, data sources, recruitment periods, and population restrictions. We collected sepsis and AKI definitions, prediction time points, predictor collection windows, and outcome observation windows. Modeling data included model types, variable selection methods, and the composition of development and evaluation datasets. Performance data

included AUROCs, confidence intervals or standard errors, threshold-specific performance, calibration, and decision curve information. Age and sex data were checked against the full texts. Age was retained in its original form as a mean, standard deviation, median, interval, or age category. Sex was retained using the reported category, count, percentage, and denominator. When only outcome-specific groups were reported, these were presented separately; subgroup medians and overlapping cohorts were not combined into overall values.

Methodological quality assessment

The updated Prediction model Risk Of Bias ASsessment Tool for regression and artificial intelligence methods (PROBAST + AI) was used to assess concerns about model development quality and risk of bias in performance evaluation separately [11]. Homogeneous assessment units were defined according to the population, predictor and outcome definitions, development process, and evaluation dataset. We assessed four domains: participants and data sources, predictors, outcome, and analysis, together with applicability in the first three domains. Answers to signaling questions and domain judgments were recorded separately. Domain-level and overall judgments were classified as low, high, or unclear according to the tool, without calculating a total score. Insufficient information was recorded as no information (NI) and judged in light of its potential consequences; it did not automatically lead to high concern about development quality or high risk of bias. Judgments were linked to specific models and cohorts rather than assigning the worst judgment within a report to all its assessments. We distinguished post hoc recommendation of the best-performing fixed candidate model after holdout evaluation from the use of evaluation data in training, feature optimization, or tuning. Sensitivity analyses examined the influence of post hoc model recommendation.

Statistical analysis

Model discrimination was synthesized using random-effects meta-analysis. AUROCs were logit-transformed, between-study variance (τ^2) was estimated by restricted maximum likelihood (REML), and 95% confidence intervals were calculated using the modified Hartung–Knapp method. Pooled estimates and their intervals were back-transformed to the AUROC scale [10, 14]. Quantitative syntheses were stratified by clinical target, prediction horizon, measure definition, model role, and assessment stage. Validation performance, performance evaluation affected by optimization, and apparent performance in the development data were analyzed separately. Each stratum included at least 2 reports, with one representative estimate per report. Pooling different models served as an exploratory assessment within the same clinical task.

Reported standard errors matched to the model and evaluation dataset were used when available and transformed to the logit scale using the delta method. Otherwise, standard errors were derived from matched reported 95% confidence intervals. For records without precision information but with verifiable event and non-event counts from the same binary-outcome evaluation dataset, standard errors were approximated using the Hanley–McNeil method [15]. The procedures used to approximate precision from boundary confidence intervals are described in the statistical methods in Additional file 1. The main

analysis included estimates with reported or approximated standard errors; sensitivity analyses restricted the source of standard errors to assess robustness.

Between-study heterogeneity was quantified using I^2, and approximate 95% prediction intervals were reported when estimable. Leave-one-report-out and influence analyses were conducted for the main validation results. Sensitivity analyses examined prediction timing, validation design, clinical phenotype, risk of bias, and post hoc model recommendation. Reports with potentially overlapping patient sources were grouped using databases, study centers, and recruitment periods; an additional sensitivity analysis retained one representative estimate per source group. Rules for selecting representative estimates, prediction interval calculations, and the detailed specifications of exploratory multilevel analyses are provided in Additional file 1.

Sensitivity, specificity, calibration slopes and intercepts, Brier scores, and Hosmer–Lemeshow test results were synthesized descriptively. Individual study results were reported when a clinical subset contained fewer than 2 reports or when performance measures and their precision information were not sufficiently comparable. All statistical analyses were conducted in R 4.4.0 using metafor 4.8.0.

Certainty of evidence

Drawing on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) conceptual framework for prediction model discrimination, we conducted an exploratory certainty assessment of pooled validation performance, targeting pooled discrimination above chance (AUROC > 0.5) [16]. Each body of evidence started at high certainty and was assessed across five domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias. Certainty was downgraded according to the seriousness of concerns and classified as high, moderate, low, or very low. Domain judgments considered the PROBAST + AI assessments, contributions of individual reports to pooled estimates, discrimination estimates, and sensitivity analyses.

The assessment covered 5 bodies of evidence combining different models at the clinical-task level and 1 body of evidence concerning the SOFA score. For evidence combining different models, we used an exploratory adaptation informed by GRADE principles (GRADE-informed); these findings are presented as a supplementary analysis.

Language model support

OpenAI Codex (OpenAI; GPT-6) was used to support literature assessment, data extraction and quality assessment, R code preparation, translation, language editing, and document and figure formatting. Two reviewers independently verified the extracted data, quality assessments, and analysis outputs.

Results

Study selection and characteristics

The searches identified 2243 records: 2229 from PubMed/MEDLINE, Embase, Web of Science, and the Cochrane Library, and 14 from ClinicalTrials.gov. After removing 672 duplicates, 1571 records underwent title and abstract screening, of which 1443 were excluded. The remaining 128 reports underwent full-text assessment. At that stage, 89 reports were excluded because of an ineligible population (57), an ineligible outcome (23), or duplicate reporting (9). The final review included 39 reports [17–55], all providing performance data suitable for stratified quantitative synthesis. Figure 1 shows the selection process.

The included reports were published between 2011 and 2026, with 25 published between 2024 and 2026. Participants were adults with sepsis or SA-AKI in the ICU. Data came primarily from US public critical care databases and Chinese hospital cohorts. Of the reports, 20 used the Medical Information Mart for Intensive Care (MIMIC)-III and/or MIMIC-IV databases; 6 of these also used the eICU Collaborative Research Database (eICU-CRD). Five reports contributed eligible eICU data. In Zhang H et al. [36], the eICU cohort was excluded because adult eligibility could not be established. Some studies combined databases or hospital cohorts for external validation, and a small number also used the Salzburg Intensive Care database (SICdb) in Austria. Sample sizes varied substantially, from hospital cohorts of approximately one hundred patients to database cohorts of tens of thousands. By clinical task, 24 reports addressed AKI occurrence or identification and 16 addressed mortality in patients with SA-AKI; 1 report included both outcomes. AKI outcomes included any AKI and severe AKI defined as KDIGO stages 2–3, assessed within 24 hours, 48 hours, 7 days, or during hospitalization. Mortality outcomes were mainly in-hospital, 28-day, and 30-day mortality. Modeling approaches included logistic regression, random forests, gradient boosting, XGBoost, and CatBoost. The included models also comprised nomograms and clinical risk scores. We extracted 314 model performance records covering apparent performance in the development data, internal validation, and external validation. Table 1 provides a concise overview of study and demographic characteristics. All 39 reports provided age data for at least one eligible cohort or outcome group, and 38 provided sex data; 3 reported age categories only. Complete study-level baseline information and its correspondence to specific cohorts are provided in Additional file 1, Table S1.

Table 1
Overview of study characteristics and demographic data in the included reports

Characteristic	Category or reporting basis	Reports (%)
Evidence base	Included reports	39 (100.0)
Publication years	2011–2026; including reports published in 2024–2026	25 (64.1)
Study design	Retrospective cohort or case-control studies	30 (76.9)
	Prospective cohort studies	7 (17.9)
	Retrospective analyses of prospective cohorts	2 (5.1)
Patient-source region	China / United States / multinational / unclear	17/14/7/1
Clinical task	AKI occurrence or identification	24 (61.5)
	Mortality risk in patients with SA-AKI	16 (41.0)
	Both clinical tasks	1 (2.6)
Public databases	Use of MIMIC-III and/or MIMIC-IV	20 (51.3)
	Of these, also used eICU-CRD	6 (15.4)
Baseline age	Extractable for at least one eligible cohort/subgroup	39 (100.0)
	Continuous age reported / age categories only	36/3
Baseline sex	Extractable for at least one eligible cohort/subgroup	38 (97.4)

Note: The denominator is 39 reports; only report counts are given for patient-source regions and age-reporting formats. Clinical tasks may overlap, and data-source categories are not mutually exclusive. Age or sex was counted as available when numerical data were reported for at least one eligible cohort or outcome subgroup. This does not imply separate reporting for every validation cohort. Database counts describe use in the original reports: 6 reports used eICU-CRD, but only 5 contributed eligible eICU data; the eICU cohort in Zhang H et al. [36] was excluded. Different statistical formats and overlapping patient sources were not combined into an overall age summary or sex proportion. Complete report-level and cohort-level information is provided in Additional file 1, Table S1.

Methodological quality and risk of bias

The PROBAST + AI assessment covered 45 model development groups and 116 performance assessment groups. For model development, concerns about quality were low in 1 group (2.2%), unclear in 29 (64.4%), and high in 15 (33.3%). The only development group with low concern about quality was from Luo et al. [31]. For performance evaluation, 61 groups (52.6%) were at high risk of bias and 55 (47.4%) at unclear risk; none had an overall low risk of bias.

The analysis domain had the highest proportion of high-risk judgments. Among the 116 performance assessment groups, 55 (47.4%) were at high risk and 58 (50.0%) at unclear risk in this domain.

Insufficient information mainly concerned sample selection, prediction time points, handling of missing data, and independence between model development and performance evaluation data. Applicability concerns were low in 31 development groups (68.9%) and 91 performance assessment groups (78.4%). Group-level and domain-level assessments are summarized in Additional file 1, Tables S3, S5, and S12 and Figures S4–S5.

Model validation performance

Validation evidence included internal and external validation results and was synthesized by clinical task. For any AKI within 7 days, pooled AUROC across 5 reports was 0.828 (95% CI 0.700–0.908). I^2 was 89.1%, and the 95% prediction interval was 0.486–0.960. For any AKI within 48 hours, pooled AUROC across 3 reports was 0.725 (95% CI 0.631–0.803). I^2 was 70.8%, and the 95% prediction interval was 0.555–0.848.

For in-hospital mortality, pooled AUROC across 7 reports was 0.778 (95% CI 0.747–0.806). I^2 was 67.4%, and the 95% prediction interval was 0.703–0.838. This analysis included general SA-AKI populations, older adults with persistent SA-AKI, patients receiving kidney replacement therapy, and patients with diabetes. Table 2 and Fig. 2 present these main validation results.

Table 2
Pooled discrimination in the main clinical validation strata

Clinical target	Reports / source blocks	AUROC (95% CI)	95% prediction interval	I^2, %
Any AKI (7 days)	5/5	0.828 (0.700–0.908)	0.486–0.960	89.1
Any AKI (48 hours)	3/2	0.725 (0.631–0.803)	0.555–0.848	70.8
In-hospital mortality	7/5	0.778 (0.747–0.806)	0.703–0.838	67.4

Among other validation outcomes, 2 reports on 28-day mortality yielded a pooled AUROC of 0.791 (95% CI 0.637–0.890), with I^2 of 0%. The 4 reports on AKI during hospitalization without a fixed observation window yielded a pooled AUROC of 0.804 (95% CI 0.684–0.885). I^2 was 89.0%, and the 95% prediction interval was 0.526–0.938 (Additional file 1, Table S2). A supplementary analysis of the same Sequential Organ Failure Assessment (SOFA) score across 2 reports yielded a pooled AUROC of 0.705 (95% CI 0.193–0.960) for 28-day mortality (Additional file 1, Table S6).

Model performance at other assessment stages

For apparent performance in the development data, the pooled AUROC was 0.902 (95% CI 0.618–0.981) across 5 reports on any AKI within 7 days. The pooled AUROC was 0.806 (95% CI 0.711–0.875) for in-hospital mortality across 5 reports. For 28-day mortality, it was 0.875 (95% CI 0.661–0.962) across 4 reports. For severe AKI within 7 days, 2 reports yielded a pooled AUROC of 0.922 (95% CI 0.888–0.946).

The analysis of 30-day mortality also included only 2 reports and produced a pooled AUROC of 0.926. Its 95% confidence interval extended from below 0.001 to above 0.999, indicating limited precision.

For performance evaluation affected by optimization or selection, 5 reports on in-hospital mortality yielded a pooled AUROC of 0.776 (95% CI 0.743–0.805). Any AKI within 7 days and severe AKI within 7 days were each addressed by only 2 reports. Pooled AUROCs were 0.759 (95% CI 0.100–0.989) and 0.869 (95% CI 0.112–0.997), respectively, both with wide confidence intervals. Complete results by assessment stage are provided in Table 3 and Additional file 1, Table S2.

Table 3
Pooled discrimination at other evaluation stages

Evaluation stage	Clinical target	Reports / source blocks	AUROC (95% CI)
Performance evaluation affected by optimization or selection	Any AKI during hospitalization without a standardized observation window	2/1	0.798 (0.641–0.898)
Apparent performance in the development data	Any AKI (48 hours)	3/1	0.756 (0.638–0.844)
Performance evaluation affected by optimization or selection	In-hospital mortality	5/1	0.776 (0.743–0.805)
Apparent performance in the development data	Any AKI during hospitalization without a standardized observation window	7/6	0.817 (0.761–0.863)
Apparent performance in the development data	Any AKI (24 hours)	2/2	0.822 (0.242–0.985)
Apparent performance in the development data	Any AKI (7 days)	5/5	0.902 (0.618–0.981)
Performance evaluation affected by optimization or selection	Any AKI (7 days)	2/2	0.759 (0.100–0.989)
Apparent performance in the development data	In-hospital mortality	5/1	0.806 (0.711–0.875)
Apparent performance in the development data	28-day mortality	4/4	0.875 (0.661–0.962)
Apparent performance in the development data	Severe AKI (7 days)	2/2	0.922 (0.888–0.946)
Performance evaluation affected by optimization or selection	Severe AKI (7 days)	2/2	0.869 (0.112–0.997)
Apparent performance in the development data	30-day mortality	2/2	0.926 (< 0.001–>0.999)

Sensitivity analyses

The in-hospital mortality sensitivity analysis excluded 3 reports involving post hoc model recommendation. The remaining 4 reports yielded a pooled AUROC of 0.777 (95% CI 0.736–0.813). This was close to the main estimate. Restriction to the 3 reports concerning general SA-AKI populations yielded a pooled AUROC of 0.785 (95% CI 0.729–0.831). In leave-one-report-out analyses of any AKI within 7 days, Zhang et al. [40] had a substantial influence on heterogeneity. Excluding this report reduced the pooled AUROC from 0.828 to 0.798 (95% CI 0.721–0.858).

A timing-restricted sensitivity analysis required predictors to be explicitly measured before outcome occurrence. No eligible representative records remained for any AKI within 7 days or AKI during hospitalization without a fixed observation window. Only Zhao et al. [50] remained for any AKI within 48 hours. None of these strata could therefore be meta-analyzed. Restricting the source of precision information also reduced the poolable evidence: only 1 report remained for any AKI within 48 hours when using reported confidence intervals or standard errors alone. Similarly, only 1 report remained for 28-day mortality after excluding the heavily weighted report by Fan et al. [52]. Approximating precision from boundary confidence intervals affected only the apparent development performance analysis for 28-day mortality that included Zhao et al. [25]. The main validation results were unchanged.

After retaining representative estimates within groups of potentially overlapping patient sources, 32 reports could still contribute to at least one quantitative synthesis. Excluding the selected high-risk-of-bias estimates without selecting replacement models reduced this number to 17; combining this restriction with the source restriction reduced it to 14. These are counts of unique reports across analysis strata. The main source, risk-of-bias, and clinical restriction analyses are presented in Fig. 3, Table 4, and Additional file 1, Tables S9–S10.

Table 4
Restriction-based sensitivity analyses in key validation strata

Clinical task	Restriction	Reports	AUROC (95% CI) / status
Any AKI (7 days)	Predictors explicitly measured before outcome occurrence	0	Cannot be pooled
Any AKI (48 hours)	Predictors explicitly measured before outcome occurrence	1	Single report; not pooled
In-hospital mortality	General SA-AKI population	3	0.785 (0.729–0.831)
In-hospital mortality	Exclude records with post hoc model recommendation	4	0.777 (0.736–0.813)
28-day mortality	Exclude Fan et al. (2023)	1	Single report; not pooled

Sensitivity of prediction intervals to the calculation method varied by outcome. With the Riley-type approximation, the 95% prediction interval for any AKI within 48 hours widened to 0.223–0.960. For AKI

during hospitalization without a fixed observation window, it widened to 0.413–0.960. Both crossed 0.5. The interval for in-hospital mortality was 0.699–0.841, close to the main analysis. For any AKI within 7 days, it was 0.428–0.969 (Additional file 1, Table S11). The exploratory multilevel analysis did not reach the effective degrees-of-freedom threshold of 4 used in this study. Inferential confidence intervals and P values were therefore not reported for that analysis.

Threshold-specific performance, calibration, and prediction error

Extractable threshold-specific performance data were available from 6 reports (15.4%), comprising 25 records across different models, evaluation datasets, and time points. The corresponding 2×2 contingency tables, sensitivities, and specificities are presented by model and evaluation dataset in Additional file 1, Table S13. No pooled threshold-specific performance estimate was produced across clinical tasks.

Extractable numerical data on calibration, overall prediction error, or goodness-of-fit tests were available from 16 reports (41.0%). These comprised 88 records: 55 Brier scores, 26 Hosmer–Lemeshow test results, 4 calibration slopes, 2 calibration intercepts, and 1 mean absolute calibration error. Numerical reporting of calibration slopes and intercepts was limited to 2 reports and 1 report, respectively. Because the estimands and precision information were not sufficiently comparable, these measures were synthesized descriptively. Reporting frequencies are summarized in Additional file 1, Table S4.

Certainty of evidence

The exploratory assessment targeted a pooled validation AUROC above 0.5. Of the 6 bodies of evidence, 4 were rated as moderate certainty, 1 as low certainty, and 1 as very low certainty. Evidence concerning candidate models for any AKI within 48 hours, any AKI within 7 days, in-hospital mortality, and 28-day mortality was downgraded by 1 level for risk of bias. Each was rated as moderate certainty. Evidence for AKI during hospitalization without a fixed observation window was downgraded by 2 levels for risk of bias and rated as low certainty. Evidence for the SOFA score in predicting 28-day mortality was downgraded by 1 level for risk of bias and 2 levels for imprecision, resulting in very low certainty.

Of these, 5 bodies of evidence combining different models were assessed using an exploratory GRADE-informed adaptation, while 1 concerned the same Sequential Organ Failure Assessment (SOFA) score. These ratings refer only to certainty that the corresponding pooled discrimination exceeds chance. The reports contributing to each body of evidence, domain-level judgments, and reasons for the ratings are provided in Additional file 1, Tables S6–S8.

Discussion

Main findings

This review included 39 reports. It synthesized model performance for SA-AKI occurrence or identification in adults with sepsis in ICUs, and for mortality in patients with SA-AKI. Pooled validation AUROCs were 0.828 for any AKI within 7 days, 0.725 for any AKI within 48 hours, and 0.778 for in-hospital mortality. None of the performance assessment groups had a low overall risk of bias. Restricting analyses to predictors explicitly measured before outcome occurrence substantially reduced the AKI evidence available for synthesis. Evidence on calibration, performance at specified thresholds, and clinical utility was limited. The discrimination results remain insufficient to establish early warning of future AKI or clinical decision-making value.

Comparison with previous reviews

Previous systematic reviews have also found some discriminative ability among SA-AKI models. Huang et al. [7] included 15 studies and reported a pooled C-statistic of 0.817 (95% CI 0.781–0.847) for SA-AKI occurrence. However, heterogeneity was high, and the external validation subgroup contained only 3 studies. The review of SA-AKI mortality prediction by Lv et al. [8] included 8 studies. The pooled C-statistic in validation datasets was 0.73 (95% CI 0.71–0.74). This estimate combined in-hospital and 28-day mortality, and most studies relied primarily on internal validation. In broader AKI populations, Xiao et al. [9] similarly observed high average discrimination but noted limited external or prospective validation and inconsistent reporting of calibration and clinical utility.

Our findings are consistent with those reviews and further distinguish clinical targets, observation windows, and model assessment stages. SA-AKI occurrence, severe AKI, and mortality are different prediction tasks, and 48-hour, 7-day, and unspecified in-hospital observation windows address different clinical questions. Differences in populations, outcome definitions, model composition, and validation designs preclude using AUROC differences between reviews or strata to judge model superiority. Our syntheses of different models primarily describe the evidence within a given clinical task; pooled results still require interpretation in relation to individual models and their evaluation populations.

Influence of prediction timing and validation design

The ADQI definition includes AKI occurring within 7 days of sepsis onset [1]. This window defines the temporal association between AKI and sepsis; it does not imply that a model can provide 7 days of advance warning. If a model uses information from the first 24 hours after ICU admission, predictors can include creatinine, urine output, or other clinical measures. When the AKI observation window also starts at admission, those predictors may overlap with outcome ascertainment. A high AUROC in this setting may partly reflect recognition of kidney dysfunction that is already present, which serves a different clinical purpose from risk prediction before AKI develops.

The timing-restricted analyses illustrate this problem. No representative records for the 7-day AKI validation stratum met the requirement that predictors were explicitly measured before outcome occurrence, and only 1 report remained in the 48-hour stratum. Neither stratum could be meta-analyzed. For studies with overlapping or unclear predictor and outcome windows, performance is therefore more

appropriately interpreted as discrimination of occurrence or identification within the specified observation window. This finding indicates limited support from existing designs and reporting for early warning; it does not imply that all models lack predictive potential. The 95% prediction interval of 0.486–0.960 for 7-day AKI also indicates that the pooled point estimate may mask substantial between-study variation.

Validation design and case mix also affect clinical interpretation. Our validation strata included internal holdout and other validation designs. In-hospital mortality studies covered general SA-AKI populations, older adults, patients with persistent AKI, those receiving kidney replacement therapy, and patients with diabetes. Changes in case mix can alter discrimination even when model parameters remain unchanged [10]. The pooled results therefore do not establish stable performance of a fixed model across centers and clinical phenotypes. Post hoc model recommendation after holdout evaluation of fixed candidates remains susceptible to selection optimism. This differs from using evaluation data in training or tuning. In the in-hospital mortality stratum, excluding 3 representative records with identified post hoc recommendation yielded a pooled AUROC of 0.777. This was close to the main estimate of 0.778, indicating little change in the point estimate under this restriction.

Clinical implications and evidence gaps

Clinical decision-making also requires accurate risk estimates and net benefit at relevant decision thresholds. AUROC primarily measures how well a model ranks risk between patients with different outcomes; it does not independently establish agreement between predicted probabilities and observed risks [6]. In this review, 16 reports provided extractable calibration measures, overall prediction error, or goodness-of-fit test results, but most concerned Brier scores or Hosmer–Lemeshow tests. Only 2 reports provided extractable calibration slopes and 1 provided an intercept. The Brier score measures overall prediction error, and a nonsignificant Hosmer–Lemeshow test does not establish good calibration [6]. The available evidence therefore permits only a limited assessment of the accuracy of absolute risk estimates.

Threshold-specific performance data came from only 6 reports and concerned different models, clinical targets, and evaluation time points, limiting the basis for a widely applicable alert threshold. For models intended to prompt closer monitoring or support risk assessment, evaluation should specify the consequences of missed cases and false alerts. It should also consider sensitivity, specificity, and intervention burden at relevant thresholds. External validation should evaluate discrimination, calibration, and net benefit across applicable decision thresholds [56]. A high AUROC in existing data does not establish that model-guided care reduces AKI incidence or mortality.

The GRADE-informed assessment was post hoc and exploratory. Moderate certainty was assigned to 4 clinical-task bodies of evidence. The ratings concern only whether pooled validation AUROC exceeded 0.5. Pooling different models by clinical task extends beyond the framework for validation evidence concerning a fixed model. These ratings do not establish adequate warning time, accurate individual risk estimates, improved patient outcomes, or a basis for recommending a specific model.

Strengths and limitations

This review organized performance evidence by clinical task, prediction horizon, and assessment stage. PROBAST + AI was used to assess concerns about model development quality separately from risk of bias in performance evaluation. Sensitivity analyses examined how pooled findings depended on potential overlap in patient sources, prediction timing, clinical phenotype, and precision information. We also summarized performance at specified thresholds, calibration, and overall prediction error.

During the analysis stage, eligibility was restricted to reports able to contribute to at least one clinically and methodologically compatible quantitative synthesis. This may have excluded studies lacking poolable data despite their clinical value and introduced selective inclusion; conclusions should therefore be limited to the quantitative evidence included. In the primary studies, incomplete reporting of sample selection, prediction time points, handling of missing data, and separation of development from evaluation also limited some bias judgments. An unclear overall risk of bias should not be interpreted as low risk.

Possible patient overlap across reports is another limitation. Of the 39 reports, 20 used MIMIC-III and/or MIMIC-IV, and some also used eICU, so patient overlap between reports is possible. Reports were grouped into blocks of potentially overlapping patient sources using databases, centers, and recruitment periods. This helps examine dependence but cannot resolve patient-level overlap or ensure independence between blocks. The multilevel robust analysis did not yield sufficient effective degrees of freedom for inference, so interval estimates must still be interpreted in light of this dependence.

Each validation stratum included only 2–7 reports, and some findings were sensitive to an individual report or assumptions about precision. The 48-hour AKI stratum could not be pooled using only matched, reported precision information; only 1 report remained for 28-day mortality after excluding the report carrying most of the weight. Alternative prediction interval calculations further showed that uncertainty with few studies depends on the approximation method. Similarity between some sensitivity analyses and the main analysis therefore does not imply equal stability across all findings. Pooling different models, clinical heterogeneity, and limited calibration data further preclude algorithm rankings or firm conclusions about the clinical usability of individual models. Selective reporting and publication bias also cannot be excluded.

Future research

Future studies should begin with a defined clinical use and specify the prediction time point, the time at which each predictor becomes available, and the outcome observation window. Studies of early warning for incident AKI should distinguish AKI already present at prediction from subsequent AKI and report a lead time that could support clinical action. Handling of missing data, feature selection, and tuning during development should remain appropriately separated from the final evaluation. After development, fixed models should undergo temporal and cross-center validation using independent patient sources to examine applicability across case mixes [11, 56].

Evaluations should report discrimination, calibration, and performance at clinically meaningful thresholds, and compare models with existing clinical scores in the same independent population. Following external and prospective validation, clinical impact studies are needed to assess how model-assisted decisions affect monitoring, alert burden, resource use, and patient outcomes.

Conclusions

Among the 39 included reports, pooled AUROCs in the three main validation strata ranged from 0.725 to 0.828. None of the performance assessment groups had a low overall risk of bias, and evidence that predictors preceded AKI onset was limited. Possible patient overlap and sparse calibration and threshold-performance data further limit interpretation. Individual models require external validation in independent populations, clearly defined prediction time points, and assessment of calibration and clinical net benefit before clinical implementation.

Abbreviations

ADQI
Acute Disease Quality Initiative
AKI
acute kidney injury
AKIN
Acute Kidney Injury Network
AUROC
area under the receiver operating characteristic curve
CHARMS
Checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies
CI
confidence interval
eICU-CRD
eICU Collaborative Research Database
GRADE
Grading of Recommendations Assessment, Development and Evaluation
ICU
intensive care unit
KDIGO
Kidney Disease:Improving Global Outcomes
MIMIC
Medical Information Mart for Intensive Care
NI
no information
PROBAST + AI

Prediction model Risk Of Bias ASsessment Tool, updated for regression and artificial intelligence methods
PROSPERO
International Prospective Register of Systematic Reviews
REML
restricted maximum likelihood
RIFLE
Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease
SA-AKI
sepsis-associated acute kidney injury
SE
standard error
SICdb
Salzburg Intensive Care database
SOFA
Sequential Organ Failure Assessment
TRIPOD-SRMA
Transparent reporting of multivariable prediction models for individual prognosis or diagnosis:checklist for systematic reviews and meta-analyses

Declarations

Ethics approval and consent to participate

Not applicable. This systematic review used published aggregate data and did not involve recruitment of participants or access to identifiable individual-level data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Author Contribution

JY contributed to study conception and design, methodology, statistical analysis, and drafting the manuscript. MQ contributed to literature screening, data extraction, and methodological quality assessment. JZ contributed to data verification, statistical review, and critical revision of the manuscript. HW contributed to study supervision and critical revision of the manuscript. All authors read and approved the manuscript and its submission to BMC Nephrology.

Acknowledgements

Not applicable.

Data Availability

Data analyzed in this review were derived from the published reports identified in the reference list. Extracted data and analysis code are available from the corresponding author on reasonable request. Study characteristics, statistical methods, and supplementary results are provided in Additional file 1.

References

1. Zarbock A, Nadim MK, Pickkers P, Gomez H, Bell S, Joannidis M, et al. Sepsis-associated acute kidney injury: consensus report of the 28th Acute Disease Quality Initiative workgroup. Nat Rev Nephrol. 2023;19(6):401–17. 10.1038/s41581-023-00683-3.
2. Takeuchi T, Flannery AH, Liu LJ, Ghazi L, Cama-Olivares A, Fushimi K, et al. Epidemiology of sepsis-associated acute kidney injury in the ICU with contemporary consensus definitions. Crit Care. 2025;29(1):128. 10.1186/s13054-025-05351-5.
3. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney Int Suppl. 2012;2(1):1-138. Available from: https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-AKI-Guideline-English.pdf
4. Feng Y, Wang AY, Jun M, Pu L, Weisbord SD, Bellomo R, et al. Characterization of Risk Prediction Models for Acute Kidney Injury: A Systematic Review and Meta-analysis. JAMA Netw Open. 2023;6(5):e2313359. 10.1001/jamanetworkopen.2023.13359.
5. Cama-Olivares A, Braun C, Takeuchi T, O'Hagan EC, Kaiser KA, Ghazi L, et al. Systematic Review and Meta-Analysis of Machine Learning Models for Acute Kidney Injury Risk Classification. J Am Soc Nephrol. 2025;36(10):1969–83. 10.1681/ASN.0000000702.
6. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. 10.1186/s12916-019-1466-7.
7. Huang M, Wu S, Shen ZH, Zhang SJ, Fu YS, Wu J. Prediction models for sepsis-associated acute kidney injury: a systematic review and meta-analysis. Front Endocrinol (Lausanne). 2026;17:1854549. 10.3389/fendo.2026.1854549.

8. Lv X, Liu D, Chen X, Chen L, Wang X, Xu X, et al. Machine learning for the prediction of mortality in patients with sepsis-associated acute kidney injury: a systematic review and meta-analysis. BMC Infect Dis. 2024;24(1):1454. 10.1186/s12879-024-10380-6.
9. Xiao H, Liang J, Li G, Yang Y, Li T, Chen X, et al. AI Models for Predicting Acute Kidney Injury (AKI) and Post-AKI Mortality: Systematic Review and Meta-Analysis. J Med Internet Res. 2026;28:e104269. 10.2196/104269.
10. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. 10.1136/bmj.i6460.
11. 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:e082505. 10.1136/bmj-2024-082505.
12. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744. 10.1371/journal.pmed.1001744.
13. Snell KIE, Levis B, Damen JAA, Dhiman P, Debray TPA, Hooft L, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). BMJ. 2023;381:e073538. 10.1136/bmj-2022-073538.
14. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol. 2015;15:99. 10.1186/s12874-015-0091-1.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. 10.1148/radiology.143.1.7063747.
16. Foroutan F, Mayer M, Guyatt G, Riley RD, Mustafa R, Kreuzberger N, et al. GRADE concept paper 8: judging the certainty of discrimination performance estimates of prognostic models in a body of validation studies. J Clin Epidemiol. 2024;170:111344. 10.1016/j.jclinepi.2024.111344.
17. Zhou J, Bai Y, Wang X, Yang J, Fu P, Cai D, et al. A simple risk score for prediction of sepsis associated-acute kidney injury in critically ill patients. J Nephrol. 2019;32(6):947–56. 10.1007/s40620-019-00625-y.
18. Deng F, Peng M, Li J, Chen Y, Zhang B, Zhao S. Nomogram to predict the risk of septic acute kidney injury in the first 24 h of admission: an analysis of intensive care unit data. Ren Fail. 2020;42(1):428–36. 10.1080/0886022X.2020.1761832.
19. Fan C, Ding X, Song Y. A new prediction model for acute kidney injury in patients with sepsis. Ann Palliat Med. 2021;10(2):1772–8. 10.21037/apm-20-1117.
20. Yang S, Su T, Huang L, Feng LH, Liao T. A novel risk-predicted nomogram for sepsis associated-acute kidney injury among critically ill patients. BMC Nephrol. 2021;22(1):173. 10.1186/s12882-021-02379-x.
21. Ma J, Deng Y, Lao H, Ouyang X, Liang S, Wang Y, et al. A nomogram incorporating functional and tubular damage biomarkers to predict the risk of acute kidney injury for septic patients. BMC

Nephrol. 2021;22(1):176. 10.1186/s12882-021-02388-w.

22. Yue S, Li S, Huang X, Liu J, Hou X, Wang Y, et al. Construction and Validation of a Risk Prediction Model for Acute Kidney Injury in Patients Suffering from Septic Shock. Dis Markers. 2022;2022:9367873. 10.1155/2022/9367873.
23. Zhang S, Tang S, Rong S, Zhu M, Liu J, Hu Q, et al. [Construction of a predictive model for early acute kidney injury risk in intensive care unit septic shock patients based on machine learning]. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue. 2022;34(3):255–9. 10.3760/cma.j.cn121430-20211126-01790.
24. Jiang W, Zhang C, Yu J, Shao J, Zheng R. Development and validation of a nomogram for predicting in-hospital mortality of elderly patients with persistent sepsis-associated acute kidney injury in intensive care units: a retrospective cohort study using the MIMIC-IV database. BMJ Open. 2023;13(3):e069824. 10.1136/bmjopen-2022-069824.
25. Zhao L, Liu Y, Chen M, Chen L, Zhou S, Bai X, et al. [Construction of anomogram for predicting the prognosis of patients with sepsis-associated acute kidney injury]. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue. 2023;35(12):1255–61. 10.3760/cma.j.cn121430-20230813-00621.
26. Zhang J, Qi M, Ma L, Zhang K, Liu D, Liu D. [Construction and validation of a risk nomogram for sepsis-associated acute kidney injury in intensive care unit]. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue. 2024;36(8):801–7. 10.3760/cma.j.cn121430-20240221-00150.
27. Li C, Zhao K, Ren Q, Chen L, Zhang Y, Wang G, et al. Development and validation of a model for predicting in-hospital mortality in patients with sepsis-associated kidney injury receiving renal replacement therapy: a retrospective cohort study based on the MIMIC-IV database. Front Cell Infect Microbiol. 2024;14:1488505. 10.3389/fcimb.2024.1488505.
28. Ge X, Chen W, Shi J, Zhang J, Tai H, Zhang Y, et al. Prediction of Moderate-to-Severe Sepsis-Associated Acute Kidney Injury Using a Dual-Timepoint Machine Learning Model: Development, Multiregional Validation, and Clinical Deployment Study. J Med Internet Res. 2025;27:e73840. 10.2196/73840.
29. Zhao L, Zhang T, Li X, Chen L, Zhou S, Meng Z, et al. Development and Validation of a Nomogram for Predicting Acute Kidney Injury in Septic Patients. J Inflamm Res. 2024;17:5653–62. 10.2147/JIR.S470773.
30. Zhou W, Zhang N, Zhao T, Ma Q, Ma X. [Predictive value of inflammatory indicator and serum cystatin C for the prognosis of patients with sepsis-associated acute kidney injury]. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue. 2025;37(3):275–9. 10.3760/cma.j.cn121430-20240908-00759.
31. Luo S, Lai J, Mo L, Shen X, Fang R. Prediction of hospital mortality in sepsis-associated acute kidney injury using a machine-learning approach: a multicenter study using SHAP interpretability analysis. Clin Kidney J. 2026;19(1):sfaf372. 10.1093/ckj/sfaf372.
32. Li YJ, Hang WL, Ren M, Jiang YX, Li HQ, Xue P. Nomogram for predicting 28-day mortality in bacterial pneumonia with sepsis and acute kidney injury: A study based on the Medical Information Mart for Intensive Care-IV database. SAGE Open Med. 2026;14:20503121261421279. 10.1177/20503121261421279.

33. Lai JM, Zheng Y, Liang TY, Huang J, Quan YY. Risk factors and nomogram prediction model for prognosis in sepsis with acute kidney injury. BMC Nephrol. 2026;27(1):437. 10.1186/s12882-026-05039-0.
34. Zheng Y, Gao J, Hua T, Dong W, Yang M. Risk stratification for in-hospital mortality in sepsis-associated acute kidney injury patients receiving continuous renal replacement therapy: an interpretable, externally validated machine learning study. Ren Fail. 2026;48(1):2677246. 10.1080/0886022X.2026.2677246.
35. Tan R, Li Y, Mai D, Liu J, Wei Y, Wang C, et al. Risk prediction of sepsis-associated acute kidney injury: development, validation of a machine learning model with multicenter data. BMC Med Inf Decis Mak. 2026;26(1):357. 10.1186/s12911-026-03620-z.
36. Zhang H, Zhang J. Construction and validation of a machine learning-based prediction model for in-hospital acute kidney injury in patients with lung cancer complicated with sepsis: clinical and nursing applications. J Thorac Dis. 2026;18(6):613. 10.21037/jtd-2026-0558.
37. Wang Y, Shen H, Wu S, Yang S, Guo F, Yang M, et al. Development and Validation of an Interpretable Machine Learning Model for Predicting In-Hospital Mortality in Diabetic Patients with Sepsis-Associated Acute Kidney Injury. J Intensive Care Med. 2026;8850666261479295. 10.1177/08850666261479295.
38. Dong Y, Du M, Maimaiti T, Li J. Urinary microRNAs in sepsis-associated acute kidney injury: A LASSO-logistic regression approach for risk prediction model development. Am J Med Sci. 2026;372(2):154–63. 10.1016/j.amjms.2026.06.001.
39. Ma W, Wang J, Zhang J, Guo X, Li W, Wang H, et al. An interpretable prediction model for severe sepsis-associated acute kidney injury in older ICU patients with sepsis: a prospective multicenter cohort study. BMC Geriatr. 2026;26(1):977. 10.1186/s12877-026-07882-0.
40. Zhang J, Tian P, Wang J, Li W, Cui N. Unveiling the immuno-inflammatory endotype of sepsis-associated acute kidney injury: a machine learning-based approach. Kidney Res Clin Pract. 2026. 10.23876/j.krcp.26.112.
41. Liu M, Kou C, Lin Y, Zhang W. UDFSSF Model is Superior to SOFA Score in Predicting 30-day Mortality of Septic Patients With Acute Kidney Injury. Am Surg. 2026;92(5):1476–87. 10.1177/00031348251412266.
42. Yuan L, Zhou Y, Wang R, Huang X, Tang R, Yan F. Unveiling the role of sTLR2: A novel biomarker for predicting septic-associated AKI. Cytokine. 2024;184:156798. 10.1016/j.cyto.2024.156798.
43. Huang D, Yang Z, Qiu L, Lin J, Cheng X. The predictive value of renal vascular resistance index and serum biomarkers for sepsis-associated acute kidney injury: a retrospective study. BMC Nephrol. 2025;26(1):208. 10.1186/s12882-025-04131-1.
44. Li N, Zhang X, Wan P, Yu M, Min J. Combination of Urinary Neutrophil Gelatinase-associated Lipocalin, Kidney Injury Molecular-1, and Angiotensinogen for the Early Diagnosis and Mortality Prediction of Septic Acute Kidney Injury. Comb Chem High Throughput Screen. 2024;27(7):1033–45. 10.2174/0113862073263073231011060142.

45. Song J, Wu W, He Y, Lin S, Zhu D, Zhong M. Value of the combination of renal resistance index and central venous pressure in the early prediction of sepsis-induced acute kidney injury. J Crit Care. 2018;45:204–8. 10.1016/j.jcrc.2018.03.016.
46. Fu Y, He C, Bai Y, Zhang N, Zhao H. [Value of the combination of renal resistive index and central venous pressure to predict septic shock induced acute kidney injury]. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue. 2020;32(4):473–7. 10.3760/cma.j.cn121430-20191014-00062.
47. Plataki M, Kashani K, Cabello-Garza J, Maldonado F, Kashyap R, Kor DJ, et al. Predictors of acute kidney injury in septic shock patients: an observational cohort study. Clin J Am Soc Nephrol. 2011;6(7):1744–51. 10.2215/CJN.05480610.
48. Zhao H, Chen Y, Feng F, Dong CM. Development and validation of a nomogram for individualized prediction of acute kidney injury in patients with sepsis. Yi Xue Yan Jiu Za Zhi. 2024;53(9):69–73. 10.11969/j.issn.1673-548X.2024.09.014. . [in Chinese].
49. Shi J, Han H, Chen S, Liu W, Li Y. Machine learning for prediction of acute kidney injury in patients diagnosed with sepsis in critical care. PLoS ONE. 2024;19(4):e0301014. 10.1371/journal.pone.0301014.
50. Zhao CC, Nan ZH, Li B, Yin YL, Zhang K, Liu LX, et al. Development and validation of a novel risk-predicted model for early sepsis-associated acute kidney injury in critically ill patients: a retrospective cohort study. BMJ Open. 2025;15(1):e088404. 10.1136/bmjopen-2024-088404.
51. Li X, Wu R, Zhao W, Shi R, Zhu Y, Wang Z, et al. Machine learning algorithm to predict mortality in critically ill patients with sepsis-associated acute kidney injury. Sci Rep. 2023;13(1):5223. 10.1038/s41598-023-32160-z.
52. Fan Z, Jiang J, Xiao C, Chen Y, Xia Q, Wang J, et al. Construction and validation of prognostic models in critically Ill patients with sepsis-associated acute kidney injury: interpretable machine learning approach. J Transl Med. 2023;21(1):406. 10.1186/s12967-023-04205-4.
53. Gao T, Nong Z, Luo Y, Mo M, Chen Z, Yang Z, et al. Machine learning-based prediction of in-hospital mortality for critically ill patients with sepsis-associated acute kidney injury. Ren Fail. 2024;46(1):2316267. 10.1080/0886022X.2024.2316267.
54. Tang J, Huang J, He X, Zou S, Gong L, Yuan Q, et al. The prediction of in-hospital mortality in elderly patients with sepsis-associated acute kidney injury utilizing machine learning models. Heliyon. 2024;10(4):e26570. 10.1016/j.heliyon.2024.e26570.
55. Dong L, Liu P, Qi Z, Lin J, Duan M. Development and validation of a machine-learning model for predicting the risk of death in sepsis patients with acute kidney injury. Heliyon. 2024;10(9):e29985. 10.1016/j.heliyon.2024.e29985.
56. 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:e074820. 10.1136/bmj-2023-074820.

Figures

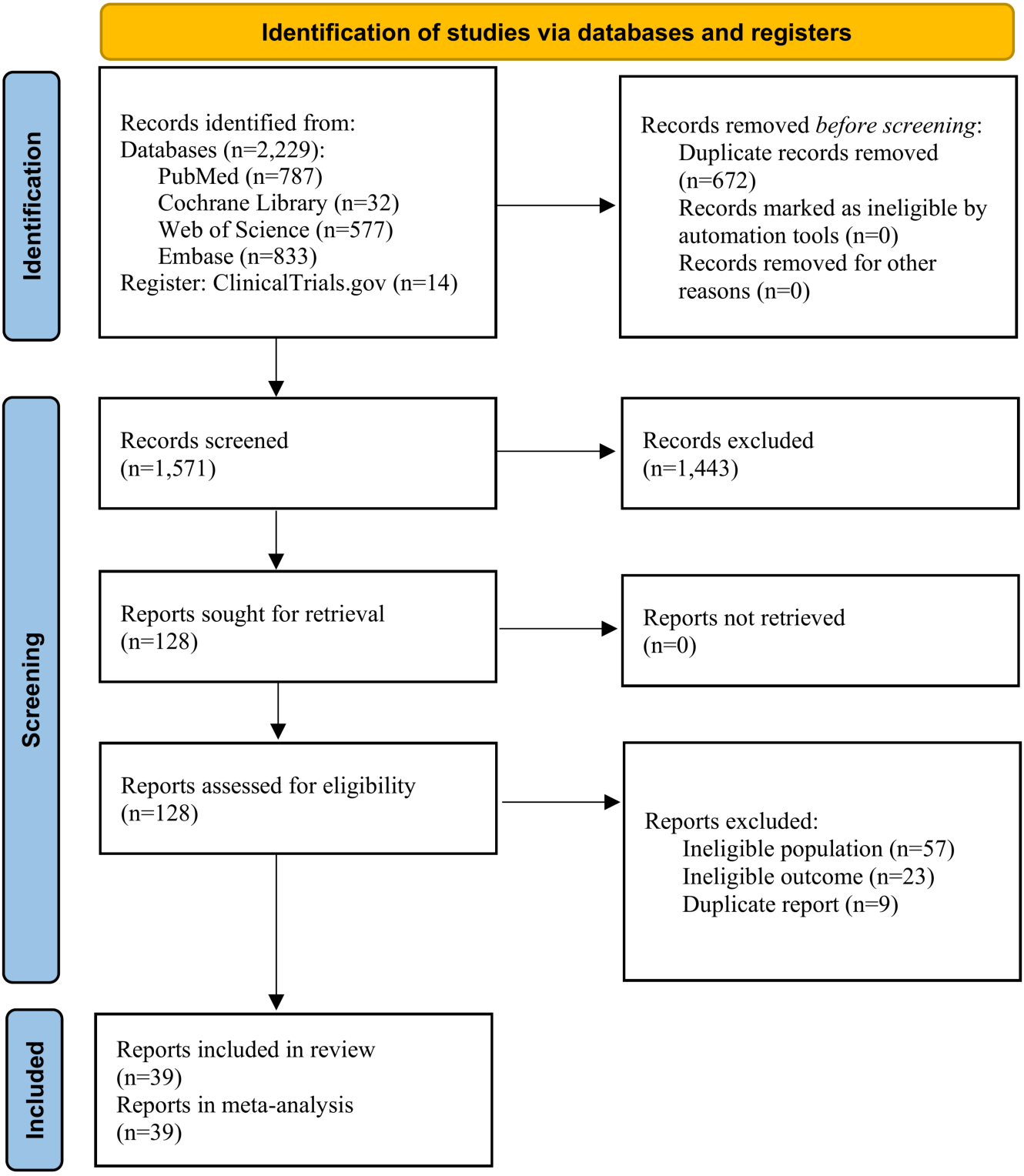

Figure 1

Literature search and selection flow diagram

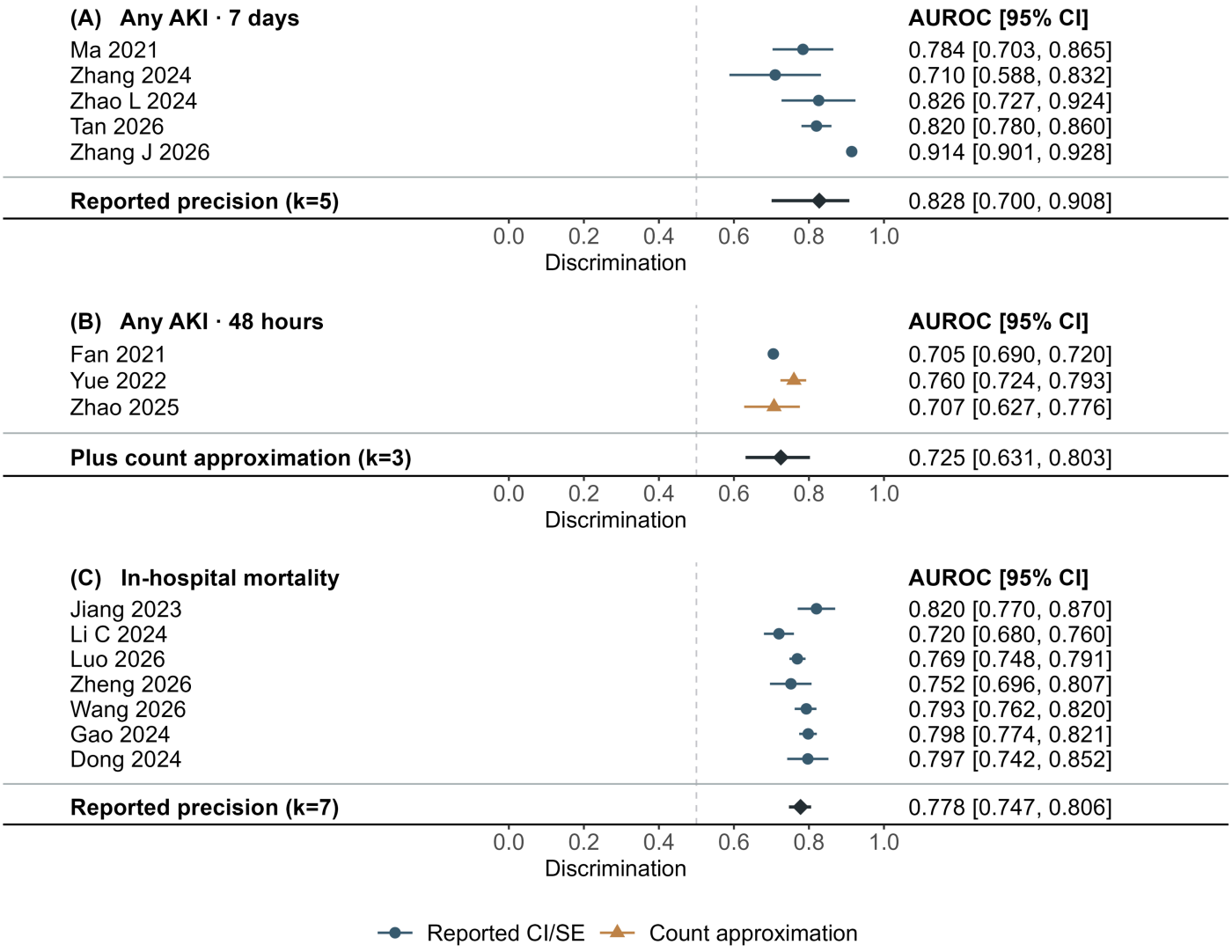

Figure 2

Validation discrimination for AKI and in-hospital mortality

Representative validation performance for (A) any AKI within 7 days, (B) any AKI within 48 hours, and (C) in-hospital mortality. Points and horizontal lines show report-level AUROCs and 95% CIs; diamonds show pooled estimates and 95% CIs. Circles denote reported precision and triangles denote precision approximated from event and non-event counts.

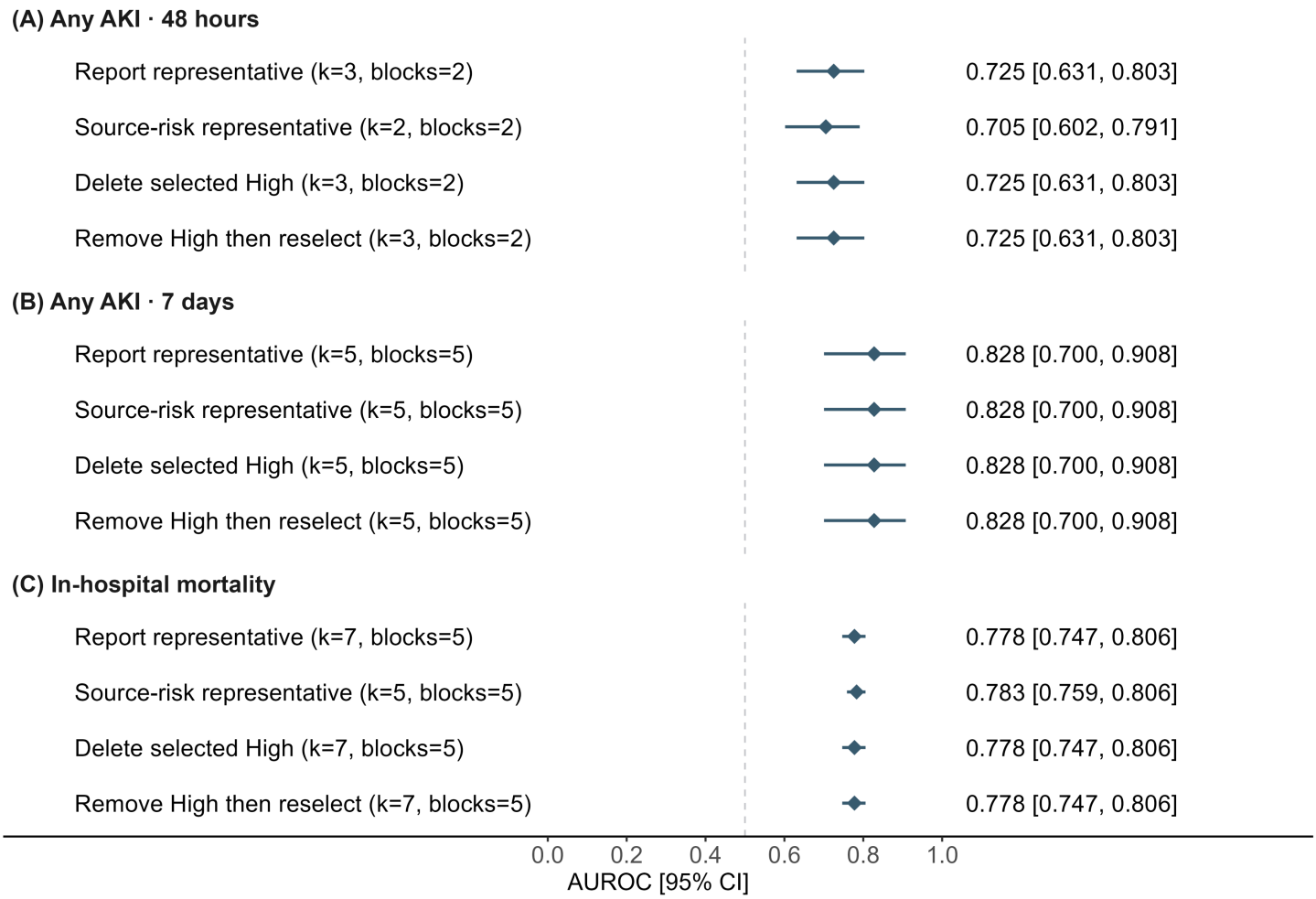

Figure 3

Sensitivity analyses of validation discrimination

Sensitivity analyses for (A) any AKI within 48 hours, (B) any AKI within 7 days, and (C) in-hospital mortality. Points and horizontal lines show pooled AUROCs and 95% CIs after restrictions based on potential patient-source overlap and risk of bias. Different restrictions may select different representative models. Unclear risk of bias is not classified as low risk.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1Supplementarymaterial.docx