Artificial Intelligence Models for Predicting Acute Kidney Injury in Adult Critical Care: A Systematic Review

Sonali Tripathi,[1] Jagdish Prasad Sunda[2]

[1]Department of Anesthesia, Chhindwara Institute of Medical Sciences, Chhindwara, Madhya Pradesh, India
[2]Deputy Director, DMHS, Jaipur, Rajasthan, India

Keywords. acute kidney injury, critical care, intensive care unit, artificial intelligence, machine learning, prediction model, calibration, decision curve analysis, AUROC, clinical utility

Introduction. Acute kidney injury (AKI) is frequent in adult critical care and is associated with increased morbidity, mortality, and long-term kidney consequences. Artificial intelligence (AI) and machine-learning (ML) prediction models may enable earlier risk identification, but clinical adoption depends on robust reporting beyond discrimination, including calibration and clinical utility.
Methods. We performed a PubMed-based systematic review (English; last 5 years) of AI/ML models predicting AKI in adult critical care/ICU populations. Records were screened and full texts were assessed for eligibility. For synthesis within the journal page limit, we restricted the final evidence set to studies reporting a complete usability-oriented outcome set: AKI definition, model type(s), prediction horizon, validation approach, AUROC/AUC, calibration, and decision curve analysis (DCA)/net benefit.
Results. Of 357 records screened, 230 full texts were assessed, and 31 studies met complete-reporting criteria and were included. The most common model families included logistic regression baselines, random forest, gradient boosting/XGBoost, and deep learning. Reported AUROC/AUC values ranged from 0.64 to 1.00 (median 0.90), with the best-performing models typically reporting AUROC ≥ 0.95. External or temporal validation was reported in 74% (23/31). By design, 100% of included studies reported calibration (e.g., calibration plot and/or Brier score and/or Hosmer–Lemeshow) and 100% reported clinical utility using DCA/net benefit.
Conclusions. Among usability-focused studies, AI/ML models show generally strong discrimination for AKI prediction in adult critical care, but reporting and validation practices remain heterogeneous. Standardized AKI definitions, transparent validation, calibration reporting, and decision-analytic evaluation are essential to support safe implementation.

RJCCN 2026; 2: 103-14
www.rjccn.org
DOI: 10.66224/rjccn.2.2.42

INTRODUCTION

Background and Rationale

Acute kidney injury (AKI) is one of the most frequent and consequential complications in adult critical care, affecting more than half of intensive care unit (ICU) patients in large multinational data when complete consensus criteria are applied, and it is strongly associated with higher mortality

Please cite this article as: Tripathi S, Prasad Sunda J. Artificial Intelligence Models for Predicting Acute Kidney Injury in Adult Critical Care: A Systematic Review. RJCCN. 2026;2(2):103-14.

and worse kidney function at hospital discharge.[1] Beyond short-term outcomes, AKI is increasingly recognized as a "gateway" event that can accelerate progression to chronic kidney disease (CKD) and contribute to long-term morbidity and mortality even after apparent recovery.[2] Because AKI often develops dynamically during evolving critical illness, clinicians need tools that can support earlier recognition, risk stratification, and timely preventive actions (e.g., hemodynamic optimization, nephrotoxin stewardship, fluid strategy, and monitoring intensity).

Prediction models are therefore clinically important in the ICU. They can help identify high-risk patients earlier than traditional rule-based triggers and enable more targeted monitoring and interventions. Over the last decade, the increasing availability of electronic health records (EHRs), bedside monitoring streams, and high-dimensional laboratory/physiologic data has accelerated the development of artificial intelligence (AI) and machine learning (ML) approaches for clinical risk prediction.[3] In AKI prediction specifically, ML methods may flexibly model nonlinear relationships, complex interactions, and time-varying patterns that are difficult to capture with conventional approaches, potentially improving performance in heterogeneous ICU populations.

Nevertheless, discrimination is not the sole measure that can be used to evaluate the clinical utility of prediction models. Albeit the area under the receiver operating characteristic curve (AUROC) is commonly reported, it does not tell whether the risks under prediction are well calibrated (i.e. whether the predicted risks approximate the observed rates) or whether a model offers better clinical decision-making at plausible risk levels.[4] Decision-analytic approaches, including decision curve analysis (DCA) and net benefit, make a direct relationship between model predictions and downstream clinical actions and harms/benefits, and give a more practice-relevant utility estimate.[5,6] Transparent reporting standards are also required to make prediction model research interpretable and reproducible; the TRIPOD statement includes the core reporting items necessary to assess the validity and applicability,[7] whereas risk-of-bias instruments like PROBAST can be used to conduct a structured assessment of internal validity and clinical relevance.[8]

The AI/ML AKI prediction in adult critical care has a scattered evidence base. The definitions of AKI (e.g., KDIGO vs older consensus systems), prediction horizons, patients, and the time of predictors measured are usually different across settings, making prediction cross-setting comparisons difficult.[9–11] Another method of inconsistent reporting that further complicates the work with the interpretation of model readiness to bedside use is inconsistent reporting (especially of calibration and clinical utility).

Study Objectives

The main aim of the presented systematic review is to provide a summary of AI/ML models to predict AKI in adult critical care and outline the discrimination performance (AUROC) in included studies. Secondary goals include defining AKI and predicting horizons, reporting validation strategies (internal or external) and measuring reporting of calibration and clinical utility (with DCA/net benefit where possible). Lastly, we attempt to find common areas of reporting deficits and comment on what this means to clinical adoption and standards of future research in critical care nephrology.

MATERIALS AND METHODS

Protocol and Reporting Standard

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement.[12] A review protocol was not registered in PROSPERO prior to study initiation.

Eligibility Criteria

Population. We included studies involving adults (≥ 18 years) managed in critical care settings, including ICU or explicitly described critically ill adult cohorts. Studies that exclusively enrolled pediatric or neonatal populations were excluded. Studies conducted entirely outside critical care settings (e.g., general wards only, outpatient settings, and elective non-critical perioperative cohorts without ICU-level illness) were excluded.

Prediction Task. Eligible studies developed and/or validated prediction models to estimate the risk

of AKI. We considered both prognostic prediction (future AKI within a defined horizon) and early-detection/near-term prediction approaches, but studies focusing solely on AKI diagnosis after meeting criteria (without a predictive intent) were excluded.

Model Types and Scope. We included models described as artificial intelligence, machine learning, deep learning, or related algorithmic prediction approaches, as well as statistical prediction models reported alongside or as baselines within AI/ML studies. Models could be derived from structured EHR data, physiologic time-series, laboratory data, or combined sources. Studies that did not present a prediction model (e.g., biomarker-only association studies without a predictive model) were excluded.

Study Designs. We included original research studies such as retrospective or prospective cohorts, registry analyses, database studies, and clinical trials that reported AKI prediction models in adult critical care. We excluded narrative reviews, systematic reviews/meta-analyses, editorials, commentaries, letters without original model results, conference abstracts without full text, and case reports/case series without predictive model development or validation.

Outcomes and "Complete Reporting Set." To ensure clinical interpretability and bedside relevance within the journal page limit, the final qualitative synthesis was restricted to studies that reported a complete usability-oriented outcome set ("complete reporting set"), defined a priori as:

1. an explicit AKI outcome definition (e.g., KDIGO, RIFLE, AKIN, or clearly operationalized creatinine/urine output criteria)
2. model type(s)
3. prediction horizon (time window for AKI prediction)
4. validation approach (internal and/or external/temporal)
5. discrimination (AUROC/AUC) plus
6. calibration reporting (e.g., calibration plot, Brier score, calibration slope/intercept, or equivalent)
7. Clinical utility reporting using decision curve analysis (DCA), net benefit, or equivalent decision-analytic assessment

Studies lacking one or more of these elements were considered eligible at full-text stage but were excluded from the final synthesis due to incomplete reporting.

Overlapping Cohorts. When multiple publications analyzed overlapping cohorts with similar modeling objectives, the most complete and/or most recent report with the most comprehensive outcome reporting was retained, and the others were excluded to avoid double counting.

Information Sources

PubMed was searched as the primary information source. The search was performed with filters for English language and publication in the last five years (last search date: 23/12/2025). No additional databases were searched.

Search Strategy

The search combined three concept blocks: 1) acute kidney injury, 2) critical care/ICU populations, and 3) artificial intelligence/machine learning prediction models. The complete PubMed search string is provided in Appendix.

Study Selection Process

All records retrieved from PubMed were imported into Zotero for management and screening. Duplicate detection was performed within Zotero, and duplicates were removed if identified. Screening proceeded in three stages:

1. Title and abstract screening to identify potentially relevant studies and prioritize full-text retrieval
2. Full-text assessment for topic eligibility (adult critical care population and AKI prediction model study)
3. Application of the "complete reporting set" filter to restrict the final synthesis to studies reporting discrimination, calibration, and decision-analytic clinical utility in addition to core model descriptors

Screening was conducted by a single reviewer using a structured decision framework and standardized exclusion reasons; internal consistency checks were performed by re-reviewing uncertain cases and verifying eligibility against predefined criteria.

Data Extraction

A standardized extraction framework was used to collect: bibliographic information (first author, year), population and setting descriptors, AKI definition, model type(s), prediction horizon, validation approach, discrimination (AUROC/AUC), calibration reporting (and metrics where extractable), and clinical utility reporting (DCA/net benefit). When studies reported multiple models, the best-performing or primary model (as described by authors) was captured, and key ranges were noted where relevant.

Missing or non-extractable items were recorded as not reported (NR). When studies reported outcomes using different denominators (e.g., development vs validation cohorts), values were extracted using the denominators as reported by the original authors. Zotero and spreadsheet-based templates were used to manage citations and extraction.

Risk of Bias Assessment

Prediction model studies are best assessed using the Prediction model Risk Of Bias Assessment Tool (PROBAST), which evaluates risk of bias and applicability across participant selection, predictors, outcomes, and analysis domains[8]. In this review, we used a structured PROBAST-aligned heuristic assessment to summarize risk of bias across the included studies, focusing primarily on outcome definition clarity, avoidance of data leakage, handling of missing data, validation strategy, and transparency of model evaluation. Overall risk of bias was categorized as low, unclear, or high based on the balance of concerns across domains.

Synthesis Approach

We conducted a narrative synthesis supported by tabulation of study characteristics, performance reporting, and risk of bias. Meta-analysis was not performed due to heterogeneity in patient case-mix, AKI definitions, prediction horizons, modeling methods, validation strategies, and reporting of calibration and decision-analytic outcomes.

RESULTS

Study Selection

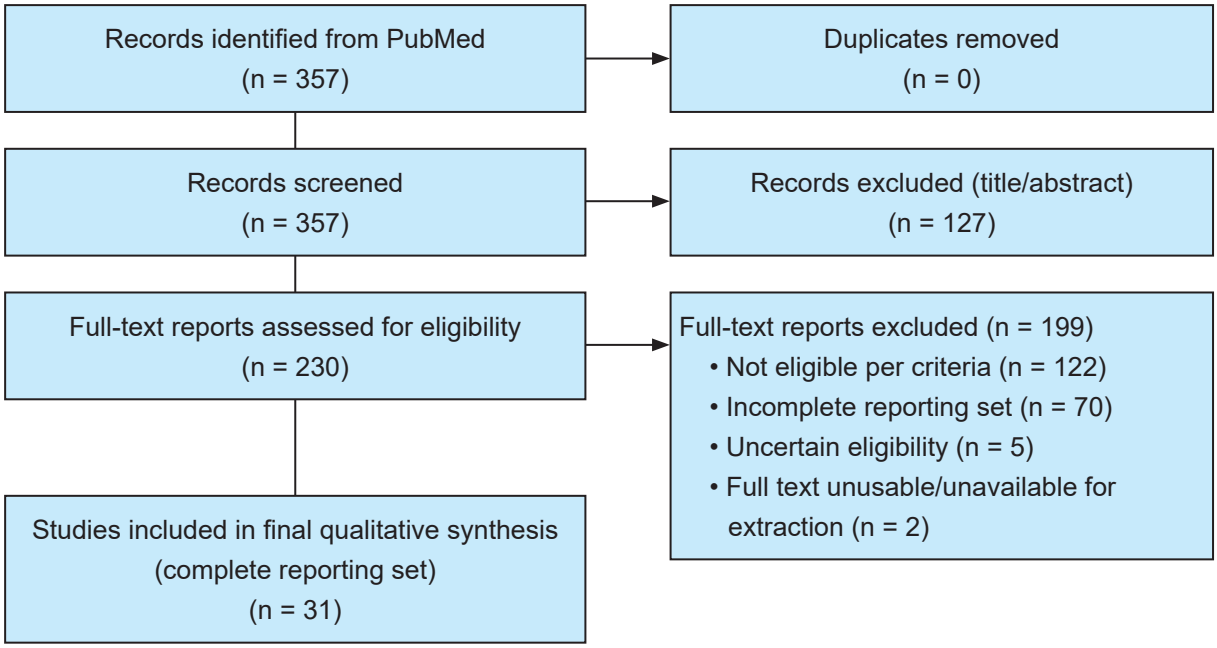
The PubMed search identified 357 records, all of which were screened at title/abstract (duplicates: 0). after title/abstract screening, 127 records were excluded, and 230 reports were sought and assessed in full text. Of these, 122 full-text reports were excluded for not meeting the review eligibility criteria, most commonly because they were not an AI/ML prediction model (n = 50), involved pediatric/neonatal populations (n = 25), were not original research (editorial/letter/protocol, n = 19), were not an AKI prediction model (n = 16), were reviews/meta-analyses (n = 7), were not ICU/critical care (n = 3), or did not use AKI as the outcome (n = 2). Five additional reports were classified as "maybe/uncertain" and set aside. This yielded 103 studies eligible at full-text decision, with 101 PDFs available for extraction.

To create a usability-focused synthesis within the journal page limit, we applied a predefined "complete reporting set" filter (core prediction descriptors plus calibration and decision-analytic utility). Seventy studies were excluded at this stage, most commonly because they lacked both calibration and clinical utility reporting (n = 41) or reported calibration without decision curve analysis/net benefit (n = 28); one study reported calibration and clinical utility but had non-extractable validation details (n = 1). Ultimately, 31 studies met the complete reporting set and were included in the final qualitative synthesis (Figure).

Characteristics of Included Studies

The final included set comprised 31 studies published between 2021 and 2025 (Table 1). Reporting of setting and cohort structure varied across studies; however, all were conducted in adult critical care/ICU populations and evaluated AI/ML prediction of AKI. AKI outcome definitions were most commonly KDIGO-based (27/31), while two studies used creatinine/urine output criteria without clearly naming a standard definition, and two studies referenced multiple/combined definitions (e.g., KDIGO with RIFLE/AKIN).

Prediction horizons were heterogeneous but most frequently short-term: 24-hour prediction windows were reported in 21/31 studies and 48-hour windows in 16/31; a smaller subset reported alternative horizons (e.g., 6 to 12 hours, 2 to 4 days, or longer windows such as 28 days). Validation



PRISMA Flow Diagram for Study Selection

approaches were mixed: external or temporal validation was reported in 23/31 (74%), while cross-validation/bootstrapping was reported in 25/31, and internal split/holdout validation in 16/31.

Model Performance and Reporting Quality

Across included studies, commonly represented model families included logistic regression baselines (30/31), random forests (28/31), gradient boosting/XGBoost variants (XGBoost: 26/31; boosting family: 24/31), and deep learning (21/31); several studies also evaluated SVM and decision-tree approaches.

Discrimination. All included studies reported AUROC/AUC by design (Table 2). The primary AUROC/AUC ranged from 0.64 to 1.00, with a median of 0.904 (interquartile range approximately 0.85 to 0.96). Reported discrimination commonly clustered in the high-0.8 to mid-0.9 range, with some studies reporting very high values (≥ 0.95) for selected models and horizons.

Validation. Although internal validation was common (cross-validation/bootstrapping and/or split-sample approaches), nearly three-quarters of included studies reported external or temporal validation (74%), supporting greater generalizability than development-only reports; nonetheless, the nature of external validation varied and was not always clearly described in extractable text.

Calibration. All included studies reported calibration in some form (a requirement of the complete reporting set). Among the included studies, Brier score reporting was detected in 11/31, and Hosmer–Lemeshow testing in 4/31; other calibration reporting was frequently described generically as "calibration" (e.g., calibration plots or narrative calibration statements).

Clinical Utility. By the complete reporting set definition, all 31 studies reported decision-analytic clinical utility (e.g., decision curve analysis and/or net benefit) to support evaluation of practical usefulness at clinically relevant thresholds.

Risk of Bias

Overall, risk-of-bias concerns were concentrated primarily in the analysis domain, while outcome definition concerns were comparatively lower within the included set (all studies reported AKI definitions sufficiently to satisfy the complete reporting filter) (Table 3). Common issues included the predominance of retrospective designs (risk of selection bias), incomplete clarity about the timing and availability of predictors relative to AKI onset

Table 1. Characteristics of Included Studies and Core Reporting Elements

First author	Year	Title	AKI definition	Prediction horizon	Model type(s)	Validation
Yue, Suru	2022	Machine learning for the prediction of acute kidney injury in patients with sepsis	KDIGO	24h/48h	DL, XGB, RF, GBM, SVM, LR, DT	CV/Boot
Wang, Yongbin	2024	Construction and evaluation of a mortality prediction model for patients with acute kidney injury undergoing continuous renal replacement therapy based on machine learning algorithms	KDIGO	24h/48h	DL, XGB, RF, GBM, LGBM, SVM, LR, DT	CV/Boot; Ext; Int
Li, Mingxia	2024	Machine Learning for Predicting Risk and Prognosis of Acute Kidney Disease in Critically Ill Elderly Patients During Hospitalization: Internet-Based and Interpretable Model Study	KDIGO	2d/48h/4d	XGB, RF, GBM, LGBM, LR	CV/Boot; Ext; Int
Yu, Shuangjiang	2025	A machine learning-based prediction model for sepsis-associated delirium in intensive care unit patients with sepsis-associated acute kidney injury	KDIGO	24h/3d	XGB, RF, GBM, SVM, LR, DT	CV/Boot; Ext; Int
Gottlieb, Eric R.	2022	Machine Learning for Acute Kidney Injury Prediction in the Intensive Care Unit	KDIGO	12h/24h/48h	DL, XGB, RF, GBM, SVM, LR, DT	Ext
Wang, Geng	2023	Machine learning-based models for predicting mortality and acute kidney injury in critical pulmonary embolism	KDIGO	24h/6h	XGB, RF, GBM, LR	CV/Boot; Ext; Int
Zuo, Fei	2025	Construction and validation of risk prediction models for renal replacement therapy in patients with acute pancreatitis	KDIGO	24h	DL, XGB, RF, GBM, SVM, LR	CV/Boot; Int
Li, Xunliang	2023	Machine learning algorithm to predict mortality in critically ill patients with sepsis-associated acute kidney injury	KDIGO	20h/24h/48h	DL, XGB, RF, GBM, SVM, LR, DT	Ext
Li, Xunliang	2024	Interpretable machine learning model for predicting acute kidney injury in critically ill patients	KDIGO	24h/48h/6h	DL, XGB, RF, GBM, SVM, LR, DT	Ext; Int
Jiang, Meng	2023	Explainable machine learning model for predicting furosemide responsiveness in patients with oliguric acute kidney injury	KDIGO	12h/18h/24h	DL, XGB, RF, SVM, LR, DT	CV/Boot; Int
Huang, Chun-Te	2023	Federated machine learning for predicting acute kidney injury in critically ill patients: a multicenter study in Taiwan	KDIGO; RIFLE	24h/3d/30h	DL, XGB, RF, GBM, LR	CV/Boot; Ext; Int; Temp
Xu, Yang	2025	Development and validation of a cardiac surgery-associated acute kidney injury prediction model using the MIMIC-IV database	KDIGO	0d/24h/48h	RF, LR	Int
Qin, Huayang	2025	Development of a machine learning-based prediction model for acute kidney injury associated with respiratory failure in the intensive care unit	KDIGO	28d/48h/6h	DL, XGB, RF, GBM, LGBM, SVM, LR	CV/Boot; Ext; Int
Zhong, Lei	2025	Construction and evaluation of prediction model for renal function recovery in acute kidney injury patients undergoing continuous renal replacement therapy based on machine learning algorithms	KDIGO	14d/24h/365d	DL, XGB, RF, GBM, LGBM, SVM, LR, DT	CV/Boot; Ext
Wang, Ruoran	2025	A machine learning predictive model for acute kidney injury among aneurysmal subarachnoid hemorrhage patients	KDIGO	2d/4d/48h	XGB, RF, GBM, SVM, LR	CV/Boot; Ext
Fan, Tingting	2023	Predicting the risk factors of diabetic ketoacidosis-associated acute kidney injury: A machine learning approach using XGBoost	KDIGO	24h	DL, XGB, GBM, SVM, LR	CV/Boot; Int

Table 1. Continued

First author	Year	Title	AKI definition	Prediction horizon	Model type(s)	Validation
Sun, Meina	2025	Predicting in-hospital mortality in patients with alcoholic cirrhosis complicated by severe acute kidney injury: development and validation of an explainable machine learning model	KDIGO	24h	DL, XGB, RF, GBM, LGBM, SVM, LR, DT	CV/Boot; Ext
Zhong, Lei	2024	Risk prediction models for successful discontinuation in acute kidney injury undergoing continuous renal replacement therapy	KDIGO	24h/365d/48h	DL, XGB, RF, GBM, LGBM, SVM, LR, DT	CV/Boot; Ext
Wang, Tsai-Jung	2024	Predictive approach for liberation from acute dialysis in ICU patients using interpretable machine learning	KDIGO	3d/30d/48h	DL, XGB, RF, GBM, LR	CV/Boot; Int
Li, Le	2024	Machine learning for the prediction of 1-year mortality in patients with sepsis-associated acute kidney injury	KDIGO	48h/4d/6h	XGB, RF, GBM, LGBM, CatB, LR, DT	Ext; Int
Zhang, Li	2025	Prediction of acute kidney injury in intensive care unit patients based on interpretable machine learning	KDIGO	24h	DL, XGB, GBM, LGBM, SVM, LR, DT	CV/Boot; Ext
Li, Mingxia	2022	Development and deployment of interpretable machine-learning model for predicting in-hospital mortality in elderly patients with acute kidney disease	KDIGO	48h/5d/5h	DL, XGB, RF, GBM, SVM, LR	CV/Boot; Ext
Dong, Lei	2024	Development and validation of a machine-learning model for predicting the risk of death in sepsis patients with acute kidney injury	KDIGO	24h/6h/72h	DL, XGB, RF, CatB, SVM, LR, DT	CV/Boot; Ext; Int
Liu, Xiaolong	2024	Machine learning-based model to predict severe acute kidney injury after total aortic arch replacement for acute type A aortic dissection	KDIGO; RIFLE; AKIN	2d/3d/48h	DL, RF, SVM, LR, DT	CV/Boot; Ext
Chang, Hsin-Hsiung	2022	Predicting Mortality Using Machine Learning Algorithms in Patients Who Require Renal Replacement Therapy in the Critical Care Unit	Creatinine/ UO criteria mentioned (unspecified)	1d/24h	DL, XGB, RF, LR, DT	CV/Boot; Ext
He, Jiawei	2021	Application of Machine Learning to Predict Acute Kidney Disease in Patients With Sepsis Associated Acute Kidney Injury	KDIGO	3d/48h/7d	DL, XGB, RF, GBM, LR, DT	CV/Boot; Ext
Cox, Eline G. M.	2023	External Validation of Mortality Prediction Models for Critical Illness Reveals Preserved Discrimination but Poor Calibration	Creatinine/ UO criteria mentioned (unspecified)	1h/2d/24h	RF, LR	Ext
Wang, Hongnian	2025	An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records	KDIGO	1h/2h/6d	DL, XGB, RF, GBM, LGBM, CatB, SVM, LR, DT	CV/Boot
Ji, Wenwen	2025	A machine learning model for predicting 28-day mortality in ICU patients with community-acquired pneumonia and acute kidney injury	KDIGO	24h/28d/6h	RF, GBM, SVM, LR	CV/Boot; Ext; Int
Ma, Xia	2025	From glycemic variability to digital signal biomarker: a prognostic and precision medicine framework for sepsis-associated acute kidney injury	KDIGO	24h/28d/48h	RF	CV/Boot
Luo, Tiangui	2025	Clinical impact and safety of continuous renal replacement therapy in critically ill patients with solid tumors and acute kidney injury: a retrospective cohort analysis	KDIGO	0h/12h/18h	XGB, GBM, LR	CV/Boot; Ext; Int

Abbreviations: AKI, acute kidney injury; AUROC, area under the receiver operating characteristic curve; NR, not reported; DL, deep learning; XGB, XGBoost; RF, random forest; LR, logistic regression; Ext, external validation; Temp, temporal validation; CV/Boot, cross-validation/ bootstrapping; Int, internal split/holdout.

Table 2. Discrimination Performance of Included Models (AUROC/AUC) and Key Notes

First author	Year	AUROC/AUC (primary)	AUROC/AUC (range)	Calibration (reported)
Yue, Suru	2022	0.943	0.600 to 0.943	Calibration mentioned
Wang, Yongbin	2024	0.954	0.504 to 0.954	Brier score; Calibration mentioned
Li, Mingxia	2024	0.963	0.500 to 0.963	Calibration mentioned
Yu, Shuangjiang	2025	0.91	0.500 to 0.910	Calibration mentioned
Gottlieb, Eric R.	2022	0.89	0.660 to 0.890	Calibration mentioned
Wang, Geng	2023	0.904	0.586 to 0.904	Calibration mentioned
Zuo, Fei	2025	1.0	0.715 to 1.000	Brier score; Calibration mentioned
Li, Xunliang	2023	0.832	0.572 to 0.832	Calibration mentioned
Li, Xunliang	2024	0.824	0.630 to 0.824	Calibration mentioned
Jiang, Meng	2023	1.0	0.610 to 1.000	Brier score; Calibration mentioned
Huang, Chun-Te	2023	0.977	0.547 to 0.977	Calibration mentioned
Xu, Yang	2025	0.755	0.700 to 0.755	Calibration mentioned; Hosmer–Lemeshow
Qin, Huayang	2025	0.89	0.890 to 0.890	Brier score; Calibration mentioned
Zhong, Lei	2025	0.915	0.685 to 0.915	Brier score; Calibration mentioned
Wang, Ruoran	2025	1.0	0.500 to 1.000	Calibration mentioned
Fan, Tingting	2023	0.835	0.518 to 0.835	Calibration mentioned
Sun, Meina	2025	0.903	0.580 to 0.903	Brier score; Calibration mentioned
Zhong, Lei	2024	0.902	0.638 to 0.902	Brier score; Calibration mentioned
Wang, Tsai-Jung	2024	0.95	0.700 to 0.950	Brier score; Calibration mentioned
Li, Le	2024	0.964	0.565 to 0.964	Brier score; Calibration mentioned
Zhang, Li	2025	0.64	0.543 to 0.640	Calibration mentioned
Li, Mingxia	2022	0.921	0.527 to 0.921	Calibration mentioned
Dong, Lei	2024	0.941	0.511 to 0.941	Calibration mentioned
Liu, Xiaolong	2024	0.963	0.734 to 0.963	Calibration mentioned
Chang, Hsin-Hsiung	2022	0.854	0.514 to 0.854	Calibration mentioned; Hosmer–Lemeshow
He, Jiawei	2021	1.0	0.500 to 1.000	Calibration mentioned
Cox, Eline G. M.	2023	0.88	0.500 to 0.880	Brier score; Calibration mentioned
Wang, Hongnian	2025	0.821	0.553 to 0.821	Calibration mentioned; Hosmer–Lemeshow
Ji, Wenwen	2025	0.755	0.568 to 0.755	Calibration mentioned
Ma, Xia	2025	0.845	0.578 to 0.845	Brier score; Calibration mentioned
Luo, Tiangui	2025	0.86	0.600 to 0.860	Calibration mentioned; Hosmer–Lemeshow

Note. Evidence snippets are brief text fragments indicating where AUROC/AUC values were reported in the PDF.

(risk of "data leakage"), and inconsistent reporting of missing data handling and model calibration procedures beyond brief mentions. Although external validation was present in a substantial proportion of included studies, validation quality and transportability were variably described, and few studies provided enough detail to fully judge implementation readiness.

In the PROBAST-aligned heuristic summary, 24/31 studies were judged moderate/unclear overall and 7/31 were judged high risk of bias, largely driven by analysis-related limitations.

DISCUSSION

Principal Findings

In this systematic review of AI/ML prediction models for AKI in adult critical care, we deliberately focused the final synthesis on studies that reported a predefined "complete reporting set" that included not only discrimination (AUROC/AUC) but also calibration and decision-analytic clinical utility (DCA/net benefit). Within this higher-reporting subset, discrimination was generally favorable (AUROC/AUC range: 0.64 to 1.00), suggesting that many models can separate patients at higher versus lower risk of AKI under study conditions. However, our screening also highlighted an important reality: even within a rapidly growing literature, only a minority of full texts ultimately provided the combination of outcome definition, horizon anchoring, validation detail, calibration, and utility evaluation needed to judge bedside readiness. External or temporal validation was present in a substantial proportion of included

Table 3. Risk of Bias Assessment Summary (Prediction-Model Heuristic / PROBAST-aligned)

First author	Year	Analysis	Overall (heuristic)
Yue, Suru	2022	High	High
Wang, Yongbin	2024	Moderate	Moderate
Li, Mingxia	2024	Moderate	Moderate
Yu, Shuangjiang	2025	Moderate	Moderate
Gottlieb, Eric R.	2022	Moderate	Moderate
Wang, Geng	2023	Moderate	Moderate
Zuo, Fei	2025	High	High
Li, Xunliang	2023	Moderate	Moderate
Li, Xunliang	2024	Moderate	Moderate
Jiang, Meng	2023	High	High
Huang, Chun-Te	2023	Moderate	Moderate
Xu, Yang	2025	Unclear	Moderate
Qin, Huayang	2025	Moderate	Moderate
Zhong, Lei	2025	Moderate	Moderate
Wang, Ruoran	2025	Moderate	Moderate
Fan, Tingting	2023	High	High
Sun, Meina	2025	Moderate	Moderate
Zhong, Lei	2024	Moderate	Moderate
Wang, Tsai-Jung	2024	High	High
Li, Le	2024	Moderate	Moderate
Zhang, Li	2025	Moderate	Moderate
Li, Mingxia	2022	Moderate	Moderate
Dong, Lei	2024	Moderate	Moderate
Liu, Xiaolong	2024	Moderate	Moderate
Chang, Hsin-Hsiung	2022	Moderate	Moderate
He, Jiawei	2021	Moderate	Moderate
Cox, Eline G. M.	2023	Moderate	Moderate
Wang, Hongnian	2025	High	High
Ji, Wenwen	2025	Moderate	Moderate
Ma, Xia	2025	High	High
Luo, Tiangui	2025	Moderate	Moderate

Note. Judgments are PROBAST-aligned heuristics and can be finalized with full PROBAST item-level review if required.

studies, yet it remained inconsistent across the broader evidence base and often lacked enough detail to confidently infer transportability across ICUs and health systems. These findings align with long-standing concerns that prediction-model literature can overemphasize discrimination while underreporting calibration, clinical utility, and implementation-facing details.[7,8]

Interpretation and Clinical Meaning

Discrimination metrics such as AUROC are useful but incomplete. AUROC does not tell clinicians whether predicted probabilities are *numerically accurate* at clinically actionable thresholds, which is essential when model output is used to trigger interventions.[4] Calibration addresses this gap: a model with acceptable AUROC may still systematically overestimate or underestimate risk, leading to alarm fatigue, missed opportunities for prevention, or inappropriate escalation. For ICU deployment—where AKI risk evolves quickly—calibration should ideally be assessed with calibration plots and quantitative measures (e.g., Brier score, calibration slope and intercept), and recalibration should be reported when models are transported to new settings.[4]

Decision curve analysis and net benefit further advance the evaluation by connecting predictions to clinical consequences across a range of thresholds, which is more aligned with ICU decision-making than AUROC alone.[5,6] DCA can clarify whether a model is likely to improve decision-making compared with "treat all" or "treat none" strategies, especially when the downstream action (e.g., enhanced monitoring, nephrology consultation, hemodynamic optimization, avoidance of nephrotoxins) carries costs and potential harms. In practice, DCA reporting may also reveal that models with similar AUROC values can differ substantially in utility depending on calibration, event rates, and threshold selection.

Heterogeneity in AKI definitions and prediction horizons remains a major barrier to cross-study comparison and implementation. KDIGO provides standardized criteria and is the most widely used contemporary definition,[9] yet older systems (RIFLE, AKIN) and operational variants persist, and studies may differ in whether they use creatinine, urine output, or both.[10,11] Similarly, prediction horizons vary widely (e.g., 6 to 12 h, 24 to 48 h, several days), and horizon definitions are often poorly anchored to clinically meaningful time-zero events (ICU admission, initiation of vasopressors, onset of sepsis, etc.). These differences influence event prevalence and "difficulty" of prediction, which can inflate or depress AUROC and complicate benchmarking.

Finally, several mechanisms can explain why models may perform well during development yet degrade in deployment: dataset shift (differences in case-mix, measurement practices, lab ordering, and AKI ascertainment), temporal drift (changes in ICU protocols), missing-data patterns, and

unintentional leakage when predictors include information recorded after the true prediction time[8]. These issues are particularly relevant in ICU EHR-based modeling where timestamp alignment and predictor availability can be ambiguous.

Clinical Implications

The results of our study would be useful in guiding future studies of AKI prediction models in adult critical care. Originally, the investigators are expected to standardize AKI endpoints by KDIGO and report the presence or absence of creatinine and urine output criteria, including definitions of baseline creatinine and ascertainment windows.[9] Second, the prediction horizons are to be set with pre-defined and clearly anchored time-even predictors to a clinically interpretable time-zero, and clear statements made concerning the availability of the predictors at the prediction time. Third, Calibration reporting ought to be standard practice and it should contain calibration plots and quantitative measures (e.g., Brier score, slope/intercept) as well as recalibration strategies in cases where models are externally validated or transported.[4] Fourth, external or temporal validation is an aspect that should be held with a minimum expectation of any claims of generalizability, and validation cohorts must be described comprehensively enough to conclude about comparability and transportability.[7,8] Lastly, decision-analytic assessment (DCA/net benefit) is to be included to show that the model enhances decision-making within reasonable boundaries, rather than just that it predicts AKI.[5]

Clinically, critical care nephrology teams are able to view the existing AI/ML AKI models as potentially clinical risk stratification tools, although this should be approached with great caution. The models to be implemented should be based on externally-validated models, clearly calibrated, and tested on clinical utility in similar ICUs. Local piloting with observable performance drift and recalibration requirements are essential, since different centers differ both in terms of ICU population and practice.

Research Gaps and Future Directions

Key priorities include prospective impact studies that test whether model-guided interventions improve patient-centered outcomes (AKI severity, renal recovery, need for dialysis, ICU length of stay, mortality) rather than only predictive performance. Multicenter evaluations with transparent transportability analyses are needed to address heterogeneity across ICUs. Reporting adherence should be strengthened using emerging extensions and frameworks for prediction models (e.g., TRIPOD-AI and PROBAST-AI concepts), and investigators should more consistently report calibration, decision-analytic utility, and model updating/recalibration strategies.[7,8] Finally, fairness and subgroup performance assessment (by age, sex, comorbidity burden, and ethnicity where available) should be reported because AKI risk and care processes are not uniform across populations, and inequitable model performance could amplify disparities.

Strengths and Limitations

This review has several strengths. We used systematic methods, focused specifically on adult critical care, and applied a usability-oriented synthesis strategy to emphasize evidence that is closest to implementation needs. The compact evidence set supports a clearer narrative and tables suitable for journal page limits. Limitations include reliance on a single database (PubMed), which may miss some engineering- or conference-focused studies, and potential publication bias toward models with favorable performance. Our decision to apply a complete reporting filter improves interpretability but excludes many otherwise eligible studies; therefore, our synthesis should be interpreted as describing the best-reported and most implementation-relevant segment of the literature rather than the entire field. We did not perform meta-analysis due to heterogeneity in populations, horizons, and reporting. Finally, risk-of-bias assessment was summarized using a PROBAST-aligned heuristic approach rather than full PROBAST item-level adjudication, which should be considered when interpreting the certainty of evidence.

CONCLUSION

Artificial intelligence and machine-learning

prediction models for AKI in adult critical care demonstrate generally promising discrimination, but their readiness for bedside adoption depends on more than AUROC alone. Across the broader literature, incomplete reporting of prediction horizons, validation strategies, calibration, and clinical utility limits interpretability and slows translation into clinical workflows. In a usability-focused subset of studies that reported discrimination, calibration, and decision-analytic utility, model performance was often strong; however, heterogeneity in AKI definitions and outcome windows, variable external validation practices, and analysis-domain risks of bias remain important barriers to implementation. Future ICU AKI prediction research should standardize AKI endpoints (preferably KDIGO), clearly anchor prediction time and horizon, routinely report calibration with quantitative measures and recalibration strategies, and prioritize external/temporal validation alongside decision-analytic evaluation (DCA/net benefit). These steps—ideally paired with prospective impact studies—are essential for moving from promising algorithms to safe, clinically meaningful decision support in critical care nephrology.

ACKNOWLEDGEMENT

Full PubMed Search Strategy

Database: PubMed (National Library of Medicine)

Last search date: [23/12/2025]

Filters applied: in the last 5 years, English.

PubMed search string:

("Acute Kidney Injury"[Mesh] OR "acute kidney injury"[tiab] OR AKI[tiab]) AND

("Intensive Care Units"[Mesh] OR ICU[tiab] OR "critical care"[tiab] OR "critically ill"[tiab]) AND ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh]

OR "artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab]

OR "neural network*"[tiab] OR "algorithm*"[tiab] OR "prediction model*"[tiab])

AI Tools

Artificial intelligence tools (ChatGPT) were used to support language refinement, organization of tables, and improvement of manuscript readability. No AI tool was used for data analysis, study selection, or interpretation of results. All scientific decisions and conclusions were made by the authors.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

FUNDING

No external funding was received.

REFERENCES

1. Hoste EAJ, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. Intensive Care Med. 2015;41(8):1411–23.
2. Chawla LS, Eggers PW, Star RA, Kimmel PL. Acute Kidney Injury and Chronic Kidney Disease as Interconnected Syndromes. N Engl J Med. 2014;371(1):58–66.
3. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ digital medicine. 2018;1(1):18.
4. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38.
5. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. bmj [Internet]. 2016 [cited 2025 Dec 25];352. Available from: https://www.bmj.com/content/352/bmj.i6.abstract.
6. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Making. 2006;26(6):565–74.
7. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery. 2015;102(3):148–58.
8. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51–8.
9. KDIGO. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl. 2012;2:1.
10. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care. 2007;11(2):R31.
11. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, the ADQI workgroup. Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care. 2004;8(4):R204.

12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj [Internet]. 2021 [cited 2025 Dec 25];372. Available from: https://www.bmj.com/content/372/bmj.n71.short.

Correspondence to:
Jagdish Prasad Sunda, MD
Deputy Director, DMHS, Jaipur, Rajasthan, India
ORCID: 0009-0000-0405-0185
E-mail: dr.jpsunda@gmail.com

Received January 2026
Revised March 2026
Accepted March 2026