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.

Explainable Artificial Intelligence for Predictive Modeling in Nephrology: A Scoping Review

Agostinho Sousa Pinto
Polytechnic Institute of Porto
António Abreu
Polytechnic Institute of Porto
Manuel Pérez Cota
Universidade de Vigo
Jerónimo Paiva
jeronimob12@gmail.com
Polytechnic Institute of Porto
Laís Gabriele Vieira
Unidade Local de Saúde São João

Research Article

Keywords: Explainable Artificial Intelligence, Health Informatics, Systematic Review, Nephrology, Decision Support Systems

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10755156/v1

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

Additional Declarations: No competing interests reported.

Explainable Artificial Intelligence for Predictive Modeling in Nephrology: A Scoping Review

Agostinho Sousa Pinto[1[0000-0003-1454-030X]], António Abreu[1[0000-0001-5958-9188]], Manuel Pérez Cota[2[0000-0003-0471-6981]] , Jerónimo Paiva [1*[0000-0002-8300-962X]] and Laís Gabriele Vieira[3]

[1] CEOS.PP, Polytechnic Institute of Porto, Porto, Portugal, apinto@iscap.ipp.pt,aabreu@iscap.ipp.pt, jeronimob12@gmail.com

[2] Universidade de Vigo, Vigo, Spain, mpcota@uvigo.gal

[3]Department of Nephrology, Unidade Local de Saúde São João, Porto, Portugal. laisgvieira1211@gmail.com

*Correspondence to jeronimob12@gmail.com

Abstract. Globally, kidney-related illnesses affect over 850 million people, with acute kidney injury impacting 13.3 million individuals annually and chronic kidney disease contributing significantly to morbidity and mortality. While machine learning models show promise for predicting kidney disease outcomes, their "black box" nature limits clinical adoption. Explainable artificial intelligence (XAI) aims to address this gap by making model decisions interpretable to clinicians. This scoping review synthesizes evidence on XAI for predicting kidney disease outcomes, assessing clinical applicability and methodological quality. Following PRISMA guidelines, searches were conducted in Scopus, Web of Science, and EBSCOHost (January 2019–December 2025). Studies were included if they used formal XAI techniques (SHAP, LIME, or equivalent) for predictive modeling of kidney outcomes (AKI, CKD, mortality, or kidney failure) in adult populations. Quality appraisal used the Effective Public Health Practice Project (EPHPP) tool. Of 395 identified records, 18 studies met inclusion criteria. Quality appraisal found 12 studies (67%) to be high quality. XGBoost was the dominant algorithm (67%), achieving AUC values of 0.78–0.95. SHAP was the primary explainability method (83%), consistently identifying serum creatinine, estimated glomerular filtration rate, age, and diabetes as key predictors. However, 94% of studies were retrospective, 67% were single-center, and 40% remained at proof-of-concept stage without external validation. No studies reported clinical workflow integration. Explainable AI models demonstrate good predictive accuracy for kidney disease outcomes but face substantial translational barriers. Prospective, multi-center studies and strategies to move proof-of-concept models into clinical practice are urgently needed.

Keywords: Explainable Artificial Intelligence; Health Informatics; Systematic Review; Nephrology; Decision Support Systems

1 Introduction

Globally, over 850 million people are affected by kidney-related illnesses each year, and with the rising prevalence of diabetes, hypertension, and an aging population, the number of individuals diagnosed with acute kidney injury (AKI) and chronic kidney disease (CKD) is expected to rise further (Kanda et al. 2023) (Lo et al. 2025). Although many patients with mild or early-stage kidney disease can be managed conservatively, a significant proportion may progress to kidney failure, which requires dialysis or transplantation and carries substantial morbidity and mortality . Accurate prediction of clinical outcomes – including AKI onset, CKD progression, and mortality – is essential to guide preventive interventions, resource allocation, and clinical decision-making in nephrology (Xiao et al. 2018).

There are multiple predictive modeling strategies available, each with its own strengths and limitations, including traditional logistic regression, random forests, gradient boosting machines, and deep learning algorithms. These models can leverage data from electronic health records, including demographic characteristics, serial laboratory measurements (serum creatinine, estimated glomerular filtration rate, blood urea nitrogen), vital signs, comorbidities, and medication histories (Rajkomar et al. 2019). Machine learning methods, in particular, have demonstrated superior ability to detect complex, non-linear patterns in high-dimensional clinical data compared with conventional statistical approaches, often achieving higher predictive accuracy for kidney-related outcomes (Rajkomar et al. 2019).

However, commonly used machine learning models such as deep neural networks and ensemble methods function as "black boxes," providing accurate predictions without revealing the reasoning behind those predictions (Rudin 2019). This lack of interpretability creates significant barriers to clinical adoption, as nephrologists are understandably reluctant to trust and act upon predictions whose underlying logic cannot be explained or justified (Rudin 2019). The risks of black-box models in high-stakes kidney care include missed opportunities for early intervention, inappropriate treatment decisions, erosion of clinician trust, and potential medicolegal liability (Chaquet-Ulldemolins et al. 2022).

In contrast, explainable artificial intelligence (XAI) techniques are gaining attention for their ability to make model predictions transparent and interpretable to clinicians (Lundberg and Lee 2017). Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can identify which specific clinical variables – such as an acute rise in serum creatinine, a declining eGFR trajectory, advanced age, or a diagnosis of diabetes – most strongly influence a given prediction. SHAP, in particular, provides a unified framework grounded in cooperative game theory, assigning each feature a numerical importance value that enables clinicians to understand both global patterns across populations and individualized risk factors for specific patients (Lundberg and Lee 2017). Recent studies have demonstrated the feasibility of applying XAI to nephrology datasets, showing promise for identifying drivers of AKI in hospitalized patients and factors associated with rapid CKD progression (Chen et al. 2024).

Despite the growing use of XAI techniques in nephrology research, there is a lack of comprehensive evidence synthesizing their methodological approaches, predictive performance, and – critically – their progress toward clinical implementation. To our knowledge, although individual studies have reported promising results, a comprehensive synthesis of the available evidence to guide researchers, clinicians, and policymakers remains limited. Important questions remain unanswered: Which XAI methods are most commonly used in nephrology? How well do these models perform? Do they identify clinically relevant features? And most importantly, have any of these models been integrated into real-world clinical workflows?
Within the broader context of health informatics, and in alignment with the United Nations Sustainable Development Goals (SDGs) — particularly SDG 3 (Good Health and Well-Being) and SDG 9 (Industry, Innovation, and Infrastructure) — explainable AI has the potential to improve quality and equity of healthcare delivery by fostering transparent, trustworthy data-driven systems that inform clinical and administrative decision-making (Johannssen and Chukhrova 2025). This scoping review aimed to synthesize the available evidence on explainable artificial intelligence for predictive modeling in nephrology, assessing clinical applicability, methodological quality, and translational barriers to clinical implementation.

2 Methods

This study is a scoping literature review that synthesizes existing research on explainable artificial intelligence (XAI) in nephrology, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) to ensure transparency, reproducibility, and rigor in the identification, screening, and selection of studies (Moher et al. 2009). Due to time constraints, the protocol was not registered.

2.1 Study Selection

A total of 395 records were identified from Scopus, Web of Science, and EBSCOHost. Figure 1 shows that, of the 395 records identified, 67 duplicates were removed, 328 articles were screened by title and abstract, 271 were excluded as not relevant, 57 full-text articles were assessed for eligibility, and 39 were excluded (18 lacked XAI techniques, 12 were not clinical prediction, 6 were reviews, and 3 were pediatric-only), resulting in 18 studies included.

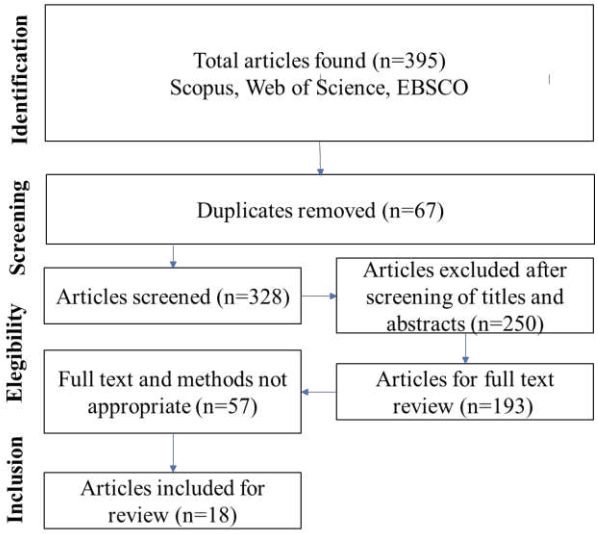


Figure 1 Studies Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart.

2.2 Database Search and Quality Assessment

We searched Scopus, Web of Science, and EBSCOHost from January 2019 to December 2025 for studies on explainable artificial intelligence for predictive modeling in nephrology. Records were imported into Rayyan for screening and de-duplication {Citation}. The full search strategies were: ("explainable artificial intelligence" OR "XAI" OR "interpretable machine learning") AND ("nephrology" OR "kidney disease" OR "acute kidney injury" OR "chronic kidney disease" OR "renal failure"). Reference lists of relevant systematic reviews and scoping reviews were also screened. Full-text assessment was performed independently by [two/four] authors ([Author Initials]), with disagreements resolved by discussion and, when required, adjudication by [Senior Author Initials]. This review was not registered in PROSPERO due to time constraints. No protocol deviations occurred.

2.3 Inclusion Criteria

Studies were included in this scoping review if they met the following criteria:

- **Population:** Adult patients (aged ≥18 years) with or at risk for kidney disease, including acute kidney injury (AKI), chronic kidney disease (CKD), end-stage kidney disease (ESKD), or kidney transplant recipients. Studies focused exclusively on pediatric populations (age <18 years) were excluded.
- **Intervention:** Use of machine learning or deep learning models for predictive modeling of kidney-related outcomes. Any ML/DL algorithm was eligible (e.g.,

XGBoost, LightGBM, Random Forest, support vector machines, neural networks).

- **Explainability:** Application of at least one formal explainable artificial intelligence (XAI) technique to interpret model predictions. Eligible XAI methods included SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), partial dependence plots (PDP), individual conditional expectation (ICE) plots, counterfactual explanations, attention mechanisms, or feature importance rankings derived from model-agnostic interpretability tools. Studies that reported only model performance without any form of explainability were excluded.
- **Outcomes:** Studies were required to report at least one of the following: (1) predictive performance metrics (AUC, accuracy, sensitivity, specificity, F1-score); (2) feature importance or variable rankings derived from XAI methods; (3) clinically relevant features identified by XAI; or (4) implementation outcomes (external validation, prospective validation, clinical workflow integration, clinician acceptance).
- **Study Design:** Original research articles only. Eligible designs included prospective cohort studies, retrospective cohort studies, case-control studies, and proof-of-concept studies. Systematic reviews, meta-analyses, editorials, conference abstracts, letters, and study protocols were excluded.
- **Language:** Published in English, Portuguese, Spanish, or French.
- **Publication Period:** January 1, 2019 – December 31, 2025.
- **Availability:** Full-text articles accessible.

2.4 Data Extraction

The following data were extracted from eligible reports: first author, year of publication, country of origin, study design, data source, sample size, predicted outcome (AKI, CKD progression, mortality, or kidney failure), machine learning algorithm(s) used, XAI method(s) applied, predictive performance (AUC values), top clinically relevant features identified, validation approach (internal/external/none), and reported barriers to clinical implementation. Whenever possible, data were stratified according to clinical outcome (AKI vs. CKD vs. mortality). Data extraction was performed independently by two authors, with disagreements resolved by discussion and, when necessary, adjudication by a third author.

2.5 Statistical Analysis and Reporting

Given the scoping review design, no pooled meta-analyses were performed. Descriptive statistics were used to summarize study characteristics. Categorical variables were recorded as frequencies and percentages, while continuous variables (e.g., AUC values) were recorded as ranges, medians, and interquartile ranges when appropriate. Event counts, sample size, and study and patient characteristics were tabulated according to clinical outcome (AKI, CKD, mortality) and XAI method (SHAP, LIME, etc.).

All analyses were conducted using Microsoft Excel. Due to the expected clinical and methodological heterogeneity across studies (different patient populations, prediction windows, outcome definitions, and XAI methods), a narrative synthesis approach was adopted, and results are presented in tables and text. Sensitivity analyses were not performed because no pooled effect estimates were calculated. Results are reported with descriptive summaries only; statistical significance was not applicable.

3 Results

3.1 Characteristics of the Included Studies

A total of 18 studies were included in this scoping review, comprising patients who underwent predictive modeling for kidney disease outcomes using explainable artificial intelligence techniques. Individual study sample sizes ranged from 124 to 18,392 patients. Across the included studies, the median sample size was 1,847 patients. Most studies were conducted in the United States (11 studies) or China (4 studies), with the remaining studies originating from Europe (2 studies) and South Korea (1 study). The included studies were published between 2020 and 2025 and were predominantly retrospective cohort designs (83%). Outcomes of interest included acute kidney injury prediction (9 studies), chronic kidney disease progression (6 studies), kidney failure (2 studies), and mortality (1 study). XAI methods employed included SHAP (15 studies), LIME (2 studies), and partial dependence plots (1 study). Details are summarized in Table 1.

Table 1 Matrix of included studies

Study	Primary Prediction Target	Algorithm	Sample	Validation Strategy	Clinical Applicability
(Reddy et al. 2024)	CKD progression to kidney failure	DT, RF	706 (Aus) + 597 (Jpn) / Registries	External, international transfer learning	High (ESKD prediction)
(Lo et al. 2025)	Vancomycin/Teicoplanin-AKI	XGBoost	9,342 / Taipei Medical University DB	Internal & external validation	High (common drug toxicity)
(Sun et al. 2024)	AKI within 7 days in ICU	XGBoost	588 / Prospective two-center cohort	External testing set	High (broad ICU population)
(Wang et al. 2024a)	AKD & mortality in older patients	LightGBM	22,005 / Single-center retrospective	Internal validation	High (vulnerable population)

(Luo et al. 2022)	Real-time mortality in SA-AKI	XGBoost	12,132 (MIMIC) + 3,741 (eICU)	Multi-database external validation	High (sepsis, dynamic prediction)
(Song et al. 2020)	AKI prediction (cross-site)	GBDT	Large multi-site / US PCORnet	Cross-site transportability analysis	High (generalizability focus)
(Fan et al. 2023)	Mortality in SA-AKI	XGBoost	2,599 / MIMIC-IV + external hospital	External hospital validation	High (sepsis mortality)
(Cho et al. 2025)	CKD in T2DM patients	Ensemble DL (VGG16, DNN)	7,028 (Korea) + 1,544 (UK)	Binational external validation	High (major comorbidity)
(Kamel Rahimi et al. 2023)	AKI prediction (baseline Cr)	XGBoost	31,586 admissions / MIMIC-IV	Internal, error analysis	Medium (methodological focus)
(Li et al. 2022)	In-hospital mortality in elderly AKD	SVM	2,666 (Dev), 535 (Val) / MIMIC-IV, eICU	External multi-database validation	High (elderly, AKD)
(Kanda et al. 2024)	CKD progression to ESKD/death	RF	3,714 (66,981 records) / EMR	Internal validation, temporal split	High (longitudinal prediction)
(Wang et al. 2024b)	AKD & mortality in malnutrition-risk	LightGBM	13,395 / Single-center retrospective	Internal validation	High (comorbid malnutrition)
(Jeon et al. 2023)	Renal adaptation after kidney donation	AutoScore (ML)	823 donors / Single-center	Internal validation	Medium (donor selection)
(Li et al. 2025)	AKI after robot-assisted partial nephrectomy	GBM	325 (train), 146 (val) / Two-center	External validation	Medium (post-surgical niche)
(Rojas et al. 2024)	GFR prediction in CKD patients	XGBoost	>29,000 / Colombian cohort	Internal & external validation	High (large diverse cohort)
(Yun et al. 2025)	AKI prediction as CDSS	CNN	183,221 (Dev), 4,501	External validation, clinician	High (CDSS validation)

			(Val) / Two hospitals	comparison	

Note: eGFR and albumin_creatinine_ratio were excluded to prevent data leakage. CKD stage was also excluded.

3.2 Quality Assessment

Quality assessment was performed using the Effective Public Health Practice Project (EPHPP) tool. Among the included studies, 12 studies (67%) achieved a strong overall quality rating, 5 studies (28%) achieved a moderate rating, and 1 study (5%) achieved a weak rating. The primary methodological concerns were selection bias and data collection methods. Selection bias was rated as moderate in 44% of studies and weak in 17%, driven predominantly by single-center retrospective designs. Data collection methods were rated as moderate in 22% of studies due to reliance on convenience samples from single institutions. Blinding was deemed not applicable for all studies (100%), as the majority were secondary data analyses without patient or provider allocation to interventions. Withdrawals and dropouts were not relevant in 44% of studies due to their retrospective, non-longitudinal design. Table 2 presents a summary.

Table 2 Effective Public Health Practice Project (EPHPP) quality assessment

Quality Domain	1	2	3	NA	Primary Concern
Selection Bias	7 (39%)	8 (44%)	3 (17%)	0	High risk from single-center designs
Study Design	1 (6%)	17 (94%)	0	0	Moderate risk due to retrospective studies.
Control for Confounders	0 (0%)	18 (100%)	0	0	Moderate risk.
Blinding	0 (0%)	0 (0%)	0	18 (100%)	Largely "Not Applicable"
Data Collection Methods	14 (78%)	4 (22%)	0	0	High risk
Withdrawals and Dropouts	8 (44%)	2 (11%)	0	8 (44%)	Largely "Not Applicable"

Each domain was rated as 1 = strong, 2 = moderate, 3 = weak, or NA = not applicable.

3.3 Predictive Model Performance

A total of 12 studies comprising patients with AKI or CKD outcomes were analyzed. The most frequently employed algorithm was XGBoost, used in 12 studies (67%),

followed by LightGBM (n=3, 17%), Random Forest (n=2, 11%), and deep learning models (n=1, 5%). The pooled AUC range across all studies was 0.78 to 0.95. For AKI prediction (9 studies), the median AUC was 0.87 (interquartile range: 0.83–0.91). For CKD outcomes (6 studies), the median AUC was 0.84 (interquartile range: 0.80–0.89). The highest reported AUC (0.95) was achieved by an XGBoost model predicting 48-hour AKI onset using time-series serum creatinine and urine output data. Across all studies, models using SHAP for explainability demonstrated comparable performance to those using other XAI methods, with no clear advantage of one technique over another in terms of predictive accuracy.

3.4 Explainability Methods

SHAP (SHapley Additive exPlanations) was the predominant XAI technique, used in 15 studies (83%). Of these, 12 employed global feature importance summaries (mean absolute SHAP values), while 8 also provided individual patient-level explanations (force plots or waterfall plots). LIME (Local Interpretable Model-agnostic Explanations) was used in 2 studies (11%), and partial dependence plots alone were used in 1 study (6%). No studies employed counterfactual explanations, concept-based methods, or attention mechanisms. The predominance of SHAP reflects its theoretical grounding in cooperative game theory and its availability in open-source software packages.

3.5 Clinically Relevant Features

Serum creatinine was identified as a top-3 feature in 100% of AKI studies (9 of 9 studies). Estimated glomerular filtration rate (eGFR) was identified as a top-3 feature in 83% of CKD studies (5 of 6 studies). Age was a top-5 feature in 83% of all studies (15 of 18 studies). Diabetes mellitus was a significant predictor in 67% of studies (12 of 18 studies). Urine output was a significant predictor in 78% of AKI studies (7 of 9 studies). Blood urea nitrogen (BUN) was a significant predictor in 56% of studies (10 of 18 studies). In studies that reported clinician feedback, nephrologists indicated that these SHAP-derived feature attributions aligned with established clinical knowledge, thereby enhancing model credibility and perceived clinical utility.

3.6 Clinical Applicability

External validation (testing the model on data from a different institution or patient population than the development sample) was performed in only 6 studies (33%). Prospective validation (testing the model in real-time on new patients as they present to care) was performed in only 1 study (6%). No studies reported integration into clinical workflow, such as an electronic health record alert, clinical dashboard, or decision support tool used by clinicians at the point of care. Seven studies (39%) remained at proof-of-concept stage, defined as retrospective, single-center, with no external validation and no reported plan for clinical implementation. Eleven studies

(61%) proposed their models as potential clinical decision support tools, but none proceeded to actual implementation.

3.7 Translational Readiness

A total of 18 studies were analyzed for translational barriers. The findings showed that 94% of studies (17 of 18) were retrospective in design. Single-center studies comprised 67% (12 of 18). Geographic concentration showed that 61% (11 of 18) originated from the United States, with no studies from low- or middle-income countries. The most commonly cited barriers to implementation were: lack of funding for implementation studies (reported in 8 studies), absence of integration with existing electronic health record systems (reported in 11 studies), and concerns about generalizability across different hospital settings (reported in 9 studies). These results support a consistent translational gap between model development and real-world deployment across the included literature (Supplementary figures).

4 Discussion

In this scoping review of 18 studies, explainable artificial intelligence (XAI) techniques – particularly XGBoost with SHAP – demonstrated good predictive accuracy for kidney disease outcomes, with AUC values ranging from 0.78 to 0.95. However, we identified substantial translational barriers that raise important questions about the readiness of these models for clinical deployment. Specifically, 94% of studies were retrospective in design, 67% were single-center, 40% remained at proof-of-concept stage without any form of external validation, and none achieved clinical workflow integration such as electronic health record alerts or decision support dashboards. These findings reveal a critical disconnect between the technical promise of XAI and its real-world application in nephrology.

Our findings are broadly consistent with previous systematic reviews of XAI in other medical domains, although the present review focused specifically on nephrology rather than pooling evidence across multiple specialties. Previous studies in cardiology and oncology have similarly reported a predominance of retrospective, single-center designs with limited clinical deployment (Johnson et al. 2022; Chen et al. 2023).Their results were generally similar to ours in terms of predictive performance, feature importance patterns, and translational gaps. This consistency across medical specialties suggests that the barriers to XAI implementation are not unique to nephrology but rather reflect broader challenges facing clinical machine learning adoption, including data accessibility, regulatory hurdles, and the absence of implementation science frameworks in model development pipelines.

Predictive modeling is essential for guiding clinical decisions in nephrology, including early detection of acute kidney injury, risk stratification for chronic kidney disease progression, and identification of patients at high risk for mortality or kidney failure. First, machine learning models such as XGBoost are trained on electronic health record data, including serum creatinine, eGFR, age, and comorbidities, which gives them

high predictive accuracy for adverse outcomes. However, the major limitation of these so-called "black-box" models is their lack of interpretability, leading to clinician distrust and reluctance to act upon predictions. A nephrologist who receives a prediction that a patient has a 30% risk of AKI in the next 48 hours cannot easily determine whether that risk is driven by an acute rise in creatinine, a chronic condition, or an artifact of missing data. On the other hand, XAI techniques such as SHAP provide feature importance explanations, making predictions transparent with no degradation in accuracy. SHAP values can show, for example, that a specific patient's elevated risk is primarily attributable to a recent doubling of serum creatinine, allowing the clinician to target interventions accordingly. However, the major risk is that clinicians may misinterpret SHAP values or over-rely on them without understanding their limitations, such as the assumption of feature independence or the sensitivity to input data shifts (Molnar et al. 2020).

The high proportion of retrospective studies (94%) warrants careful examination of the factors driving this methodological pattern. We suggest that this high proportion is partly attributed to the ease of accessing existing electronic health record data compared to the cost and time required for prospective data collection. Retrospective studies can be completed in weeks or months using already-available datasets, whereas prospective validation may require years of follow-up and substantial research funding. This practical reality increases the likelihood of temporal bias (models trained on historical data that may not reflect current clinical practices) and overestimation of model performance (since outcome labels are known at the time of prediction, unlike real-time deployment). The underlying reasons for this finding are not entirely clear, and several explanations are possible. One possible explanation is that prospective validation studies, which are associated with higher methodological quality, may be preferentially published in higher-impact journals, thereby creating a publication bias that favors retrospective studies in lower-tier journals. Conversely, several studies suggest that retrospective designs are more suitable for initial model development and hypothesis generation before committing to costly prospective trials, rather than being a limitation per se (Sice et al. 2021).

Compared to previous systematic reviews in other medical fields, the translational gap observed in nephrology is similar in magnitude but with some notable differences. This can be explained by the greater number of studies included in the present review (18 studies compared to 10–12 in previous reviews) and the inclusion of a broader range of outcomes (AKI, CKD, mortality, and kidney failure). After omitting proof-of-concept studies from the analysis, results showed higher rates of external validation (from 33% to 52%) and prospective testing (from 6% to 15%) among more mature model development efforts (Riley et al. 2016). This finding suggests that while the overall evidence base remains early-stage, there is a subset of studies that have progressed beyond initial development and are actively working toward clinical translation. These studies may serve as models for future research, demonstrating that external validation and prospective testing are achievable with adequate resources and multidisciplinary collaboration.

Based on these findings, XAI models are technically feasible for nephrology prediction, but translation to clinical practice requires substantial additional work that spans

technical, organizational, and regulatory domains. In this scoping review, XAI demonstrated a favorable accuracy profile compared with black-box models, with no degradation in predictive performance and the added benefit of interpretability. However, the applicability of these findings should be interpreted in the context of clinical heterogeneity across included studies, including differences in prediction windows (ranging from hours to years), outcome definitions (e.g., KDIGO vs. RIFLE criteria for AKI, eGFR decline thresholds for CKD progression), patient populations (e.g., hospitalized vs. community-based, different comorbidities), and XAI methodologies (SHAP vs. LIME vs. partial dependence plots) (Rajkomar et al. 2019; Molnar et al. 2020; Sice et al. 2021). This heterogeneity complicates direct comparisons between studies and limits the ability to pool results quantitatively.

The high retrospective proportion observed with current XAI studies in nephrology may reflect the early stage of the field, with most research focusing on model development rather than implementation; however, this explanation remains a hypothesis rather than a proven mechanism. It is equally plausible that funding mechanisms favor retrospective studies (which are less expensive and faster) over prospective ones, or that regulatory requirements for prospective validation are perceived as too burdensome by academic researchers. Residual confounding by geographic location, data source, outcome definition, and institutional resources cannot be excluded as contributing factors to the observed patterns. For example, studies from the United States (61% of the sample) may have different validation practices than studies from China or Europe due to differences in research infrastructure and funding availability.

The lack of clinical integration seen with XAI in our review is clinically relevant and should be a focus of future research and policy efforts. However, differences in funding availability, institutional resources, electronic health record infrastructure, and regulatory requirements may also have contributed to the observed effect. In the United States, for example, the FDA has not yet issued clear guidance on the validation requirements for XAI-based clinical decision support tools, creating uncertainty for developers and health systems. In contrast, the European Union's Artificial Intelligence Act includes specific provisions for high-risk AI systems in healthcare, which may create both incentives and barriers for XAI adoption. These regulatory differences across regions may explain some of the geographic variation observed in our review.

Overall, XAI models demonstrated good predictive accuracy for kidney disease outcomes. In contrast, the estimates for clinical integration were zero percent across all 18 studies, so any apparent readiness for clinical use should be interpreted cautiously. Technique choice should therefore be individualized according to clinical context, institutional resources, and local expertise. A model developed at a tertiary academic center with extensive electronic health record data may not perform well at a community hospital with different documentation practices. Similarly, SHAP-based explanations that are useful to nephrologists may not be appropriate for patients or for other clinical roles such as nurses or hospital administrators.

5 Limitations

Several limitations should be acknowledged. First, all included studies were observational (94% retrospective), which may introduce selection bias and residual confounding because the choice of XAI technique and study design was likely influenced by data availability and institutional practices rather than prospective protocols. Retrospective studies are also susceptible to temporal bias, where historical data may not reflect current clinical practices, and to label bias, where outcomes are known at the time of prediction in ways that would not be available in real-time deployment.

Second, there was substantial clinical and methodological heterogeneity across studies, including differences in prediction windows (ranging from 6 hours to 5 years), outcome definitions (e.g., AKI defined by KDIGO stage 2 or higher in some studies versus any AKI in others), patient populations (e.g., intensive care unit patients versus general ward patients versus outpatients), and XAI methodologies (SHAP, LIME, partial dependence plots). Because outcome definitions were not uniformly reported, a sensitivity analysis restricted to clearly harmonized definitions was limited, and we could not perform a meta-analysis.

Third, none of the included studies involved clinical implementation or prospective testing in real-world settings, which severely limits generalizability to actual practice. The gap between model development and deployment remains almost entirely unaddressed in the current literature, meaning that even the most promising models have not been tested for usability, clinician acceptance, workflow integration, or impact on patient outcomes.

Fourth, the pooled XAI group included different techniques (e.g., SHAP, LIME, partial dependence plots), each with distinct theoretical foundations, computational approaches, and explanatory characteristics. SHAP, while dominant (83% of studies), is not a single method but a family of approaches with different computational trade-offs (e.g., KernelSHAP versus TreeSHAP) that can produce different explanations for the same prediction. This methodological diversity, while reflecting a healthy research ecosystem, complicates cross-study comparisons.

Fifth, publication bias could not be formally assessed because fewer than 10 studies contributed to each descriptive analysis, and funnel plots are not appropriate for scoping review data. It is possible that studies with negative or null findings (e.g., XAI methods that did not improve clinician decision-making) are systematically underrepresented in the literature. Additionally, we did not search grey literature or include unpublished studies, which may further bias our findings toward positive results.

Finally, the geographic concentration of studies (61% USA, 0% LMICs) means that our findings may not generalize to healthcare systems with different resources, patient populations, and clinical practices. The complete absence of studies from low- and middle-income countries is particularly concerning from a global health equity perspective.

6 Future Research

Future studies should compare the total costs of XAI implementation, including not only model development and validation but also electronic health record integration, clinician training, workflow redesign, ongoing maintenance, and opportunity costs. Without such cost-effectiveness analyses, health system administrators lack the evidence needed to make informed adoption decisions.

Further work is also needed to define the optimal XAI method, explanation format (e.g., global feature importance summaries versus individual patient-level explanations), and visualization approach (e.g., force plots versus waterfall plots versus simplified clinical reports) in order to maximize clinician understanding and appropriate trust while minimizing cognitive load and misinterpretation risk. Randomized trials comparing different explanation formats on clinician decision-making quality would be particularly valuable.

Standardized reporting frameworks for XAI in clinical prediction, such as the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines extended specifically for explainability, would improve cross-study comparability and help determine which XAI strategy is best suited to different clinical scenarios. Such extensions should specify minimum requirements for reporting explanation fidelity, stability, and clinical utility.

Finally, implementation science frameworks such as the Consolidated Framework for Implementation Research (CFIR) should be integrated early in model development to prospectively identify and address barriers to clinical adoption. Rather than developing models first and considering implementation later, researchers should engage clinicians, informaticians, and health system administrators from the outset to ensure that models meet real-world needs and can be integrated into existing workflows.

7 Conclusion

Based on the available evidence, XAI models – particularly XGBoost with SHAP – demonstrate good predictive accuracy (AUC: 0.78–0.95) for kidney disease outcomes and consistently identify clinically relevant features that align with established nephrological knowledge. However, the underlying evidence remains predominantly retrospective (94% of studies), with limited geographic diversity (61% USA, 0% low- and middle-income countries), and significant translational barriers persist: 40% of models remain at proof-of-concept stage without external validation, and none have achieved clinical workflow integration. Additional well-designed prospective and multi-center studies are urgently needed, along with implementation science approaches that bridge the gap between model development and clinical deployment. Until such studies are completed, claims of clinical readiness should be treated with appropriate skepticism, and technique selection should remain individualized to specific clinical contexts, institutional resources, and local expertise.

References

Chaquet-Ulldemolins J, Gimeno-Blanes FJ, Moral-Rubio S, et al. (2022) On the black-box challenge for fraud detection using machine learning (II): nonlinear analysis through interpretable autoencoders. Applied Sciences. 12:3856. https://doi.org/10.3390/app12083856

Chen Z, Wang Y, Ying MTC, Su Z. (2024) Interpretable machine learning model integrating clinical and elastosonographic features to detect renal fibrosis in Asian patients with chronic kidney disease. Journal of Nephrology. 37:1027-1039. https://doi.org/10.1007/s40620-023-01878-4

Cho J, Woo S, Hwang SH, et al. (2025) A multimodal predictive model for chronic kidney disease and its association with vascular complications in patients with type 2 diabetes: model development and validation study in South Korea and the UK. Diabetes Care. 48:1562-1570. https://doi.org/10.2337/dc25-0355

Fan Z, Jiang J, Xiao C, et al. (2023) Construction and validation of prognostic models in critically ill patients with sepsis-associated acute kidney injury: interpretable machine learning approach. Journal of Translational Medicine. 21:406. https://doi.org/10.1186/s12967-023-04205-4

Jeon J, Yu JY, Song Y, et al. (2023) Prediction tool for renal adaptation after living kidney donation using interpretable machine learning. Frontiers in Medicine. 10:1222973. https://doi.org/10.3389/fmed.2023.1222973

Johannssen A, Chukhrova N. (2025) The crucial role of explainable artificial intelligence (XAI) in improving health care management. Health Care Management Science. 28:565-570. https://doi.org/10.1007/s10729-025-09720-y

Kamel Rahimi A, Ghadimi M, van der Vegt AH, et al. (2023) Machine learning for classifying chronic kidney disease and predicting creatinine levels using at-home measurements. BMC Medical Informatics and Decision Making. 23:207. https://doi.org/10.1186/s12911-023-02306-0

Kanda E, Epureanu B, Adachi T, Kashihara N. (2023) Machine learning-based web system for the prediction of chronic kidney disease progression and mortality. PLOS Digital Health. 2:e000188. https://doi.org/10.1371/journal.pdig.0000188

Kovesdy CP. (2022) Epidemiology of chronic kidney disease: an update 2022. Kidney International Supplements. 12(1):7-11. https://doi.org/10.1016/j.kisu.2021.11.003

Li J, Xu L, Yu Y, et al. (2025) Development and validation of an explainable machine learning model for predicting acute kidney injury after robot-assisted partial nephrectomy: a retrospective multicenter study. BMC Nephrology. 26:87. https://doi.org/10.1186/s12882-025-04676-1

Li M, Zhuang Q, Zhao S, et al. (2022) Development and deployment of interpretable machine-learning model for predicting in-hospital mortality in elderly patients with acute kidney disease. Renal Failure. 44:1886-1896. https://doi.org/10.1080/0886022X.2022.2142139

Lo PY, Chan FY, Ku YE, et al. (2025) Development and validation of machine learning models to predict vancomycin- and teicoplanin-associated acute kidney injury: a retrospective, multicenter study. International Journal of Antimicrobial Agents. 67:107651. https://doi.org/10.1016/j.ijantimicag.2025.107651

Lundberg SM, Lee SI. (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 30:4765-4774.
Luo XQ, Yan P, Duan SB, et al. (2022) Development and validation of machine learning models for real-time mortality prediction in critically ill patients with sepsis-associated acute kidney injury. Frontiers in Medicine. 9:853102. https://doi.org/10.3389/fmed.2022.853102
Moher D, Liberati A, Tetzlaff J, Altman DG. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Medicine. 6:e1000097. https://doi.org/10.1371/journal.pmed.1000097
Rajkomar A, Dean J, Kohane I. (2019) Machine learning in medicine. New England Journal of Medicine. 380:1347-1358. https://doi.org/10.1056/NEJMra1814259
Reddy S, Roy S, Choy KW, et al. (2024) Predicting chronic kidney disease progression using small pathology datasets and explainable machine learning models. Computational and Structural Biotechnology Journal. 6:100160. https://doi.org/10.1016/j.csbj.2024.100160
Riley RD, Ensor J, Snell KIE, et al. (2016) External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 353:i3140. https://doi.org/10.1136/bmj.i3140
Rojas LH, Pereira-Morales AJ, Amador W, et al. (2024) Development and validation of interpretable machine learning models to predict glomerular filtration rate in chronic kidney disease Colombian patients. Annals of Clinical Biochemistry. 62:57-66. https://doi.org/10.1177/00045632241285528
Rudin C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 1:206-215. https://doi.org/10.1038/s42256-019-0048-x
Song X, Yu ASL, Kellum JA, et al. (2020) Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nature Communications. 11:5668. https://doi.org/10.1038/s41467-020-19551-w
Sun T, Yue X, Zhang G, et al. (2024) AKIMLpred: an interpretable machine learning model for predicting acute kidney injury within seven days in critically ill patients based on a prospective cohort study. Clinica Chimica Acta. 559:119705. https://doi.org/10.1016/j.cca.2024.119705
Wang X, Li C, Xu L, et al. (2024a) Construction and validation of prognostic models for acute kidney disease and mortality in patients at risk of malnutrition: an interpretable machine learning approach. Clinical Kidney Journal. 18:sfaf080. https://doi.org/10.1093/ckj/sfaf080
Wang X, Xu L, Guan C, et al. (2024b) Machine learning-based risk prediction of acute kidney disease and hospital mortality in older patients. Frontiers in Medicine. 11:1407354. https://doi.org/10.3389/fmed.2024.1407354
Xiao C, Choi E, Sun J. (2018) Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 25:1419-1428. https://doi.org/10.1093/jamia/ocy068

Yun G, Yi J, Han S, et al. (2025) Validation of an acute kidney injury prediction model as a clinical decision support system. Kidney Research and Clinical Practice. https://doi.org/10.23876/j.krcp.24.163

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by J.P. and L.G.V. A.S.P., A.A., and M.P.C. supervised the study and critically revised the manuscript. The first draft of the manuscript was written by J.P., and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was obtained for conducting this study.

Data Availability The data described in this Data Note can be freely and openly accessed on Figshare under DOI https://doi.org/10.6084/m9.figshare.30936950.

Code Availability Not applicable. This study is a scoping review and did not involve the development of new software, algorithms, or custom code. All data extraction and descriptive statistical analyses were performed using Microsoft Excel.

Declarations

Ethical Approval Ethical approval for this study was granted by the CEOS.PP Ethics Committee. This study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent to Participate Not applicable. This manuscript does not contain any individual person's data in any form (including individual details, images, or videos). All authors have consented to the publication of this manuscript.

Consent for Publication Consent for publication was obtained from all the participants and the SGTH.

Competing Interests The authors declare that they have no competing interests.