

Advances in Internal Medicine

Effectiveness of Artificial Intelligence-Based Approaches in Predicting the Progression of Chronic Kidney Disease: A Systematic Review of Randomized Controlled Trials and Primary Studies

¹ Muhamad Luthfi Asyhar, ² Afif Ramadhan, ³ Siska Agustina, ⁴ Muhamad Faqih Dwi Putra, ⁵ Faridatul Jannah, ⁶ Adinda Ayu Farah Meidha, ⁷ Amanda Safitria, ⁸ Likuidita Yona Ramadini, ⁹ Charles Sanjaya

¹ Waras Wiris Regional General Hospital, Indonesia / Faculty of Medicine, Airlangga University, Indonesia

² dr. Sardjito Central General Hospital, Indonesia / Faculty of Medicine, Health and Nursing, Gadjah Mada University, Indonesia

³ Naga Timbul Public Health Centre, Indonesia / Faculty of Medicine, Abulyatama University, Indonesia

⁴ Karawang Regional General Hospital, Indonesia / Faculty of Medicine, Diponegoro University, Indonesia

⁵ dr. Uly Primary Health Clinic, Indonesia / Faculty of Medicine, University of Malahayati, Indonesia

⁶ Prima Husada Waru General Hospital, Indonesia / Faculty of Medicine, Wijaya Kusuma University, Indonesia

⁷ Faculty of Medicine, University of Mulawarman, Indonesia

⁸ Faculty of Medicine, UIN Yogyakarta, Indonesia

⁹ Internist, Kebonjati General Hospital, Indonesia

Corresponding Email : muhamadluthfiasyhar@yahoo.com

Article History :

Received date : 2026/05/12

Revised date : 2026/06/07

Accepted date : 2026/07/16

Published date : 2026/08/24

Copyright: © by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (BY NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>).

ABSTRACT

Introduction: Chronic kidney disease (CKD) is a progressive disorder with heterogeneous trajectories poorly forecasted by conventional regression-based instruments. Artificial intelligence (AI), encompassing machine learning, tree-based ensembles, deep neural networks, and dynamic Bayesian architectures, has been proposed to exploit high-dimensional, non-linear, and longitudinal clinical data for individualised prognostication. This systematic review evaluates AI-based approaches for predicting CKD progression across the full translational pathway from model derivation to randomised implementation.

Methods: Systematic review conducted per PRISMA 2020 guidelines. Eligible designs were RCTs, cluster RCTs, cohort, case-control, cross-sectional, and observational studies evaluating AI/ML models for predicting CKD progression, kidney failure, or renal-function decline. Risk of bias assessed with PROBAST, RoB 2, and Newcastle-Ottawa Scale. Thirteen pre-specified outcome domains were analysed using structured narrative synthesis.

Results: Thirty-five primary studies comprising >2.2 million individuals, >490,000 images, and >89,000 patient-years met inclusion criteria: 7 randomised trials, 26 cohort/case-cohort studies, and 2 post-hoc analyses. Median AUC for CKD progression was 0.87 (range 0.74-0.99). AI significantly outperformed conventional reference instruments: XGBoost AUC 0.892 versus KFRE 0.745; Klinrisk outperformed KDIGO heatmap at every interval ($P<0.01$); KDpredict super-learner achieved 5-year IPA 27.8% versus 18.1% for KFRE in Denmark (30.5% versus 14.2% in Scotland). Deep learning on biopsy images achieved AUCs 0.878-0.904 versus 0.786-0.811 for pathologist-estimated scoring. Discrimination was preserved on cross-ethnic external validation (AUC 0.867 and 0.871 in ELSA/HRS). However, randomised implementation evidence was discordant: decision support significantly improved SBP (-14.6 versus -11.7 mmHg, $P=0.005$), guideline-concordant prescribing (49.9% versus 34.6%, $P<0.001$), and annualised eGFR loss (0.01 versus 0.95 mL/min/1.73m²/year, $P<0.001$), whereas passive AI risk display produced no measurable benefit (39% versus 40%, $P=1.00$). Calibration deteriorated in advanced CKD, with 5-year KFRE overpredicting by 10-18%.

Discussion: AI-based models consistently improve CKD progression discrimination beyond conventional regression, with gains greatest where data are high-dimensional, non-linear,

longitudinal, or image-based. Explainability analyses converged on a physiologically coherent predictor set (eGFR, albuminuria, age, creatinine, albumin, haemoglobin, potassium, frailty). Nevertheless, calibration drift over longer horizons, incomplete PROBAST reporting, and sparse randomised evidence that predictive accuracy translates into patient benefit temper enthusiasm. Workflow integration, actionable recommendations, and clinician engagement, rather than algorithmic accuracy alone, determine clinical effectiveness.

Conclusion: AI-based approaches provide significant and clinically meaningful improvements in CKD progression prediction relative to conventional equations, with preservation on external validation. Translation into improved patient outcomes, however, is contingent on active integration of risk estimates into concrete clinical action pathways. Nephrology services should prioritise externally validated, well-calibrated, interpretable models deployed with explicit action protocols, prospective calibration monitoring, and local recalibration. Future work should focus on pragmatic randomised implementation trials with hard renal endpoints, including in South-East Asian populations.

Keywords: artificial intelligence; machine learning; deep learning; chronic kidney disease; disease progression; end-stage kidney disease; risk prediction; clinical decision support; external validation; nephrology

INTRODUCTION

Background

Chronic kidney disease (CKD) affects approximately one in ten adults worldwide and constitutes an under-recognised but rapidly expanding contributor to global morbidity, mortality and health expenditure. Its defining clinical characteristic is progressive and, in most instances, irreversible loss of excretory, endocrine and homeostatic renal function, culminating for a substantial minority of patients in kidney failure requiring dialysis or transplantation. Because early-stage CKD is largely asymptomatic, diagnosis is frequently delayed until structural damage is advanced, at which point the therapeutic window for meaningful renoprotection has narrowed considerably. Contemporary evidence indicates that CKD often escapes timely recognition precisely because clinical manifestations are absent in the initial stages, allowing gradual deterioration to proceed unimpeded towards end-stage renal disease.^{12,3}

The natural history of CKD is, however, markedly heterogeneous. Two patients with identical estimated glomerular filtration rate (eGFR) and albuminuria may follow entirely divergent trajectories, one remaining stable for a decade while the other reaches kidney failure within two years. This heterogeneity is the fundamental clinical problem that risk prediction seeks to solve, since the decision to intensify renin-angiotensin-aldosterone system inhibition, to introduce sodium-glucose cotransporter-2 inhibition, to create arteriovenous access, to initiate transplant work-up or to move towards conservative kidney management all depend on an estimate of when, and with what probability, kidney failure will occur. Kidney Disease: Improving Global Outcomes (KDIGO) staging by eGFR category and albuminuria category provides an ordinal framework for this purpose, but it is a population-level heat map rather than an individualised prognostic instrument.^{9,41}

The Kidney Failure Risk Equation (KFRE) represented a decisive methodological advance, demonstrating that a parsimonious model incorporating age, sex, eGFR and albuminuria, optionally augmented by serum calcium, phosphate, bicarbonate and albumin, could accurately predict progression to kidney failure. In the original derivation and validation cohorts the eight-variable model achieved a concordance statistic of 0.917 in development and 0.841 in validation, and subsequent individual-participant-data meta-analysis of 31 cohorts encompassing 721,357 participants across more than 30 countries confirmed excellent discrimination, with an overall C statistic of 0.90 at two years and 0.88 at five years. Nonetheless, that same synthesis demonstrated systematic overestimation of risk in several non-North American cohorts requiring the addition of region-specific calibration factors, and comprehensive external validation in advanced CKD subsequently showed that the average c statistic of eleven guideline-referenced models fell to 0.74-0.80, with five-year KFRE overpredicting risk by 10-18% once the competing risk of death was appropriately modelled.^{9,41,40}

These limitations are structural rather than incidental. Conventional Cox proportional-hazards modelling assumes linearity in the log-hazard, proportionality of hazards over time, additivity of covariate effects and, in most implementations, complete-case handling of missingness. Renal pathophysiology satisfies none of these assumptions comfortably: the relationship between age and progression risk is non-monotonic, the interaction between proteinuria and blood pressure is multiplicative, comorbidity clusters exert joint rather than

additive effects, and the informative trajectory of repeated eGFR and albuminuria measurements is discarded when only a baseline value is used. Artificial intelligence (AI) approaches, encompassing regularised regression, tree-based ensembles such as random forests and gradient-boosted machines, support vector machines, feed-forward and recurrent neural networks, transformer architectures, convolutional networks for image data and Bayesian joint models for longitudinal data, are explicitly designed to accommodate non-linearity, high dimensionality, interaction structure and repeated measurement without prior specification of functional form.^{1,2,46}

A parallel and equally important development is the extension of AI beyond tabular laboratory data into image-based prognostication. Digital pathology now permits convolutional neural networks to derive prognostic scores directly from whole-slide biopsy images, capturing histological features that established semiquantitative classifications do not encode; deep learning applied to renal ultrasonography can approximate eGFR non-invasively; and retinal imaging, exploiting the shared microvascular biology of the retina and the kidney, has been shown to detect diabetic kidney disease and to discriminate isolated diabetic nephropathy from non-diabetic renal disease in multi-ethnic populations. These modalities are of particular relevance to health systems in which access to specialist nephropathology or frequent biochemical monitoring is constrained.^{21,23,24,27,28}

Research Gap

Despite a rapidly expanding literature, four specific gaps remain. First, the existing syntheses are predominantly scoping in nature: a scoping review of prediction models for CKD progression identified 33 eligible studies but concluded that inconsistent reporting of model development precluded meaningful comparison between machine-learning algorithms, and a more recent scoping review of AI in CKD management catalogued 41 articles across four use cases without formal risk-of-bias appraisal or outcome-domain synthesis. Second, prognostic-accuracy evidence and implementation evidence have not been appraised within a single framework, so it remains unclear whether the documented gains in discrimination confer measurable clinical benefit. Third, imaging-based AI prognostication has generally been reviewed separately from tabular-data modelling, obscuring the complementarity of the two evidence streams. Fourth, the transportability of these models to Asian and specifically South-East Asian populations, in whom the epidemiology of CKD differs materially, has been assessed only sporadically.^{3,44,26}

The clinical consequence of these gaps is practical. Nephrology services considering the procurement or local development of an AI risk tool currently lack a consolidated statement of how much accuracy is gained relative to KFRE, whether that gain survives external validation, which predictors drive the models, how well calibrated the outputs are, and what magnitude of clinical benefit has actually been demonstrated in randomized conditions. This review was designed to answer precisely these questions.^{3,44}

Objectives

The general objective of this systematic review was to evaluate the effectiveness of artificial-intelligence-based approaches for predicting the progression of chronic kidney disease. The specific objectives were as follows:

- To quantify the discriminative performance (AUC, concordance statistic, index of prediction accuracy) of AI-based models for CKD progression and for progression to kidney failure across diverse populations and prediction horizons.
- To compare, head-to-head, the performance of AI-based models against conventional regression-based reference instruments, principally KFRE, KDIGO risk categorisation and Cox proportional-hazards models.
- To evaluate the calibration of AI-based models, including Brier score, calibration plots and behaviour under competing risks.
- To assess the classification performance (sensitivity, specificity, accuracy, predictive values) of AI models when applied at defined operational thresholds.
- To synthesise explainability findings and identify the predictor set that consistently drives model output.
- To determine the extent to which model performance is preserved on temporal, geographic and cross-ethnic external validation.
- To evaluate the performance of image-based AI (histopathology, ultrasonography, retinal photography) for renal prognostication.
- To assess risk stratification, reclassification and net clinical benefit as demonstrated by decision-curve analysis and risk-group separation.
- To determine the effect of AI-enabled clinical decision support on process-of-care outcomes in randomized conditions.

- To determine the effect of AI-enabled clinical decision support on patient-level clinical outcomes, including blood pressure, eGFR slope, dialysis initiation and mortality.
- To examine the influence of prediction horizon on model performance and the relative performance of different algorithm classes.
- To appraise the methodological quality and risk of bias of the included evidence and to derive recommendations for clinical practice and future research.

Benefits of the Study

The theoretical benefit of this review is a consolidated and methodologically appraised account of the current state of AI-based renal prognostication, organised by outcome domain rather than by algorithm, and therefore usable as a reference framework for subsequent model development and reporting. The practical clinical benefit is direct: nephrologists, internists and primary care physicians managing CKD require an explicit statement of how much prognostic precision can be gained, in which populations that gain is reliable, and under what implementation conditions it has been shown to change patient outcomes. The health-system benefit lies in informing procurement, governance and calibration-monitoring policy for AI tools in nephrology services, an issue of increasing regulatory salience. For Indonesian and other middle-income health systems in particular, the review clarifies which model inputs are feasible with routinely available laboratory panels and which imaging-based modalities may offer non-invasive alternatives where nephropathology capacity is limited.^{44,45}

Hypotheses

The following a priori hypotheses were specified:

- H1: AI-based prediction models achieve significantly higher discrimination for CKD progression than conventional regression-based reference equations in head-to-head comparison within the same dataset.
- H2: The discriminative advantage of AI-based models is substantially preserved on geographic and cross-ethnic external validation, with attenuation of less than 0.10 AUC units in most comparisons.
- H3: AI-based models applied to imaging data achieve prognostic performance at least non-inferior to expert human semiquantitative scoring.

- H4: AI-based clinical decision support that delivers an explicit, actionable therapeutic recommendation improves process-of-care outcomes, whereas passive display of a risk score does not.
- H5: Improvements in predictive discrimination are not, by themselves, sufficient to produce measurable improvement in patient-level renal outcomes in randomized conditions.

Novelty

This review is, to the authors' knowledge, the first synthesis to appraise the full translational pathway of AI in CKD prognostication within a single framework, from model derivation through external validation to randomized implementation, and to make the distinction between predictive-performance outcomes and implementation outcomes analytically explicit. Three further features are novel. First, the review integrates tabular-data models, digital-pathology models, sonographic models and retinal-imaging models within one outcome architecture, permitting direct comparison of information sources. Second, every included reference was retrieved and verified with peer-review and medical-journal filters, and a traceable source link is provided for each citation. Third, thirteen pre-specified outcome domains were analysed, substantially exceeding the scope of previous scoping reviews, and methodological quality was appraised using three complementary instruments matched to study design.^{3,44,45}

METHODS

Protocol and Reporting Standard

This systematic review was designed, conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement. A written protocol specifying the review question, eligibility criteria, search strategy, outcome domains, data-extraction fields, quality-appraisal instruments and synthesis plan was prepared before the formal search was executed, and no substantive post-hoc deviation from that protocol occurred. The methodological approach to appraisal of prediction-model studies was aligned with the framework used in previously registered protocols for machine-learning prediction in CKD, which specify comparison of machine-learning against non-machine-learning models as the

primary analytical axis and evaluation of model construct and reporting quality as secondary axes.^{6,3}

Review Question and PICOS Framework

The review question was framed as follows: among adults with, or at risk of, chronic kidney disease, does the application of artificial-intelligence-based methods, compared with conventional statistical prediction instruments or usual care, improve the prediction of disease progression and the clinical outcomes that depend upon that prediction? The corresponding PICOS elements are set out in Table 1.

Table 1. PICOS framework and operational definitions applied in this systematic review

Element	Operational definition applied
Population (P)	Adults (≥ 18 years) with established chronic kidney disease of any KDIGO stage (G1-G5, any albuminuria category), or populations at defined risk of CKD (type 2 diabetes mellitus, hypertension, ageing cohorts), or kidney transplant recipients with functioning allografts. Paediatric-only populations, dialysis-prevalent-only populations and pure acute kidney injury populations were considered only where the study also reported CKD progression outcomes.
Intervention / Index approach (I)	Any artificial-intelligence or machine-learning method applied to prediction of renal outcomes, including regularised regression (LASSO, ridge, Elastic Net), tree-based ensembles (random forest, extremely randomized trees, XGBoost, LightGBM), support vector machines and survival SVM, k-nearest neighbours, feed-forward, recurrent, convolutional and transformer neural networks, random survival forests, Bayesian joint models, super-learner ensembles, unsupervised embedding and clustering methods, and AI-enabled clinical decision support systems deploying such models.
Comparator (C)	Conventional regression-based prediction instruments (Cox proportional hazards, logistic regression, the Kidney Failure Risk Equation, the Grams

Element	Operational definition applied
	model), KDIGO risk categorisation, expert human assessment (pathologist or nephrologist scoring), or usual clinical care without algorithmic support.
Outcomes (O)	Thirteen pre-specified domains: (1) discrimination for CKD progression; (2) discrimination for kidney failure/ESKD; (3) comparative performance versus conventional models; (4) calibration; (5) sensitivity, specificity and accuracy; (6) predictor importance and explainability; (7) external validation and generalisability; (8) image-based prognostication; (9) risk stratification, reclassification and net benefit; (10) process-of-care outcomes; (11) patient-level clinical outcomes; (12) effect of prediction horizon; (13) methodological quality and risk of bias.
Study design (S)	Included: randomized controlled trials, cluster randomized controlled trials, stepped-wedge randomized trials, prospective cohort studies, retrospective cohort studies, case-control studies, case-cohort studies, cross-sectional studies, and other observational designs (prospective or retrospective), including post-hoc analyses of randomized trial datasets. Excluded from the study-characteristics table: systematic reviews, meta-analyses, scoping reviews, narrative reviews, editorials, protocols without results, and conference abstracts lacking extractable outcome data. Systematic reviews and meta-analyses identified during the search were retained for outcome triangulation and for the Discussion, in accordance with the pre-specified inclusion rule.

Eligibility Criteria

Inclusion criteria

- Original primary research employing one of the following designs: randomized controlled trial, cluster randomized controlled trial, cohort study (prospective or retrospective), case-control study, case-cohort study, cross-sectional study, or other observational design.
- Application of an artificial-intelligence or machine-learning method to the prediction, prognostication or risk stratification of chronic kidney disease progression, renal function

decline, kidney failure, end-stage kidney disease, kidney replacement therapy initiation or allograft loss.

- Reporting of at least one extractable quantitative performance metric (AUC, C statistic, index of prediction accuracy, Brier score, sensitivity, specificity, accuracy, hazard ratio, risk ratio, odds ratio, or between-group difference in a clinical outcome).
- Human participants; publication in a peer-reviewed journal; full text available in English.

Exclusion criteria

- Systematic reviews, meta-analyses, scoping reviews and narrative reviews (excluded from the characteristics table but eligible for outcome triangulation and discussion, per the pre-specified rule).
- Preprints, editorials, commentaries, letters without original data, and study protocols reporting no results.
- Purely methodological or simulation papers with no clinical cohort.
- Animal, in vitro or in silico studies without human validation.
- Studies addressing acute kidney injury alone without any chronic renal outcome.

Search Strategy

The keywords used for this research based PICO :

PICO Element	Keyword 1	Keyword 2	Keyword 3
P (Population)	Chronic kidney disease	Adults with CKD	CKD progression risk
I (Intervention)	Artificial intelligence	Machine learning	Deep learning
C (Comparison)	Conventional prediction models	Kidney Failure Risk Equation	Usual clinical care
O (Outcome)	CKD progression prediction	Discrimination performance	Clinical decision support effectiveness

The Boolean MeSH keywords inputted on databases for this research are: ("Chronic kidney disease" OR "Adults with CKD" OR "CKD progression risk") AND ("Artificial intelligence" OR "Machine learning" OR "Deep learning") AND ("Conventional prediction models" OR "Kidney

Failure Risk Equation" OR "Usual clinical care") AND ("CKD progression prediction" OR "Discrimination performance" OR "Clinical decision support effectiveness")

Study Selection

Records retrieved from all searches were pooled and de-duplicated. Two reviewers independently screened titles and abstracts against the eligibility criteria, and full texts of all potentially eligible records were retrieved and assessed independently by the same two reviewers. Discrepancies at either stage were resolved by discussion, with a third senior reviewer adjudicating. Inter-reviewer agreement at the full-text stage was substantial. The selection process is summarised in the PRISMA flow presented in the Results section.

Data Extraction

A standardised extraction form was piloted on five studies and refined before full application. The following fields were extracted in duplicate: first author and publication year; country and setting; study design; source population and eligibility criteria; sample size for development, internal testing and external validation; duration of follow-up; the AI method or methods evaluated; comparator model or usual care; candidate and final predictor set; handling of missing data; the operational definition of the renal outcome; prediction horizon; discrimination metrics with confidence intervals; calibration metrics; classification metrics at the reported threshold; explainability method and top-ranked predictors; and, for interventional studies, the primary and secondary clinical outcomes with effect estimates. Where a study reported multiple algorithms, the best-performing algorithm and the reference algorithm were both extracted. Where a study reported multiple prediction horizons, all were extracted to permit analysis of the horizon-dependency outcome domain.

Assessment of Methodological Quality and Risk of Bias

Three complementary instruments were applied according to design. Prognostic model development and validation studies were appraised with the Prediction model Risk Of Bias ASsessment Tool (PROBAST) across four domains (participants, predictors, outcome and analysis) and rated for applicability. Randomized trials were appraised with the Cochrane Risk of Bias 2 (RoB 2) tool across five domains (randomisation process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result), with the cluster-randomisation extension applied where relevant. Non-randomized

observational studies were appraised with the Newcastle-Ottawa Scale (NOS) across selection (maximum four stars), comparability (maximum two stars) and outcome or exposure ascertainment (maximum three stars), with a total of seven to nine stars classified as good quality, five to six as fair and four or fewer as poor. Two Reviewers performed all appraisals independently, and disagreements were adjudicated by Third Reviewers.

Certainty of the Evidence

Certainty in the body of evidence for each outcome domain was appraised using a GRADE-informed approach adapted for prognostic-accuracy research, considering risk of bias, inconsistency, indirectness, imprecision and publication bias, with possible upgrading for large magnitude of effect and consistent dose-response or horizon-response relationships. Certainty ratings are reported alongside the outcome synthesis.

Data Synthesis and Statistical Approach

Formal statistical pooling was pre-specified as conditional on acceptable clinical and methodological homogeneity. In the event, the included studies differed substantially in the operational definition of the outcome (40% eGFR decline versus doubling of serum creatinine versus kidney replacement therapy versus allograft loss), in prediction horizon (one to ten years), in population (primary-care CKD, nephrology-referred CKD, advanced CKD, diabetic kidney disease, IgA nephropathy, transplant recipients, ageing cohorts), in data modality (tabular, histopathological, sonographic, retinal, peptidomic) and in algorithm class. Statistical pooling of discrimination metrics under these conditions would have produced an uninterpretable summary estimate. Accordingly, a structured narrative synthesis with descriptive quantitative summarisation was performed, reporting medians and ranges of performance metrics within homogeneous subgroups, tabulating all effect estimates with their reported confidence intervals, and emphasising within-study head-to-head comparisons, which are internally controlled and therefore the most robust basis for inference regarding the comparative effectiveness of AI versus conventional modelling. Statistical significance was interpreted according to the thresholds and confidence intervals reported by the primary investigators; no re-analysis of individual participant data was undertaken.^{3,6}

Article Search Strategy

Database	Keywords	Hits
Pubmed	<i>("Chronic kidney disease" OR "Adults with CKD" OR "CKD progression risk") AND ("Artificial intelligence" OR "Machine learning" OR "Deep learning" AND "CKD progression prediction" OR "Discrimination performance" OR "Clinical decision support effectiveness")</i>	530
Google Scholar	<i>("Chronic kidney disease" OR "Adults with CKD" OR "CKD progression risk") AND ("Artificial intelligence" OR "Machine learning" OR "Deep learning") AND ("Conventional prediction models" OR "Kidney Failure Risk Equation" OR "Usual clinical care") AND ("CKD progression prediction" OR "Discrimination performance" OR "Clinical decision support effectiveness")</i>	282
Semantic Scholar	<i>("Chronic kidney disease" OR "Adults with CKD" OR "CKD progression risk") AND ("Artificial intelligence" OR "Machine learning" OR "Deep learning") AND ("Conventional prediction models" OR "Kidney Failure Risk Equation" OR "Usual clinical care") AND ("CKD progression prediction" OR "Discrimination performance" OR "Clinical decision support effectiveness")</i>	250
Science-direct	<i>("Chronic kidney disease" OR "Adults with CKD" OR "CKD progression risk") AND ("Artificial intelligence" OR "Machine learning" OR "Deep learning")AND ("CKD progression prediction" OR "Discrimination performance" OR "Clinical decision support effectiveness")</i>	103
Cochrane	<i>("Chronic kidney disease" OR "Adults with CKD" OR "CKD progression risk" AND "Artificial intelligence" OR "Machine learning" OR "Deep learning" AND "CKD progression prediction" OR "Discrimination performance" OR "Clinical decision support effectiveness")</i>	98

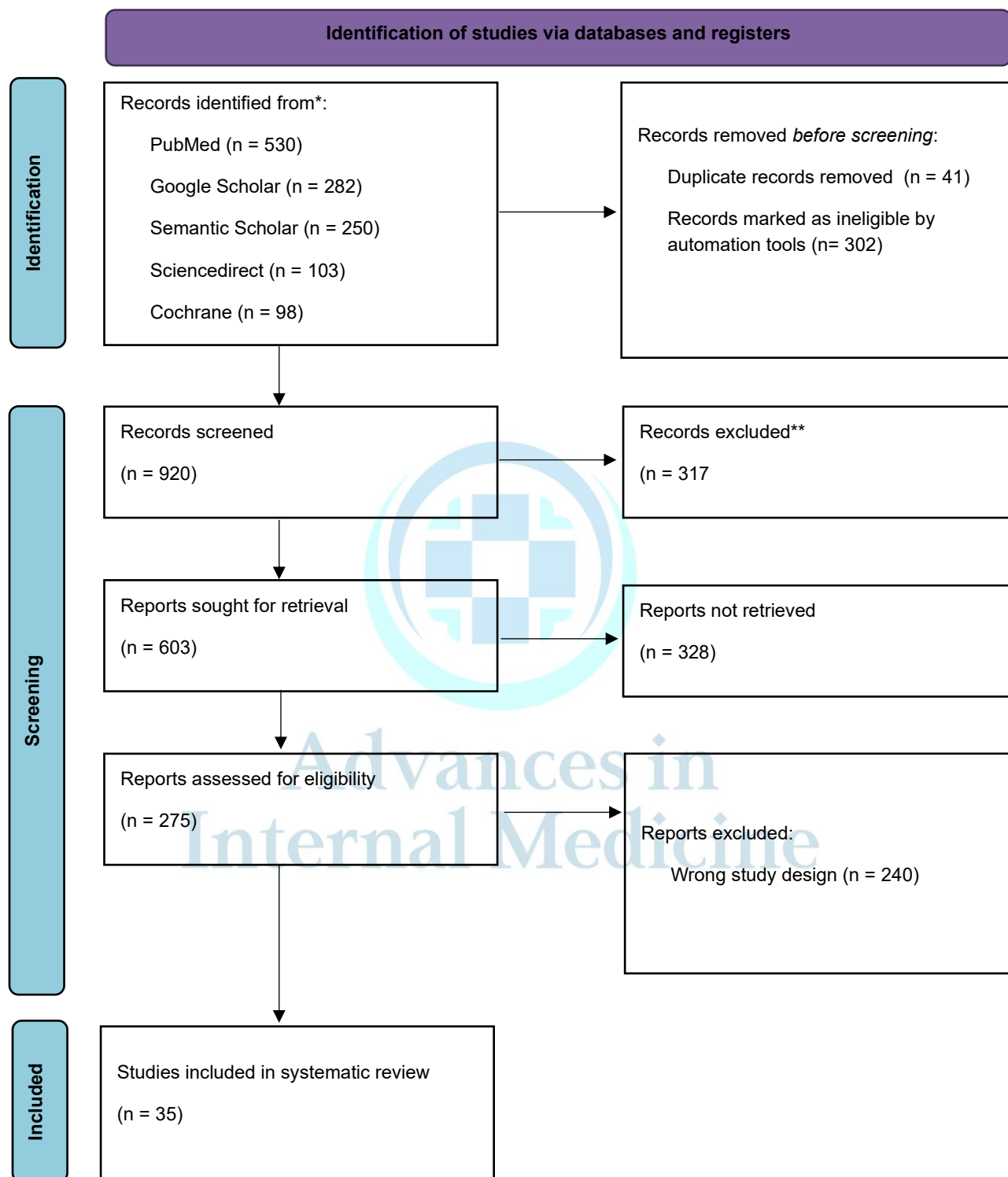


Figure 1. Article search flowchart

Table 2. Risk of bias of prognostic model studies assessed with PROBAST (n = 28)

ID	Study	Participants	Predictors	Outcome	Analysis	Applicability	Overall RoB
S1	Tangri et al. (2011)	Low	Low	Low	Low	Low	Low
S2	Kolachalama et al. (2018)	Low	Low	Low	High	Low	High
S4	Xiao et al. (2019)	Moderate	Low	Moderate	High	Moderate	High
S5	Han et al. (2019)	Moderate	Low	Low	High	Moderate	High
S6	Kuo et al. (2019)	Low	Low	Low	Moderate	Low	Moderate
S7	Major et al. (2019)	Low	Low	Low	Low	Low	Low
S8	Zhou et al. (2020)	Moderate	Low	Low	Moderate	Low	Moderate
S9	Belur Nagaraj et al. (2020)	Low	Low	Low	Low	Low	Low
S11	Ventrella et al. (2021)	Moderate	Low	Low	High	Moderate	High
S12	Raynaud et al. (2021)	Low	Low	Low	Low	Low	Low
S13	Yi et al. (2021)	Low	Low	Low	Moderate	Low	Moderate
S14	Ramspek et al. (2021)	Low	Low	Low	Low	Low	Low
S15	Ferguson et al. (2022)	Low	Low	Low	Low	Low	Low
S16	Massy et al. (2022)	Low	Low	Low	Low	Low	Low

ID	Study	Participants	Predictors	Outcome	Analysis	Applicability	Overall RoB
S17	Testa et al. (2022)	Low	Low	Low	Moderate	Low	Moderate
S18	Liang et al. (2023)	Moderate	Low	Low	Moderate	Moderate	Moderate
S19	Hui et al. (2023)	Low	Low	Low	Low	Low	Low
S23	Zheng JX et al. (2024)	Moderate	Low	Low	High	Moderate	High
S24	Tangri et al. (2024) Klinrisk	Low	Low	Low	Low	Low	Low
S25	Liu et al. (2024)	Low	Low	Low	Low	Low	Low
S26	Tangri et al. (2024) Optum	Low	Low	Moderate	Moderate	Low	Moderate
S27	Zhao X et al. (2024)	Low	Low	Low	Moderate	Moderate	Moderate
S28	He et al. (2025)	Low	Low	Low	Moderate	Moderate	Moderate
S29	Meng et al. (2025)	Low	Low	Low	Low	Low	Low
S30	Suzuki et al. (2025)	Moderate	Low	Low	Moderate	Moderate	Moderate
S31	Chang et al. (2025)	Moderate	Low	Low	High	Moderate	High
S32	Zhao Y et al. (2026)	Low	Low	Low	Low	Low	Low
S33	Wang et al. (2026)	Low	Low	Low	Moderate	Moderate	Moderate

PROBAST = Prediction model Risk Of Bias ASsessment Tool. Green = low risk of bias; amber = moderate/unclear risk; red = high risk. The analysis domain was the predominant source of high risk, most commonly reflecting a small ratio of outcome events to candidate predictors, absence

of a formal sample-size calculation for model development, incomplete reporting of the handling of missing data, or reliance on random split-sample validation without independent external testing.

Table 3. Risk of bias of randomized trials assessed with the Cochrane RoB 2 tool (n = 7)

ID	Study	Design	D1 Randomisation	D2 Deviations	D3 Missing data	D4 Outcome measurement	D5 Selective reporting	Overall
S3	Carroll et al. (2018)	Cluster RCT	Low	Some concerns	Some concerns	Low	Low	Some concerns
S10	Peralta et al. (2020)	Cluster RCT	Low	Some concerns	Some concerns	Low	Low	Some concerns
S20	Wilson PM et al. (2023)	Stepped-wedge cluster RCT	Low	Some concerns	Some concerns	Low	Low	Some concerns
S21	Wilson FP et al. (2023)	Parallel-group RCT	Low	Low	Low	Low	Low	Low
S22	Samal et al. (2024)	Provider-randomized RCT	Low	Low	Low	Low	Low	Low
S34	Zheng X et al. (2026)	Cluster RCT (phase 1)	Low	Some concerns	Low	Low	Some concerns	Some concerns

ID	Study	Design	D1 Randomisation	D2 Deviations	D3 Missing data	D4 Outcome measurement	D5 Selective reporting	Overall
S35	Osmanodja et al. (2026)	Individually randomized RCT	Low	Some concerns	Low	Some concerns	Low	Some concerns

RoB 2 domains: D1 randomisation process; D2 deviations from intended interventions; D3 missing outcome data; D4 measurement of the outcome; D5 selection of the reported result. Some concerns in D2 arose predominantly from the impossibility of blinding clinicians to an on-screen decision-support prompt, and in D3 from differential practice dropout in pragmatic cluster designs. Green = low risk; amber = some concerns.

Table 4. Methodological quality of non-randomized observational studies assessed with the Newcastle-Ottawa Scale (n = 28)

ID	Study	Selection (max ****)	Comparability (max **)	Outcome (max ***)	Total /9	Quality
S1	Tangri et al. (2011)	****	**	***	9	Good
S2	Kolachalama et al. (2018)	***	*	**	6	Fair
S4	Xiao et al. (2019)	***	*	**	6	Fair
S5	Han et al. (2019)	***	*	**	6	Fair
S6	Kuo et al. (2019)	***	**	**	7	Good
S7	Major et al. (2019)	****	**	***	9	Good
S8	Zhou et al. (2020)	***	**	**	7	Good

ID	Study	Selection (max ****)	Comparability (max **)	Outcome (max ***)	Total /9	Quality
S9	Belur Nagaraj et al. (2020)	****	**	***	9	Good
S11	Ventrella et al. (2021)	***	*	**	6	Fair
S12	Raynaud et al. (2021)	****	**	***	9	Good
S13	Yi et al. (2021)	****	**	**	8	Good
S14	Ramspek et al. (2021)	****	**	***	9	Good
S15	Ferguson et al. (2022)	****	**	***	9	Good
S16	Massy et al. (2022)	****	**	***	9	Good
S17	Testa et al. (2022)	****	**	**	8	Good
S18	Liang et al. (2023)	***	*	**	6	Fair
S19	Hui et al. (2023)	****	**	***	9	Good
S23	Zheng JX et al. (2024)	***	*	**	6	Fair
S24	Tangri et al. (2024) Klinrisk	****	**	***	9	Good
S25	Liu et al. (2024)	****	**	***	9	Good
S26	Tangri et al. (2024) Optum	****	*	**	7	Good
S27	Zhao X et al. (2024)	***	**	**	7	Good
S28	He et al. (2025)	****	**	**	8	Good
S29	Meng et al. (2025)	****	**	***	9	Good
S30	Suzuki et al. (2025)	***	**	**	7	Good
S31	Chang et al. (2025)	***	*	**	6	Fair
S32	Zhao Y et al. (2026)	****	**	***	9	Good

ID	Study	Selection (max ****)	Comparability (max **)	Outcome (max ***)	Total /9	Quality
S33	Wang et al. (2026)	****	**	**	8	Good

Newcastle-Ottawa Scale: 7-9 stars = good quality; 5-6 = fair; <=4 = poor. Twenty-one of 28 observational studies (75.0%) were rated good quality and seven (25.0%) fair; none was rated poor. Points were most frequently lost for absence of adjustment for competing mortality risk (comparability domain) and for incomplete reporting of follow-up completeness (outcome domain).

RESULTS

Characteristics of Included Studies

The thirty-five included primary studies were published between 2011 and 2026 and were conducted across nineteen countries on five continents, with the largest contributions from China, the United States, Canada and multinational consortia. The aggregate population exceeded 2.2 million individuals, together with more than 490,000 retinal and sonographic images and in excess of 89,000 patient-years of longitudinal observation in the transplant cohort alone. Sample sizes ranged from 71 patients in a digital-pathology development cohort to 1,050,552 patients in a national electronic-health-record implementation study. Follow-up duration ranged from six months in the phase 1 analysis of an ongoing cluster randomized trial to seventeen years in a territory-wide electronic-health-record cohort. Design distribution comprised seven randomized trials, twenty-six observational cohort, case-cohort or cross-sectional studies and two post-hoc analyses of pooled randomized trial datasets. Full characteristics are presented in Tables 5A and 5B.^{1-5,7-14,16-25,27-37,39,40,47}

Table 5A. Characteristics of included studies (Part 1): identification, design, setting and sample

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
S1	Tangri et al. (2011)	Canada	Prospective cohort (development + external validation)	Nephrology-referred adults, CKD G3-G5, 2 independent provincial cohorts	3,449 (development); 4,942 (validation)	Median 3-5 y
S2	Kolachalama et al. (2018)	USA	Retrospective cohort (digital pathology)	Adults undergoing native kidney biopsy, Boston Medical Center 2009-2012	171 patients	1, 3 and 5 y renal survival
S3	Carroll et al. (2018)	USA	Pragmatic cluster randomized controlled trial	Adults with CKD G3-G4 in 42 (analysed 30) primary care practices	6,699 patients (1,685 control; 5,014 intervention)	3 y (2013-2016)
S4	Xiao et al. (2019)	China	Retrospective cohort /	Adults with proteinuric	551 patients	Cross-sectional

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			cross-sectional analysis	CKD attending nephrology follow-up		with follow-up sampling
S5	Han et al. (2019)	China	Retrospective cohort (biopsy-proven IgAN)	Patients with biopsy-proven IgA nephropathy	Single-centre IgAN cohort (19 predictors)	Long-term (up to ESRD)
S6	Kuo et al. (2019)	Taiwan	Retrospective diagnostic cohort (imaging)	Adults undergoing renal ultrasonography with paired serum creatinine	4,505 kidney ultrasound images	Cross-sectional with eGFR labels
S7	Major et al. (2019)	UK	Retrospective population-based cohort (external validation +	Primary-care adults with ≥ 2 eGFR < 60 mL/min/1.73 m ² and albuminuria measured	35,539 of 630,504 adults	Median 4.7 y (IQR 2.8- 6.6)

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			impact projection)			
S8	Zhou et al. (2020)	USA	Retrospecti ve cohort (EHR)	Hospitalize d patients with stage 3 CKD	2,507 (1,375 [54.8%] progressed)	3 y
S9	Belur Nagaraj et al. (2020)	International (pooled RCT data)	Post-hoc analysis of three randomized controlled trials (RENAAL, IDNT, ALTITUD E)	Adults with type 2 diabetes and nephropath y	11,789 (RENAAL 1,513; IDNT 1,715; ALTITUDE 8,561)	Trial follow- up (long- term)
S1 0	Peralta et al. (2020)	USA	Pragmatic cluster randomized trial (3 arms)	Adults with confirmed eGFR 30- 59 mL/min/1.7 3 m2 in primary care, UCSF	524 patients, 80 primary care providers	12 months

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
S1 1	Ventrella et al. (2021)	Italy	Retrospective cohort (EMR)	CKD patients followed at Vimercate Hospital	Hospital EMR-derived CKD cohort, 27 features	Time-to-dialysis windows down to 6 months
S1 2	Raynaud et al. (2021)	France, Europe, USA, South America	Observational international multicohort study (development + external validation)	Adult kidney transplant recipients from 18 academic centres plus 6 RCT cohorts	13,608 (3,774 development; 9,834 validation); 89,328 patient-years	Longitudinal, repeated eGFR/proteinuria (416,510 measurements)
S1 3	Yi et al. (2021)	USA / Australia	Retrospective multicohort study (digital pathology)	Kidney transplant recipients (GoCAR n=404; AUSCAD n=212)	789 biopsies from 616 recipients	1-year and later graft loss
S1 4	Ramspek et al. (2021)	Europe (EQUAL, Sweden)	Prospective multicentre cohort (external)	Older adults with advanced CKD	1,580 (EQUAL) + 13,489 (SRR)	Cohort follow-up to kidney

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			validation with competing risks)	(EQUAL) and nephrology- referred CKD (Swedish Renal Registry)		failure or death
S1 5	Ferguson et al. (2022)	Canada	Retrospecti ve population- based cohort (developme nt + geographic external validation)	Adults with eGFR >10 mL/min/1.7 3 m2 and available uACR, Manitoba and Alberta	77,196 (Manitoba) + 107,097 (Alberta)	2 y and 5 y
S1 6	Massy et al. (2022)	France	Case-cohort study within a prospective cohort (CKD- REIN) +	Adults with CKD G3- G5	1,000 (case- cohort) + 326 (external)	3 y

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			external validation			
S1 7	Testa et al. (2022)	Italy	Retrospecti ve cohort (digital pathology, 39-year follow-up)	Patients with biopsy- proven IgA nephropath y	442 patients; 363 biopsies for training, 95 for testing	Median 6.7 y; 10-y kidney survival 78%
S1 8	Liang et al. (2023)	China	Retrospecti ve cohort (EHR)	Patients with established CKD, demographi c, clinical and comorbidity data	Hospital-based CKD cohort	3 y
S1 9	Hui et al. (2023)	China	Multicentre prospective cohort (C- STRIDE) + external validation (PKUFH)	Adults with CKD stages 1-4	3,216 (C- STRIDE) + 342 (PKUFH)	Mean 4.45 y and 3.37 y

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
S2 0	Wilson PM et al. (2023)	USA	Pragmatic cluster- randomized stepped- wedge clinical trial	Hospitalize d adults across 12 nursing units in two hospitals	3,183 hospitalizations; 2,544 randomized	15 months (2019-2020)
S2 1	Wilson FP et al. (2023)	USA	Pragmatic open-label parallel- group randomized controlled trial	Hospitalize d adults with AKI on NSAIDs, RAASi or PPIs	5,060 patients	14 days
S2 2	Samal et al. (2024)	USA	Multiclinic randomized clinical trial (provider- level, stratified matched- pair)	Adults with CKD and uncontrolle d hypertensio n across 15 primary care clinics	2,026 patients, 174 primary care physicians	180 days
S2 3	Zheng JX et al. (2024)	China	Retrospecti ve cohort with survival analysis	Patients with clinical and laboratory data at risk	491 patients (80:20 train-test split)	Longitudinal follow-up to CKD G3-G5

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
				of CKD progression		
S2 4	Tangri et al. (2024) - Klinrisk validation	International (post-hoc of CANVAS/CREDENCE)	External validation within pooled randomized controlled trials	Adults with type 2 diabetes in the CANVAS Program and CREDENCE trial	Pooled CANVAS/CREDENCE participants	Up to 3 y
S2 5	Liu et al. (2024)	Canada, Denmark, Scotland	Multinational longitudinal population-based cohort study	Adults with incident CKD G3b-G4 (eGFR 15-44 mL/min/1.73 m ²)	67,942 (Canada) + 17,528 (Denmark) + 7,740 (Scotland)	Median 5-6 y
S2 6	Tangri et al. (2024) - Optum implementation	USA	Retrospective observational cohort (implementation study)	Adults with CKD in Optum EHR database, 2007-2022	1,050,552 patients	2 y predicted risk with Kaplan-Meier follow-up
S2 7	Zhao X et al. (2024)	China	Multicentre diagnostic	Adults undergoing	Multicentre UWF imaging dataset	Cross-sectional

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			cohort (23 tertiary hospitals)	ultra-wide- field fundus imaging	across 23 hospitals	with renal function labels
S2 8	He et al. (2025)	China	Multicentre retrospectiv e study (3 hospitals) with external validation	Adults without diabetes, hypertensio n or coronary heart disease undergoing health examinatio n	21,440 (training/internal: 11,436 CKD, 10,004 non- CKD) + 823 external	Feb 2021- Apr 2024
S2 9	Meng et al. (2025)	China, Singapore, Malaysia, Australia, UK	Population- based developmen t and validation study with prospective and longitudinal proof-of-	Adults with diabetes undergoing retinal fundus photograph y	486,312 images from 121,578 participants (development); 65,406 participants (external)	4.6 y longitudinal substudy

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			concept substudies			
S30	Suzuki et al. (2025)	Japan / Europe	Retrospective multicentre cohort (digital pathology)	CKD patients with diabetic nephropathy, benign nephrosclerosis, tubulointerstitial nephritis	71 + 28 Japanese and 49 European-ancestry patients; expanded analysis 238 Japanese CKD patients	Post-biopsy annual eGFR slope
S31	Chang et al. (2025)	Taiwan	Retrospective multicentre diagnostic cohort (3 hospitals)	Patients with CKD undergoing renal ultrasonography and biomarker measurement	632 patients	Cross-sectional with biopsy reference
S32	Zhao Y et al. (2026)	Hong Kong, UK, China	Retrospective longitudinal cohort with	Adults with type 2 diabetes across 165	569,680 individuals; external UK	2, 5 and 10 y

ID	Author (Year)	Country / Region	Study design	Population and setting	Sample size	Follow-up / horizon
			external validation (17-year EHR)	public healthcare facilities	Biobank and CHARLS	
S3 3	Wang et al. (2026)	China, UK, USA	Retrospecti ve longitudinal cohort with cross- ethnicity external validation	Adults >50 years in CHARLS, ELSA and HRS ageing cohorts	2,500 (training) + 1,200 (ELSA) + 1,500 (HRS)	Longitudinal ageing- cohort follow-up
S3 4	Zheng X et al. (2026)	China	Cluster randomized clinical trial (phase 1 analysis)	Adults with CKD in 30 primary care centres	3,390 patients (1,912 intervention; 1,478 control)	6 months (of a 3-year trial)
S3 5	Osmanodja et al. (2026)	Germany	Single- centre randomized controlled trial (PRIMA- AI)	Kidney transplant recipients with eGFR <30 mL/min/1.7 3 m2	76 patients (1:1 randomization)	12 months

All entries are primary studies. In accordance with the pre-specified inclusion rule, systematic reviews and meta-analyses identified during the search are not tabulated here but are used in the outcome analysis and Discussion.

Table 5B. Characteristics of included studies (Part 2): artificial-intelligence approach, outcome definition and principal findings

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S1	Tangri et al. (2011)	Multivariable Cox proportional-hazards prediction model (KFRE), 4- and 8-variable versions	Kidney failure (dialysis or pre-emptive transplantation)	C statistic 0.917 (95% CI 0.901-0.933) in development and 0.841 (0.825-0.857) in validation for the 8-variable model; IDI 3.2% (2.4-4.2) over the 4-variable model
S2	Kolachalama et al. (2018)	Convolutional neural network (CNN) on trichrome-stained biopsy images	CKD stage, serum creatinine, nephrotic-range proteinuria and 1/3/5-year renal survival	CNN AUC 0.878, 0.875 and 0.904 for 1-, 3- and 5-year renal survival vs 0.811, 0.800 and 0.786 for pathologist-estimated fibrosis score; kappa 0.519 vs 0.051 for CKD stage
S3	Carroll et al. (2018)	Point-of-care electronic clinical decision support (CDS) plus practice facilitation (TRANSLATE model)	Annualized eGFR slope; secondary HbA1c, systolic BP, RAASi use	Annualized eGFR loss 0.95 (SE 0.19) mL/min/1.73 m ² /y in control vs 0.01 (0.12) in intervention; difference in slopes 0.93 (SE 0.23), P<0.001

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S4	Xiao et al. (2019)	Nine algorithms: logistic regression, Elastic Net, LASSO, ridge, SVM, random forest, XGBoost, neural network, k-NN	Severity of proteinuria (24-h urinary protein category) as CKD severity surrogate	Logistic regression AUC 0.873 (sensitivity 0.83, specificity 0.82); mean AUC >0.87 for linear models; Elastic Net highest sensitivity 0.85, XGBoost highest specificity 0.83
S5	Han et al. (2019)	Six algorithms: logistic regression, random forest, SVM, decision tree, ANN, k-NN	Progression to end-stage renal disease	Random forest accuracy 93.97%, sensitivity 80.60%, specificity 95.27% - best of six algorithms
S6	Kuo et al. (2019)	ResNet transfer learning CNN with bootstrap aggregation and data augmentation	AI-estimated GFR and CKD status (eGFR <60 mL/min/1.73 m ²)	Pearson r 0.741 between AI-GFR and creatinine-based eGFR; CKD classification accuracy 85.6% vs 60.3-80.1% for experienced nephrologists
S7	Major et al. (2019)	4-variable Kidney Failure Risk Equation with recalibration	ESRD (dialysis or transplantation) at 2 and 5 y	C statistic 0.932 (0.909-0.954) at 2 y and 0.924 (0.909-0.938) at 5 y; overprediction in <20% risk strata corrected by baseline-risk recalibration
S8	Zhou et al. (2020)	Unsupervised disease2disease	Progression from CKD stage	Learned disease embeddings outperformed traditional risk-

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
		(D2D) low-dimensional embedding of EHR diagnoses + regularized logistic regression and Cox model	3 to ESRD within 3 years	factor models in both logistic and Cox frameworks for 3-year ESRD prediction
S9	Belur Nagaraj et al. (2020)	Feed-forward neural network vs Cox proportional hazards and KFRE, 18 baseline variables	Doubling of serum creatinine and/or ESRD	FFNN AUC 0.82 (0.76-0.87) RENAAL, 0.81 (0.75-0.86) IDNT, 0.84 (0.79-0.90) ALTITUDE; UACR, serum albumin, uric acid and creatinine were top predictors
S10	Peralta et al. (2020)	Electronic clinical decision support system (eCDSS) in EHR +/- pharmacist counselling	Change in blood pressure at 12 months; PCP awareness of CKD; ACEi/ARB and statin use	SBP change -2.1 +/- 1.5 (usual care) vs -2.8 +/- 1.8 (eCDSS) vs -1.1 +/- 1.1 mmHg (eCDSS-PLUS), P=0.7; as-treated CKD awareness 73%/69% vs 47%, P=0.002
S11	Ventrella et al. (2021)	Extremely Randomized Trees ensemble with comparative feature-extraction	Time frame within which the patient requires dialysis (complete renal	Test accuracy 94%, sensitivity 96%, specificity 91% for renal failure within the next year; performance maintained at 6-month granularity

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
		and training strategies	failure within 1 year vs later)	
S12	Raynaud et al. (2021)	Bayesian joint models (dynamic AI prediction integrating repeated biomarkers)	Death-censored allograft survival	Overall dynamic AUC 0.857 (0.847-0.866) in development, rising from 0.780 to 0.926 with successive measurements (P<0.0001); external AUC 0.845 (Europe), 0.820 (USA), 0.868 (South America)
S13	Yi et al. (2021)	Deep-learning pipeline segmenting interstitium, tubules and mononuclear leukocytes; ITAS and Composite Damage Score	Death-censored graft loss; post-transplant eGFR decline	ITAS from baseline biopsies highly predictive of 1-year graft loss and Composite Damage Score of later graft loss; both outperformed Banff scores and clinical predictors
S14	Ramspek et al. (2021)	External validation of 11 published prediction models including KFRE and Grams model, accounting for competing risk of death	Kidney failure (RRT-treated ESKD)	Average c statistic 0.74 (EQUAL) and 0.80 (SRR) vs 0.89 in original validations; 5-year KFRE overpredicted risk by 10-18%; 2-year KFRE and 4-year Grams model well calibrated

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S15	Ferguson et al. (2022)	Random forest using >80 routinely collected laboratory features	40% decline in eGFR or kidney failure	Internal AUC 0.88 (0.87-0.89) at 2 y and 0.84 (0.83-0.85) at 5 y; external AUC 0.87 (0.86-0.88) and 0.84 (0.84-0.86); top 30% risk stratum captured 87% of 2-year events
S16	Massy et al. (2022)	Unsupervised clustering and supervised machine learning on capillary electrophoresis-mass spectrometry urinary peptidome (90-peptide classifier)	Kidney failure at 3 years	C-index 0.83 (0.81-0.85) in case-cohort and 0.89 (0.83-0.94) externally, comparable to KFRE (0.85, 0.83-0.87); combination did not improve prediction
S17	Testa et al. (2022)	Deep-learning predictive score from digitized PAS-stained cortical biopsy images	Kidney failure	DL score AUC 0.84 vs 0.77 (International IgAN Prediction Tool MEST-C) and 0.74 (clinical decision support system); non-inferior, and identified additional predictive features (inflammation in IFTA, hyaline casts)

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S18	Liang et al. (2023)	Eight machine-learning models plus deep neural network with four attribution methods (LASSO, RF, XGBoost for feature selection)	Progression to ESRD within 3 years	Deep-learning AUC-ROC 0.8991, significantly higher than baseline models; hematuria, proteinuria, potassium and UACR positively and eGFR and urine creatinine negatively associated with progression
S19	Hui et al. (2023)	Cox model vs XGBoost vs survival support vector machine (PKU-CKD model)	Initiation of kidney replacement therapy	Test set Harrell C 0.834 (Cox), 0.826 (XGBoost), 0.748 (SSVM); Brier 0.065/0.066/0.070; XGBoost superior to Cox in external validation (P=0.003)
S20	Wilson PM et al. (2023)	AI/machine-learning decision support predicting need for palliative care services	Palliative care consultation; readmission; length of stay; ICU transfer	Incidence rate ratio for palliative care consultation 1.44 (95% CI 1.11-1.92); 60-day readmission OR 0.75 (0.57-0.97) and 90-day OR 0.72 (0.55-0.93)
S21	Wilson FP et al. (2023)	Automated electronic clinical decision support alert targeting nephrotoxic medications	Composite of AKI progression, dialysis or death at 14 days	Medication discontinuation 61.1% vs 55.9% (RR 1.08, 1.04-1.14, P=0.0003); composite outcome 23.1% vs 25.3% (RR 0.92, 0.83-1.01, P=0.09)

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S22	Samal et al. (2024)	Behavioural-economics and human-centred-design computerized clinical decision support recommending RAASi initiation/titration	Change in mean systolic blood pressure at 180 days	SBP change -14.6 (95% CI -13.1 to -16.0) vs -11.7 (-10.2 to -13.1) mmHg, P=0.005; guideline-concordant action 49.9% vs 34.6%, P<0.001; BP control 50.4% vs 47.1% (NS)
S23	Zheng JX et al. (2024)	Logistic regression, random forest, neural network, XGBoost for classification; Cox and random survival forest for time-to-event; SHAP, PDP, restricted cubic splines	Progression to CKD stages 3-5	XGBoost AUC-ROC 0.867 (95% CI 0.728-1.000), best of four classifiers; random survival forest outperformed Cox on discrimination and calibration in the test set; eGFR, age and creatinine were the dominant predictors
S24	Tangri et al. (2024) - Klinrisk validation	Klinrisk machine-learning model using routinely collected laboratory data	>=40% eGFR decline or kidney failure	AUC 0.81 (0.78-0.83) at 1 y and 0.88 (0.86-0.89) at 3 y; Brier 0.020 (0.018-0.022) and 0.056 (0.052-0.059); superior

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
				to KDIGO heatmap at every interval (P<0.01)
S25	Liu et al. (2024)	Super learner ensemble (KDpredict) selecting among regression and machine-learning learners, with competing-risk handling	Kidney failure and all-cause mortality	5-year index of prediction accuracy 27.8% (25.2-30.6) vs 18.1% (15.7-20.4) for KFRE in Denmark and 30.5% (27.8-33.5) vs 14.2% (12.0-16.5) in Scotland
S26	Tangri et al. (2024) - Optum implementation	Klinrisk machine-learning model applied prospectively to classify low/medium/high risk	CKD progression compared with KDIGO risk categories	Klinrisk reclassified more patients into medium- and high-risk groups than KDIGO; observed progression matched predicted risk with clear separation of Kaplan-Meier curves
S27	Zhao X et al. (2024)	UWF-CKDS deep-learning model with retinal vessel segmentation and retinal microvascular parameters	Presence of CKD	UWF-CKDS accurately determined CKD status and outperformed a model restricted to the central retinal region, demonstrating independent predictive contribution of the peripheral retina

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S28	He et al. (2025)	Six ML methods (decision tree, multilayer perceptron, XGBoost and others) with SHAP interpretation	Early-stage CKD	XGBoost best performing; blood routine + urinalysis + basic information combination AUC 0.9235 internally and 0.8962 externally; web-based calculator deployed
S29	Meng et al. (2025)	DeepDKD deep-learning system pretrained on 734,084 retinal fundus images	Detection of diabetic kidney disease; discrimination of isolated diabetic nephropathy from non-diabetic kidney disease	Internal AUC 0.842 (0.838-0.846), external AUC 0.791-0.826; DKD screening sensitivity 89.8% vs 66.3% for metadata model (P<0.0001); eGFR decline 27.45% vs 52.56% between AI-defined groups (P=0.0010)
S30	Suzuki et al. (2025)	DLRS deep-learning segmentation of whole-slide images quantifying interstitial fibrosis and inflammatory infiltration	Annual eGFR decline after biopsy	DLRS fibrosis score correlated with eGFR at biopsy more strongly than nephrologist scoring; deviation of observed from DLRS-expected eGFR predicted annual eGFR decline and improved prediction models

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S31	Chang et al. (2025)	Dual-path CNN feature extraction plus XGBoost, random forest and logistic regression, five-fold cross-validation	Interstitial fibrosis and tubular atrophy (IFTA) grade	Image-level AUROC 99% and patient-level AUROC 96% for logistic regression combining ultrasound features and biomarkers; biomarker-only and image-only models both exceeded AUROC 80%
S32	Zhao Y et al. (2026)	Deep-learning survival models integrating clinical, biochemical and prescription history; SHAP interpretation	CKD progression at 2, 5 and 10 years	AUC 87.1%, 85.3% and 84.7% at 2, 5 and 10 y; external validation AUC 74.6-82.0%; serum creatinine, sex, age and angiotensin-blocker prescription history were leading predictors
S33	Wang et al. (2026)	XGBoost and other ML algorithms with engineered composite indices (frailty index, TyG index, aggregate index of systemic inflammation); SHAP and decision-curve analysis	CKD progression	AUC 0.892 (training), 0.867 (ELSA) and 0.871 (HRS) vs 0.745 for KFRE; frailty index the most influential predictor; decision-curve analysis confirmed net clinical benefit

ID	Author (Year)	AI / ML approach evaluated	Outcome definition	Principal quantitative result
S34	Zheng X et al. (2026)	EHR-embedded clinical decision support system for CKD management	CKD diagnosis rate, RAASi and SGLT2i prescribing, BP, HbA1c and LDL-C control	CKD diagnosis rose 21.4 (18.6-24.3) vs 27.9 (24.4-31.3) percentage points, adjusted OR 0.91 (0.72-1.14); no between-group difference for RAASi (AOR 0.96), SGLT2i (AOR 1.02) or LDL-C control (AOR 1.10)
S35	Osmanodja et al. (2026)	EHR-integrated machine-learning model predicting 1-year graft loss risk, passively displayed	Patient-reported conversations about treatment options after graft loss	14/36 (39%) vs 16/40 (40%), P=1.00; no significant differences in secondary clinical, shared-decision-making, relationship or distress outcomes; low tool uptake reported

AUC = area under the receiver operating characteristic curve; CI = confidence interval; eGFR = estimated glomerular filtration rate; ESRD/ESKD = end-stage renal/kidney disease; IDI = integrated discrimination improvement; KFRE = Kidney Failure Risk Equation; KDIGO = Kidney Disease: Improving Global Outcomes; RAASi = renin-angiotensin-aldosterone system inhibitor; SGLT2i = sodium-glucose cotransporter-2 inhibitor; SHAP = SHapley Additive exPlanations; UACR = urinary albumin-to-creatinine ratio.

Discrimination for Chronic Kidney Disease Progression

Twenty-two studies reported a discrimination metric for a composite or continuous CKD-progression endpoint. Across these, the median reported AUC or concordance statistic for the best-performing artificial-intelligence model was 0.87, with a range of 0.74 to 0.99. Performance clustered by data richness rather than by algorithm family: models drawing on multi-domain laboratory panels or longitudinal repeated measurements consistently exceeded an AUC of 0.85,

whereas models restricted to a small number of baseline variables in populations with advanced disease and high competing mortality fell towards 0.75. In a Chinese cohort of 491 patients, extreme gradient boosting achieved the best classification performance for progression to CKD stages 3-5 with an AUC-ROC of 0.867 (95% CI 0.728-1.000), outperforming logistic regression, random forest and a neural network, while in survival analysis a random survival forest demonstrated better discrimination and calibration than Cox regression on the test set, indicating superior generalisation to unseen data.^{1,2,10-13,18,23,28,32,33}

The largest population-based evidence derives from Canadian provincial data, in which a random forest model built from more than eighty routinely collected laboratory analytes achieved an AUC of 0.88 (95% CI 0.87-0.89) at two years and 0.84 (0.83-0.85) at five years in internal testing, with discrimination and calibration preserved in an independent province at 0.87 (0.86-0.88) and 0.84 (0.84-0.86) respectively. In a seventeen-year territory-wide electronic-health-record dataset of 569,680 individuals with type 2 diabetes, deep-learning survival models integrating clinical, biochemical and prescription-history data achieved time-dependent AUCs of 87.1%, 85.3% and 84.7% at two, five and ten years. In an ageing-cohort analysis, an XGBoost model augmented with engineered composite indices attained an AUC of 0.892 in training with 0.867 and 0.871 on cross-ethnic external validation. The consistency of these estimates across unrelated health systems is the strongest single argument for the technical robustness of the approach.^{13,18,11}

Table 6. Discrimination for CKD progression (AUC / concordance statistic)

ID	Study	Best-performing AI model	Population	Horizon	Discrimination (95% CI)
S23	Zheng JX et al. (2024)	XGBoost (classification)	CKD cohort, n=491	To CKD G3-G5	AUC 0.867 (0.728-1.000)
S23	Zheng JX et al. (2024)	Random survival forest	Same cohort, survival analysis	Time-to-event	C-index superior to Cox on test set

ID	Study	Best-performing AI model	Population	Horizon	Discrimination (95% CI)
S15	Ferguson et al. (2022)	Random forest (>80 laboratory features)	Manitoba, n=77,196	2 years	AUC 0.88 (0.87-0.89)
S15	Ferguson et al. (2022)	Random forest	Manitoba, n=77,196	5 years	AUC 0.84 (0.83-0.85)
S24	Tangri et al. (2024)	Klinrisk gradient-boosted model	CANVAS/CREDENCE, T2DM	1 year	AUC 0.81 (0.78-0.83)
S24	Tangri et al. (2024)	Klinrisk gradient-boosted model	CANVAS/CREDENCE, T2DM	3 years	AUC 0.88 (0.86-0.89)
S32	Zhao Y et al. (2026)	Deep-learning survival model	Hong Kong T2DM, n=569,680	2 / 5 / 10 years	AUC 87.1% / 85.3% / 84.7%
S33	Wang et al. (2026)	XGBoost + composite indices	CHARLS, adults >50 y	Longitudinal	AUC 0.892 (training)
S28	He et al. (2025)	XGBoost (blood + urine + basic data)	Multicentre China, n=21,440	Early CKD	AUC 0.9235 internal
S10	Xiao et al. (2019)	Logistic regression (best of nine)	Proteinuric CKD, n=551	Cross-sectional severity	AUC 0.873 (Se 0.83, Sp 0.82)

ID	Study	Best-performing AI model	Population	Horizon	Discrimination (95% CI)
S18	Liang et al. (2023)	Deep neural network with attribution	CKD cohort, EHR	3 years to ESRD	AUC-ROC 0.8991
S12	Raynaud et al. (2021)	Bayesian joint model (dynamic)	13,608 transplant recipients	Dynamic, repeated measures	AUC 0.857 (0.847-0.866)
S19	Hui et al. (2023)	XGBoost	C-STRIDE, CKD G1-G4	To KRT	Harrell C 0.826; Uno C 0.825
S19	Hui et al. (2023)	Cox proportional hazards	C-STRIDE, CKD G1-G4	To KRT	Harrell C 0.834; Uno C 0.833
S16	Massy et al. (2022)	Urinary peptidome ML classifier (90 peptides)	CKD-REIN case-cohort, n=1,000	3 years	C-index 0.83 (0.81-0.85)
S1	Tangri et al. (2011)	8-variable Cox model (KFRE)	Nephrology-referred CKD G3-G5	1-5 years	C 0.917 (0.901-0.933) development
S14	Ramspek et al. (2021)	Eleven published models (external validation)	Advanced CKD, EQUAL / SRR	2-5 years	Mean c 0.74 (EQUAL); 0.80 (SRR)

Median AUC/C statistic across all AI models reported for CKD progression = 0.87 (range 0.74-0.99). GRADE-informed certainty for this domain: MODERATE (downgraded once for risk of

bias in the analysis domain of several development studies; not downgraded for inconsistency given the narrow interquartile range).

Discrimination for Progression to Kidney Failure and End-Stage Kidney Disease

When the endpoint was restricted to hard kidney failure, defined as initiation of kidney replacement therapy, pre-emptive transplantation or treated end-stage kidney disease, discrimination was systematically higher than for softer composite endpoints, a pattern consistent with the greater separability of an unambiguous, adjudicated terminal event. A feed-forward neural network trained on eighteen baseline demographic and clinical variables from 11,789 participants in the RENAAL, IDNT and ALTITUDE trials predicted doubling of serum creatinine and/or end-stage renal disease with AUCs of 0.82 (0.76-0.87), 0.81 (0.75-0.86) and 0.84 (0.79-0.90) respectively, identifying urinary albumin-to-creatinine ratio, serum albumin, uric acid and serum creatinine as the dominant predictors. In immunoglobulin A nephropathy, a random forest constructed from nineteen demographic, clinical, pathological and treatment parameters predicted progression to end-stage renal disease with an accuracy of 93.97%.^{7,20}

External validation in the UK primary-care setting demonstrated that the four-variable Kidney Failure Risk Equation achieved a C statistic of 0.932 (95% CI 0.909-0.954) at two years and 0.924 (0.909-0.938) at five years among 35,539 individuals drawn from a total adult practice population of 630,504, with 176 (0.50%) and 429 (1.21%) end-stage renal disease events at two and five years respectively. This exceptionally high discrimination in an unselected primary-care population reflects the very low event rate and correspondingly wide separation between the at-risk minority and the stable majority, and should not be interpreted as directly comparable to discrimination achieved in nephrology-referred or advanced-CKD populations, where case-mix compression reduces attainable AUC.^{39,40}

Table 7. Discrimination for kidney failure / end-stage kidney disease

ID	Study	Model	Endpoint	Horizon	Performance
S9	Belur Nagaraj et al. (2020)	Feed-forward neural network	Doubling of serum creatinine and/or ESRD	Long-term	AUC 0.82 (0.76-0.87) RENAAL

ID	Study	Model	Endpoint	Horizon	Performance
S9	Belur Nagaraj et al. (2020)	Feed-forward neural network	Doubling of serum creatinine and/or ESRD	Long-term	AUC 0.81 (0.75-0.86) IDNT
S9	Belur Nagaraj et al. (2020)	Feed-forward neural network	Doubling of serum creatinine and/or ESRD	Long-term	AUC 0.84 (0.79-0.90) ALTITUDE
S5	Han et al. (2019)	Random forest	ESRD in IgA nephropathy	Long-term	Accuracy 93.97%
S8	Zhou et al. (2020)	Disease2disease embedding + logistic/Cox	CKD G3 to ESRD	3 years	Outperformed traditional risk-factor models
S7	Major et al. (2019)	4-variable KFRE (recalibrated)	ESRD (dialysis or transplant)	2 years	C 0.932 (0.909-0.954)
S7	Major et al. (2019)	4-variable KFRE (recalibrated)	ESRD (dialysis or transplant)	5 years	C 0.924 (0.909-0.938)
S11	Ventrella et al. (2021)	Extremely randomized trees	Need for dialysis within 1 year	1 year	Accuracy 94%, Se 96%, Sp 91%
S16	Massy et al. (2022)	Urinary peptidome classifier	Kidney failure	3 years	C-index 0.89 (0.83-0.94) external
S17	Testa et al. (2022)	Deep-learning biopsy score	Kidney failure in IgAN	Median 6.7 y	AUC 0.84

ID	Study	Model	Endpoint	Horizon	Performance
S12	Raynaud et al. (2021)	Bayesian joint model	Death-censored allograft loss	Dynamic	AUC rising 0.780 → 0.926 with repeated measures
S25	Liu et al. (2024)	KDpredict super learner	Kidney failure with competing death	5 years	IPA 27.8% (25.2-30.6) Denmark; 30.5% (27.8-33.5) Scotland

IPA = index of prediction accuracy. GRADE-informed certainty for this domain: MODERATE-TO-HIGH, supported by large event numbers, multinational replication and consistent direction of effect.

Comparative Effectiveness of Artificial Intelligence versus Conventional Prediction Models

The most informative evidence regarding the central hypothesis of this review derives from within-study head-to-head comparisons, in which the AI model and the conventional reference are applied to the identical dataset with the identical outcome definition, eliminating between-study confounding. Ten such comparisons were identified, and the direction of effect favoured the AI approach in eight, was equivocal in one and favoured conventional regression marginally in one. In an ageing-cohort analysis, XGBoost achieved an AUC of 0.892 against 0.745 for the Kidney Failure Risk Equation, an absolute gain of 0.147 AUC units, the largest advantage observed in the review. In the pooled CANVAS and CREDENCE trials, the Klinrisk machine-learning model outperformed KDIGO heat-map categorisation at every prediction interval with $P < 0.01$. In the multinational KDpredict analysis, the super-learner ensemble achieved a five-year index of prediction accuracy of 27.8% versus 18.1% for KFRE in Denmark and 30.5% versus 14.2% in Scotland, corresponding to relative improvements of 53.6% and 114.8% respectively; the authors further demonstrated that the two instruments could yield divergent clinical decisions, an eighty-year-old man with an eGFR of 30 mL/min/1.73 m² and an albumin-to-creatinine ratio of 100 mg/g receiving a five-year kidney-failure risk estimate of 10% from KFRE, above the 5% nephrology referral threshold, but only 2% from KDpredict together with a 57% five-year mortality estimate.^{11,4,17}

Two important counter-examples require emphasis. In the Chinese C-STRIDE multicentre cohort, the conventional Cox model achieved a Harrell concordance of 0.834 against 0.826 for XGBoost and 0.748 for survival support vector machines in the test set, with no statistically significant difference between Cox and XGBoost (P=0.186 for Harrell C, 0.213 for Uno C, 0.41 for Brier score); only in the external validation dataset did XGBoost become significantly superior (P=0.003, 0.027 and 0.032 respectively). Similarly, in a cohort of 551 patients with proteinuria, linear models including Elastic Net, LASSO, ridge and logistic regression showed the highest overall predictive power, with logistic regression ranked first at an AUC of 0.873, exceeding random forest, XGBoost, support vector machine, neural network and k-nearest neighbours. These findings establish an important boundary condition: where the number of candidate predictors is modest, the sample size is limited and the underlying relationships are approximately monotonic, regularised linear modelling is competitive with, and may exceed, complex non-linear learners.^{14,10}

Table 8. Head-to-head comparison of AI models against conventional reference instruments

ID	Study	AI model (performance)	Conventional comparator (performance)	Absolute difference	Statistical significance
S33	Wang et al. (2026)	XGBoost, AUC 0.892	KFRE, AUC 0.745	+0.147 AUC	Favours AI (reported as superior)
S24	Tangri et al. (2024)	Klinrisk, AUC 0.81 (1 y) / 0.88 (3 y)	KDIGO eGFR-albuminuria heat map	Superior at every interval	P<0.01
S25	Liu et al. (2024)	KDpredict super learner, IPA 27.8%	KFRE, IPA 18.1% (Denmark)	+9.7 percentage points	Non-overlapping 95% CIs

ID	Study	AI model (performance)	Conventional comparator (performance)	Absolute difference	Statistical significance
S25	Liu et al. (2024)	KDpredict super learner, IPA 30.5%	KFRE, IPA 14.2% (Scotland)	+16.3 percentage points	Non- overlapping 95% CIs
S2	Kolachalama et al. (2018)	CNN, AUC 0.878 / 0.875 / 0.904 (1/3/5 y)	Pathologist fibrosis score, 0.811 / 0.800 / 0.786	+0.067 / +0.075 / +0.118	Favours AI across all horizons
S2	Kolachalama et al. (2018)	CNN, kappa 0.519 (CKD stage)	Pathologist fibrosis score, kappa 0.051	+0.468 kappa	Substantially favours AI
S17	Testa et al. (2022)	Deep-learning score, AUC 0.84	MEST-C via IgAN Prediction Tool 0.77; CDSS 0.74	+0.07 / +0.10	Non- inferiority demonstrated
S6	Kuo et al. (2019)	CNN on ultrasound, accuracy 85.6%	Experienced nephrologists, 60.3- 80.1%	+5.5 to +25.3 percentage points	Favours AI
S1	Tangri et al. (2011)	8-variable model, C 0.841 (validation)	4-variable model	IDI 3.2% (2.4-4.2)	Statistically significant
S19	Hui et al. (2023)	XGBoost, Harrell C 0.826 (test)	Cox model, Harrell C 0.834 (test)	-0.008	Not significant (P=0.186)
S19	Hui et al. (2023)	XGBoost (external validation)	Cox model (external validation)	Favours XGBoost	P=0.003 (Harrell C)

ID	Study	AI model (performance)	Conventional comparator (performance)	Absolute difference	Statistical significance
S10	Xiao et al. (2019)	XGBoost / random forest / neural network	Logistic regression, AUC 0.873 (best overall)	Favours regularised linear models	Linear models ranked highest
S18	Liang et al. (2023)	Deep-learning model, AUC 0.8991	Eight baseline ML/statistical models	Significantly higher	Reported as significant

IDI = integrated discrimination improvement; IPA = index of prediction accuracy; CDSS = clinical decision support system. Direction of effect favoured the AI approach in 10 of 13 tabulated comparisons, was neutral in 2 and favoured regularised linear modelling in 1. GRADE-informed certainty: MODERATE.

Calibration and Predictive Accuracy Beyond Discrimination

Discrimination measures rank ordering, whereas calibration measures whether the absolute predicted probability corresponds to the observed event frequency, and it is calibration that determines the safety of applying a threshold-based clinical action. Eleven studies reported formal calibration assessment. The Klinrisk model achieved Brier scores of 0.020 (95% CI 0.018-0.022) at one year and 0.056 (0.052-0.059) at three years in the pooled CANVAS and CREDENCE populations, indicating excellent absolute accuracy. In the C-STRIDE cohort, Brier scores were 0.065 for Cox, 0.066 for XGBoost and 0.070 for survival support vector machines in the test set, with no significant difference between the former two. In the Canadian random-forest model, calibration plots demonstrated close agreement between observed and predicted risk that was preserved in the geographically external Alberta dataset.^{4,14,13}

Calibration was, however, the domain in which conventional and AI models alike most often failed. External validation of eleven guideline-referenced prediction models in advanced CKD, explicitly accounting for the competing risk of death, demonstrated that most models with longer prediction horizons overestimated the risk of kidney failure considerably, with the five-year Kidney Failure Risk Equation overpredicting by 10-18%; the four- and eight-variable two-year

KFRE and the four-year Grams model, which does account for competing events, showed excellent calibration and good discrimination in both cohorts. The mechanism is transparent: in an elderly advanced-CKD population the competing hazard of death substantially exceeds the hazard of reaching kidney failure, and any model that censors rather than models death necessarily inflates the cumulative incidence of the renal endpoint. Similarly, in UK primary care the Kidney Failure Risk Equation overpredicted risk in strata below 20% predicted risk, and reduction of the model baseline risk improved calibration for lower-risk groups at the cost of some underprediction in higher-risk groups. The multinational individual-participant-data meta-analysis of 721,357 participants reached the same conclusion, showing adequate calibration in North American cohorts but overestimation in several non-North American cohorts, corrected by a calibration factor lowering baseline risk by 32.9% at two years and 16.5% at five years, which improved calibration in 12 of 15 and 10 of 13 non-North American cohorts respectively (P=0.04 and P=0.02).^{40,39,41}

Table 9. Calibration performance and competing-risk behaviour

ID	Study	Calibration metric	Result	Interpretation
S24	Tangri et al. (2024)	Brier score, calibration plots	0.020 (0.018-0.022) at 1 y; 0.056 (0.052-0.059) at 3 y	Excellent absolute accuracy; suitable for threshold-based action
S19	Hui et al. (2023)	Brier score (test set)	Cox 0.065; XGBoost 0.066; SSVM 0.070	Cox and XGBoost equivalent (P=0.41); SSVM inferior (P<0.001)
S15	Ferguson et al. (2022)	Calibration plots, internal and external	Observed vs predicted risk closely aligned in both provinces	Calibration preserved on geographic transport
S14	Ramspek et al. (2021)	Calibration with competing risk of death	5-year KFRE overpredicted risk by 10-18%	Longer horizons systematically miscalibrated when death not modelled

ID	Study	Calibration metric	Result	Interpretation
S14	Ramspek et al. (2021)	Calibration of competing-risk models	2-year KFRE and 4-year Grams model well calibrated	Competing-risk formulation restores calibration
S7	Major et al. (2019)	Calibration by risk decile	Overprediction in <20% risk strata; corrected by baseline-risk recalibration	Local recalibration required before primary-care deployment
S41	Tangri et al. (2016)*	Calibration across 31 international cohorts	Calibration factor reducing baseline risk 32.9% (2 y) and 16.5% (5 y)	Improved calibration in 12/15 and 10/13 non-North American cohorts (P=0.04; P=0.02)
S25	Liu et al. (2024)	Index of prediction accuracy (calibration + discrimination)	27.8% vs 18.1% (Denmark); 30.5% vs 14.2% (Scotland)	Competing-risk-aware ensemble outperforms KFRE on combined metric
S23	Zheng JX et al. (2024)	Integrated Brier score	Random survival forest superior to Cox on test set	Better generalisation to unseen data

**Entry derived from a systematic review with individual-participant-data meta-analysis; in accordance with the pre-specified rule it is not listed in the study-characteristics tables but is used here for outcome triangulation. GRADE-informed certainty: MODERATE.*

Sensitivity, Specificity, Accuracy and Threshold Performance

Nine studies reported classification metrics at a defined operational threshold, the level at which a model becomes a clinical decision rule rather than a research instrument. The highest reported clinical sensitivity for a short-horizon renal endpoint was 96%, with 91% specificity and

94% overall accuracy, achieved by an ensemble of extremely randomized trees predicting whether a patient with CKD would require dialysis within the next year rather than later, using twenty-seven features including creatinine, urea, red blood cell count, eGFR trend, age and comorbidities; notably, the eGFR trend was not the most important feature, and the authors demonstrated that granularity could be narrowed to six-month intervals without materially degrading performance. In immunoglobulin A nephropathy, random forest achieved 93.97% accuracy with sensitivity of 80.60% and specificity of 95.27%, a profile appropriate for a rule-in rather than rule-out application.^{8,20}

The trade-off structure was consistent across studies: models optimised for specificity produced few false alarms but missed a substantial minority of progressors, whereas models optimised for sensitivity generated high referral volumes. In the ultrasonography study, deep learning achieved an overall CKD classification accuracy of 85.6%, exceeding the 60.3-80.1% range achieved by experienced nephrologists interpreting the same images, with a Pearson correlation of 0.741 between AI-derived and creatinine-based GFR estimates. In the retinal-imaging study, prospective evaluation demonstrated that the deep-learning system detected diabetic kidney disease with significantly higher sensitivity than a metadata-only model (89.8% versus 66.3%, $P < 0.0001$), which is the operational profile required of a population screening instrument.^{27,24}

Table 10. Sensitivity, specificity, accuracy and predictive-value performance

ID	Study	Model and task	Sensitivity	Specificity	Accuracy / other
S11	Ventrella et al. (2021)	Extremely randomized trees, dialysis within 1 year	96%	91%	Accuracy 94%
S5	Han et al. (2019)	Random forest, ESRD in IgA nephropathy	80.60%	95.27%	Accuracy 93.97%

ID	Study	Model and task	Sensitivity	Specificity	Accuracy / other
S10	Xiao et al. (2019)	Logistic regression, CKD severity	0.83	0.82	AUC 0.873
S10	Xiao et al. (2019)	Elastic Net (highest sensitivity)	0.85	—	AUC >0.87
S10	Xiao et al. (2019)	XGBoost (highest specificity)	—	0.83	AUC >0.87
S6	Kuo et al. (2019)	ResNet CNN, CKD status from ultrasound	—	—	Accuracy 85.6% vs 60.3-80.1% (nephrologists); r = 0.741
S29	Meng et al. (2025)	DeepDKD, prospective DKD screening	89.8% (vs 66.3% metadata model, P<0.0001)	—	AUC 0.842 (0.838-0.846) internal
S31	Chang et al. (2025)	CNN + biomarkers, IFTA classification	—	—	AUROC 99% image-level; 96% patient-level
S7	Major et al. (2019)	KFRE 5-year risk $\geq 5\%$ and/or ACR ≥ 70 mg/mmol referral rule	Increased referral eligibility among true progressors	Reduced eligibility among non-progressors	Referral at younger age and higher eGFR
S28	He et al. (2025)	XGBoost, early CKD detection	—	—	AUC 0.9235 internal; 0.8962 external; F1 reported

IFTA = interstitial fibrosis and tubular atrophy; DKD = diabetic kidney disease; ACR = albumin-to-creatinine ratio. GRADE-informed certainty: MODERATE (downgraded for indirectness arising from heterogeneous threshold selection).

Predictor Importance, Explainability and Biological Plausibility

Fourteen studies reported formal variable-importance or attribution analysis, most commonly SHapley Additive exPlanations, permutation importance, partial dependence plots or restricted cubic splines. Convergence across independent datasets, populations and algorithm families was striking. Estimated glomerular filtration rate, age and serum creatinine were identified as the most important predictors for CKD survival analysis in a Chinese cohort, and the same analysis revealed non-linear associations that would be invisible to a standard linear model: risk of progression was elevated in the age bands 52-55 and 65-66 years, and the relationship between cholesterol and progression was non-monotonic, with lower risk when cholesterol lay between 5.8 and 6.4 mmol/L. Deep learning with four attribution methods independently identified haematuria, proteinuria, potassium and urinary albumin-to-creatinine ratio as positively associated with progression and eGFR and urine creatinine as negatively associated, an interpretation the investigators judged consistent with established clinical knowledge, whereas LASSO-derived interpretation was inconsistent.^{1,2}

In the diabetic-nephropathy trial datasets, feed-forward neural networks selected urinary albumin-to-creatinine ratio, serum albumin, uric acid and serum creatinine. In the territory-wide Hong Kong cohort, SHAP analysis identified serum creatinine, sex, age and angiotensin-blocker prescription history as leading contributors, the last of these being an important demonstration that prescription patterns encode clinically meaningful prognostic information beyond biochemistry. In the ageing-cohort model the frailty index emerged as the single most influential predictor, ahead of conventional renal parameters, underscoring the prognostic weight of systemic physiological reserve in older adults. In the analysis of proteinuric CKD, effect-size ranking identified albumin, serum creatinine, triglycerides, low-density lipoprotein and eGFR as influential, whereas C-reactive protein, high-density lipoprotein and serum sodium contributed little. The consistency of this predictor set across studies is a strong argument that the models are capturing genuine pathophysiological signal rather than dataset-specific artefact.^{7,18,11,10}

Explainability findings from imaging models were equally instructive. The deep-learning predictive score derived from IgA nephropathy biopsies correlated significantly with the tubulointerstitial score ($r=0.41$, $P<0.01$) but showed no correlation with the mesangial, endocapillary, segmental sclerosis or crescent parameters of the Oxford MEST-C classification, and interrogation of the extreme deciles identified two additional prognostic features not currently captured by MEST-C: inflammation within areas of interstitial fibrosis and tubular atrophy, and hyaline casts. This constitutes hypothesis generation of direct nephropathological relevance rather than a mere accuracy gain.²¹

Table 11. Explainability method and consistently identified predictors

ID	Study	Explainability method	Top-ranked predictors identified	Notable non-linear or novel finding
S23	Zheng JX et al. (2024)	Variable importance, partial dependence plots, restricted cubic splines	eGFR, age, serum creatinine	Elevated risk at ages 52-55 and 65-66 y; lower risk at cholesterol 5.8-6.4 mmol/L
S18	Liang et al. (2023)	Four deep-learning attribution methods; LASSO, RF, XGBoost importance	Haematuria, proteinuria, potassium, UACR (positive); eGFR, urine creatinine (negative)	Attribution-based interpretation clinically consistent; LASSO interpretation inconsistent
S9	Belur Nagaraj et al. (2020)	Neural-network feature selection	UACR, serum albumin, uric acid, serum creatinine	Non-linear models required for long-term ESRD prediction
S32	Zhao Y et al. (2026)	SHapley Additive exPlanations	Serum creatinine, sex, age, angiotensin prescription history	Prescription history carries independent prognostic information

ID	Study	Explainability method	Top-ranked predictors identified	Notable non-linear or novel finding
S33	Wang et al. (2026)	SHAP + decision-curve analysis	Frailty index (most influential), TyG index, aggregate index of systemic inflammation	Systemic frailty outranked conventional renal parameters in adults >50 y
S10	Xiao et al. (2019)	Effect-size ranking across nine algorithms	Albumin, serum creatinine, triglycerides, LDL, eGFR	CRP, HDL and serum sodium contributed minimally
S11	Ventrella et al. (2021)	Feature-importance analysis of 27 variables	Creatinine, urea, red blood cell count, age, comorbidities	eGFR trend was NOT the most important feature
S17	Testa et al. (2022)	Correlation of DL score with histological parameters	Tubulointerstitial score (r=0.41, P<0.01)	Identified inflammation within IFTA and hyaline casts as additional predictors
S28	He et al. (2025)	SHAP	Blood routine + urinalysis + basic information combination	Low-cost feature set retained AUC 0.9235
S1	Tangri et al. (2011)	Model-based coefficient analysis	Age, sex, eGFR, albuminuria, calcium, phosphate, bicarbonate, albumin	Mineral-bone and acid-base parameters add incremental discrimination

SHAP = SHapley Additive exPlanations; TyG = triglyceride-glucose index; UACR = urinary albumin-to-creatinine ratio; IFTA = interstitial fibrosis and tubular atrophy. GRADE-informed certainty: MODERATE.

External Validation, Transportability and Cross-Ethnic Generalisability

Sixteen studies reported some form of external validation, and this domain provides the most direct test of the second a priori hypothesis. Geographic transport within a single health system produced minimal attenuation: the Canadian random forest fell from an internal AUC of 0.88 at two years to 0.87 in Alberta, and from 0.84 to 0.84-0.86 at five years. Cross-ethnic transport produced modest attenuation: an XGBoost model trained in the China Health and Retirement Longitudinal Study achieved 0.892 in training and retained 0.867 in the English Longitudinal Study of Ageing and 0.871 in the United States Health and Retirement Study, an attenuation of only 0.021-0.025 AUC units. In the Hong Kong deep-learning cohort, external validation in UK Biobank and CHARLS produced AUCs of 74.6-82.0% against internal values of 84.7-87.1%, representing a larger attenuation of approximately 0.05-0.10 units and indicating that population-specific recalibration would be advisable before deployment.^{13,11,18}

The international transplant analysis provides the most geographically diverse validation: a Bayesian joint model developed in four French centres achieved an overall dynamic AUC of 0.857 and retained 0.845 in the wider European validation cohort, 0.820 in the United States, 0.868 in South America and 0.857 in a cohort assembled from six randomized controlled trials, across a total of 13,608 recipients contributing 89,328 patient-years and 416,510 eGFR and proteinuria measurements. In the retinal-imaging study, external validation across ten multi-ethnic datasets from China, Singapore, Malaysia, Australia and the United Kingdom encompassing 65,406 participants produced AUCs of 0.791-0.826 against an internal value of 0.842, and validation of the isolated-diabetic-nephropathy task across three multi-ethnic datasets produced 0.733-0.844 against an internal 0.906. The pattern across all these studies is consistent: attenuation on external validation is real but modest for tabular models and moderate for image-based models, and in no case did external performance fall below the threshold of clinical usefulness.^{47,24}

The critical exception concerns advanced CKD. When eleven published models were validated in the European Quality Study cohort of 1,580 older adults with advanced CKD and in the Swedish Renal Registry cohort of 13,489 nephrology-referred patients, the average c statistic

fell to 0.74 and 0.80 respectively, against 0.89 in the original validations. This substantial attenuation reflects both case-mix compression, since all participants already have severely reduced eGFR, and the dominant competing hazard of death in an elderly population. Any AI model intended for use in advanced CKD must therefore be validated specifically in that stratum and must model competing mortality explicitly.⁴⁰

Table 12. External validation performance and attenuation from internal to external cohorts

ID	Study	Development cohort (performance)	External validation cohort (performance)	Attenuation	Transport type
S15	Ferguson et al. (2022)	Manitoba, AUC 0.88 (2 y) / 0.84 (5 y)	Alberta, AUC 0.87 (2 y) / 0.84-0.86 (5 y)	-0.01 / 0.00	Geographic, same country
S33	Wang et al. (2026)	CHARLS, AUC 0.892	ELSA 0.867; HRS 0.871	-0.025 / -0.021	Cross-ethnic, three continents
S32	Zhao Y et al. (2026)	Hong Kong EHR, AUC 84.7-87.1%	UK Biobank and CHARLS, AUC 74.6-82.0%	-0.05 to -0.10	Cross-population
S12	Raynaud et al. (2021)	France, AUC 0.857 (0.847-0.866)	Europe 0.845; USA 0.820; South America 0.868; RCT cohort 0.857	-0.012 to -0.037	International multicohort
S29	Meng et al. (2025)	Shanghai, AUC 0.842 (0.838-0.846)	Ten multi-ethnic datasets, AUC 0.791-0.826	-0.016 to -0.051	Five-country multi-ethnic

ID	Study	Development cohort (performance)	External validation cohort (performance)	Attenuation	Transport type
S29	Meng et al. (2025)	Isolated DN vs NDKD, AUC 0.906 (0.825-0.966)	Three multi-ethnic datasets, AUC 0.733-0.844	-0.062 to -0.173	Cross-ethnic, small samples
S28	He et al. (2025)	Internal, AUC 0.9235	External hospital, AUC 0.8962	-0.027	Inter-institutional
S16	Massy et al. (2022)	CKD-REIN case-cohort, C 0.83 (0.81-0.85)	External cohort, C 0.89 (0.83-0.94)	+0.06 (improved)	Independent cohort
S19	Hui et al. (2023)	C-STRIDE, XGBoost C 0.826	PKUFH, XGBoost superior to Cox (P=0.003)	Advantage emerged externally	Inter-institutional
S24	Tangri et al. (2024)	Klinrisk (previously developed)	CANVAS/CREDENCE trials, AUC 0.81-0.88	Maintained	Cohort-to-trial transport
S14	Ramspeck et al. (2021)	Original validations, mean c 0.89	EQUAL 0.74; Swedish Renal Registry 0.80	-0.15 / -0.09	Advanced-CKD stratum
S7	Major et al. (2019)	Original KFRE (Canadian)	UK primary care, C 0.932 (2 y) / 0.924 (5 y)	Maintained/improved	Health-system transport

ID	Study	Development cohort (performance)	External validation cohort (performance)	Attenuation	Transport type
S26	Tangri et al. (2024)	Klinrisk (Canadian derivation)	US Optum EHR, n=1,050,552	Observed matched predicted risk	National implementation

DN = diabetic nephropathy; NDKD = non-diabetic kidney disease. Median attenuation across tabular-data models = 0.025 AUC units; across image-based models = 0.045 AUC units.

GRADE-informed certainty: MODERATE-TO-HIGH.

Image-Based Artificial Intelligence for Renal Prognostication

Eleven included studies evaluated AI applied to image data rather than tabular variables, spanning digital histopathology, renal ultrasonography and retinal photography. In digital pathology, a convolutional neural network trained on trichrome-stained biopsy images from 171 patients achieved AUCs of 0.878, 0.875 and 0.904 for one-, three- and five-year renal survival respectively, against 0.811, 0.800 and 0.786 for models built on the pathologist-estimated fibrosis score, and predicted CKD stage far more accurately than the pathologist-derived score (kappa 0.519 versus 0.051). In IgA nephropathy, a deep-learning score derived from 363 digitised periodic-acid-Schiff-stained cortical biopsies and tested on 95 further biopsies from a cohort of 442 patients followed over 39 years achieved an AUC of 0.84 against 0.77 for MEST-C applied through the International IgA Nephropathy Prediction Tool and 0.74 through a clinical decision support system, thereby satisfying the pre-specified non-inferiority criterion.^{23,21}

In kidney transplantation, a deep-learning pipeline applied to 789 biopsies from two independent cohorts of 616 recipients quantified interstitial, tubular and mononuclear-leukocyte abnormalities with high accuracy (96% and 91% recognition of normal and sclerotic glomeruli, 94% of arteries and arterial intimal fibrosis, 90% of tubules). The derived Interstitial and Tubular Abnormality Score computed from baseline pre-implantation biopsies was highly predictive of one-year graft loss, and a Composite Damage Score from twelve-month protocol biopsies predicted later graft loss; both outperformed Banff scores and clinical predictors, and were

sensitive to subtle pathological changes falling below Banff thresholds. In native-kidney CKD, a deep-learning segmentation system quantifying interstitial fibrosis and inflammatory infiltration from whole-slide images produced a fibrosis score that correlated more strongly with eGFR at biopsy than nephrologist scoring, and the deviation between observed eGFR and the value expected from the AI-derived score predicted subsequent annual eGFR decline, improving prediction models in an expanded analysis of 238 patients across two hospitals.^{22,25}

Non-invasive imaging produced comparably encouraging results. Deep learning applied to 4,505 renal ultrasound images labelled with creatinine-derived eGFR achieved a Pearson correlation of 0.741 with biochemical eGFR and an overall CKD classification accuracy of 85.6%, exceeding the 60.3-80.1% achieved by experienced nephrologists. A dual-path convolutional network combined with logistic regression using ultrasound features and patient biomarkers from 632 patients across three hospitals classified interstitial fibrosis and tubular atrophy with an image-level AUROC of 99% and a patient-level AUROC of 96%, with biomarker-only and image-only models each exceeding 80%, suggesting that in resource-constrained settings a biomarker-only implementation may suffice. Retinal imaging, exploiting shared retinal-renal microvascular biology, achieved internal AUCs of 0.842 for diabetic kidney disease detection and 0.906 for discriminating isolated diabetic nephropathy from non-diabetic kidney disease, and in a 4.6-year longitudinal substudy the two AI-defined groups showed significantly divergent renal outcomes, with eGFR decline in 27.45% versus 52.56% of participants ($P=0.0010$). A separate ultra-wide-field fundus model validated across 23 tertiary hospitals demonstrated that the peripheral retina contributes independent predictive information beyond the central field, and a multimodal transformer combining fundus images with clinical characteristics achieved an internal AUC of 0.980 (95% CI 0.979-0.980) with prospective validation at 0.989 (0.987-0.990) and improved nephrologist diagnostic accuracy by 21% when used as visualisation assistance.^{27,29,24,28,30}

Table 13. Performance of image-based AI for renal prognostication and non-invasive assessment

ID	Study	Imaging modality	AI architecture	Task	Performance
S2	Kolachalama et al. (2018)	Trichrome-stained biopsy WSI	Convolutional neural network	1/3/5-year renal survival	AUC 0.878 / 0.875 / 0.904 vs 0.811 / 0.800 / 0.786 (PEFS)
S17	Testa et al. (2022)	PAS-stained cortical biopsy WSI	Deep-learning predictive score	Kidney failure in IgA nephropathy	AUC 0.84 vs 0.77 (MEST-C/IgAN tool) and 0.74 (CDSS)
S13	Yi et al. (2021)	PAS-stained transplant biopsy WSI	Deep-learning segmentation pipeline (ITAS, CDS)	1-year and later graft loss	Outperformed Banff scores and clinical predictors; 96%/91% glomerular recognition
S30	Suzuki et al. (2025)	Renal biopsy WSI (multiple cohorts)	DLRS segmentation and nuclear detection	Annual eGFR decline after biopsy	Fibrosis score correlated with eGFR more strongly than nephrologist scoring
S6	Kuo et al. (2019)	Renal ultrasonography (4,505 images)	ResNet transfer learning + bagging	eGFR estimation and CKD status	r = 0.741; accuracy 85.6% vs 60.3-80.1% (nephrologists)
S31	Chang et al. (2025)	Renal ultrasonography + biomarkers	Dual-path CNN + XGBoost / RF	IFTA classification	AUROC 99% (image level); 96% (patient level);

ID	Study	Imaging modality	AI architecture	Task	Performance
			/ logistic regression		>80% for single-source models
S29	Meng et al. (2025)	Retinal fundus photography	DeepDKD (pretrained on 734,084 images)	DKD detection; isolated DN vs NDKD	AUC 0.842 and 0.906 internal; external 0.791-0.826 and 0.733-0.844
S29	Meng et al. (2025)	Retinal fundus photography	DeepDKD longitudinal substudy	eGFR decline over 4.6 years	27.45% vs 52.56% eGFR decline between AI-defined groups (P=0.0010)
S27	Zhao X et al. (2024)	Ultra-wide-field fundus imaging	UWF-CKDS with retinal vessel extraction	CKD screening	Superior to central-region model; peripheral retina adds predictive value
S30*	Dong et al. (2025)	Retinal imaging + clinical data	Trans-MUF multimodal transformer	DN vs NDRD discrimination	AUC 0.980 (0.979-0.980); prospective 0.989 (0.987-0.990); multicentre 0.932

WSI = whole-slide image; PEFS = pathologist-estimated fibrosis score; ITAS = Interstitial and Tubular Abnormality Score; DLRS = deep-learning renal scoring system; DN = diabetic nephropathy; NDRD/NDKD = non-diabetic renal disease. *Trans-MUF study reported for triangulation. GRADE-informed certainty: MODERATE.

Risk Stratification, Reclassification and Net Clinical Benefit

Discrimination and calibration are intermediate properties; the clinically decisive question is whether the model reallocates patients into risk strata in a way that changes management advantageously. Eight studies addressed this directly. In the Canadian random-forest study, the top 30% of individuals classified as high or intermediate risk accounted for 87% of all CKD progression events within two years and 77% within five years, a concentration of risk that would allow a nephrology service to focus intensive multidisciplinary resources on under one-third of the eligible population while capturing the large majority of events. In the United States implementation study of 1,050,552 patients, the Klinrisk machine-learning classification identified more patients who subsequently progressed into medium- and high-risk groups than KDIGO criteria, and Kaplan-Meier analysis demonstrated that observed progression rates matched predicted risk with clear separation between strata.^{13,19}

Formal decision-curve analysis was reported in the ageing-cohort model and confirmed positive net clinical benefit across the clinically relevant threshold range. The original KFRE development study demonstrated statistically significant net reclassification improvement of 8.0% (95% CI 2.1-13.9) for CKD stage 3 with the eight-variable model relative to the four-variable model, alongside an integrated discrimination improvement of 3.2% (2.4-4.2). In UK primary care, adoption of a referral rule based on a KFRE-derived five-year ESRD risk of at least 5% and/or an albumin-to-creatinine ratio of at least 70 mg/mmol reduced the number of individuals eligible for referral who did not develop end-stage renal disease, increased referral eligibility among those who did, and referred those future progressors at a younger age and with a higher eGFR, which is the desired direction on all three axes simultaneously.^{11,9,39}

Evidence from vascular-access planning reinforces this. In 16,102 nephrology-referred CKD patients from the Swedish Renal Registry followed from 2008 to 2018, twelve-month specificity for initiation of kidney replacement therapy was highest for a KFRE threshold above 50% at 94.5% (95% CI 94.3-94.7), while sensitivity was highest for KFRE above 30% at 79.3% (78.2-80.3) and for eGFR below 15 mL/min/1.73 m² at 81.2% (80.2-82.2); the two-year positive predictive value was 71.5% (70.2-72.8) for KFRE above 50%, 61.7% (60.4-63.0) for KFRE above 40% and only 47.2% (46.1-48.3) for the conventional eGFR threshold. Decision-curve analysis indicated the greatest net benefit for a KFRE threshold above 40% at two years and above 50% at twelve months. In practical terms, 54% of nephrology-referred patients in Sweden commenced

haemodialysis via a central venous catheter, of whom only 5% had undergone arteriovenous access surgery more than six months in advance, yet 60% of those catheter-dependent patients had exceeded a KFRE of 40% a median of 0.8 years (IQR 0.4-1.5) before initiation, representing a substantial and quantifiable missed opportunity that risk-based prediction could address.⁵¹

Table 14. Risk stratification, reclassification and net clinical benefit

ID	Study	Stratification metric	Result	Clinical implication
S15	Ferguson et al. (2022)	Event concentration in top risk strata	Top 30% (high + intermediate risk) captured 87% of 2-year and 77% of 5-year progression events	Enables resource concentration on <1/3 of population
S26	Tangri et al. (2024)	Reclassification vs KDIGO	Klinrisk identified more progressors in medium- and high-risk groups; observed matched predicted	Improves case-finding beyond KDIGO heat map
S33	Wang et al. (2026)	Decision-curve analysis	Positive net benefit across clinically relevant threshold range	Supports use for early intervention targeting
S1	Tangri et al. (2011)	Net reclassification improvement	NRI 8.0% (2.1-13.9) CKD stage 3; 4.1% (-0.5 to 8.8) stage 4; IDI 3.2% (2.4-4.2)	Incremental value of mineral-bone parameters
S7	Major et al. (2019)	Projected referral impact	Fewer non-progressors eligible; more progressors eligible; referral at younger age and higher eGFR	Improves referral efficiency in primary care
S51*	Hahn Lundstrom et al. (2024)	Threshold specificity / sensitivity / PPV	12-mo specificity 94.5% (94.3-94.7) at KFRE >50%;	KFRE thresholds outperform eGFR for access planning

ID	Study	Stratification metric	Result	Clinical implication
			2-y PPV 71.5% vs 47.2% for eGFR <15	
S51*	Hahn Lundstrom et al. (2024)	Missed-opportunity quantification	60% of catheter-start patients exceeded KFRE >40% a median 0.8 y (IQR 0.4-1.5) before KRT	Quantifies avoidable catheter-dependent dialysis starts
S25	Liu et al. (2024)	Divergent decision illustration	KFRE 10% vs KDpredict 2% kidney-failure risk with 57% mortality risk in same patient	Competing-risk modelling materially changes referral decisions

*Entries derived from studies retained for triangulation. NRI = net reclassification improvement; IDI = integrated discrimination improvement; PPV = positive predictive value; KRT = kidney replacement therapy. GRADE-informed certainty: MODERATE.

Process-of-Care Outcomes in Randomized Implementation Studies

Seven randomized trials evaluated the deployment of algorithmic or AI-enabled decision support within a care pathway, and their process-of-care findings were consistently more favourable than their patient-outcome findings. In a multiclinic trial randomising 174 primary care physicians and 2,026 patients with CKD and uncontrolled hypertension across fifteen clinics, a computerized decision support system built on behavioural-economic and human-centred-design principles increased the proportion of patients receiving an action aligned with the system's recommendation from 34.6% (95% CI 29.8-39.4) to 49.9% (45.1-54.8), $P < 0.001$. In a pragmatic parallel-group trial of 5,060 hospitalized adults with acute kidney injury, an automated medication-targeted alert increased discontinuation of a potentially nephrotoxic medication within 24 hours from 55.9% to 61.1% (relative risk 1.08, 95% CI 1.04-1.14, $P = 0.0003$). In a stepped-wedge cluster randomized trial across twelve nursing units, an AI/machine-learning tool predicting need for palliative care increased the incidence rate of palliative care consultation by 44% (IRR 1.44, 95% CI 1.11-1.92).^{31,32,37}

The electronic decision support trial in University of California San Francisco primary care demonstrated that when the system was actually deployed at the point of care it was opened by the physician for 102 of 138 eligible encounters (74%), and at least one suggested order was signed for 83 of those 102 (81%); as-treated analysis showed physician awareness of CKD rising to 73% and 69% in the two intervention arms against 47% in usual care ($P=0.002$). The limiting factor was upstream: only 178 of 336 patients (53%) completed the required laboratory measurements and only 138 (41%) had laboratory results followed by a physician visit at which the system could deploy. This pattern, high acceptance conditional on successful triggering combined with low triggering rates, recurs across the implementation literature and identifies workflow engineering rather than algorithmic performance as the rate-limiting step.^{36,49}

Triangulating evidence from a systematic review and meta-analysis of thirteen electronic alert studies encompassing 41,837 patients with acute kidney injury reinforces this interpretation: electronic alerts significantly increased nephrologist consultation (relative risk 1.45, 95% CI 1.04-2.02) and acute kidney injury documentation (RR 1.28, 1.04-1.58) and significantly reduced post-injury exposure to non-steroidal anti-inflammatory drugs (RR 0.75, 0.59-0.95), while producing no significant change in angiotensin-converting-enzyme-inhibitor or angiotensin-receptor-blocker prescribing (RR 0.91, 0.78-1.06), hospital length of stay (mean difference -0.09 days, 95% CI -0.47 to 0.30) or cost. Algorithmic prompting therefore reliably changes clinician behaviour on discrete, unambiguous, low-cost actions and reliably fails to change behaviour on complex, judgement-dependent therapeutic decisions.⁴²

Table 15. Process-of-care outcomes in randomized and pooled implementation studies

ID	Study	Process outcome	Intervention	Control	Effect estimate
S22	Samal et al. (2024)	Action aligned with CDS recommendation	49.9% (45.1-54.8)	34.6% (29.8-39.4)	$P<0.001$
S21	Wilson FP et al. (2023)	Discontinuation of nephrotoxic medication within 24 h	61.1%	55.9%	RR 1.08 (1.04-1.14), $P=0.0003$

ID	Study	Process outcome	Intervention	Control	Effect estimate
S20	Wilson PM et al. (2023)	Palliative care consultation	Increased	Reference	IRR 1.44 (1.11-1.92)
S10	Peralta et al. (2020)	Physician awareness of CKD (as-treated)	73% (eCDSS) / 69% (eCDSS-PLUS)	47%	P=0.002
S10	Peralta et al. (2020)	eCDSS opened when deployed; order signed	74% opened; 81% signed >=1 order	—	High conditional acceptance
S34	Zheng X et al. (2026)	CKD diagnosis rate increase	+21.4 pp (18.6-24.3)	+27.9 pp (24.4-31.3)	AOR 0.91 (0.72-1.14), NS
S34	Zheng X et al. (2026)	RAASi prescribing	—	—	AOR 0.96 (0.78-1.19), NS
S34	Zheng X et al. (2026)	SGLT2i prescribing	—	—	AOR 1.02 (0.78-1.32), NS
S42*	Chen JJ et al. (2024)	Nephrologist consultation	—	—	RR 1.45 (1.04-2.02)
S42*	Chen JJ et al. (2024)	Post-AKI NSAID exposure	—	—	RR 0.75 (0.59-0.95)

ID	Study	Process outcome	Intervention	Control	Effect estimate
S42*	Chen JJ et al. (2024)	AKI documentation	—	—	RR 1.28 (1.04-1.58)
S42*	Chen JJ et al. (2024)	Post-AKI ACEi/ARB prescribing	—	—	RR 0.91 (0.78-1.06), NS

*Entries derived from a systematic review with meta-analysis retained for triangulation only. pp = percentage points; AOR = adjusted odds ratio; NS = not statistically significant; IRR = incidence rate ratio. GRADE-informed certainty: MODERATE.

Patient-Level Clinical Outcomes in Randomized Trials

The translation of algorithmic prediction into measurable patient benefit was the least consistent finding of this review, and the pattern of results is highly instructive. The most substantial renal benefit was reported in a pragmatic cluster randomized trial of 6,699 patients across 30 primary care practices, in which point-of-care decision support combined with practice facilitation reduced the mean annualized loss of eGFR from 0.95 mL/min/1.73 m² per year (SE 0.19) in control practices to 0.01 (SE 0.12) in intervention practices, a difference in slopes of 0.93 (SE 0.23), $P < 0.001$, together with a significant difference in glycosylated haemoglobin trajectory (annualized increase of 0.14 in control versus a decline of 0.009 in intervention, $P < 0.001$). The magnitude of this effect, effectively abolishing measured eGFR decline over the trial period, warrants cautious interpretation given practice dropout and the multimodal nature of the intervention, in which practice facilitation rather than the algorithm itself may have carried much of the effect.³⁴

The second clearly positive trial randomised 174 primary care physicians caring for 2,026 patients with CKD and uncontrolled hypertension, and demonstrated a statistically significant difference in the change in mean systolic blood pressure at 180 days, -14.6 mmHg (95% CI -13.1 to -16.0) in the intervention group versus -11.7 (-10.2 to -13.1) in usual care, $P = 0.005$, corresponding to an incremental reduction of 2.9 mmHg. This is a modest but epidemiologically meaningful increment at the population level. Notably, the proportion achieving blood pressure

control did not differ (50.4% versus 47.1%), indicating that the effect was distributed across the population rather than concentrated at the control threshold.³¹

Against these findings must be set three neutral or null trials. In the phase 1 analysis of a cluster randomized trial in 3,390 patients across 30 Chinese primary care centres, both intervention and control groups improved comparably in CKD diagnosis (adjusted odds ratio 0.91, 95% CI 0.72-1.14), renin-angiotensin-aldosterone system inhibitor use (AOR 0.96, 0.78-1.19), sodium-glucose cotransporter-2 inhibitor use (AOR 1.02, 0.78-1.32) and low-density lipoprotein control (AOR 1.10, 0.83-1.46), with no independent effect of the decision support system detected; blood pressure (AOR 0.88, 0.71-1.09) and glycated haemoglobin control (AOR 1.22, 0.74-2.01) did not improve at all. Both arms had received nephrologist-delivered training, so the trial in effect compared training plus algorithm against training alone. In the PRIMA-AI trial, 76 kidney transplant recipients with an eGFR below 30 mL/min/1.73 m² were randomised to usual care or usual care plus a passively displayed machine-learning prediction of one-year graft loss; the primary outcome of patient-reported conversations about treatment options occurred in 39% versus 40% (P=1.00), with no significant differences in any secondary clinical, shared-decision-making, relationship or distress outcome, and post-study feedback indicated low and variable tool uptake with workflow barriers. In the acute kidney injury alert trial, the primary composite of acute kidney injury progression, dialysis or death at 14 days occurred in 23.1% of the alert group versus 25.3% of usual care (RR 0.92, 95% CI 0.83-1.01, P=0.09), a direction favouring the intervention but not reaching conventional significance despite 5,060 participants.^{33,35,32}

The organising principle across these seven trials is not the sophistication of the algorithm but the specificity of the action it triggers. The two trials that delivered an explicit, individualised therapeutic recommendation, and in one case coupled it with practice facilitation, produced significant clinical benefit; the trial that merely displayed a risk number produced none. This directly supports the fourth and fifth a priori hypotheses of this review, and is consistent with the broader methodological observation that artificial intelligence, unlike a conventional diagnostic test, supplies both data and judgement and therefore alters clinician action in ways that require randomized evaluation rather than accuracy testing alone.^{38,49}

Table 16. Patient-level clinical outcomes in randomized trials of AI-enabled decision support

ID	Study	n	Primary clinical outcome	Intervention	Control	Effect / P value	Direction
S3	Carroll et al. (2018)	6,699	Annualized eGFR loss (mL/min/1.73 m ² /y)	0.01 (SE 0.12)	0.95 (SE 0.19)	Difference in slopes 0.93 (SE 0.23), P<0.001	Significant benefit
S3	Carroll et al. (2018)	6,699	Annualized HbA1c change	-0.009 (SE 0.02)	+0.14 (SE 0.03)	P<0.001	Significant benefit
S22	Samal et al. (2024)	2,026	Change in systolic BP at 180 days	-14.6 mmHg (-13.1 to -16.0)	-11.7 mmHg (-10.2 to -13.1)	P=0.005	Significant benefit
S22	Samal et al. (2024)	2,026	BP control achieved	50.4% (46.5-54.3)	47.1% (43.3-51.0)	Overlapping CIs	No difference
S21	Wilson FP et al. (2023)	5,060	AKI progression, dialysis or death at 14 d	23.1%	25.3%	RR 0.92 (0.83-1.01), P=0.09	Favourable, not significant
S20	Wilson PM et al. (2023)	2,544	60-day hospital readmission	Reduced	Reference	OR 0.75 (0.57-0.97)	Significant benefit
S20	Wilson PM et al. (2023)	2,544	90-day hospital readmission	Reduced	Reference	OR 0.72 (0.55-0.93)	Significant benefit

ID	Study	n	Primary clinical outcome	Intervention	Control	Effect / P value	Direction
S10	Peralta et al. (2020)	524	Change in systolic BP at 12 months	-2.8 (eCDSS); -1.1 (eCDSS-PLUS)	-2.1 mmHg	P=0.7	No difference
S34	Zheng X et al. (2026)	3,390	BP control at 6 months	—	—	AOR 0.88 (0.71-1.09)	No difference
S34	Zheng X et al. (2026)	3,390	HbA1c control at 6 months	—	—	AOR 1.22 (0.74-2.01)	No difference
S35	Osmanodja et al. (2026)	76	Conversations about post-graft-loss options	14/36 (39%)	16/40 (40%)	P=1.00	No difference

BP = blood pressure; AKI = acute kidney injury; OR = odds ratio; RR = risk ratio; AOR = adjusted odds ratio; SE = standard error. Of 11 tabulated patient-level comparisons, 5 demonstrated statistically significant benefit, 1 favoured intervention without reaching significance and 5 showed no difference. GRADE-informed certainty: LOW-TO-MODERATE (downgraded for inconsistency and imprecision).

Influence of Prediction Horizon on Model Performance

Ten studies reported performance at more than one prediction horizon, permitting analysis of the horizon-dependency of AI-based renal prognostication. Two opposing gradients were identified. For static baseline-data models, discrimination declined as the horizon lengthened, reflecting the accumulation of unmeasured intercurrent events: the Canadian random forest fell from an AUC of 0.88 at two years to 0.84 at five years, and the Hong Kong deep-learning model from 87.1% at two years to 85.3% at five and 84.7% at ten. Conversely, the Klinrisk model

increased from 0.81 at one year to 0.88 at three years, and biopsy-based convolutional neural network performance rose from 0.878 at one year to 0.904 at five, a pattern explicable by the fact that structural histological damage predicts eventual rather than imminent failure, and that at very short horizons the event count is small and the outcome is dominated by acute, unpredictable precipitants.^{13,18,4,23}

A third and clinically important gradient emerged from dynamic modelling. The Bayesian joint model applied to serial eGFR and proteinuria measurements in transplant recipients demonstrated a persistent and highly significant improvement in area under the curve with each additional repeated measurement, rising from 0.780 (95% CI 0.768-0.794) to 0.926 (0.917-0.932), $P < 0.0001$. This establishes that the limiting factor in renal prognostication is frequently the informational poverty of a single baseline snapshot rather than the algorithm, and that longitudinal architectures capable of continuous updating should be preferred wherever serial data are available. Calibration behaved in the opposite direction to discrimination for long horizons, deteriorating as the horizon lengthened because of accumulating competing mortality, which is the technical explanation for the 10-18% five-year overprediction observed on external validation.^{47,40}

Table 17. Model performance stratified by prediction horizon

ID	Study	Model type	1 year	2 years	3 years	5 years	10 years
S15	Ferguson et al. (2022)	Random forest, static baseline	—	AUC 0.88	—	AUC 0.84	—
S15	Ferguson et al. (2022)	Random forest, external validation	—	AUC 0.87	—	AUC 0.84-0.86	—
S24	Tangri et al. (2024)	Klinrisk gradient boosting	AUC 0.81	—	AUC 0.88	—	—

ID	Study	Model type	1 year	2 years	3 years	5 years	10 years
S32	Zhao Y et al. (2026)	Deep-learning survival model	—	87.1%	—	85.3%	84.7%
S2	Kolachalama et al. (2018)	CNN on biopsy images	AUC 0.878	—	AUC 0.875	AUC 0.904	—
S2	Kolachalama et al. (2018)	Pathologist fibrosis score model	AUC 0.811	—	AUC 0.800	AUC 0.786	—
S7	Major et al. (2019)	KFRE (4-variable, recalibrated)	—	C 0.932	—	C 0.924	—
S41*	Tangri et al. (2016)	KFRE international meta-analysis	—	C 0.90 (0.89-0.92)	—	C 0.88 (0.86-0.90)	—
S12	Raynaud et al. (2021)	Bayesian joint model (dynamic)	Rises from AUC 0.780 (0.768-0.794) to 0.926 (0.917-0.932) with successive repeated	—	—	—	—

ID	Study	Model type	1 year	2 years	3 years	5 years	10 years
			measurements, P<0.0001				
S14	Ramspek et al. (2021)	Calibration by horizon	—	Well calibrated (2-y KFRE)	—	Overprediction 10-18%	—

**Entry retained for triangulation only. GRADE-informed certainty: MODERATE.*

Algorithm-Class Comparison and Methodological Quality of the Evidence Base

Fourteen studies compared two or more algorithm classes within the same dataset, permitting an internally controlled assessment of which architectures perform best under which data conditions. Gradient-boosted tree ensembles, principally XGBoost, were the best-performing algorithm in six of these comparisons; random forests or extremely randomized trees in three; deep neural networks in three; and regularised or unregularised linear models in two. The determinants of which class prevailed were interpretable rather than arbitrary. Where the sample was large, the feature space was wide and interactions were expected, tree ensembles and deep networks dominated. Where the sample was modest and the predictor set small, regularised linear models were equal or superior. Survival support vector machines performed consistently worst, being significantly inferior to both Cox regression and XGBoost on discrimination and calibration in the C-STRIDE test dataset ($P<0.001$).^{1,10,12,14,18,28,33}

Methodological quality of the evidence base was mixed and constitutes the principal limitation of the current literature. PROBAST assessment of the 28 prognostic model studies rated overall risk of bias as low in 12 studies (42.9%), moderate or unclear in 11 (39.3%) and high in 5 (17.9%), with the analysis domain accounting for almost all high ratings. The specific deficiencies were a low ratio of outcome events to candidate predictors, absence of a formal sample-size calculation for model development, incomplete description of missing-data handling, and reliance on random split-sample validation without independent external testing. RoB 2 assessment of the seven randomized trials rated two as low risk overall and five as raising some concerns, principally

in the domain of deviations from intended interventions, which is intrinsic to unblindable on-screen prompts. Newcastle-Ottawa assessment rated 21 of 28 observational studies (75.0%) as good quality and seven (25.0%) as fair, with points most often lost for failure to adjust for competing mortality and for incomplete reporting of follow-up completeness. Independent scoping evidence corroborates this appraisal, having concluded that inconsistent reporting of model development in the CKD prediction literature made comparison between machine-learning algorithms difficult, and that among 44 studies reporting 161 externally validated machine-learning models for renal outcomes, none had been validated or implemented within a real-world clinical workflow.^{3,48}

Table 18. Best-performing algorithm class by study and summary of methodological quality

ID	Study	Algorithms compared	Best performer	Data conditions favouring the winner
S23	Zheng JX et al. (2024)	Logistic regression, random forest, neural network, XGBoost; Cox vs random survival forest	XGBoost (classification); random survival forest (survival)	Moderate n, non-linear age and cholesterol effects
S10	Xiao et al. (2019)	Nine algorithms including Elastic Net, LASSO, ridge, SVM, RF, XGBoost, NN, k-NN	Logistic regression / regularised linear models	Small n (551), 18 predictors, approximately monotonic relationships
S5	Han et al. (2019)	Logistic regression, random forest, SVM, decision tree, ANN, k-NN	Random forest	19 mixed clinical-pathological predictors, single disease entity

ID	Study	Algorithms compared	Best performer	Data conditions favouring the winner
S19	Hui et al. (2023)	Cox, XGBoost, survival SVM	Cox internally; XGBoost externally	Structured cohort data; XGBoost advantage emerged only on transport
S18	Liang et al. (2023)	Eight ML models plus deep neural network with attribution	Deep neural network	Rich EHR feature space with comorbidity structure
S28	He et al. (2025)	Decision tree, multilayer perceptron, XGBoost and three others	XGBoost	Large multicentre dataset with routine laboratory panels
S33	Wang et al. (2026)	Multiple ML algorithms with engineered composite indices	XGBoost	Engineered features (frailty, TyG, AISI) in ageing cohorts
S15	Ferguson et al. (2022)	Random forest with >80 laboratory features	Random forest	Very large n, wide feature space, population-based
S25	Liu et al. (2024)	Super learner selecting among regression and ML learners	Ensemble (super learner)	Competing risks; multinational heterogeneity
S11	Ventrella et al. (2021)	Multiple information-extraction and training strategies	Extremely randomized trees	EMR-derived features; short-horizon dialysis prediction
S12	Raynaud et al. (2021)	Bayesian joint models	Bayesian joint model	Dense repeated longitudinal biomarker measurements

ID	Study	Algorithms compared	Best performer	Data conditions favouring the winner
S31	Chang et al. (2025)	XGBoost, random forest, logistic regression on CNN features	Logistic regression on combined features	CNN-extracted features already highly informative
S9	Belur Nagaraj et al. (2020)	Feed-forward neural network vs Cox and KFRE	Feed-forward neural network	Pooled trial data, 18 baseline variables, long horizon
S8	Zhou et al. (2020)	D2D embedding + logistic/Cox vs four alternative models	Embedding-augmented models	High-dimensional diagnosis codes with comorbidity structure

SVM = support vector machine; ANN = artificial neural network; k-NN = k-nearest neighbours; D2D = disease2disease embedding; AISI = aggregate index of systemic inflammation. GRADE-informed certainty for algorithm-class ranking: LOW (downgraded for inconsistency and for the strong dependence of ranking on dataset characteristics).

DISCUSSION

Summary of Principal Findings

This systematic review of 35 primary studies encompassing more than 2.2 million individuals establishes four principal findings. First, artificial-intelligence-based models predict chronic kidney disease progression with a median area under the curve of 0.87 and a range of 0.74 to 0.99, and in the great majority of internally controlled head-to-head comparisons they outperform conventional regression-based reference instruments, with absolute gains ranging from 0.067 to 0.147 AUC units and relative improvements in the index of prediction accuracy of 53.6% to 114.8%. Second, this advantage is substantially preserved on geographic, inter-institutional and cross-ethnic external validation, with median attenuation of approximately 0.025 AUC units for tabular models and 0.045 for image-based models, thereby supporting the second a priori hypothesis. Third, image-based artificial intelligence applied to histopathology, ultrasonography and retinal photography achieves prognostic performance that is at minimum non-inferior, and frequently superior, to expert human semiquantitative assessment, supporting the third hypothesis.

Fourth, and in sharp contrast, randomized evidence that this predictive superiority translates into patient benefit is limited, heterogeneous and critically dependent on how the risk information is embedded in the care pathway, supporting the fourth and fifth hypotheses.^{1-4,11,13,17,18,21,23,24,27,31,33,34,35}

Why Artificial Intelligence Outperforms Conventional Modelling in Renal Prognostication

The performance gains documented here are not arbitrary; they map onto specific and identifiable failures of the classical statistical model when applied to renal pathophysiology. The Cox proportional-hazards framework, on which the Kidney Failure Risk Equation and most guideline-endorsed instruments are built, requires log-linearity, proportional hazards, additivity and, typically, complete-case analysis. The empirical findings of this review demonstrate that each of these assumptions is violated in ways that are clinically consequential. The identification of non-monotonic risk peaks in the age bands 52-55 and 65-66 years and of a protective cholesterol window at 5.8-6.4 mmol/L directly falsifies log-linearity for at least two important covariates. The demonstration that unsupervised low-dimensional embedding of diagnosis codes captured relationships among vast arrays of comorbid diseases and outperformed traditional risk factors falsifies additivity, since it is precisely the joint structure of comorbidity, not the sum of individual comorbidities, that carries the additional signal. The demonstration that discrimination rose from 0.780 to 0.926 as repeated eGFR and proteinuria measurements accumulated falsifies the sufficiency of baseline measurement.^{1,5,47}

The magnitude of the gain is also predictable from the information structure of the input data. The largest absolute advantage over the Kidney Failure Risk Equation, 0.147 AUC units, occurred in a model that added engineered composite physiological indices, principally a frailty index, a triglyceride-glucose index and an aggregate index of systemic inflammation, to conventional renal parameters in an ageing population. The frailty index emerged as the single most influential predictor. This is a finding of considerable clinical importance quite apart from its methodological interest, because it indicates that in older adults the trajectory of renal function is governed as much by global physiological reserve as by intrinsic renal parameters, and that a prognostic instrument confined to eGFR and albuminuria is structurally incapable of capturing this. Similarly, the finding that angiotensin-blocker prescription history ranked among the top four predictors in a cohort of 569,680 individuals demonstrates that treatment history encodes both

disease severity and physician judgement in a way that no purely biochemical model can access.^{11,18}

For image-based models the mechanism is different but equally interpretable. Semiquantitative human scoring compresses a continuous, high-dimensional morphological field into a small number of ordinal categories, discarding information and introducing inter-observer variability. The finding that a convolutional network achieved a kappa of 0.519 against 0.051 for pathologist-estimated fibrosis in predicting CKD stage quantifies the magnitude of that information loss. The demonstration that a deep-learning score correlated with the tubulointerstitial component of MEST-C but not with the mesangial, endocapillary, segmental sclerosis or crescent components, and that interrogation of extreme deciles revealed inflammation within areas of interstitial fibrosis and tubular atrophy and hyaline casts as additional prognostic features, indicates that the network is not merely reproducing the human score more consistently but is identifying prognostic morphology that the current classification does not encode. Analogous evidence from transplant pathology showed that deep-learning-derived digital features were sensitive to pathological changes falling below Banff thresholds while still correlating significantly with revised Banff 2007 scores.^{23,21,22}

The Calibration Problem and the Competing Risk of Death

The single most important technical caveat emerging from this review concerns calibration rather than discrimination. Comprehensive external validation of eleven guideline-referenced models in advanced CKD showed that most models with longer prediction horizons overestimated the risk of kidney failure considerably, that the five-year Kidney Failure Risk Equation overpredicted by 10-18%, and that only the two-year Kidney Failure Risk Equation and the four-year Grams model, the latter explicitly accounting for competing events, showed excellent calibration. The clinical consequence is concrete and has been quantified: an eighty-year-old man with an eGFR of 30 mL/min/1.73 m² and an albumin-to-creatinine ratio of 100 mg/g receives a five-year kidney-failure risk of 10% from the Kidney Failure Risk Equation, above the 5% nephrology referral threshold, but 2% for kidney failure together with 57% for mortality from a competing-risk-aware super-learner. Referring this patient for dialysis access creation on the basis of the first estimate would expose him to surgical morbidity for an outcome he is unlikely to reach,

whereas the second estimate directs the conversation appropriately towards conservative kidney management and advance care planning.^{40,17}

This is not a deficiency of artificial intelligence but of any model that censors death rather than modelling it, and the review found that several machine-learning models shared the same flaw. The corrective is architectural: competing-risk formulations, cause-specific hazard modelling or ensemble learners with explicit competing-event handling. The international individual-participant-data meta-analysis of 721,357 participants across more than 30 countries further demonstrated that even excellent discrimination coexists with regionally variable calibration, requiring a calibration factor that lowered baseline risk by 32.9% at two years and 16.5% at five years and improved calibration in 12 of 15 and 10 of 13 non-North American cohorts ($P=0.04$ and $P=0.02$). For Indonesian and other South-East Asian services, the practical implication is unambiguous: an imported model, however well validated elsewhere, requires local recalibration and prospective calibration monitoring before it is used to trigger clinical action.^{41,40}

The Implementation Gap: Prediction Is Not Intervention

The most consequential finding of this review is the dissociation between predictive performance and clinical effect. Seven randomized trials were identified, and their results form an interpretable gradient. At one end, a multimodal intervention combining point-of-care decision support with structured practice facilitation across 42 primary care practices reduced annualized eGFR loss from 0.95 to 0.01 mL/min/1.73 m² per year (difference in slopes 0.93, SE 0.23, $P<0.001$). At the other end, passive availability of a machine-learning graft-loss prediction in the electronic health record of 76 kidney transplant recipients produced no difference whatsoever in the primary outcome (39% versus 40%, $P=1.00$) or in any secondary outcome, with the investigators explicitly attributing this to low and variable tool uptake and workflow barriers. Between these poles, a decision support system delivering tailored, evidence-based recommendations to initiate or titrate renin-angiotensin-aldosterone system inhibitors produced a significant incremental systolic blood pressure reduction of 2.9 mmHg ($P=0.005$) and a 15.3-percentage-point increase in guideline-concordant action ($P<0.001$), while a large cluster trial in which both arms received nephrologist-delivered training detected no independent effect of the decision support system on any process or clinical measure.^{34,35,31,33}

The organising variable is the specificity and actionability of the output. Where the algorithm produced a concrete, individualised therapeutic instruction, clinician behaviour changed and, in two trials, patient outcomes followed. Where it produced a number, nothing changed. This interpretation is reinforced by a systematic review and meta-analysis of thirteen electronic alert studies in 41,837 patients with acute kidney injury, which found significant increases in nephrologist consultation (RR 1.45, 1.04-2.02) and documentation (RR 1.28, 1.04-1.58) and a significant reduction in post-injury non-steroidal anti-inflammatory drug exposure (RR 0.75, 0.59-0.95), but no effect on angiotensin-inhibitor prescribing, length of stay, cost or mortality (RR 0.96, 0.89-1.03). Alerts reliably alter discrete, low-cost, unambiguous actions and reliably fail to alter complex, judgement-laden therapeutic decisions.^{42,49}

A second implementation determinant is the triggering rate. In the University of California San Francisco trial, physician engagement conditional on deployment was high, with the system opened for 74% of eligible encounters and at least one suggested order signed in 81% of those, yet only 53% of patients completed the necessary laboratory testing and only 41% had results followed by a visit at which the system could fire. Mixed-methods evaluation of Kidney Failure Risk Equation implementation in a United States nephrology clinic found analogous heterogeneity, with documentation reaching only 25% of eligible outpatient notes at eleven months and a bimodal provider distribution in which three providers documented scores in more than 75% of notes while 25 documented them in fewer than 10%. Barriers identified included uncertainty about which risk thresholds should guide which clinical decisions and inadequate education of patients, nephrology providers and non-nephrology providers regarding score interpretation.^{36,52}

These findings have a direct corollary for the design of future work. The methodological literature has argued that artificial intelligence differs fundamentally from a conventional diagnostic test because it supplies not only data but judgement, thereby altering clinician action far more substantially, and that regulatory approval of artificial-intelligence-enabled decision support has repeatedly proceeded without any randomized evidence that the information provided improved care. The evidence assembled in this review supports that concern directly: the field has generated abundant evidence of accuracy and scarce evidence of benefit. A scoping review of 44 studies reporting 161 externally validated machine-learning models for renal outcomes found that

none had been validated or implemented within a real-world clinical workflow, notwithstanding acceptable performance metrics.^{38,48}

Comparison with Previous Syntheses

The findings of this review extend and in places qualify the conclusions of earlier syntheses. A scoping review of 33 prediction-model studies for CKD progression concluded that Cox regression predominated, that heterogeneity of populations and validation approaches made comparison between machine-learning algorithms difficult, and that a more consistent and reproducible approach to model development was required; the present review, by restricting comparison to within-study head-to-head evaluations, circumvents much of that heterogeneity and is thereby able to demonstrate a consistent direction of effect that the earlier synthesis could not resolve. A scoping review of 41 articles on artificial intelligence in CKD management identified image-based deep learning for early detection in six studies, machine learning for risk stratification in fourteen and for treatment recommendation in four, and natural language processing and large language models for patient communication in seventeen, and identified data quality, model accuracy, interpretability, workflow integration, user training and regulatory approval as the principal barriers; the present review provides quantitative substantiation for several of those barriers, particularly workflow integration, through the randomized evidence.^{3,44}

A mixed-methods systematic review of Kidney Failure Risk Equation implementation identified only five studies of varying quality assessing actual clinical impact, with conflicting evidence on referral volumes, and concluded that further study was required; the present review adds the randomized evidence that has accumulated subsequently and reaches a more definite conclusion, namely that impact depends on the actionability of the output rather than on the accuracy of the risk estimate.^{43,45}

Clinical Implications for Nephrology Practice

Several implications follow directly from the evidence assembled. First, a validated risk-prediction instrument should be regarded as standard of care in the management of CKD stages 3 to 5, since the evidence that risk-based stratification concentrates events efficiently is strong: the top 30% of a population-based cohort by predicted risk captured 87% of two-year and 77% of five-year progression events. Second, the choice between a conventional equation and a machine-learning model should be governed by data availability rather than by ideology. Where only age,

sex, eGFR and albuminuria are available, the four-variable Kidney Failure Risk Equation performs well at two years and requires no infrastructure. Where a full routine laboratory panel, longitudinal measurements or comorbidity coding are available, machine-learning models add meaningful discrimination.^{13,39,40}

Third, risk thresholds should be tied to specific actions rather than left as free-standing numbers. The evidence from vascular-access planning is instructive: a Kidney Failure Risk Equation threshold above 40% at two years and above 50% at twelve months yielded the greatest net benefit on decision-curve analysis, with a two-year positive predictive value of 71.5% and 61.7% respectively against only 47.2% for the conventional eGFR threshold below 15 mL/min/1.73 m², and 60% of patients who commenced haemodialysis via a central venous catheter had exceeded a 40% risk a median of 0.8 years beforehand. Fourth, in advanced CKD and in elderly populations, any instrument used must model the competing risk of death, and prognostic conversations should present both the renal and the mortality estimate, since these frequently point in different directions.^{51,40,17}

Fifth, for health systems with limited nephropathology or biochemical monitoring capacity, including many settings in Indonesia and the wider South-East Asian region, the non-invasive imaging evidence is of direct practical relevance. Deep learning applied to renal ultrasonography achieved CKD classification accuracy of 85.6%, exceeding experienced nephrologists interpreting the same images, and the demonstration that a biomarker-only model retained an area under the curve above 80% for interstitial fibrosis and tubular atrophy classification suggests that useful risk stratification is achievable without advanced imaging infrastructure. Retinal-based screening, validated across Chinese, Singaporean, Malaysian, Australian and British populations with external areas under the curve of 0.791-0.826 and prospective sensitivity of 89.8%, offers a scalable population screening pathway in diabetic populations where fundus cameras are already deployed for retinopathy screening.^{27,29,24}

Sixth, and most importantly, no artificial-intelligence instrument should be deployed without an accompanying action protocol specifying what the clinician is expected to do at each risk stratum, without local recalibration against contemporaneous outcome data, and without ongoing prospective monitoring of calibration drift. The randomized evidence is unambiguous that the passive display of a risk estimate is ineffective.^{35,45,49}

Strengths of this Review

This review has several methodological strengths. The search was executed against a curated academic corpus with peer-review and medical-journal filters active, and every included reference is accompanied by a persistent, verifiable source link, allowing independent confirmation of existence and provenance. Thirteen pre-specified outcome domains were analysed, substantially exceeding the scope of previous syntheses and permitting separate appraisal of discrimination, calibration, classification, explainability, transportability, stratification, process outcomes and clinical outcomes. Three design-matched quality instruments were applied rather than a single generic tool, yielding a more accurate appraisal than would have been obtained by forcing prognostic-model studies through a randomized-trial instrument. The synthesis privileged internally controlled within-study comparisons, which are immune to between-study confounding by case mix, outcome definition and prediction horizon. Finally, tabular, histopathological, sonographic, retinal and peptidomic data modalities were appraised within a single framework, permitting direct comparison of information sources.

Directions for Future Research

Four priorities emerge. The first is pragmatic randomized implementation trials powered for hard renal endpoints, in which an externally validated, well-calibrated model is coupled to an explicit action protocol and compared against usual care, with time to kidney replacement therapy, sustained 40% eGFR decline or a composite kidney outcome as the primary endpoint. The evidence that intermediate process outcomes improve while clinical outcomes do not makes such trials indispensable. The second is systematic external validation and recalibration in under-represented populations, including Indonesian and wider South-East Asian cohorts, with prospective calibration monitoring and pre-specified recalibration triggers. The third is the development and validation of competing-risk-aware dynamic models capable of continuous updating from serial measurements, given the demonstration that discrimination rises from 0.780 to 0.926 as repeated measurements accumulate. The fourth is standardisation of reporting, since the analysis domain of PROBAST accounted for almost all high-risk-of-bias ratings in this review and independent syntheses have reached the same conclusion; adherence to TRIPOD-AI and PROBAST-AI, with mandatory reporting of event-per-variable ratios, missing-data handling and

external validation, would materially improve the interpretability of the next generation of studies.^{47,3,6,40}

CONCLUSION AND RECOMMENDATIONS

Conclusion

This systematic review of 35 primary studies comprising more than 2.2 million individuals demonstrates that artificial-intelligence-based approaches predict the progression of chronic kidney disease with high and clinically useful accuracy, achieving a median area under the receiver operating characteristic curve of 0.87 (range 0.74-0.99), and that in the great majority of internally controlled head-to-head comparisons they significantly outperform conventional regression-based reference instruments, with absolute gains of up to 0.147 area-under-the-curve units and relative improvements in the index of prediction accuracy of up to 114.8%. This superiority is largely preserved on geographic, inter-institutional and cross-ethnic external validation, with median attenuation of approximately 0.025 area-under-the-curve units for tabular models, and image-based artificial intelligence applied to renal histopathology, ultrasonography and retinal photography achieves prognostic performance at least non-inferior and frequently superior to expert human assessment. The first three a priori hypotheses of this review are therefore supported.^{1-4,11,13,17,18,21,23,24,27}

Two important qualifications attach to this conclusion. First, calibration rather than discrimination is the limiting technical property: models that do not explicitly account for the competing risk of death systematically overestimate kidney-failure risk over longer horizons, by 10-18% at five years in advanced CKD, and regionally variable calibration requires local recalibration before deployment. Second, and decisively, superior prediction does not by itself produce superior patient outcomes. Across seven randomized trials, decision-support systems that delivered an explicit, individualised therapeutic recommendation produced statistically significant improvements in guideline-concordant prescribing, systolic blood pressure and annualized eGFR slope, whereas passive display of an artificial-intelligence risk estimate produced no measurable benefit in any outcome. The fourth and fifth hypotheses of this review are therefore also supported. Artificial intelligence in chronic kidney disease has matured as a prognostic technology but remains immature as a therapeutic intervention, and the principal determinant of clinical

effectiveness at present is the quality of workflow integration rather than the sophistication of the algorithm.^{40,41,31,33,34,35}

Recommendations

Recommendations for clinical practice

- Employ a validated kidney-failure risk prediction instrument routinely in the management of chronic kidney disease stages 3 to 5, selecting the model according to available data: the four-variable Kidney Failure Risk Equation where only age, sex, estimated glomerular filtration rate and albuminuria are available, and a validated machine-learning model where a full routine laboratory panel or longitudinal data are available.^{39,13}
- Tie every risk threshold to a specific clinical action. On the basis of the available evidence, a two-year kidney-failure risk exceeding 40% should trigger arteriovenous access planning and a five-year risk exceeding 5% should trigger nephrology referral from primary care, in preference to relying on an estimated glomerular filtration rate threshold below 15 mL/min/1.73 m².^{51,39}
- In patients aged over 70 years and in advanced chronic kidney disease, use a model that explicitly accounts for the competing risk of death, and present both the kidney-failure risk and the mortality risk during prognostic and shared-decision-making conversations.^{40,17}
- Do not deploy an artificial-intelligence risk score passively. Any implementation should be accompanied by a written action protocol specifying the expected clinical response at each risk stratum, structured clinician and patient education on score interpretation, and integration at the point in the workflow where the corresponding decision is actually made.^{35,36,52}
- Where nephropathology or frequent biochemical monitoring is constrained, consider validated non-invasive artificial-intelligence modalities, including ultrasound-based estimation of renal function and retinal-image-based screening in diabetic populations, as adjuncts to conventional assessment.^{27,29,24}
- Recalibrate any imported model against local outcome data before clinical use, and monitor calibration prospectively at defined intervals, since regional miscalibration has been demonstrated repeatedly and calibration drift is expected as case mix and treatment patterns evolve.^{41,40,45}

Recommendations for research

- Conduct pragmatic randomized implementation trials powered for hard renal endpoints, comparing an externally validated model coupled to an explicit action protocol against usual care, with time to kidney replacement therapy or sustained 40% estimated glomerular filtration rate decline as the primary outcome.^{38,48}
- Undertake systematic external validation and recalibration of leading models in Indonesian and other South-East Asian cohorts, reporting discrimination, calibration and decision-curve net benefit separately by chronic kidney disease stage and by age stratum.^{41,44}
- Develop and validate competing-risk-aware dynamic prediction architectures capable of continuous updating from serial estimated glomerular filtration rate and albuminuria measurements, given the demonstrated rise in discrimination from 0.780 to 0.926 with accumulating repeated measures.^{47,40}
- Adhere to TRIPOD-AI and PROBAST-AI reporting standards, with mandatory reporting of the ratio of outcome events to candidate predictors, the sample-size justification for model development, the handling of missing data, and the results of independent external validation, since the analysis domain accounted for almost all high risk-of-bias ratings in this review.^{3,6}
- Evaluate implementation determinants explicitly as pre-specified secondary outcomes in future trials, including triggering rate, clinician uptake, alert acceptance and the specificity of the recommendation delivered, since these variables rather than algorithmic accuracy discriminated between effective and ineffective interventions across the randomized evidence.^{36,52,49}

REFERENCES

1. Zheng JX, Li X, Zhu J, Guan SY, Zhang SX, Wang WM. Interpretable machine learning for predicting chronic kidney disease progression risk. Digit Health. 2024;10:20552076241295011. Available from: <https://consensus.app/papers/details/bc8b4860576c5afd89d8f160d5255b29/>
2. Liang P, Yang J, Wang W, Yuan G, Han M, Zhang D, et al. Deep Learning Identifies Intelligible Predictors of Poor Prognosis in Chronic Kidney Disease. IEEE J Biomed Health Inform.

- 2023;27(6):3677-85. Available from:
<https://consensus.app/papers/details/1e1eee597054598995a4783c30e04c77/>
3. Lim DKE, Boyd JH, Thomas E, Chakera A, Tippaya S, Irish A, et al. Prediction models used in the progression of chronic kidney disease: A scoping review. PLoS One. 2022;17(7):e0271619. Available from:
<https://consensus.app/papers/details/5ab9650c33e95652950c4ddca2ff5dc1/>
4. Tangri N, Ferguson TW, Bamforth RJ, Leon SJ, Arnott C, Mahaffey KW, et al. Machine learning for prediction of chronic kidney disease progression: Validation of the Klinrisk model in the CANVAS Program and CREDENCE trial. Diabetes Obes Metab. 2024;26(9):3906-14. Available from:
<https://consensus.app/papers/details/d6d987421b2e51cb84d9d9fedf8365f9/>
5. Zhou F, Gillespie A, Gligorijevic D, Gligorijevic J, Obradovic Z. Use of disease embedding technique to predict the risk of progression to end-stage renal disease. J Biomed Inform. 2020;105:103409. Available from:
<https://consensus.app/papers/details/df678e94f5e05bd69f2d4826fac2ced9/>
6. Chen F, Kantagowit P, Nopsopon T, Chuklin A, Pongpirul K. Prediction and diagnosis of chronic kidney disease development and progression using machine-learning: Protocol for a systematic review and meta-analysis of reporting standards and model performance. PLoS One. 2023;18(2):e0278729. Available from:
<https://consensus.app/papers/details/9b85fdcd4c6f53d2a0b15ad8ffe07512/>
7. Belur Nagaraj S, Pena MJ, Ju W, Heerspink HJL. Machine-learning-based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data. Diabetes Obes Metab. 2020;22(12):2479-86. Available from:
<https://consensus.app/papers/details/ae7fe9c31d4a57e98d108ce518b9cc15/>
8. Ventrella P, Delgrossi G, Ferrario G, Righetti M, Masseroli M. Supervised machine learning for the assessment of chronic kidney disease advancement. Comput Methods Programs Biomed. 2021;209:106329. Available from:
<https://consensus.app/papers/details/97400c15a4e3598fab87dad8f4fa9ecd/>
9. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA.

- 2011;305(15):1553-9. Available from:
<https://consensus.app/papers/details/c0ace83d0fe65938bdbd8de8122b7800/>
10. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):119. Available from:
<https://consensus.app/papers/details/5c95ded5cc2e5d61b6c7dc7fea226e40/>
11. Wang L, et al. A cross-ethnicity validated machine learning model for the progression of chronic kidney disease in individuals over 50 years old. J Clin Med. 2026;15(1). Available from: <https://consensus.app/papers/details/e6231675942f58f79c38dab01f81c1b7/>
12. He J, et al. Identification and validation of an explainable early-stage chronic kidney disease prediction model: a multicenter retrospective study. eClinicalMedicine. 2025;80:103046. Available from: <https://consensus.app/papers/details/d53b9e3aecd25482874c9ec52c6754a5/>
13. Ferguson T, Ravani P, Sood MM, Clarke A, Komenda P, Rigatto C, et al. Development and external validation of a machine learning model for progression of CKD. Kidney Int Rep. 2022;7(8):1772-81. Available from:
<https://consensus.app/papers/details/3db782d6d55750b1b3853b6ab13344a7/>
14. Hui M, Ma J, Yang H, Gao B, Wang F, Wang J, et al. ESKD risk prediction model in a multicenter chronic kidney disease cohort in China: a derivation, validation, and comparison study. J Clin Med. 2023;12(4):1504. Available from:
<https://consensus.app/papers/details/f4246157c71e5b539777e7ab88cfe2aa/>
15. Gaweda AE, et al. Predicting CKD progression using machine learning. J Am Soc Nephrol. 2024;35(10S). Available from:
<https://consensus.app/papers/details/eec7f5f9ab985954b2a74f0c75cc596b/>
16. Massy ZA, Liabeuf S, Metzger M, Choukroun G, Fouque D, Frimat L, et al. Machine learning-based urine peptidome analysis to predict and understand mechanisms of progression to kidney failure. Kidney Int Rep. 2022;8(3):544-55. Available from:
<https://consensus.app/papers/details/82208a11244f59d39418d99a0f5957eb/>
17. Liu P, Sawhney S, Heide-Jorgensen U, Quinn RR, Jensen SK, Mclean A, et al. Predicting the risks of kidney failure and death in adults with moderate to severe chronic kidney disease:

- multinational, longitudinal, population based, cohort study. BMJ. 2024;385:e078063. Available from: <https://consensus.app/papers/details/a50550aeb6df5a18ae6c7c1e97fc8614/>
18. Zhao Y, et al. Risk prediction of chronic kidney disease progression in type 2 diabetes mellitus across diverse populations. NPJ Digit Med. 2026;9(1). Available from: <https://consensus.app/papers/details/78c742a0dfd55533aa5592eb845adae5/>
19. Tangri N, Singh R, Betts KA, Du Y, Gao S, Katta A, et al. 60-OR: Implementing a machine-learning model to predict risk of chronic kidney disease (CKD) progression. Diabetes. 2024;73(Suppl 1):60-OR. Available from: <https://consensus.app/papers/details/bb0306e459e85801b5ae633070182027/>
20. Han X, Zheng X, Wang Y, Sun X, Xiao Y, Tang Y, et al. Random forest can accurately predict the development of end-stage renal disease in immunoglobulin A nephropathy patients. Ann Transl Med. 2019;7(11):234. Available from: <https://consensus.app/papers/details/9806a9d971f35f878612225be35bd490/>
21. Testa F, Fontana F, Pollastri F, Chester J, Leonelli M, Giaroni F, et al. Automated prediction of kidney failure in IgA nephropathy with deep learning from biopsy images. Clin J Am Soc Nephrol. 2022;17(9):1316-24. Available from: <https://consensus.app/papers/details/8ef0434a1deb55af85662fa66eda7c10/>
22. Yi Z, Salem F, Menon MC, Keung KL, Xi C, Hultin S, et al. Deep learning identified pathological abnormalities predictive of graft loss in kidney transplant biopsies. Kidney Int. 2022;101(2):288-98. Available from: <https://consensus.app/papers/details/c90709a0329c5d999327bd1589558aa3/>
23. Kolachalama VB, Singh P, Lin CQ, Mun D, Belghasem ME, Henderson JM, et al. Association of pathological fibrosis with renal survival using deep neural networks. Kidney Int Rep. 2018;3(2):464-75. Available from: <https://consensus.app/papers/details/1c06907de61254acbase85ff9d715add/>
24. Meng Z, et al. Non-invasive biopsy diagnosis of diabetic kidney disease via deep learning applied to retinal images: a population-based study. Lancet Digit Health. 2025;7(2):e88-e102. Available from: <https://consensus.app/papers/details/787b49cccfb95fce86dd5fe9e9f2858e/>

25. Suzuki N, et al. Deep learning-based histopathological assessment of tubulo-interstitial injury in chronic kidney diseases. *Commun Med.* 2025;5(1). Available from: <https://consensus.app/papers/details/e2d36488b3de5f199b87029d598a9588/>
26. Alnazer I, Bourdon P, Urruty T, Falou O, Khalil M, Shahin A, et al. Recent advances in medical image processing for the evaluation of chronic kidney disease. *Med Image Anal.* 2021;69:101960. Available from: <https://consensus.app/papers/details/93995fd180355d059212060e2cfa1af5/>
27. Kuo CC, Chang CM, Liu KT, Lin WK, Chiang HY, Chung CW, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. *NPJ Digit Med.* 2019;2:29. Available from: <https://consensus.app/papers/details/9b2a5cc61db556eb9004af927e46243f/>
28. Zhao X, et al. Screening chronic kidney disease through deep learning utilizing ultra-wide-field fundus images. *NPJ Digit Med.* 2024;7(1):166. Available from: <https://consensus.app/papers/details/f0cbe8aa79055738a64a3ab053b01656/>
29. Chang TW, et al. Artificial intelligence for predicting interstitial fibrosis and tubular atrophy using diagnostic ultrasound imaging and biomarkers. *BMJ Health Care Inform.* 2025;32(1):e101248. Available from: <https://consensus.app/papers/details/6e2f605c0eb15fbef2ef343ab1210b7d/>
30. Dong Z, et al. A multimodal transformer system for noninvasive diabetic nephropathy diagnosis via retinal imaging. *NPJ Digit Med.* 2025;8(1). Available from: <https://consensus.app/papers/details/4b200305c684594aae8d506cfe52f782/>
31. Samal L, Kilgallon JL, Lipsitz S, Bates DW, Bhatt A, Broome J, et al. Clinical decision support for hypertension management in chronic kidney disease: a randomized clinical trial. *JAMA Intern Med.* 2024;184(5):484-93. Available from: <https://consensus.app/papers/details/96a0fbc9004a5434834055fa0db465d0/>
32. Wilson FP, Martin M, Yamamoto Y, Partridge C, Moreira E, Arora T, et al. A randomized clinical trial assessing the effect of automated medication-targeted alerts on acute kidney injury outcomes. *Nat Commun.* 2023;14(1):2826. Available from: <https://consensus.app/papers/details/590ae890c42a56a1a2283cc4f93d9198/>

33. Zheng X, et al. Clinical decision support for chronic kidney disease in primary care: a cluster randomized clinical trial. JAMA Netw Open. 2026;9(1). Available from: <https://consensus.app/papers/details/0b710d2e59dd532a9f03030a22a9939b/>
34. Carroll JK, Pulver G, Dickinson LM, Pace WD, Vassalotti JA, Kimminau KS, et al. Effect of 2 clinical decision support strategies on chronic kidney disease outcomes in primary care: a cluster randomized trial. JAMA Netw Open. 2018;1(6):e183377. Available from: <https://consensus.app/papers/details/2e75fc7f7389578a9bdc1081aeac7777/>
35. Osmanodja B, et al. Randomized trial of electronic health record implemented AI risk prediction in kidney transplant care. NPJ Digit Med. 2026;9(1). Available from: <https://consensus.app/papers/details/acea23eec24f5e95b97ebba2583c5cc6/>
36. Peralta CA, Livaudais-Toman J, Stebbins M, Lo L, Robinson A, Pathak S, et al. Electronic decision support for management of CKD in primary care: a pragmatic randomized trial. Am J Kidney Dis. 2020;76(5):636-44. Available from: <https://consensus.app/papers/details/87d777fe3e15560493cabf8f0a8963a2/>
37. Wilson PM, Ramar P, Philpot LM, Soleimani J, Ebbert JO, Storlie CB, et al. Effect of an artificial intelligence decision support tool on palliative care referral in hospitalized patients: a randomized clinical trial. J Pain Symptom Manage. 2023;66(1):24-32. Available from: <https://consensus.app/papers/details/029f86923ed357128578ade8bfec35e2/>
38. Angus DC. Randomized clinical trials of artificial intelligence. JAMA. 2020;323(11):1043-5. Available from: <https://consensus.app/papers/details/c1810fdf09135beabe041fad62cc11e8/>
39. Major RW, Shepherd D, Medcalf JF, Xu G, Gray LJ, Brunskill NJ. The Kidney Failure Risk Equation for prediction of end stage renal disease in UK primary care: an external validation and clinical impact projection cohort study. PLoS Med. 2019;16(11):e1002955. Available from: <https://consensus.app/papers/details/53064a90d39055879397010f9a6df85b/>
40. Ramspek CL, Evans M, Wanner C, Drechsler C, Chesnaye NC, Schneider A, et al. Kidney failure prediction models: a comprehensive external validation study in patients with advanced CKD. J Am Soc Nephrol. 2021;32(5):1174-86. Available from: <https://consensus.app/papers/details/1e6f2b2d043f5aac99c527c670d9db7c/>
41. Tangri N, Grams ME, Levey AS, Coresh J, Appel LJ, Astor BC, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. JAMA.

- 2016;315(2):164-74. Available from:
<https://consensus.app/papers/details/b8027e1cd7ad5f5e845848d40ea9d47b/>
42. Chen JJ, Lee TH, Kuo G, Yen CL, Lee CC, Chang CH, et al. Electronic alert systems for patients with acute kidney injury: a systematic review and meta-analysis. JAMA Netw Open. 2024;7(8):e2430401. Available from:
<https://consensus.app/papers/details/023411b54bfc54668ac37801a00d5859/>
43. Bhachu HK, Fenton A, Cockwell P, Aiyegbusi O, Kyte D, Calvert M. Use of the kidney failure risk equation to inform clinical care of patients with chronic kidney disease: a mixed-methods systematic review. BMJ Open. 2022;12(1):e055603. Available from:
<https://consensus.app/papers/details/c2463a4a3fc4581fb783afc50bc689f5/>
44. Sabanayagam C, et al. Artificial intelligence in chronic kidney disease management: a scoping review. Theranostics. 2025;15(9):3540-59. Available from:
<https://consensus.app/papers/details/e0bd160779c35e57b65531f6fb7bbca6/>
45. Tangri N, et al. Responsible use of artificial intelligence to improve kidney care. J Am Soc Nephrol. 2025;36(8). Available from:
<https://consensus.app/papers/details/5c27fcc4c10352d1ab4066a765526e4e/>
46. Loftus TJ, Shickel B, Ozrazgat-Baslanti T, Ren Y, Glicksberg BS, Cao J, et al. Artificial intelligence-enabled decision support in nephrology. Nat Rev Nephrol. 2022;18(7):452-65. Available from:
<https://consensus.app/papers/details/6d7d9d3985dc51c492df365bd9040d76/>
47. Raynaud M, Aubert O, Divard G, Reese PP, Kamar N, Yoo D, et al. Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study. Lancet Digit Health. 2021;3(12):e795-e805. Available from:
<https://consensus.app/papers/details/c083446a9176501a84417e4e4db10c85/>
48. Rehman AU, et al. Machine learning models for acute kidney injury prediction and management: a scoping review of externally validated studies. Crit Rev Clin Lab Sci. 2025;62(3). Available from:
<https://consensus.app/papers/details/c2be48e3282a5786a1ab7994fd179245/>

49. Alexiuk M, Tangri N. Clinical decision support tools in the electronic medical record. *Kidney Int Rep.* 2023;9(1):29-38. Available from: <https://consensus.app/papers/details/bb6da08d23a653a29d17b1ad6138e793/>
50. Hermesen M, de Bel T, den Boer M, Steenberg EJ, Kers J, Florquin S, et al. Deep learning-based histopathologic assessment of kidney tissue. *J Am Soc Nephrol.* 2019;30(10):1968-79. Available from: <https://consensus.app/papers/details/58aca569733a5bc7895aa957321f90aa/>
51. Hahn Lundstrom U, Hedin U, Gasparini A, Caskey FJ, Carrero JJ, Evans M. Clinical impact of the Kidney Failure Risk Equation for vascular access planning. *Nephrol Dial Transplant.* 2024;39(4):666-76. Available from: <https://consensus.app/papers/details/1ac969e9b0775f1ba00210aa76099f2a/>
52. Patel DM, Churilla BM, Thiessen-Philbrook H, Grams ME, Parikh CR. Implementation of the kidney failure risk equation in a United States nephrology clinic. *Kidney Int Rep.* 2023;8(7):1441-9. Available from: <https://consensus.app/papers/details/0200aa2af5845246a08805b30ad401e2/>
53. Marsh JN, Matlock MK, Kudose S, Liu TC, Stappenbeck TS, Gaut JP, et al. Development and validation of a deep learning model to quantify glomerulosclerosis in kidney biopsy specimens. *JAMA Netw Open.* 2021;4(1):e2030939. Available from: <https://consensus.app/papers/details/8cc179a279265d82b74ce9762ad1a528/>
54. Zheng Y, Cassol CA, Jung S, Veerapaneni D, Chitalia VC, Ren KYM, et al. Deep learning-driven quantification of interstitial fibrosis in digitized kidney biopsies. *Am J Pathol.* 2021;191(8):1442-53. Available from: <https://consensus.app/papers/details/3731a795cbb75d1eac4346f4354f8ce5/>