ORIGINAL ARTICLE

Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal

Iacopo Vagliano1, Nicholas C. Chesnaye2, Jan Hendrik Leopold1, Kitty J. Jager2, Ameen Abu-Hanna1 and Martijn C. Schut1

1Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands and 2ERA Registry, Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

Correspondence to: Iacopo Vagliano; E-mail: i.vagliano@amsterdamumc.nl

ABSTRACT

Background. The number of studies applying machine learning (ML) to predict acute kidney injury (AKI) has grown steadily over the past decade. We assess and critically appraise the state of the art in ML models for AKI prediction, considering performance, methodological soundness, and applicability.

Methods. We searched PubMed and ArXiv, extracted data, and critically appraised studies based on the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD), Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS), and Prediction Model Risk of Bias Assessment Tool (PROBAST) guidelines.

Results. Forty-six studies from 3166 titles were included. Thirty-eight studies developed a model, five developed and externally validated one, and three studies externally validated one. Flexible ML methods were used more often than deep learning, although the latter was common with temporal variables and text as predictors. Predictive performance showed an area under receiver operating curves ranging from 0.49 to 0.99. Our critical appraisal identified a high risk of bias in 39 studies. Some studies lacked internal validation, whereas external validation and interpretability of results were rarely considered. Fifteen studies focused on AKI prediction in the intensive care setting, and the US-derived Medical Information Mart for Intensive Care (MIMIC) data set was commonly used. Reproducibility was limited as data and code were usually unavailable.

Conclusions. Flexible ML methods are popular for the prediction of AKI, although more complex models based on deep learning are emerging. Our critical appraisal identified a high risk of bias in most models: Studies should use calibration measures and external validation more often, improve model interpretability, and share data and code to improve reproducibility.

LAY SUMMARY

The number of studies applying machine learning (ML) to predict acute kidney injury (AKI) has grown steadily over the past decade. We assessed and critically appraised the state of the art in ML models for AKI prediction, considering performance, methodological soundness, and applicability. Forty-six studies from 3166 titles were included. Thirty-eight studies developed a model, five developed and externally validated one, and three studies externally...
Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal

The number of studies applying machine learning (ML) to predict acute kidney injury (AKI) has grown steadily over the last decade. We assess and critically appraise the state of the art in ML models for AKI prediction.

Keywords: acute kidney injury, clinical prediction models, critical appraisal, machine learning, systematic review

INTRODUCTION

Acute kidney injury (AKI) has a substantial impact on the global burden of kidney disease, with a global estimate of 13.3 million cases in 2017 [1, 2] and 1.7 million deaths each year globally [3, 4]. Early recognition, risk assessment, and care of AKI are suboptimal and contribute to disease progression, high health care costs, and poor patient outcomes [5, 6]. To assist physicians with risk assessment of AKI, prediction models have been developed across various patient populations with varying degrees of predictive accuracy [7, 8]. Models being built using machine learning (ML), which are mathematical models to make decisions and predictions based on data sets, have become popular [9]. ML differs from standard regression modelling (including models that tend to be parametric and their extensions, semiparametric or with a relatively low number of parameters—e.g. logistic regression and Cox models) in the high volume of data that can be used as input and the computational effort required for analysis [9, 10].

Recently, we have seen rapid growth in ML models for AKI prediction [12–30]. The sudden rise of such a novel and immediately popular modeling paradigm raises questions about how well these models perform, the soundness of their methodology, and whether the models are applicable to clinical settings (e.g. populations and availability of predictors).

Systematic reviews on AKI prediction are plentiful [12–29]. We are aware of a single review of AKI prediction using ML models [30], which assessed whether ML models outperform logistic regression for predicting AKI. This review did not perform any critical appraisal. In contrast, we review and critically appraise ML models for the prediction of AKI in terms of performance, methodological soundness, and clinical applicability.
MATERIALS AND METHODS

The protocol for this study was registered in the online PROSPERO database (CRD42022304868). We followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines [31].

Study identification

We used PubMed (pubmed.ncbi.nlm.nih.gov) and ArXiv (arxiv.org) for our search. We searched title or abstract with the string (Clinical OR medical) AND (predict*) AND (AKI OR AKF OR AKD OR ARI OR ARF OR ARD OR 'acute kidney injury' OR 'acute kidney failure' OR 'acute renal failure' OR 'acute renal insufficiency'). The search was conducted on March 1, 2021.

Study inclusion

We included studies that (i) developed or validated prediction models for AKI and (ii) used ML models. We excluded studies that focused on identifying or analyzing individual predictors instead of model development or validation. We excluded studies that used only standard regression models, gray literature, and informal publications (commentaries, letters to the editor, editorials, and meeting abstracts).

Study selection

Pilot selection and extraction were conducted by I.V., N.C.C. and J.H.L. to validate and refine the research question, the inclusion criteria, and the data-extraction form. Subsequently, we selected full-text papers based on abstract screening and divided them equally among I.V., N.C.C., and J.H.L. At least two researchers reviewed a quarter of the included studies to ensure an adequate level of inter-reviewer agreement. Discrepancies between reviewers were resolved by discussion.

Data extraction

We created a data-extraction form (Supplementary Table S1) based on the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklists [32,33]. We included items regarding specific aspects of the models (prediction time window and duration of follow-up), the type of data, the methods used for model interpretability, and the availability of data and code. I.V., N.C.C., and J.H.L. performed the data extraction.

Critical appraisal

We assessed potential biases in the included studies by using the Prediction Model Risk of Bias Assessment Tool (PROBAST) [34]. PROBAST distinguishes among different aspects that may generate bias: (i) the use of unsuitable data, (ii) participant selection, (iii) definition or assessment of predictors, (iv) outcome definition and its relation to the predictors, and (v) incorrect data analysis. The latter pertains to the handling of missing data, validation, and use of proper performance measures. To define common criteria for rating bias and applicability, I.V., N.C.C., J.H.L., and A.A.H. first reviewed and discussed one study. I.V., N.C.C., and J.H.L. then completed the critical appraisal. At least two researchers reviewed a quarter of the included studies to ensure inter-reviewer agreement. Disagreement between reviewers was resolved by discussion.

RESULTS

Literature search

We retrieved 3166 titles through our search (Fig. 1). Fifty-four were selected for full-text screening, and 46 studies were finally included. Most of these studies were published over the past 2 years (Fig. 2). Thirty-eight studies (82%) developed a model, five (11%) developed and externally validated one, and three (7%) externally validated one.

General study characteristics

Outcome

Thirty-two studies (70%) defined AKI as the outcome (distinguishing only between patients with and without AKI), and six studies (13%) focused on postoperative AKI. Other outcomes included the severity of AKI (10 studies [22%]), the progression of AKI (1 study [2%]), late AKI (AKI occurring after resuscitation or first 48 hours, 1 study [2%]), pre-existing AKI on arrival (1 study [2%]), hospital-acquired AKI (1 study [2%]), community-acquired AKI (1 study [2%]), drug-induced AKI (1 study [2%]), perioperative AKI (1 study [2%]), and cardiac surgery–associated AKI (1 study [2%]).

Definition and prevalence of AKI

The Kidney Disease Improving Global Outcomes (KDIGO) criteria [1] defined AKI in 36 studies (78%), whereas 4 studies (8%) used Acute Kidney Injury Network [35] 2 studies (4%) use Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease (RIFLE) criteria [36]. 1 study (2%) used codes from the International Classification of Diseases, Ninth Revision [37] and 1 study (2%) used the National Health Service England algorithm [38] together with KDIGO. The prevalence of AKI ranged from 0.5% (general hospital population) [38] to 72.7% (patients who underwent aortic arch surgery) [39] and was not reported in three studies (7%).

Type of prediction model

Figure 2 shows the number of studies over time, grouped by the type of model, and that deep-learning models emerged around 2017. Figure 3A shows the wide variety of models used in the selected studies. We distinguish between (i) flexible ML models, which tend to be nonparametric or are ‘parameter-rich’ models, such as decision trees and random forests, and (ii) deep-learning models, which are based on neural networks, have multiple levels of representation and rely on simple, nonlinear modules to transform the representation at one level into a more abstract representation. The most common models were random forest (17 studies [37%]) and gradient-boosted trees (9 studies [20%]). Among deep-learning models, recurrent neural networks were the most frequent (6 studies [13%]). Figure 3B illustrates the type of model used by data type.

Type and origin of data

The vast majority of studies used clinical variables with a single measurement (28 studies [61%]), whereas 13 studies (28%) used clinical variables with repeated measurements, 3 studies (7%)

...
used clinical variables with repeated measurements together with clinical notes, and 2 studies (4%) combined their data with external data. Twenty-seven studies used data from their own center. The Medical Information Mart for Intensive Care (MIMIC) data set, an openly available, intensive care–specific data set from the United States, was widely used (10 studies [28%]) [40]. Supplementary Figure S1 includes more details.

Model predictive performance

The predictive performance of clinical models is assessed through discrimination and calibration. The former is the ability of a predictive model to separate data into classes (e.g. correctly distinguishing between patients with and without AKI). The latter measures the agreement between predicted and observed outcomes [41].

Figure 4 summarizes the performance measures used for evaluating the models. Area under the receiver operating characteristic curve (AUROC) was the most used discrimination measure (41 studies [89%]). Calibration was rarely assessed (3 studies [7%]). Table 1 summarizes the reported performance measures for each study. AUROC varied from 0.49 to 0.99. Random forest was often the best performing model (12 studies [26%]) within the study. Other best performing models included recurrent neural networks (RNNs) (6 studies [13%]) and gradient-boosted trees (5 studies [11%]).

Twenty-three studies (50%) compared the performance of ML models to standard regression models. In all 23, logistic regression was used as a comparator, but least-angle regression (LARS) (one study [2%]), linear regression (one study [2%]), and multivariate adaptive regression splines (MARS) (one study [2%]) were used, as well. Logistic regression was the best performer (outperforming support vector machine, random for-
Model validation

The most common methods for internal validation were cross-validation (20 studies [43%]) and the separation of data in a training and a test set (19 studies [41%]). External validation of the model in a different population was performed in eight studies (17%). Supplementary Figure S2 provides further information.

Critical appraisal

Assessment of bias

Table 2 shows the result of the critical appraisal with PROBAST. The vast majority of studies were identified as having a high risk of bias [39 studies (85%)] because of how the analysis was performed: Calibration was not assessed [35 studies (76%)], and missing data were not optimally handled [21 studies (46%)]. One study (2%) had a risk of bias because of the selection of participants, one (2%) because of predictors, and three (7%) because of the outcome definition. One study (2%) had an unclear risk of bias for the predictors, one (2%) for the outcome. Concerns for applicability in clinical practice were raised by four studies (8%) because the predictors the model used were unavailable at the time of prediction. Two studies (4%) showed unclear applicability, one because of predictors, the other because of the outcome. Two studies (4%) that only externally validated a model were included in the critical appraisal but with high concerns for applicability because their main goal was to compare model performance with clinicians.

Data pre-processing

Twelve studies (26%) did not specify whether missing data were present or how they were treated, and five (11%) did not use any imputation method. In the studies that did handle missing data, mean and carry-forward imputation were the most common methods [six studies (13%)]. Four studies (8%) applied the multivariate imputation by chained equations (MICE) [42] method (Supplementary Figure S3). Another relevant aspect of model development concerns variable selection. Twelve studies (26%) used all the available variables, eight (16%) used expert opinion to pre-select variables, six (13%) used the least absolute shrinkage and selection operator (LASSO) [43], and five (11%) selected variables based on existing literature. Five studies (11%) did not specify whether variable selection was used. Supplementary Figure S4 contains more details.

Interpretability

Interpretability reflects the degree to which a human can consistently predict the model’s output [44]. Interpretability was rarely addressed [13 studies (28%). The most popular method to improve interpretability was providing the variable importance [seven studies (15%)]. Additional methods, used by a single study each (2%), were Shapley Additive Explanations [45], regression coefficients, contributions of variables to the predicted probability, statistical testing and manual evaluation to identify discriminant predictors, predicting future trajectories for clinically relevant biomarkers, and using a more interpretable logistic regression (fewer predictors) alongside the best model.

Applicability and reproducibility

Thirty studies (65%) were performed in tertiary care hospitals (Supplementary Figure S5). Twenty-eight studies (61%) included data from a single center. The number of study sites ranged from 1 to 1239 and was unspecified in 6 studies (13%). The intensive care unit (ICU) patient population was most frequently studied (15 studies [33%]), followed by the general hospital, surgery, and cardiac surgery populations [12, 8, and 5 studies, respectively (26%, 17%, 11%)]. The study population size ranged from 50 to 1841 951 [median, 23 246 [interquartile range: 4485–52 686]]. The duration of follow-up ranged from 24 to 1000 hours and was omitted in 11 studies (24%). The prediction window ranged from the time of admission to 7 days, but eight studies (17%) failed to specify it. Regarding reproducibility, few studies shared code [five studies [11%]] or data [nine studies [19%]].

DISCUSSION

Findings

We reviewed and critically appraised ML models for the prediction of AKI in terms of performance, methodological soundness, and clinical applicability. Models were mostly developed for the ICU population, followed by the general hospital and (cardiac) surgery populations. Although deep-learning models have emerged since 2014, more traditional, flexible ML methods (random forest and gradient-boosted trees) are still widely used to predict AKI. Prediction models typically include clinical predictors at baseline and, to a lesser extent, repeated measures. Although all studies provided model discrimination, equally important measures of calibration were rare. Most models were not externally validated. Our critical appraisal demonstrated a high
| Study | Settings | AUROC | Other measures | Best model | Comparison with | Validation |
|-------|----------|--------|----------------|------------|----------------|------------|
| S01 [54] | Any AKI 48 h ahead | 0.863–0.921 | PR AUPRC: 0.173–0.297 | RNN | Gradient-boosted trees, logistic regression | Internal |
| S01 | AKI 2–3 48 h ahead | 0.870–0.957 | PR AUPRC: 0.167–0.387 | PR AUC: 0.245–0.487 | | |
| S01 | AKI 3 48 h ahead | 0.930–0.980 | | | | |
| S02 [55] | Unstructured and structured features | 0.673–0.835 | F-measure: 0.091–0.542 | SVM | | Internal |
| S02 | Structured features | 0.657–0.812 | F-measure: 0.233–0.501 | Random forest | | |
| S02 | Unstructured features | 0.750–0.774 | F-measure: 0.066–0.495 | Logistic regression | | |
| S03 [56] | MIMIC eICU | 0.743–0.893 | 0.812–0.871 | | | Internal |
| S04 | AKI | 0.817–0.834 | F-measure: 0.283–0.430 | | | |
| S05 [57] | AKI | 0.499–0.867 | PR AUPRC: 0.063–0.332 | | | |
| S07 [58] | AKI stage sCr | 0.499–0.867 | PR AUPRC: 0.063–0.332 | | | |
| S08 [59] | AKI stage 1/sCr | 0.675–0.707 | Accuracy: 0.500–0.820 | | | |
| S09 [38] | Onset | 0.762–0.841 | 0.650–0.760 | Gradient-boosted trees | | Internal |
| S10 | AKI stage ≥1 | 0.730 | | | | |
| S11 [61] | AKI | 0.870 | 0.930 | | | |
| S12 [62] | AKI 7-days | 0.840–0.870 | Accuracy: 0.570–0.810 | | | |
| S15 [63] | AKI | 0.690–0.760 | | | | |
| S18 [64] | AKI stage ≥1 | 0.746–0.758 | | | | |
| S20 [65] | At 24 h from admission | 0.621–0.664 | Ensemble (of all techniques) | | | |

Table 1. Overview of the results reported by the studies and their settings
### Table 1. Continued

| Study | Settings | AUROC | Other measures | Best model | Comparison with | Validation |
|-------|----------|-------|----------------|------------|----------------|------------|
| S22 [66] | All features | 0.797–0.827 | Accuracy: 0.744–0.767 | Generalized additive model | Logistic regression, naïve Bayes, SVM | Internal |
| Feature selection with LASSO | 0.797–0.824 | Accuracy: 0.744–0.767 |
| Feature extraction with 5 principal components | 0.819–0.858 | Accuracy: 0.741–0.777 |
| S23 [67] | AKI data from admission | 0.751–0.765 | Random forest | AdaBoost, logistic regression | Internal |
| AKI data 24 h before admission | 0.732–0.747 |
| AKI data 7 days before admission | 0.733–0.747 |
| AKI data 15 days before admission | 0.733–0.742 |
| AKI data 30 days before admission | 0.732–0.747 |
| S26 [68] | AKI within first 48 h | 0.716–0.769 | PR AUPRC: 0.430–0.479 | RNN (LSTM) | RNN (GRU) | Internal |
| S32 [69] | Late AKI within first 24 h | 0.740–0.800 | Accuracy: 0.733 | CART | Geographical |
| Postoperative AKI | 0.730–0.890 |
| S38 [70] | AKI | 0.750–0.800 | Random forest | Bayesian model averaging SVMs, logistic regression | Internal |
| AKI data before admission | 0.890–0.940 |
| S40 [72] | On admission | 0.750–0.800 | AKIpredictor | Physicians | External |
| First morning | 0.890–0.950 |
| First 24 h | 0.890–0.950 |
| S41 [73] | AKI | 0.560–0.920 | Accuracy: 0.800–1.00 | KNN | Only KNN but using different predictors | Internal |
| S42 [74] | Any AKI | 0.882 | Random forest | | Internal |
| AKI stage ≥2 | 0.878 |
| S43 [75] | AKI before onset | 0.687–0.744 | F-measure: 0.261–0.330 | Ensemble (logistic regression and random forest) | Logistic regression, random forest, naïve Bayes, Bayesian network | Internal |
| AKI within the stay | 0.676–0.734 | F-measure: 0.253–0.318 |
| AKI within first 30 days | 0.720–0.764 | F-measure: 0.184–0.316 |
| AKI within first 5 days | 0.600–0.764 | F-measure: 0.047–0.184 |
| S44 [76] | AKI | 0.772–0.796 | Accuracy: 0.724–0.744 | MLP | Logistic regression, random forest | Internal |
| S46 [77] | AKI | 0.550–0.780 | Gradient-boosted trees | Decision trees, random forest, gradient-boosted trees, SVM, MLP, deep-belief networks | Internal |
| S48 [78] | AKI | 0.573–0.809 | Accuracy: 0.575–0.813 | Random forest | Preselected random forest comparing it with gradient-boosted trees, bayesian networks, SVM, logistic regression, naïve Bayes, KNN, deep learning (unspecified) | Internal |
| | | | F-measure: 0.628–0.833 |
| | | | 0.589–0.809 | Random forest + local and global pattern detection | Only random forest (using 3 different pattern-detection variants) and last recorded value |
| | | | F-measure: 0.634–0.833 | | |
| Study | Settings | AUROC | Other measures | Best model | Comparison with | Validation |
|-------|----------|-------|----------------|------------|----------------|------------|
| S49 [79] | AKI 0 days ahead | 0.745–0.875 | F1: 0.745–0.875 | KNN | AdaBoost, logistic regression, random forest | Internal |
| | AKI 1 day ahead | | F1: 0.686–0.759 | | | |
| | AKI 2 days ahead | | F1: 0.605–0.695 | | | |
| | AKI 3 days ahead | | F1: 0.588–0.654 | | | |
| | AKI 4 days ahead | | F1: 0.590–0.659 | | | |
| | AKI 5 days ahead | | F1: 0.572–0.646 | | | |
| S50 [80] | Hospital-acquired AKI 24–96 h ahead | 0.552–0.791 | Accuracy: 0.648–0.736 | Recurrent additive network | Logistic regression, SVM | Internal |
| | | | F1: 0.403–0.644 | | | |
| S52 [81] | AKI | 0.720–0.960 | Accuracy: 0.730–0.900 | RNN | KDIGO | External |
| | | | F1: 0.660–0.900 | | | |
| S53 [82] | AKI | 0.580–0.824 | PR AUPRC: 0.137–0.264 | F-GAM | Decision trees, logistic regression, random forest, gradient-boosted stumps, SVM, deep learning (unspecified) | Internal |
| | | | | | Logistic regression, random forest, gradient-boosted trees | |
| S54 [83] | Unstructured and structured features | 0.660–0.700 | RNN | | | Internal |
| | Structured features | 0.700–0.709 | | | | |
| | Unstructured features | 0.720–0.775 | | | | |
| S56 [84] | AKI | 0.650–0.790 | MySurgeryRisk physicians predictor | Physicians | External |
| | AKI data before admission | 0.750 | | | Internal |
| | AKI data before and on admission | 0.770 | | | |
| | AKI data before admission and first 24 h | 0.800 | | | |
| | AKI data before admission and first 24 h and radio-contrast 1 week before | 0.820 | | | |
| S59 [86] | AKI | 0.738–0.988 | CNN | Decision trees, logistic regression, random forest, RNN | Internal |
| S60 [87] | AKI | 0.745–0.901 | AUPRC: 0.747–0.907 | RNN | Physicians | Internal |
| | | | Accuracy: 0.711–0.846 | | | |
| | | | F1: 0.673–0.848 | | | |
| S61 [88] | AKI | 0.690–0.70 | SVM | Logistic regression, random forest, SVM, KNN, AdaBoost | Internal |
| S62 [89] | AKI 24 h ahead | 0.530–0.810 | ETSM | | Geographical |
| | AKI 48 h ahead | 0.520–0.780 | | | |
| S63 [90] | AKI | 0.845–0.855 | Random forest | | Internal |
risk of bias in the majority of studies, with some concern regarding their applicability in clinical practice.

Performance
Random forest was often the best performing method compared with other models within the same study. RNN demonstrated promising results. The popularity and performance of the simpler, flexible ML models, such as random forest, may indicate that flexible ML methods are sufficiently effective or perhaps better than deep-learning techniques for the type of data and tasks relevant for AKI prediction. Most studies relied on baseline clinical predictors and less so on clinical notes or repeated measures. Choosing the optimal model highly depends on the type of data available. Deep learning is typically beneficial for complex data, as demonstrated by several studies incorporating predictors derived from text or repeated measures. Although the use of deep learning may improve predictive performance...
Table 1. Continued

| Study  | Settings                      | AUROC               | Other measures                                    | Best model                                      | Comparison with          | Validation |
|--------|-------------------------------|---------------------|---------------------------------------------------|------------------------------------------------|--------------------------|------------|
| S64 [91] | AKI stage ≥1                  | 0.670–0.720         | Gradient-boosted trees                            | Only 1 model                                    | Temporal                |            |
|        | AKI stage ≥2                  | 0.850–0.860         |                                                   | Gradient-boosted trees                           |                          |            |
|        | AKI stage ≥3                  | 0.910–0.920         |                                                   |                                                  |                          |            |
| S65 [92] | AKI stage ≥1                  | 0.761               | Logistic regression (LASSO)                       |                                                 | Internal                |            |
|        | AKI stage ≥2                  | 0.850               |                                                   |                                                 |                          | Internal   |
| S66 [93] | AKI stage ≥3                  | 0.910               |                                                   |                                                 |                          | Internal   |
|        | AKI stages                    | 0.761               |                                                   |                                                 |                          | Internal   |
| S67 [94] | AKI stage ≥1 within 24 h      | 0.800               | Random forest                                     |                                                 | Internal                |            |
|        | AKI stage ≥2 within 24 h      | 0.760               |                                                   |                                                 |                          | Internal   |
|        | AKI stage ≥1 within 48 h      | 0.740               |                                                   |                                                 |                          |            |
|        | AKI stage ≥2 within 48 h      | 0.810               |                                                   |                                                 |                          | Internal   |
|        | AKI stage ≥1 within 72 h      | 0.770               |                                                   |                                                 |                          | Internal   |
|        | AKI stage ≥2 within 72 h      | 0.750               |                                                   |                                                 |                          | Internal   |
| S68 [38] | AKI                           | 0.640–0.800         | Light gradient machine                            |                                                 | Internal                |            |
| S70 [95] | AKI                           | 0.560–0.710         | Bayesian networks                                 |                                                 | Internal                |            |
| S71 [96] | AKI                           | 0.728–0.755         | Bayesian networks                                 |                                                 | Internal                |            |
| S72 [97] | AKI                           | 0.812–0.835         | Bayesian networks                                 |                                                 | Internal                |            |
|        |                              | 0.682–0.782         | Deep rule forest                                  |                                                 | None                     |            |

AdaBoost: adaptive boosting; AUC: area under the curve; AUPRC: area under the precision-recall curve; CART: classification and regression trees; eICU: xx; ETSM: ensemble time-series model; F-GAM: factored-generalized additive model; GRU: gated recurrent unit; LSTM: long short-term memory; MLP: multilayer perceptron (feed-forward neural network); PR: AUPRC; sCr: serum creatinine; SGD: stochastic gradient descend; SOFA: sequential organ failure assessment; SVM: support vector machine.

in these settings, it comes at the cost of being less interpretable, which may discourage its uptake in clinical practice. Prediction models are inherently uninterpretable from a causal perspective. Interpretability in the context of prediction refers to the explicable of the predictions (i.e. how the model made the prediction) and which predictors contributed the most to the prediction (i.e. variable importance). Although some models are easier to interpret than others, making predictions understandable does not provide any information about the underlying causal mechanisms between predictors and outcome. Inferring causality from prediction models is referred to as the ‘Table 2 fallacy’ [46].

**Methodologic soundness**

We found a high risk of bias in the majority of studies, mostly because of flaws in the analysis. A common flaw was the lack of model calibration. Although model discrimination was typically assessed, calibration was often overlooked; both, however, should be reported to evaluate model performance [34]. Specific tasks call upon different performance measures. For example, benchmarking and decision-making based on individual predictions require good calibration, while identifying the most vulnerable patients mainly requires discrimination. The reviewed studies did not explain why they did or did not use specific measures. Another common flaw was the reliance on simple internal validation methods, such as splitting data in train and test sets, without correcting for optimism and overfitting. More reliable methods, such as cross-validation, should be preferred. Similarly, suboptimal methods for dealing with missing data were often used, whereas MICE provides the least biased results [47]. The two main strategies used for variable selection were the inclusion of all available variables and backward-elimination methods. There is no consensus on the best method for variable selection [48], although including all variables can avoid overfitting and selection bias [49], even though this is often impractical [48]. Finally, only two studies relied on prospective data. Although we acknowledge the difficulties associated with collecting data prospectively, retrospective data may not be representative of the patient population and are prone to selection bias, recall bias, and misclassification bias [50].
Table 2. Results of the critical appraisal with PROBAST

| Study | ROB | Applicability | Overall |
|-------|-----|---------------|---------|
|       | Participants | Predictors | Outcome | Analysis | Participants | Predictors | Outcome | ROB | Applicability |
| S01 [54] | + | + | + | + | + | + | + | + | + |
| S02 [55] | + | + | + | + | + | + | + | + | + |
| S03 [56] | + | + | + | + | + | + | + | + | + |
| S05 [57] | + | ? | ? | ? | + | - | ? | - | - |
| S07 [58] | + | + | + | - | + | ? | + | - | + |
| S08 [59] | + | + | + | - | + | + | + | - | + |
| S09 [60] | + | + | + | - | + | + | + | - | + |
| S10 | + | + | + | - | + | + | + | - | + |
| S11 [61] | + | + | + | - | + | + | + | - | + |
| S12 [62] | + | + | - | - | + | + | + | - | + |
| S15 [63] | + | + | + | + | + | + | + | + | + |
| S18 [64] | + | + | + | - | + | + | + | - | + |
| S20 [65] | + | + | + | - | + | + | + | - | + |
| S22 [66] | + | + | + | - | + | + | + | - | + |
| S23 [67] | + | + | + | - | + | + | + | - | + |
| S26 [68] | + | + | ? | ? | + | - | ? | - | - |
| S32 [69] | + | + | + | - | + | + | + | - | + |
| S38 [70] | + | + | ? | ? | + | - | ? | - | - |
| S39 [71] | + | + | + | - | + | + | + | - | + |
| S40 [72] | + | + | + | - | + | + | + | - | + |
| S41 [73] | + | + | + | - | + | + | + | - | + |
| S42 [74] | + | + | + | - | + | + | + | - | + |
| S43 [75] | + | + | + | - | + | + | + | - | + |
| S44 [76] | + | + | + | - | + | + | + | - | + |
| S46 [77] | + | + | + | - | + | + | + | - | + |
| S48 [78] | + | + | + | - | + | + | + | - | + |
| S49 [79] | + | + | + | - | + | + | + | - | + |
| S50 [80] | + | + | + | - | + | + | + | - | + |
| S52 [81] | + | + | + | - | + | + | + | - | + |
| S53 [82] | + | + | + | - | + | + | + | - | + |
| S54 [83] | + | + | + | - | + | + | + | - | + |
| S56 [84] | + | + | + | - | + | + | + | - | + |
| S58 [85] | + | + | + | - | + | + | + | - | + |
| S59 [86] | + | + | + | - | + | + | + | - | + |
| S60 [87] | + | + | + | - | + | + | + | - | + |
| S61 [88] | + | + | + | - | + | + | + | - | + |
| S62 [89] | + | + | + | - | + | + | + | - | + |
| S63 [90] | + | + | + | - | + | + | + | - | + |
| S64 [91] | + | + | + | - | + | + | + | - | + |
| S65 [92] | + | + | + | - | + | + | + | - | + |
| S66 [93] | + | + | + | - | + | + | + | - | + |
| S67 [94] | + | + | + | - | + | + | + | - | + |
| S68 [95] | + | + | + | - | + | + | + | - | + |
| S70 [96] | + | + | + | - | + | + | + | - | + |
| S71 [97] | + | + | + | - | + | + | + | - | + |
| S72 [98] | + | + | + | - | + | + | + | - | + |

The plus symbol (+) indicates low risk of bias (ROB) or low concern for applicability; the minus symbol (−) means high ROB or high concern for applicability; the question mark (?) implies unclear ROB or unclear concern for applicability.

Clinical applicability

The majority of the studies used data from a single center, implying that the model would be less generalizable to the broader patient population. Although many studies have been performed in the ICU, the MIMIC data set was often used, possibly because MIMIC is publicly available and includes complex data (repeated measures and clinical notes). Although using the same data may foster the comparison of models among studies, prediction results risk being biased toward its specific population and may be less generalizable to the broader ICU population. External validation of models was rare, further limiting the generalizability to other populations.

Reproducible research has become a pressing issue across many scientific disciplines, and sharing data and code is key [47, 51, 52]. The ability to reproduce studies is limited as data and code were usually unavailable. Even when there are commercial concerns about intellectual property, strong arguments exist for ensuring that algorithms are nonproprietary and available for scrutiny [53]. Proprietary algorithms hamper transparency and prevent external validation in different settings by independent researchers.
Challenges and opportunities
The main opportunity that ML offers for the prediction of AKI is that these models allow for a more flexible relationship between the predictors and the outcome than standard regression methods. Flexible ML models allow expression of highly nonlinear relationships between predictors and AKI. Besides the typical use of baseline predictors in most models, deep-learning models are capable of including time-updated measurements of predictors as well as text from clinical notes, with the potential of improving model performance. Deep learning, with its latent representations (e.g., a hidden layer in a neural network) can uncover complex relationships between predictors and outcome, hence improving the prediction. This advantage makes sense only if complex relationships exist and if there are sufficient data to reliably estimate model parameters. Learning such models requires managing their complexity as they are prone to overfitting.

Limitations
Our study has three main limitations. First, although comprehensive, our search strategy may have missed some relevant studies. We selected two sources (PubMed and ArXiv) that should have identified the most significant studies from the medical and ML domains (see Supplementary Section B), but we excluded studies with only standard regression models. Second, the risk of bias entails some subjective judgment, and people with different experiences of ML performance could have varying perceptions. To limit this effect, 12 were reviewed by at least two assessors. Third, PROBAST was designed for regression models. There are no clear guidelines on how to score some questions (e.g., regarding predictors and sample size) for machine learning and deep-learning models. The upcoming TRIPOD-AI and PROBAST-AI might overcome this limitation [53].

CONCLUSIONS
Relatively simple models, such as random forest and gradient-boosted trees, are still common, although more complex models based on deep learning are emerging, providing opportunities for the inclusion of temporal data and text as predictors. Although deep-learning models have the potential to improve predictions, they are also less interpretable, which may impede uptake in clinical practice—challenges that should be addressed in the future. In accordance with reporting guidelines, we encourage reporting both model discrimination and model calibration. The generalizability of prediction models should be improved through the use of multicenter data during development or external validation. Sharing data and code is encouraged to improve study reproducibility.

SUPPLEMENTARY DATA
Supplementary data are available at ckj online.

AUTHORS’ CONTRIBUTIONS
I.V. contributed to research idea, study design, methodology, extraction, analysis, and interpretation of data, writing—original draft; N.C.C. contributed to methodology, extraction, analysis, and interpretation of data, writing—original draft; J.H.L. contributed to methodology, extraction and analysis of data, writing—review & editing; A.A.H. contributed to methodology, interpretation of data, writing—review & editing; K.J.J. contributed to methodology, extraction, and interpretation of data, writing—review & editing; K.J.J. contributed to study design, methodology, extraction, and interpretation of data, writing—original draft; J.H.L.

REFERENCES
1. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract 2013;120:c179–84. https://doi.org/10.1159/000339789
2. Jager KJ, Kovesdy C, Langham R et al. A single number for advocacy and communication—worldwide more than 850 million individuals have kidney diseases. Kidney Int 2019;96:1048–50. https://doi.org/10.1016/j.kint.2019.07.012
3. Susantitaphong P, Cruz DN, Cerda J et al. Acute Kidney Injury Advisory Group of the American Society of Nephrology. World incidence of AKI: a meta-analysis. Clin J Am Soc Nephrol 2013;8:1482–93. https://doi.org/10.2215/CJN.00710113
4. Mehta RL, Cerda J, Burdmann EA et al. International Society of Nephrology’s 0by25 initiative for acute kidney injury (zero preventable deaths by 2025): a human rights case for nephrology. Lancet 2015;385:2616–43. https://doi.org/10.1016/S0140-6736(15)60126-X
5. Hoste EAJ, Kellum JA, Selby NM et al. Global epidemiology and outcomes of acute kidney injury. Nat Rev Nephrol 2018;14:607–25. https://doi.org/10.1038/s41581-018-0052-0
6. National Confidential Enquiry into Patient Outcome and Death. Adding insult to injury: a review of the care of patients who died in hospital with a primary diagnosis of acute kidney injury (acute renal failure). National Confidential Enquiry into Patient Outcome and Death, 2009
7. Matheny ME, Miller RA, Ikizler TA et al. Development of patient risk stratification models of acute kidney injury for use in electronic health records. Med Decis Making 2010;30:639–50. https://doi.org/10.1177/0272989X10364246
8. Mehran R, Aymong ED, Nikolsky E et al. A simple risk score for prediction of contrast-induced nephropathy after percutaneous coronary intervention: development and initial validation. J Am Coll Cardiol 2004;44:1393–9. https://doi.org/10.1016/j.jacc.2004.06.068
9. Coorey CP, Sharma A, Mueller S et al. Prediction modelling—part 2: using machine learning strategies to improve transplantation outcomes. Kidney Int 2021;99:817–23. https://doi.org/10.1016/j.kint.2020.08.026
10. Au EH, Francis A, Bernier-Jean A et al. Prediction modelling—part 1: regression modeling. Kidney Int 2020;97:877–84. https://doi.org/10.1016/j.kint.2020.02.007
11. Gameiro J, Branco T, Lopes JA. Artificial intelligence in acute kidney injury risk prediction. J Clin Med 2020;9:678. https://doi.org/10.3390/jcm9030678

12. Park S, Lee H. Acute kidney injury prediction models: current concepts and future strategies. Curr Opin Nephrol Hypertens 2019;28:552-9. https://doi.org/10.1097/MNH.0000000000000536

13. Hodgson LE, Selby N, Huang T-M et al. The role of risk prediction models in prevention and management of AKI. Semin Nephrol 2019;39:421–30. https://doi.org/10.1016/j.sneph.2019.06.002

14. Hodgson LE, Sarnowski A, Roderick PJ et al. Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations. BMJ Open 2017;7:e016591. https://doi.org/10.1136/bmjopen-2017-016591

15. Pozzoli S, Simonini M, Manunta P. Predicting acute kidney injury: current status and future challenges. J Nephrol 2018;31:209–23. https://doi.org/10.1007/s40620-017-0416-8

16. Wilson T, Quan S, Cheema K et al. Risk prediction models for acute kidney injury following major noncardiac surgery: systematic review. Nephrol Dial Transplant 2016;31:231–40. https://doi.org/10.1093/ndt/gfv414

17. Allen DW, Ma B, Leung KC et al. Risk prediction models for contrast-induced acute kidney injury accompanying cardiac catheterization: systematic review and meta-analysis. Can J Cardiol 2017;33:724–36. https://doi.org/10.1016/j.cjca.2017.01.018

18. Caragata R, Wysusske KH, Kruger P. Acute kidney injury following liver transplantation: a systematic review of published predictive models. Anaesth Intensive Care 2016;44:251–61. https://doi.org/10.1111/apn.12310

19. Szerlip HM, Chawla LS. Predicting acute kidney injury prognosis. Curr Opin Nephrol Hypertens 2016;25:226–31. https://doi.org/10.1097/MNH.0000000000000223

20. Safari S, Yousefifard M, Hashemi B et al. The role of scoring systems and urine dipstick in prediction of rhabdomyolysis-induced acute kidney injury: a systematic review. Iran J Kidney Dis 2016;10:1–6.

21. Lin X, Yuan J, Zhao Y et al. Urine interleukin-18 in prediction of acute kidney injury: a systemic review and meta-analysis. J Nephrol 2015;28:7–16. https://doi.org/10.1007/s40620-014-0113-9

22. de Geus HR, Betjes MG, Bakker J. Biomarkers for the prediction of acute kidney injury: a narrative review on current status and future challenges. Clin Kidney J 2012;5:102–8. https://doi.org/10.1093/ckj/sfs008

23. Liu X, Guan Y, Xu S et al. Early predictors of acute kidney injury: a narrative review. Kidney Blood Press Res 2016;41:680–700. https://doi.org/10.1007/s00447937

24. Meisner A, Kerr KF, Thiessen-Philbrook H et al. Methodological issues in current practice may lead to bias in the development of biomarker combinations for predicting acute kidney injury. Kidney Int 2016;89:429–38. https://doi.org/10.1038/ki.2015.283 ISSN 0085-2538

25. Ho J, Tangri N, Komenda P et al. Urinary, plasma, and serum biomarkers’ utility for predicting acute kidney injury associated with cardiac surgery in adults: a meta-analysis. Am J Kidney Dis 2015;66:993–1005. https://doi.org/10.1053/j.ajkd.2015.06.018

26. Mosa O, Skitek M, Jerin A. Validity of Klotho, CYR61 and YKL-40 as ideal predictive biomarkers for acute kidney injury: review study. Sao Paulo Med J 2017;135:57–65. https://doi.org/10.1590/1516-3180.2016.0099220516

27. Darmon M, Truche AS, Abdel-Nabey M et al. Early recognition of persistent acute kidney injury. Semin Nephrol 2019;39:431-41. https://doi.org/10.1016/j.sj.snmephrol.2019.06.003 ISSN 0270-9295

28. Sutherland SM, Goldstein SL, Bagshaw SM. Acute kidney injury and big data. Contrib Nephrol 2018;193:55–67. https://doi.org/10.1159/000484963

29. Song X, Liu X, Liu F et al. Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis. Int J Med Informatics 2021;151:104484. https://doi.org/10.1016/j.ijmedinf.2021.104484

30. Moher D, Liberati A, Tetzlaff J et al. PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535. https://doi.org/10.1136/bmj.b2535

31. Collins GS, Reitsma JB, Altman DG et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. https://doi.org/10.1136/bmj.g7594

32. Moons KGM, de Groot JAH, Bouwmeester W et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMES checklist. PLoS Med 2014;11:e1001744. https://doi.org/10.1371/journal.pmed.1001744

33. Moons KGM, Wolf RP, Riley RD et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–33. https://doi.org/10.7326/M18-1377

34. Bellomo R, Ronco C, Kellum JA et al. Acute Dialysis Quality Initiative Working Group. Acute renal failure—definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care 2004;8:R204–12. https://doi.org/10.1186/cc2872

35. Mehta RL, Kellum JA, Shah SV et al. Acute kidney injury network: report of an initiative to improve outcomes in acute kidney injury, Crit Care 2007;11:R31. https://doi.org/10.1186/cc5713

36. Centers for Disease Control and Prevention. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Atlanta: Centers for Disease Control and Prevention, 2002.

37. Mohamadlou H, Lynn-Palevsky A, Barton C et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. Can J Kidney Health Dis 2018;5:2054358118776326. https://doi.org/10.1177/2054358118776326

38. Lei G, Wang G, Zhang C et al. Using machine learning to predict acute kidney injury after aortic arch surgery. J Cardiothorac Vasc Anesth 2020;34:3321–8. https://doi.org/10.1053/j.jvca.2020.06.007

39. Johnson AEW, Pollard TJ, Shen L et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35

40. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer, 2009.

41. van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67. https://doi.org/10.18637/jss.v045.i03

42. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc 1996;58:267–88.
43. Kim B, Khanna R, Koyejo O. Examples are not enough, learn to criticize! Criticism for interpretability. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016;2288–96.

44. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017;4768288–77.

45. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol* 2013;177:292–8. https://doi.org/10.1093/aje/kws412

46. Camerer CF, Dreber A, Holzmeister F et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat Hum Behav* 2018;2:637–44. https://doi.org/10.1038/s41562-018-0399-z

47. Royston P, Moons KGM, Altman DG et al. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604. https://doi.org/10.1136/bmj.b604

48. Harrell FE, Jr. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001.

49. Ebrahim S, Sohani ZN, Montoya L et al. Retrospective reanalyses of randomized clinical trial data. *JAMA* 2014;312:1024–32. https://doi.org/10.1001/jama.2014.9646

50. Schumacher M, Heinze G, Sauerbrei W et al. How to aclear the black box? A graphical tool to explain black-box models. *BMC Med Res Methodol* 2015;15:143. https://doi.org/10.1186/s12874-015-0088-4

51. Ehmke HC, Fink D, Schumann M et al. Artificial intelligence algorithms for medical prediction should be nonproprietary and readily available. *JAMA Intern Med* 2019;179:731. https://doi.org/10.1001/jamainternmed.2019.0597

52. Collins GS, Dinan SM, Vangel MG et al. A comparison of risk prediction models for the prediction of hospital length of stay. *BMJ Open* 2016;6:e015570. https://doi.org/10.1136/bmjopen-2016-011866

53. Collins GS, Dinan SM, Vangel MG et al. A comparison of risk prediction models for the prediction of hospital length of stay. *BMJ Open* 2016;6:e015570. https://doi.org/10.1136/bmjopen-2016-011866

54. Tomašev N, Glorot X, Rae JW et al. Artificial intelligence algorithms for medical prediction should be nonproprietary and readily available. *JAMA Intern Med* 2019;179:731. https://doi.org/10.1001/jamainternmed.2019.0597

55. Sun M, Baron J, Dighe A et al. Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. *Stud Health Technol Inform* 2019;264:366–72. https://doi.org/10.3233/SHTI190245

56. Pan Z, Du H, Ngiam KY et al. A self-correcting deep learning approach to predict acute conditions in critical care. *arXiv:19010436*. https://doi.org/10.48550/arXiv.1901.04364

57. Parreco J, Chatooor M. Comparing machine learning algorithms for predicting acute kidney injury. *Am Surg* 2019;85:725–9.

58. Weisenthai SJ, Quill C, Farooq S et al. Predicting acute kidney injury at hospital re-entry using high-dimensional electronic health record data. *PLoS One* 2018;13:e0204920. https://doi.org/10.1371/journal.pone.0204920

59. Park N, Kang E, Park M et al. Predicting acute kidney injury in cancer patients using heterogeneous and irregular data. *PLoS One* 2018;13:e0199839. https://doi.org/10.1371/journal.pone.0199839

60. Koyner J, Carey K, Edelson D et al. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med* 2018;46:1070–7. https://doi.org/10.1097/CCM.0000000000003123

61. Bihorac A, Ozrazgat-Baslanti T, Ebadi A et al. My-SurgeryRisk: development and validation of a Machine-learning risk algorithm for major complications and death after surgery. *Ann Surg* 2018;269:652. https://doi.org/10.1097/SLA.0000000000002706

62. Adhikari N, Ozrazgat-Baslanti T, Ruppert M et al. Improved predictive models for acute kidney injury with IDEA: intraoperative data embedded analytics. *PLoS One* 2019;14:e0214904. https://doi.org/10.1371/journal.pone.0214904

63. Davis SE, Lasko TA, Chen G et al. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017;24:1052–61. https://doi.org/10.1093/jamia/occ030

64. Cronin RM, VanHouten JP, Siew ED et al. National Veterans Health Administration inpatient risk stratification models for hospital-acquired acute kidney injury. *J Am Med Inform Assoc* 2015;22:1054–71. https://doi.org/10.1093/jamia/ovu051

65. Kate RJ, Perez RM, Mazumdar D et al. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med Inform Decis Mak* 2016;16:39. https://doi.org/10.1186/s12911-016-0277-4

66. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One* 2016;11:e0155705. https://doi.org/10.1371/journal.pone.0155705

67. Cheng P, Waitman LR, Hu Y et al. Predicting inpatient acute kidney injury over different time horizons: how early and accurate? *AMIA Annu Symp Proc* 2018;2017:565–74.

68. Zhang K, Xue Y, Flores G et al. Modelling EHR timeseries by restricting feature interaction. *arXiv* 2019. https://doi.org/10.48550/arXiv.1911.06410

69. Schneider DF, Dobrowolsky A, Shakir IA et al. Predicting acute kidney injury among burn patients in the 21st century: a classification and regression tree analysis. *J Burn Care Res* 2012;33:242–51. https://doi.org/10.1097/BCR.0b013e318239cc24

70. Kerr KF, Morenz ER, Roth J et al. Developing biomarker panels to predict progression of acute kidney injury after cardiac surgery. *Kidney Int Rep* 2019;4:1677–88. https://doi.org/10.1016/j.ekir.2019.08.017

71. Zhou C, Wang R, Jiang W et al. Machine learning for the prediction of acute kidney injury and paraplegia after thoracoabdominal aortic aneurysm repair. *J Card Surg* 2020;35:89–99. https://doi.org/10.1111/jocs.14317

72. Flechet M, Falini S, Bonetti C et al. Machine learning versus physicians’ prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIPredictor. *Crit Care* 2019;23:282. https://doi.org/10.1186/s13054-019-2563-x

73. Tran NK, Sen S, Palmeri TL et al. Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: a proof of concept. *Burns* 2019;45:1350–8. https://doi.org/10.1016/j.burns.2019.03.021

74. Chiofolo C, Chbat N, Ghosh E et al. Automated continuous acute kidney injury prediction and surveillance: a random forest model. *Mayo Clin Proc* 2019;94:783–92. https://doi.org/10.1016/j.mayocp.2019.02.009

75. He J, Hu Y, Zhang X et al. Multi-perspective predictive modeling for acute kidney injury in general hospital populations
using electronic medical records. JAMIA Open 2019;2:115–22. https://doi.org/10.1093/jamiaopen/oo043
76. Zimmerman LP, Reyfman PA, Smith ADR et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. BMC Med Inf Decis Mak 2019;19:16. https://doi.org/10.1186/s12911-019-0733-z
77. Lee HC, Yoon HK, Nam K et al. Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. J Clin Med 2018;7:322. https://doi.org/10.3390/jcm7100322
78. Morid MA, Sheng ORL, Fiol GD et al. Temporal pattern detection to predict adverse events in critical care: case study with acute kidney injury. JMIIR Med Inform 2020;8:e14272. https://doi.org/10.2196/14272
79. Chen YS, Chou CY, Chen ALP. Early prediction of acquiring injury with convolutional neural network based on electronic health record data with memory networks. J Biomed Inform 2020;102:10361. https://doi.org/10.1016/j.jbi.2019.10361
80. Goodwin TR, Demner-Fushman D. A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. J Am Med Inform Assoc 2020;27:567–76. https://doi.org/10.1093/jamia/ocaa004
81. Meyer A, Zverinski D, Pfahringer B et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018;6:905–14. https://doi.org/10.1016/S2213-2600(18)30300-X
82. Cui Z, Fritz BA, King CR et al. A factored generalized additive model for clinical decision support in the operating room. AMIA Annu Symp Proc 2020;2019:343–52
83. Xu Z, Chou J, Zhang XS et al. Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. J Biomed Inform 2020;102:10361. https://doi.org/10.1016/j.jbi.2019.10361
84. Brennan M, Puri S, Ozrazgat-Baslanti T et al. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: a pilot study. Surgery 2019;165:1035–45. https://doi.org/10.1016/j.surg.2019.01.002
85. Flechet M, Güiza F, Schetz M et al. AKI predictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. Intensive Care Med 2017;43:764–73. https://doi.org/10.1007/s00134-017-4678-3
86. Wang Y, Bao J, Du J et al. Precisely predicting acute kidney injury with convolutional neural network based on electronic health record data. arXiv:200513171. https://doi.org/10.48550/arXiv.2005.13171
87. Rank N, Pfahringer B, Kempfert J et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPI J Digit Med 2020;3:139. https://doi.org/10.1093/s41746-020-00346-8
88. Al-Jefri M, Lee J, James M. Predicting acute kidney injury after surgery. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2020, 5606–9. https://doi.org/10.1109/EMBC44109.2020.9175448
89. Wang Y, Wei Y, Yang H et al. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. BMC Med Inf Decis Mak 2020;20:238. https://doi.org/10.1186/s12911-020-01245-4
90. Li Y, Chen X, Shen Z et al. Prediction models for acute kidney injury in patients with gastrointestinal cancers: a real-world study based on bayesian networks. Ren Fail 2020;42:869–76. https://doi.org/10.1080/0886022X.2020.1810068
91. Churpek MM, Carey KA, Edelson DP et al. Internal and external validation of a machine learning risk score for acute kidney injury. JAMA Netw Open 2020;3:e2012892. https://doi.org/10.1001/jamanetworkopen.2020.12892
92. Hsu CN, Liu CL, Tain YL et al. Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: development and validation study. J Med Internet Res 2020;22:e16903. https://doi.org/10.2196/16903
93. Tseng PY, Chen YT, Wang CH et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020;24:478. https://doi.org/10.1186/s13054-020-03179-9
94. Martinez DA, Levin SR, Klein EY et al. Early prediction of acute kidney injury in the emergency department with machine-learning methods applied to electronic health record data. Ann Emerg Med 2020;76:501–14. https://doi.org/10.1016/j.annemergmed.2020.05.026
95. Li Y, Xu J, Wang Y et al. A novel machine learning algorithm, bayesian networks model, to predict the high-risk patients with cardiac surgery-associated acute kidney injury. Clin Cardiol 2020;43:752–61. https://doi.org/10.1002/clc.23377
96. Li Y, Chen X, Wang Y et al. Application of group LASSO regression based bayesian networks in risk factors exploration and disease prediction for acute kidney injury in hospitalized patients with hematologic malignancies. BMC Nephrol 2020;21:162. https://doi.org/10.1186/s12882-020-01786-w
97. Kuo B, Kang Y, Wu P et al. Discovering drug-drug and drug-disease interactions inducing acute kidney injury using deep rule forests. arXiv:200702103. https://doi.org/10.1109/I49571.2020.00062