International Journal of Medical Informatics 151 (2021) 104484

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Review article

# Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis

Xuan Song [a], Xinyan Liu [a], Fei Liu [b], Chunting Wang [c,*]

[a] ICU, DongE Hospital Affiliated to Shandong First Medical University, Shandong, China
[b] Urology Department, Tai'an Traditional Chinese Medicine Hospital Affiliated to Shandong University of Traditional Chinese Medicine, Shandong, China
[c] ICU, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Shandong, China

ARTICLE INFO

*Keywords:*
Acute kidney injury
Machine learning
Artificial intelligence
Logistic regression

ABSTRACT

*Introduction:* We aimed to assess whether machine learning models are superior at predicting acute kidney injury (AKI) compared to logistic regression (LR), a conventional prediction model.
*Methods:* Eligible studies were identified using PubMed and Embase. A total of 24 studies consisting of 84 prediction models met inclusion criteria. Independent samples *t*-test was performed to detect mean differences in area under the curve (AUC) between ML and LR models. One-way ANOVA and post-hoc *t*-tests were performed to assess mean differences in AUC between ML methods.
*Results:* AUC data were similar between ML (0.736 ± 0.116) and LR (0.748 ± 0.057) models ($p = 0.538$). However, specific ML models, such as gradient boosting (0.838 ± 0.077), exhibited superior performance at predicting AKI as compared to other ML models in the literature ($p < 0.05$). Creatinine and urine output, standard variables assessed for AKI staging, were classified as significant predictors across multiple ML models, although the majority of significant predictors were unique and study specific.
*Conclusions:* These data suggest that ML models perform equally to that of LR, however ML models exhibit variable performance with some ML models displaying exceptional performance. The variability in ML prediction of AKI can be attributed, in part, to the specific ML model utilized, variable selection and processing, study and subject characteristics, and the steps associated with model training, validation, testing, and calibration.

## 1. Introduction

The success of patient care in a clinical setting is dependent on the assessment and treatment of primary conditions as well as timely identification of sequelae that occur as a result of, or accompany, the original condition. Acute kidney injury (AKI) is associated with a loss in glomerular filtration rate (GFR) and can be a primary cause for hospitalization as well as secondary to other conditions, such as sepsis, surgeries, burns, and heart conditions [1,2]. AKI has a rapid onset with classification systems (Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease [RIFLE], Acute Kidney Injury Network [AKIN], Kidney Disease: Improving Global Outcomes [KDIGO]) utilizing changes in serum creatinine and urine output as indicators for GFR dysfunction and AKI progression (stage). Importantly, the development of AKI increases the risk of morbidity (e.g., chronic kidney disease [CKD]) as well as mortality, the latter including associations between AKI-induced inflammatory responses and distal organ dysfunction [3]. AKI treatment places significant demands on hospital resources, in part, due to a prolonged length of stay [4]. Similarly, average healthcare costs increase with the development of AKI; thus, it is of great interest to predict those patients who have the highest probability of developing AKI to improve patient health and resource management while lowering healthcare costs.

Machine learning (ML, artificial intelligence) has seen a rise in popularity in healthcare environments. For example, artificial neural networks have been utilized to learn from, and eventually identify,

*Abbreviations:* AKI, acute kidney injury; GFR, glomerular filtration rate; RIFLE, risk, injury, failure, loss of kidney function, and end-stage kidney disease; AKIN, acute kidney injury network; KDIGO, kidney disease: improving global outcomes; CKD, chronic kidney disease; ML, machine learning; LR, logistic regression; HER, electronic health records; AUC, area under the curve; SD, standard deviation; BUN, blood urea nitrogen.

* Corresponding author at: Department of Intensive Care Unit, Shandong Provincial Hospital Affiliated to Shandong First Medical University, #324 Jingwu Road, Jinan, Shandong Province, 250021, China.
*E-mail address:* wcteicu@126.com (C. Wang).

https://doi.org/10.1016/j.ijmedinf.2021.104484
Received 20 February 2021; Received in revised form 10 April 2021; Accepted 6 May 2021
Available online 8 May 2021

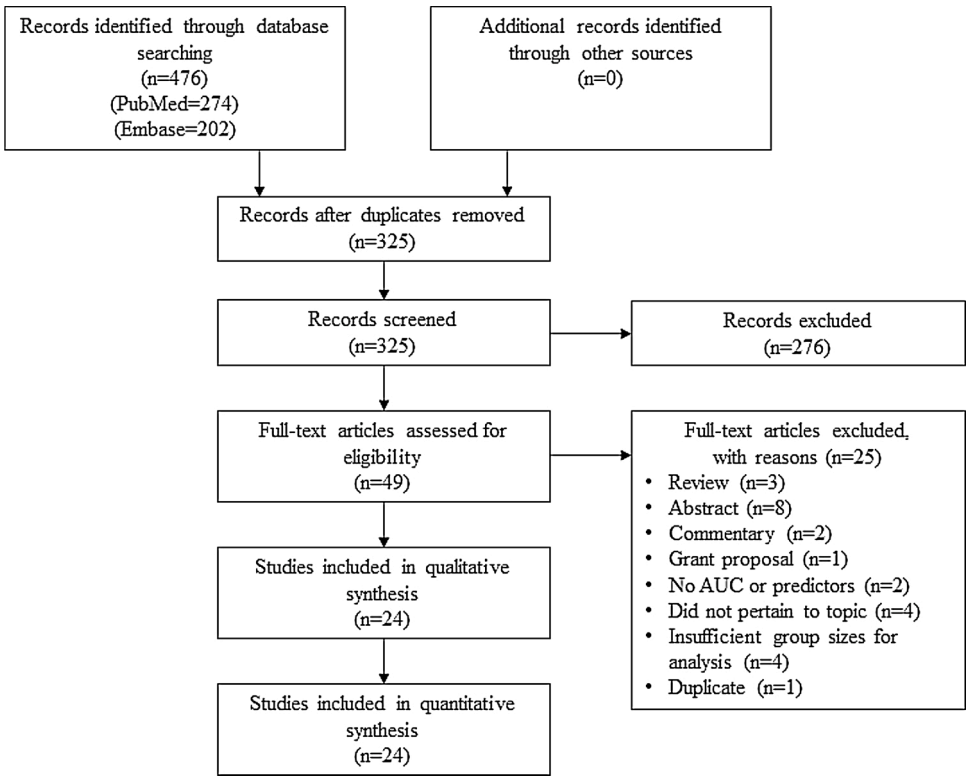


**Fig. 1.** PRISMA flow diagram of study selection.

patterns in digital images to enhance clinical diagnostic accuracy [5]. ML algorithms, such as extreme gradient boosting, have been implemented to improve prediction models for disease [6,7]. An important advantage of ML over conventional statistical methods (e.g., logistic regression [LR]) is that the various ML algorithms do not require data to conform to statistical assumptions, such as independence of observations and the avoidance of multicollinearity of independent variables. A second advantage lies in the abundance of data in the form of electronic health records (EHR) along with the computing power of modern-day computer systems. Collectively, these factors have made ML algorithms attractive options for predicting disease under real-time conditions. Similar to statistical methods, there are a large number of ML algorithms (random forest, Bayesian network, support vector machine, etc.) to choose from to develop a prediction model. However, it is unclear if one or more ML algorithms are superior to others in predicting disease.

Here, we performed a meta-analysis on the efficacy of ML algorithms to predict AKI. Moreover, we contrasted the performance of ML algorithms against conventional statistical methods such as LR to better characterize the overall performance of the models.

## 2. Methods

A systematic literature search was performed to identify studies investigating artificial intelligence used in predicting, diagnosing, or assessing acute kidney injury (AKI). A set of search terms including "acute kidney injury machine algorithm, ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("acute kidney injury" OR AKI), ("Acute Kidney Injury/diagnosis"[MAJR]) AND "Machine Learning"[MAJR], ("Acute Kidney Injury/diagnosis"[MAJR]) AND "Artificial intelligence"[MAJR], acute kidney artificial intelligence, acute kidney injury machine algorithm, acute kidney machine learning" were performed in PubMed. A single search term was searched in Embase: ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("acute kidney injury" OR AKI). Studies were included if they presented information and/or data related to AKI and artificial intelligence together. The following article types were excluded: review, editorial, correspondence, and abstract.

### 2.1. Data analysis

Data are expressed as mean ± standard deviation (SD). Due to the diversity of models under study, models were categorized by type (random forest, neural network, etc.) irrespective of variable selection procedures, number of predictor variables in the model, and number or characteristics (including treatments) of subjects in the study. Similarly, predictor variables were grouped for analysis; for example, measures (preoperative values, change over time, min/max) related to creatinine were grouped as creatinine. The majority of models (68/84, 82.9 %) were tested to predict AKI occurrence, while the remaining 17.1 % (14/82) were tested to predict specific AKI stages according to AKIN or KDIGO classification [8,9]. In the event of multiple AUC data from a single model and study (multiple tests of the same model from the same study), we prioritized the selection of predicting AKI occurrence data followed by predicting AKI stage 3, stage 2, or stage 1, respectively, in order to express optimal model performance. We did not select all AUC data from a single model and study as this could skew group data due to a high or low performing model that was run with great frequency. An independent samples *t*-test was performed to detect mean differences in AUC between ML and LR models. A one-way ANOVA was performed to assess mean differences in AUC between ML methods. Post-hoc *t*-tests were performed to identify ML models with superior performance. Statistical analyses were performed using IBM SPSS Statistics for Windows, v24 (IBM Corp., Armonk, NY). Significance was set at $p < 0.05$.

## 3. Results

A total of 476 studies were identified, and after auto-exclusion of duplicates, non-human studies, and non-English studies, a total of 325 studies remained. Of the screened studies, 276 were excluded and 49

**Table 1**
Study Characteristics.

| Author, year | N | Subject characteristics | Model type | Outcomes (top models/ predictors) |
|---|---|---|---|---|
| Cronin et al., 2015 [10] | 1,620,898 | Patient hospitalizations | LR, RF | Model performance was similar.<br>• BP meds<br>• NSAIDs<br>• Antibiotics<br>• IV fluids (48 h) |
| Kate et al., 2016 [11] | 25,521 | Hospital stays (>60 y.o.) | LR, BN, ENS, RF, SVM | ENS best at AKI prediction.<br>• Comorbidities<br>• History of AKI<br>• Heart/lung pathologies<br>• Diabetes<br>• Combination drugs |
| Thottakkara et al., 2016 [12] | 50,318 | Any major surgery | LR, BN, SVM | LR and SVM performed better than BN. |
| Davis et al., 2017 [13] | 170,675 | Any admission | LR, ANN, BN, RF | RF and ANN were superior.<br>• Antiemetics (@ adm)<br>• Vancomycin (@ adm)<br>• Age (@ adm)<br>• Mean GFR<br>• History of cancer and diabetes |
| Chen et al., 2018 [14] | 358 | Tertiary care (≥2 days) | ANN, ENS, DT, KNN, RF | ENS, DT, KNN, and RF performed similarly.<br>• Sennosides<br>• 1,2,6-hexanetriol<br>• Famotidine<br>• Benzimidazole |
| Cheng et al., 2018 [15] | 48,955 | Inpatient encounters | LR, AB, RF | RF performed best. Increase in prediction time to AKI onset reduces model performance.<br>• Medications<br>• Admission diagnosis<br>• Comorbidity |
| Huang et al., 2018 [16] | 947,091 | Percutaneous coronary intervention | LR, GB | Extreme GB performed better. |
| Koyner et al., 2018 [17] | 121,158 | Adult inpatients | GB | Extreme GB performance decreases with lower AKI stage (e.g., S1) prediction.<br>• Serum creatinine (change in)<br>• Length of hospital stay<br>• Serum creatinine (current)<br>• BUN (change in)<br>• Saturation (pO2/FiO2) |
| Lee et al., 2018 [18] | 2,010 | Heart surgery | LR, ANN, DT, GB, RF, SVM | GB exhibited superior performance.<br>• Packed RBCs (intraop)<br>• Hematocrit (preop)<br>• Creatinine (preop) |
| Lee et al., 2018 [19] | 1,211 | Liver transplant | LR, ANN, BN, DT, GB, RF, SVM | GB exhibited superior performance.<br>• Cold ischemic time<br>• Mean SvO2 (intraop)<br>• Mean cardiac index<br>• Urine output<br>• Glucose (preop) |
| Mohamadlou et al., 2018 [20] | 19,737 | Inpatient encounters | GB | Extreme GB outperformed SOFA for AKI prediction. |
| Park et al., 2018 [21] | 21,022 | Cancer | LR, RF | RF possessed greater F1, precision, and sensitivity for AKI prediction.<br>• Serum creatinine<br>• Inverse of serum creatinine |
| Adhikari et al., 2019 [22] | 2,911 | Inpatient operative procedure | RF | RF models with intraop data performed better than RF models with preop data only.<br>• Hypotension (intraop)<br>• Elevated HR (intraop) |
| Chiofolo et al., 2019 [23] | 6,530 | ICU - medical, surgical | RF | RF model performs well for continuous prediction of AKI in ICU population. |
| Flechet et al., 2019 [24] | 252 | ICU - critically ill | RF | RF performed similarly to physician in AKI prediction, less overestimation of AKI risk.<br>• Serum creatinine<br>• Urine output |
| He et al., 2019 [25] | 76,957 | Inpatient encounters | LR, BN, ENS, RF | ENS exhibited best performance. |
| Ibrahim et al., 2019 [26] | 889 | Coronary angiography | LR | ML identified markers contributed to good LR performance.<br>• History of diabetes<br>• BUN/ creatinine ratio |

(continued on next page)

**Table 1** (*continued*)

| Author, year | N | Subject characteristics | Model type | Outcomes (top models/ predictors) |
|---|---|---|---|---|
| | | | | • C-reactive protein<br>• Osteopontin<br>• CD5 antigen-like<br>• Factor VII |
| Parreco et al., 2019 [27] | 151,098 | ICU stays | LR, ANN, GB | GB possessed superior performance.<br>• Slope of minimum creatinine<br>• Creatinine (max, day 1) |
| Sun et al., 2019 [28] | 16,558 | ICU - critical care | LR, ANN, BN, RF, SVM | SVM and LR performed best<br>• Creatinine level (max)<br>• Mechanical ventilation<br>• International normalized ratio maximum<br>• Potassium level maximum<br>• Prothrombin time minimum |
| Tomasev et al., 2019 [29] | 703,782 | Inpatient, outpatient | ANN | Deep learning (ANN) exhibited 55.8 % sensitivity for AKI prediction (48 h), 2 false positives for 1 true positive |
| Tran et al., 2019 [30] | 50 | Burn | KNN | KNN is effective at AKI prediction in burn patients.<br>• Neutrophil gelatinase associated lipocalin<br>• N-terminal B-type natriuretic peptide<br>• Creatinine<br>• Urine output |
| Zhang et al., 2019 [31] | 6,682 | Low UO (0.5 ml/ kg/h) for 1 st 6 h after ICU admission | LR, GB | Extreme GB was superior to LR for volume responsive/ unresponsive AKI prediction<br>• Urinary creatinine<br>• BUN (max)<br>• Age<br>• Albumin (min)<br>• Body temperature (max) |
| Zimmerman et al., 2019 [32] | 23,950 | ICU - critical care inpatients | LR, ANN, RF | All 3 models performed well with ANN performing slightly better than LR and RF. |
| Zhou et al., 2020 [33] | 27 | Thoracoabdominal aorta aneurysm | LR, RF, SVM | RF and SVM (linear) performed best.<br>• Lactic acid (postop)<br>• BMI<br>• Infused RBCs<br>• Age<br>• Bleeding (surgery)<br>• Serum creatinine (preop) |

LR = logistic regression; AB = adaptive boosting; ANN = artificial neural network; BN = Bayesian network; DT = decision tree; ENS = ensemble; GB = gradient boosting; RF = random forest; SVM = support vector machine; BP = blood pressure; NSAID = non-steroidal anti-inflammatory drug; GFR = glomerular filtration rate; BUN = blood urea nitrogen; RBC = red blood cell; SvO2=mixed venous oxygen saturation; SOFA = Sequential Organ Failure Assessment; HR = heart rate; BMI = body mass index.

**Table 2**
Model Type.

| Model | Number of tests (% total) |
|---|---|
| Logistic regression (LR) | 21/82 (25.6 %) |
| LR | 14 |
| LASSO | 4 |
| LR, L-1 penalized | 1 |
| LR, L-2 penalized | 1 |
| LR, L-1/L-2 penalized | 1 |
| LR w/ stepwise variable selection | 1 |
| Random forest (RF) | 16/82 (19.5 %) |
| RF | 15 |
| RF, IDEA | 1 |
| RF, SMOTE | 1 |
| Artificial neural network (ANN) | 11/82 (13.4 %) |
| ANN | 3 |
| ANN, backpropagation | 1 |
| ANN, convolutional | 1 |
| ANN, deep belief network | 2 |
| ANN, deep learning | 1 |
| ANN, multilayer perceptron | 2 |
| ANN, recurrent | 1 |
| Support vector machine (SVM) | 9/82 (11.0 %) |
| SVM | 4 |
| SVM, gaussian | 1 |
| SVM, linear | 1 |
| SVM, classifier | 1 |
| SVM, least square | 1 |
| SVM, SMOTE | 1 |
| Gradient boosting (GB) | 7/82 (8.5 %) |
| GB | 4 |
| GB, extreme | 3 |
| Bayesian network (BN) | 7/82 (8.5 %) |
| BN, bayes net | 1 |
| BN, naive bayes | 6 |
| Decision tree (DT) | 5/82 (6.1 %) |
| DT | 3 |
| DT, CART | 1 |
| DT, ROSE | 1 |
| Ensemble | 3/82 (3.7 %) |
| K-nearest neighbor | 2/82 (2.4 %) |
| Adaptive boosting | 1/82 (1.2 %) |

Data are n/N (%). LASSO = Least Absolute Shrinkage and Selection Operator; IDEA = Intraoperative Data Embedded Analytics; SMOTE = Synthetic Minority Oversampling Technique; CART = Classification And Regression Tree; ROSE = computeRized prOlaps Syndrome dEtermination.

studies were selected for full-text review (PubMed 32, Embase 17). Studies classified as review (3), abstract (8), commentary (2), or grant proposal (1) were not investigated. Furthermore, studies (2) that failed to provide area under the curve (AUC) or significant predictor data were not included. Four studies were excluded as they did not pertain to the

**Table 3**
Model Output Variables.

| Variable | Number of tests |
|---|---|
| AUC | 79/82 (96.3 %) |
| Sensitivity (recall) | 27/82 (32.9 %) |
| Precision | 23/82 (28.0 %) |
| Accuracy | 20/82 (24.4 %) |
| F1 | 20/82 (24.4 %) |
| Error rate | 13/82 (15.9 %) |
| Specificity | 11/82 (13.4 %) |
| Positive predictive value (PPV) | 9/82 (11.0 %) |
| Brier score | 6/82 (7.3 %) |
| Negative predictive value (NPV) | 6/82 (7.3 %) |
| Net benefit | 2/82 (2.4 %) |
| Resolution | 2/82 (2.4 %) |
| Reliability (x10-2) | 2/82 (2.4 %) |
| Accuracy (across K values [%]) | 1/82 (1.2 %) |
| Negative likelihood ratio (NLR) | 1/82 (1.2 %) |
| Net reclassification improvement - w/ AKI | 1/82 (1.2 %) |
| Net reclassification improvement - w/o AKI | 1/82 (1.2 %) |
| Overall net reclassification improvement | 1/82 (1.2 %) |
| Positive likelihood ratio (PLR) | 1/82 (1.2 %) |

Data are n/N (%).
AUC = area under the curve; AKI = acute kidney injury.

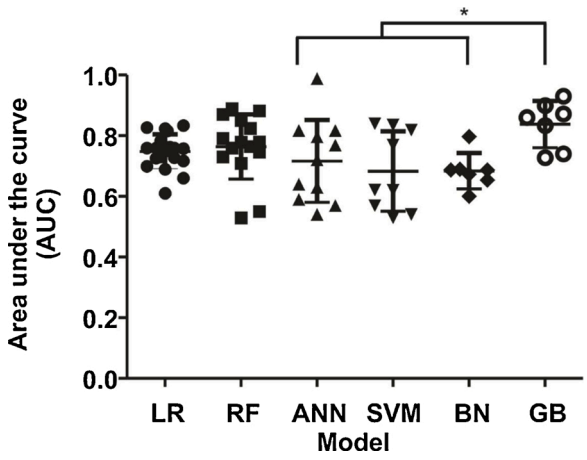


**Fig. 2. Scatter plot of model performance.** Gradient boosting (GB), consisting of both gradient boosting and extreme gradient boosting models, was significantly more effective at predicting AKI compared to all other ML models (artificial neural network, ANN; support vector machine, SVM; Bayesian network, BN; $p < 0.05$) with the exception of random forest (RF; $p = 0.152$).

topic of machine learning prediction or diagnosis of AKI or AKI-related mortality. Four studies, 2 related to ML diagnosis of AKI and 2 related to ML prediction of AKI-related mortality, were not analyzed due to insufficient group sizes. The remaining 24 studies were included in the present analysis (Fig. 1, Table 1) [10–33].

Among these 24 studies, there were a total of 82 models described, of which 25.6 % (21/82) were categorized as logistic regression with the remaining 74.4 % (61/82) categorized as ML (Table 2). An ML model was defined as any method that utilizes an algorithm to learn from the data in order to improve performance (i.e., AKI prediction) [34].

AUC data was provided in 96.3 % (79/82) of the models and were included in subsequent model performance comparisons. Table 3 contains a full list of variables used to describe model performance.

### 3.1. Model performance

To investigate model performance of LR and ML to predict AKI, mean AUC was calculated per group and analyzed. There was no difference in model performance between LR and ML as determined by AUC (LR: $0.748 \pm 0.057$, ML: $0.736 \pm 0.116$; $p = 0.538$). We next examined model performance of ML models to determine if one or more models exhibited greater performance. The following model groups had insufficient numbers for statistical comparison and were excluded from the analysis (AUC): decision tree ($0.698 \pm 0.136$), ensemble ($0.746 \pm 0.083$), k-nearest neighbor (0.824, only 1 study reported AUC), and adaptive boosting (0.715). Gradient boosting ($0.838 \pm 0.077$), consisting of both gradient boosting and extreme gradient boosting models, was significantly more effective at predicting AKI as compared to all other ML models studied ($p < 0.05$) with the exception of random forest ($0.764 \pm 0.106$, $p = 0.152$) (Fig. 2).

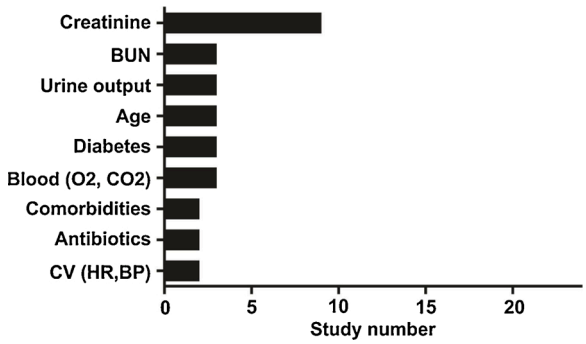


**Fig. 3. Frequency of the most common predictors for machine learning models.** Creatinine (blood or urine) was the most common significant predictor across the included studies.

### 3.2. AKI predictors

Data were examined further to determine variables that possessed the greatest predictive power in machine learning models. The number and type of predictive variables were diverse and included demographic, medical history, administrative/chart, laboratory (blood, urine, etc.), and surgical data. Seven studies (29.2 %) did not clearly indicate the most important predictors for their respective models. In cases where multiple models were performed, authors presented data on predictive variables for the best performing models. Creatinine (either blood or urine) was the most common, significant predictor observed appearing in 9 (37.5 %) studies (Fig. 3). Blood urea nitrogen (BUN), urine output, age, and diabetes were significant predictors, albeit to a lesser degree, appearing in 3 (12.5 %) studies. While creatinine was the most common predictor across studies, it represented only 16.9 % of total predictor variables (71) utilized in the studies. A list of the most common predictors for machine learning models is provided in Fig. 3.

## 4. Discussion

Here, we examined the performance of ML models to predict AKI and contrasted the performance of those ML models to LR, a conventional statistical approach used for clinical prediction models [35–37]. On average, ML models performed the same as LR models at predicting AKI with ML models displaying substantial variability across models. An in-group comparison of ML models revealed that gradient boosting, consisting of gradient boosting and extreme gradient boosting models, possessed superior performance at predicting AKI as compared to artificial neural network, support vector machine, and Bayesian network models. Variables related to creatinine were the most common, significant predictive variables described, while representing a relatively small amount of total significant predictive variables. Collectively, these data suggest that ML and LR models are effective at predicting AKI, however the specific type of model as well as variables included in the model weigh heavily on model performance.

Our finding that ML and LR models perform equally well to predict AKI was consistent with previous studies [36–38]. Christodoulou et al. found similar overall performance between LR and ML methods for

clinical risk prediction in a meta-analysis consisting of 71 studies, and they noted a high risk of bias in studies that reported better performance with ML [36]. Nusinovici et al. observed equivalent performance between LR and ML methods to predict risk of cardiovascular diseases, chronic kidney diseases, diabetes, and hypertension in an Asian cohort. LR, gradient boosting, and neural networks were ranked as the best performing models overall [38]. Lynam et al. investigated model performance of LR and 7 ML methods to discriminate between types I and II diabetes using a small set of 7 predictor variables [37]. There was no difference in type I and II diabetes classification performance between LR and ML methods. Taken together, the present results for AKI prediction along with results from the aforementioned studies suggest that LR and ML perform equally well to classify a variety of clinical conditions.

While ML and LR performed similarly overall in the present study, specific ML models demonstrated exceptional performance in predicting AKI. Gradient boosting, which in our study consisted of gradient boosting and extreme gradient boosting, performed better than all ML models studied with the exception of random forest. Although ML algorithms have inherent structural differences, the variables included in the model can drastically alter the quality of prediction. Creatinine and urine output, two variables that are utilized in AKI staging (KDIGO, AKIN), were noted to be significant predictors in multiple studies that utilized various ML models. Other variables related to kidney function such as BUN and diabetes were also noted in multiple studies. However, the majority of important predictors in the present study were unique possibly due to the nature of the subjects in the study (cardiac surgery, burn, etc.). For example, creatinine, while the most common variable observed in the models, represented a small proportion (16.9 %) of the total number of significant predictors listed. Therefore, other classes of variables, such as those related to cardiovascular and pulmonary function as well as certain medications (e.g., antibiotics, antiemetics), should be examined in more detail to determine their value to improve ML performance.

It is unclear if ML is being incorporated into the clinical setting for AKI prediction. The abundance of algorithms and variations in data type and handling can lead to a wide variety of predictive scores as evidenced by the present study. Moreover, significant challenges to the value of ML lie in the ability to predict AKI in a time frame that would be sufficient to initiate therapeutic interventions for AKI prevention. The work by Tomasev and his colleagues at DeepMind yielded the highest ML performance scores observed in the present study [29]. In this retrospective study, the artificial neural network was designed to predict disease 48 h prior to the onset of AKI. The model was able to predict 55.8 % of AKI cases at any stage with 2 false positives predicted for every 1 true positive. While measures related to creatinine were critical features of the model, the model could classify patients at a time point (48 h) prior to pathological changes in creatinine, which illustrates the potential value of such a model. Indeed, the value of ML for AKI prediction lies in the ability to handle large amounts of data and to identify seemingly benign relationships with those data as it relates to disease onset.

While not the highest performing model observed in the present study, the AKIpredictor is a public, online calculator (www.akipredictor.com) that utilizes serum creatinine as well as other patient information (age, diabetes, admission information) to calculate, upon admission to the ICU, a risk of developing AKI (stages 2 or 3) during the first week of an ICU stay [24,39]. There are other data (APACHE II score, bilirubin, maximum lactate, etc.) [24] that can be entered in on the first morning of ICU stay and after 24 h that can contribute to the model's performance. This model was prospectively evaluated and is easy to use, although it would require validation on a larger population before implementation in a clinical setting. The purpose of any ML model for AKI prediction in the clinic would be to serve as an adjunctive measure to perform real-time risk assessment of AKI and to act as an early warning system for clinicians to enhance patient care and resource management.

ML models vary in complexity from older Bayesian networks to elaborate artificial neural networks such as deep learning. In order for a ML model to successfully predict AKI, the model must be trained, validated, tested, and calibrated. All 15 studies that directly examined LR with other ML methods clearly described appropriate model training techniques (e.g., cross-validation methods), and there were no overt differences noted in the application of LR and ML methods. None of the included studies examined external validity of the models.

LR assumes that data conform to certain characteristics (e.g., independence of observations, low correlation between variables), of which violations can impact the quality of analysis. Similarly, some ML methods may perform better when applied to data that exhibit specific characteristics (e.g., large margin assumption in learning SVM models). The present results would suggest that potential violations of assumptions with respect to LR may not meaningfully impact the quality of performance as LR performed equally well to that of ML. In this case, LR models might be more advantageous than ML methods due to transparency and interpretability of LR.

AUC was the most often reported performance variable across models. However, there were 18 other performance variables that appeared in less than a third of models performed with over half of these appearing in less than 8 % of models performed. The wide variety of performance variables expressed precludes a thorough comparison of model performance aside from AUC. Future studies would add greater value by calculating other basic performance variables (e.g., accuracy, specificity, sensitivity, etc.) to allow for comprehensive model comparisons.

The limitations of the study stem from the heterogeneity of the included studies. We grouped gradient boosting and extreme gradient boosting as both models are fundamentally similar, and treated individually, could not be statistically evaluated due to insufficient numbers. Nevertheless, extreme gradient boosting is considered an enhanced model that utilizes a gradient boosting method, and it may be superior to gradient boosting in model performance for AKI prediction. More studies are needed to better characterize the performance of extreme gradient boosting algorithms to predict AKI. Support vector machine and Bayesian network models performed poorly (AUC: $<0.7$), although support vector machine data were highly variable. An important caveat to the results is that studies differed substantially in the quantity of data used to train the algorithm as well as the methods selected for data pre-processing. Thus, methods related to feature selection as well as the handling of missing or invalid data could have impacted model performance regardless of the type of model deployed. Moreover, potentially important variables may not always be collected and can depend on patient condition. A challenge moving forward is identifying meaningful variables for AKI prediction that are routinely collected. Similar to conventional statistical approaches, ML models are only as strong as the variables included in the model.

## 5. Conclusions

In conclusion, interest in ML for clinical prediction has increased due to widespread availability of both ML algorithms as well as EHR. ML performed similarly to LR in predicting AKI, while gradient boosting exhibited superior performance as compared to other ML models with the exception of random forest. Important variables associated with AKI staging such as creatinine and urine output were found to be important predictors for AKI across multiple studies. There remains substantial heterogeneity in variables included in the model, model set-up (training, validation, testing, calibration), and model application (specific condition; e.g., burn vs. general inpatient). Nevertheless, our data suggest that ML and LR are equally effective at predicting AKI.

## Ethics approval and consent to participate

Not applicable.

**What was already known on the topic**

- Logistic regression has been used to predict acute kidney injury in critical care settings.
- Interest in machine learning algorithms to predict acute kidney injury has grown; however, optimal algorithms and predictor variables are poorly understood.

**What this study added to our knowledge**

- Machine learning performance to predict acute kidney injury is variable and depends on the predictor variables included in the model as well as the type of algorithm deployed.
- While common biomarkers (creatinine, BUN) for acute kidney injury were important for model performance, there are a large number of predictor variables that have been identified and are currently being used in machine learning models.
- Machine learning performance is comparable to conventional logistic regression models in predicting acute kidney injury.

## Consent for publication

Not applicable.

## Availability of data and materials

All data generated or analyzed during this study are included in this published article.

## Funding

No external funding supported this study or preparation of the manuscript.

## Authors' contributions

CW had conception and design, XS, XL and FL performed the literature review, XS drafted the manuscript, XL performed statistical analysis. All authors critically revised the manuscript. All authors read and approved the final manuscript.

## Summary table

## Declaration of Competing Interest

The authors report no declarations of interest.

## Acknowledgement

None.

## References

[1] M.E. Pavkov, J.L. Harding, N.R. Burrows, Trends in hospitalizations for acute kidney injury - United States, 2000–2014, MMWR Morb. Mortal. Wkly. Rep. 67 (2018) 289–293.
[2] A. Clark, J.A. Neyra, T. Madni, J. Imran, H. Phelan, B. Arnoldo, S.E. Wolf, Acute kidney injury after burn, Burns 43 (2017) 898–908.
[3] D.P. Basile, M.D. Anderson, T.A. Sutton, Pathophysiology of acute kidney injury, Compr. Physiol. 2 (2012) 1303–1353.
[4] S.A. Silver, J. Long, Y. Zheng, G.M. Chertow, Cost of acute kidney injury in hospitalized patients, J. Hosp. Med. 12 (2017) 70–76.
[5] S.A. Varghese, T.B. Powell, M.G. Janech, M.N. Budisavljevic, R.C. Stanislaus, J. S. Almeida, J.M. Arthur, Identification of diagnostic urinary biomarkers for acute kidney injury, J. Investig. Med. 58 (2010) 612–620.
[6] X. Chen, L. Huang, D. Xie, Q. Zhao, EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction, Cell Death Dis. 9 (2018) 3.
[7] F. Commandeur, P.J. Slomka, M. Goeller, X. Chen, S. Cadet, A. Razipour, P. McElhinney, H. Gransar, S. Cantu, R.J.H. Miller, A. Rozanski, S. Achenbach, B. K. Tamarappoo, D.S. Berman, D. Dey, Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study, Cardiovasc. Res. (2019).
[8] J.A. Lopes, S. Jorge, The RIFLE and AKIN classifications for acute kidney injury: a critical and comprehensive review, Clin. Kidney J. 6 (2013) 8–14.
[9] A. Khwaja, KDIGO clinical practice guidelines for acute kidney injury, Nephron Clin. Pract. 120 (2012) c179–184.
[10] R.M. Cronin, J.P. VanHouten, E.D. Siew, S.K. Eden, S.D. Fihn, C.D. Nielson, J. F. Peterson, C.R. Baker, T.A. Ikizler, T. Speroff, M.E. Matheny, National Veterans Health Administration inpatient risk stratification models for hospital-acquired acute kidney injury, J. Am. Med. Inform. Assoc. 22 (2015) 1054–1071.
[11] R.J. Kate, R.M. Perez, D. Mazumdar, K.S. Pasupathy, V. Nilakantan, Prediction and detection models for acute kidney injury in hospitalized older adults, BMC Med. Inform. Decis. Mak. 16 (2016) 39.
[12] P. Thottakkara, T. Ozrazgat-Baslanti, B.B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, A. Bihorac, Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications, PLoS One 11 (2016), e0155705.
[13] S.E. Davis, T.A. Lasko, G. Chen, E.D. Siew, M.E. Matheny, Calibration drift in regression and machine learning models for acute kidney injury, J. Am. Med. Inform. Assoc. 24 (2017) 1052–1061.
[14] W. Chen, Y. Hu, X. Zhang, L. Wu, K. Liu, J. He, Z. Tang, X. Song, L.R. Waitman, M. Liu, Causal risk factor discovery for severe acute kidney injury using electronic health records, BMC Med. Inform. Decis. Mak. 18 (2018) 13.
[15] P. Cheng, L.R. Waitman, Y. Hu, M. Liu, Predicting inpatient acute kidney injury over different time horizons: how early and accurate? AMIA Annu. Symp. Proc. 2017 (2017) 565–574.
[16] C. Huang, K. Murugiah, S. Mahajan, S.X. Li, S.S. Dhruva, J.S. Haimovich, Y. Wang, W.L. Schulz, J.M. Testani, F.P. Wilson, C.I. Mena, F.A. Masoudi, J.S. Rumsfeld, J. A. Spertus, B.J. Mortazavi, H.M. Krumholz, Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: a retrospective cohort study, PLoS Med. 15 (2018), e1002703.
[17] J.L. Koyner, K.A. Carey, D.P. Edelson, M.M. Churpek, The development of a machine learning inpatient acute kidney injury prediction model, Crit. Care Med. 46 (2018) 1070–1077.
[18] H.C. Lee, H.K. Yoon, K. Nam, Y.J. Cho, T.K. Kim, W.H. Kim, J.H. Bahk, Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery, J. Clin. Med. 7 (2018).
[19] H.C. Lee, S.B. Yoon, S.M. Yang, W.H. Kim, H.G. Ryu, C.W. Jung, K.S. Suh, K.H. Lee, Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model, J. Clin. Med. 7 (2018).
[20] H. Mohamadlou, A. Lynn-Palevsky, C. Barton, U. Chettipally, L. Shieh, J. Calvert, N.R. Saber, R. Das, Prediction of acute kidney injury with a machine learning algorithm using electronic health record data, Can. J. Kidney Health Dis. 5 (2018), 2054358118776326.
[21] N. Park, E. Kang, M. Park, H. Lee, H.G. Kang, H.J. Yoon, U. Kang, Predicting acute kidney injury in cancer patients using heterogeneous and irregular data, PLoS One 13 (2018), e0199839.
[22] L. Adhikari, T. Ozrazgat-Baslanti, M. Ruppert, R. Madushani, S. Paliwal, H. Hashemighouchani, F. Zheng, M. Tao, J.M. Lopes, X. Li, P. Rashidi, A. Bihorac, Improved predictive models for acute kidney injury with IDEA: intraoperative data embedded analytics, PLoS One 14 (2019), e0214904.
[23] C. Chiofolo, N. Chbat, E. Ghosh, L. Eshelman, K. Kashani, Automated continuous acute kidney injury prediction and surveillance: a random forest model, Mayo Clin. Proc. 94 (2019) 783–792.
[24] M. Flechet, S. Falini, C. Bonetti, F. Guiza, M. Schetz, G. Van den Berghe, G. Meyfroidt, Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKI predictor, Crit. Care 23 (2019) 282.
[25] J. He, Y. Hu, X. Zhang, L. Wu, L.R. Waitman, M. Liu, Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records, JAMIA Open 2 (2019) 115–122.
[26] N.E. Ibrahim, C.P. McCarthy, S. Shrestha, H.K. Gaggin, R. Mukai, C.A. Magaret, R. F. Rhyne, J.L. Januzzi Jr., A clinical, proteomics, and artificial intelligence-driven

model to predict acute kidney injury in patients undergoing coronary angiography, Clin. Cardiol. 42 (2019) 292–298.
[27] J. Parreco, H. Soe-Lin, J.J. Parks, S. Byerly, M. Chatoor, J.L. Buicko, N. Namias, R. Rattan, Comparing machine learning algorithms for predicting acute kidney injury, Am. Surg. 85 (2019) 725–729.
[28] M. Sun, J. Baron, A. Dighe, P. Szolovits, R.G. Wunderink, T. Isakova, Y. Luo, Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements, Stud. Health Technol. Inform. 264 (2019) 368–372.
[29] N. Tomasev, X. Glorot, J.W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, A. Connell, C.O. Hughes, A. Karthikesalingam, J. Cornebise, H. Montgomery, G. Rees, C. Laing, C.R. Baker, K. Peterson, R. Reeves, D. Hassabis, D. King, M. Suleyman, T. Back, C. Nielson, J.R. Ledsam, S. Mohamed, A clinically applicable approach to continuous prediction of future acute kidney injury, Nature 572 (2019) 116–119.
[30] N.K. Tran, S. Sen, T.L. Palmieri, K. Lima, S. Falwell, J. Wajda, H.H. Rashidi, Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: a proof of concept, Burns 45 (2019) 1350–1358.
[31] Z. Zhang, K.M. Ho, Y. Hong, Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care, Crit. Care 23 (2019) 112.
[32] L.P. Zimmerman, P.A. Reyfman, A.D.R. Smith, Z. Zeng, A. Kho, L.N. Sanchez-Pinto, Y. Luo, Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements, BMC Med. Inform. Decis. Mak. 19 (2019) 16.
[33] C. Zhou, R. Wang, W. Jiang, J. Zhu, Y. Liu, J. Zheng, X. Wang, W. Shang, L. Sun, Machine learning for the prediction of acute kidney injury and paraplegia after thoracoabdominal aortic aneurysm repair, J. Card. Surg. 35 (2020) 89–99.
[34] L. Breiman, Statistical modeling: the two cultures, Stat. Sci. 16 (2001) 199–231.
[35] M.E. Shipe, S.A. Deppen, F. Farjah, E.L. Grogan, Developing prediction models for clinical use using logistic regression: an overview, J. Thorac. Dis. 11 (2019) S574–S584.
[36] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. Van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019) 12–22.
[37] A.L. Lynam, J.M. Dennis, K.R. Owen, R.A. Oram, A.G. Jones, B.M. Shields, L. A. Ferrat, Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults, Diagn. Progn. Res. 4 (2020) 6.
[38] S. Nusinovici, Y.C. Tham, M.Y. Chak Yan, D.S. Wei Ting, J. Li, C. Sabanayagam, T. Y. Wong, C.Y. Cheng, Logistic regression was as good as machine learning for predicting major chronic diseases, J. Clin. Epidemiol. 122 (2020) 56–69.
[39] M. Flechet, F. Guiza, M. Schetz, P. Wouters, I. Vanhorebeek, I. Derese, J. Gunst, I. Spriet, M. Casaer, G. Van den Berghe, G. Meyfroidt, AKIpredictor, 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. 43 (2017) 764–773.