Canadian Society of Nephrology/
Société canadienne de néphrologie
CSN/SCN

CANADIAN JOURNAL OF
KIDNEY HEALTH AND DISEASE
Journal canadien de la santé et de la maladie rénale

Original Research Article

Canadian Journal of Kidney Health and Disease
Volume 5: 1–9

DOI: 10.1177/2054358118776326
journals.sagepub.com/home/cjk
SAGE

Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data

Hamid Mohamadlou[1], Anna Lynn-Palevsky[1], Christopher Barton[2], Uli Chettipally[2,3], Lisa Shieh[4], Jacob Calvert[1], Nicholas R. Saber[1], and Ritankar Das[1]

Abstract

Background: A major problem in treating acute kidney injury (AKI) is that clinical criteria for recognition are markers of established kidney damage or impaired function; treatment before such damage manifests is desirable. Clinicians could intervene during what may be a crucial stage for preventing permanent kidney injury if patients with incipient AKI and those at high risk of developing AKI could be identified.

Objective: In this study, we evaluate a machine learning algorithm for early detection and prediction of AKI.

Design: We used a machine learning technique, boosted ensembles of decision trees, to train an AKI prediction tool on retrospective data taken from more than 300 000 inpatient encounters.

Setting: Data were collected from inpatient wards at Stanford Medical Center and intensive care unit patients at Beth Israel Deaconess Medical Center.

Patients: Patients older than the age of 18 whose hospital stays lasted between 5 and 1000 hours and who had at least one documented measurement of heart rate, respiratory rate, temperature, serum creatinine (SCr), and Glasgow Coma Scale (GCS).

Measurements: We tested the algorithm's ability to detect AKI at onset and to predict AKI 12, 24, 48, and 72 hours before onset.

Methods: We tested AKI detection and prediction using the National Health Service (NHS) England AKI Algorithm as a gold standard. We additionally tested the algorithm's ability to detect AKI as defined by the Kidney Disease: Improving Global Outcomes (KDIGO) guidelines. We compared the algorithm's 3-fold cross-validation performance to the Sequential Organ Failure Assessment (SOFA) score for AKI identification in terms of area under the receiver operating characteristic (AUROC).

Results: The algorithm demonstrated high AUROC for detecting and predicting NHS-defined AKI at all tested time points. The algorithm achieves AUROC of 0.872 (95% confidence interval [CI], 0.867-0.878) for AKI detection at time of onset. For prediction 12 hours before onset, the algorithm achieves an AUROC of 0.800 (95% CI, 0.792-0.809). For 24-hour predictions, the algorithm achieves AUROC of 0.795 (95% CI, 0.785-0.804). For 48-hour and 72-hour predictions, the algorithm achieves AUROC values of 0.761 (95% CI, 0.753-0.768) and 0.728 (95% CI, 0.719-0.737), respectively.

Limitations: Because of the retrospective nature of this study, we cannot draw any conclusions about the impact the algorithm's predictions will have on patient outcomes in a clinical setting.

Conclusions: The results of these experiments suggest that a machine learning–based AKI prediction tool may offer important prognostic capabilities for determining which patients are likely to suffer AKI, potentially allowing clinicians to intervene before kidney damage manifests.

Abrégé

Contexte: Une des principales difficultés liées au traitement de l'insuffisance rénale aiguë (IRA) est le fait que les critères cliniques diagnostiques sont des marqueurs d'une lésion ou d'une dysfonction rénale déjà établie. Il est souhaitable d'intervenir avant une telle issue. En dépistant les patients à risque d'IRA ou atteints d'IRA débutante, les cliniciens seraient en mesure d'intervenir précocement et ainsi prévenir les lésions rénales permanentes.

Objectif de l'étude: L'étude visait à évaluer un algorithme d'apprentissage automatique destiné à la prédiction des cas d'IRA et à sa détection précoce.

Type d'étude: Nous avons employé une technique d'apprentissage automatique, soit des ensembles d'arbres décisionnels amplifiés, pour entrainer un outil de prédiction de l'IRA à partir de données rétrospectives provenant de plus de 300 000 consultations auprès de patients hospitalisés.

Cadre de l'étude: Les données ont été colligées à partir des dossiers des unités d'hospitalisation du centre médical de l'université Stanford et de l'unité des soins intensifs du centre médical Beth Israel Deaconess.
Participants: Ont été inclus dans l'étude tous les patients adultes dont l'hospitalisation avait duré de 5 à 1 000 heures et pour lesquels on disposait d'au moins une mesure parmi les suivantes : pouls, rythme respiratoire, température corporelle, taux de créatinine sérique (SCr) et score de Glasgow.
Mesures: Nous avons testé l'efficacité de l'algorithme à détecter l'IRA dès son apparition, et à la prédire 12, 24, 48 et 72 heures avant qu'elle ne se manifeste.
Méthodologie: L'algorithme du NHS England a servi de référence pour tester l'efficacité de notre algorithme de prédiction et de détection de l'IRA. Nous avons également testé l'efficacité de notre algorithme à détecter l'IRA telle que définie par les *Recommandations de Bonnes Pratiques Cliniques* du KDIGO (*Kidney Disease: Improving Global Outcomes*). Nous avons utilisé la surface sous la courbe ROC (*Receiver Operating Characteristic*) pour comparer le score SOFA à l'efficacité de validation croisée tripartite de notre algorithme.
Résultats: L'algorithme a démontré une SSROC (surface sous la courbe ROC) élevée pour la détection et la prédiction de l'IRA (telle que définie par le NHS) pour tous les moments testés. En détection de la maladie à son apparition, l'algorithme a obtenu une SSROC de 0,872 (IC 95 % : 0,867-0,878). En prédiction, l'algorithme a obtenu une SSROC de 0,800 (IC 95 % : 0,792-0,809) à 12 heures, de 0,795 à 24 heures (IC 95 % : 0,785-0,804), de 0,761 (IC 95 % : 0,753-0,768) à 48 heures et de 0,728 (IC 95 % : 0,719-0,737) à 72 heures avant l'apparition des premiers symptômes.
Limites de l'étude: La nature rétrospective de l'étude ne nous permet pas de tirer de conclusions sur les conséquences qu'auront les prédictions de l'algorithme sur les résultats cliniques des patients.
Conclusion: Les résultats de nos essais laissent supposer qu'un outil de prédiction de l'IRA fondé sur l'apprentissage automatique pourrait offrir d'importantes fonctions pronostiques pour détecter les patients susceptibles de développer une IRA en vue d'une intervention précoce.

Keywords
acute kidney injury, machine learning

Received October 11, 2017. Accepted for publication March 28, 2018.

What was known before

Early detection of acute kidney injury (AKI) is necessary for clinicians to intervene in early stages of disease progression and prevent kidney damage. However, AKI can be difficult to detect before kidney damage and impaired function are present.

What this adds

The machine learning algorithm described in this study is capable of predicting an AKI up to 72 hours before onset, allowing for early clinical intervention.

Introduction

Acute kidney injury (AKI) is common, affecting 5% to 7% of all hospitalizations and causing $10 billion of additional health care–related expenditures per year through per-hospitalization excess costs of $7933.[1-3] Acute kidney injury is associated with increased mortality, end-stage renal disease, and chronic kidney disease, which can require ongoing dialysis and kidney replacement.[4-6] There exists some controversy as to how to best treat patients experiencing AKI. Standard approaches include reducing or eliminating nephrotoxic and antibiotic medications, relieving possible obstruction, and correcting electrolyte and fluid imbalances.[7,8] However, the effectiveness of these interventions may be limited by an inability to consistently identify patients with active or incipient AKI.[9] A system which identifies incipient AKI or predicts clinical manifestations of AKI with a substantial lead time may enable clinicians to better assess existing and novel interventions, and to ultimately provide more effective therapy which mitigates or avoids AKI and long-term kidney damage.

It has been recognized that early identification of AKI is desirable in hospital settings and that even small increases in serum creatinine (SCr) levels are associated with long-term

[1]Dascena, Inc, Hayward, CA, USA
[2]Department of Emergency Medicine, University of California, San Francisco, USA
[3]Kaiser Permanente South San Francisco Medical Center, CA, USA
[4]Department of Medicine, Stanford University School of Medicine, CA, USA

Corresponding Author:
Anna Lynn-Palevsky, Dascena, Inc, 22710 Foothill Boulevard, Suite #2, Hayward, CA 94541, USA.
Email: anna@dascena.com

damage and increased mortality.[2,10] Furthermore, accurate prediction of AKI onset before patients meet clinical criteria for recognition is advantageous, as such current clinical criteria represent markers of established kidney damage or impaired function.[11,12] Electronic health records present an opportunity to utilize machine learning techniques for predicting AKI and sending automated alerts for individual patients at risk of developing AKI. Several studies have assessed clinical decision support (CDS) tools for early detection of AKI, but many of these tools suffer from a variety of design and performance problems. These issues include lack of predictive ability, lack of an e-alert implementation, heavy tradeoffs between sensitivity and specificity, and restrictions to limited patient populations such as intensive care unit (ICU), postcardiac surgical, or elderly patients.[13-16]

In this paper, we describe an approach based on a machine learning algorithm (MLA)—a procedure which, in this case, identifies the statistical patterns in electronic health record data corresponding to AKI-related outcomes and the result of which is a software-based prediction tool intended to provide significant, accurate advance warning of AKI. Machine learning methods may provide advantages for AKI detection, as they can be trained to predict AKI far in advance of onset, can maintain concurrently high sensitivity and specificity, and can be customized to specific populations for increased accuracy. The machine learning method assessed in this study was that of gradient boosted trees, a method that iteratively combines the results of multiple decision trees into an overall risk prediction score.

The machine learning method assessed in this study was that of gradient boosted trees, a method that iteratively combines the results of multiple decision trees into an overall risk prediction score. The objective of this study was to assess the performance of the MLA for detecting AKI onset and predicting an impending AKI 12, 24, 48, and 72 hours before onset. In addition, we compared the performance of this prediction tool to the Sequential Organ Failure Assessment (SOFA) score.[17] The SOFA score is a commonly used disease severity scoring system which was developed to assess organ function in hospitalized patients. The SOFA score is tabulated from subscores for each of respiratory, coagulation, liver, cardiovascular, renal, and neurological systems. In past work, the SOFA score has been shown to independently predict AKI risk and outcomes, and thus serves as an important comparator for our approach.[12,18]

Materials and Methods

Data Sources

Data used in this study were drawn from the 651-bed Beth Israel Deaconess Medical Center (BIDMC; Boston, Massachusetts) and from the 613-bed Stanford University Medical Center (Stanford, California). BIDMC data were collected from the Medical Information Mart for Intensive Care III (MIMIC-III) v1.3 database.[19] This database was compiled by the MIT Laboratory for computational physiology and contains 61 532 inpatient ICU encounters collected between 2001 and 2012. The Stanford University data set contains 286 797 inpatient encounters from all hospital wards between December 2008 and May 2017.

For both data sets, we included only those patients whose hospital stays lasted between 5 and 1000 hours. Of those patients, we included only those who had at least one measurement of each required measurement (see "Imputation and Feature Creation" section), and who had age data available and were older than 18 years of age. Patients with chronic kidney disease were not excluded from the final patient population. The inclusion flowchart is presented in Figure 1.

Data collection for both data sets was passive and had no impact on patient safety. Both data sets were deidentified in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Studies performed on deidentified data constitute nonhuman subject research, thus no institutional or ethical approvals were required for this study.

Statistical Analysis

Data Processing

All data from both data sets were processed by custom database queries. The retrieved data were converted into flat .csv files, which were in turn loaded into a custom data processing code written in the programming language Python (Python Software Foundation, https://www.python.org/). This code associated each measurement or observation with a timestamp and a measurement type key. Demographics and other patient characteristics (eg, age) were stored with a similar keyed retrieval mechanism.

Imputation and Feature Creation

Beginning at the time of the first recorded patient measurement, all data were discretized into 1-hour intervals. If multiple observations of the same patient measurement were taken within a given hour, those measurements were averaged to produce a single value for that hour. This ensured that the rate at which measurements were fed into the algorithm was standardized across patients. If no measurement of a clinical variable was available for a given hour, a carry-forward imputation method was employed to fill the missing measurement with the most recently available previous measurement, a causal procedure. Details of these data processing methods have been described in a previous publication on the use of gradient boosted trees for sepsis detection and prediction.[20]

For all experiments, we generated MLA predictions using patient data on heart rate, respiratory rate, temperature, SCr,

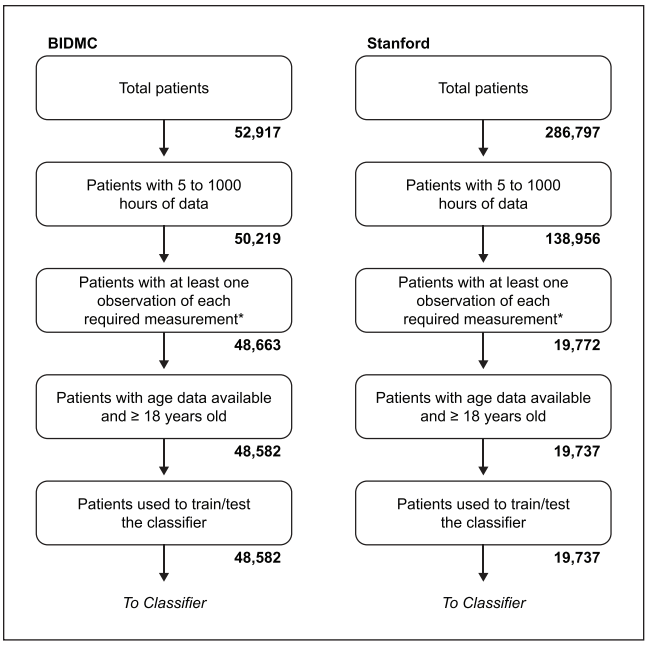

Figure 1. Inclusion criteria for patients in the BIDMC and Stanford data sets.
Note. Patients who met all inclusion criteria were included in this study. BIDMC = Beth Israel Deaconess Medical Center.
[a]Required measurements include heart rate, respiratory rate, temperature, Glasgow Coma Scale, and serum creatinine.

and Glasgow Coma Scale (GCS). These measurements were selected because they are frequently and easily collected at the bedsides, even before clinical suspicion of AKI is present. After imputation and averaging, for each prediction time, we took our causal feature vector to include the previous 5 hourly values of each of heart rate, respiratory rate, temperature, SCr, and GCS as well as the patient's age.

Gold Standard

We implemented the National Health Service (NHS) England AKI Algorithm as our gold standard.[21] This system is based on Kidney Disease: Improving Global Outcomes (KDIGO) guidelines,[22] but relies exclusively on changes in SCr levels to determine the presence and staging of AKI. The NHS algorithm is an appropriate gold standard for this work because it was designed explicitly for early AKI detection and generation of e-alerts for affected patients, and because it does not rely on urine output, which has been shown to be a poorer indicator of AKI than SCr and is subject to poor documentation, particularly in the emergency department.[23,24]

We determined the presence of AKI for adult inpatients only. Using either the lowest value from the past 0 to 7 days or the median value from the past 8 to 365 days as a baseline reference value, the ratio of current SCr levels to the reference value was calculated as in the NHS Algorithm (Supplemental Table 1). We computed these ratios using SCr measurements from the past 0 to 7 days whenever these data were available, using measurements from the past 8 to 365 days in all other cases. We determined the MLA's ability to predict stage 2 or stage 3 AKI at 0, 12, 24, 48, and 72 hours before onset.

We additionally assessed the algorithm's ability to detect stage 2 or stage 3 AKI as determined by the KDIGO criteria (Supplemental Table 2). For these experiments, we included all patients who had at least 1 recorded observation of the 5 required measurements, plus at least 1 observation of urine output (Supplemental Tables 2 and 3). The KDIGO criteria require a premorbid SCr level to be used as a baseline; for patients for whom no premorbid SCr measurements were available, a baseline SCr was estimated using the Modification of Diet in Renal Disease (MDRD) equation.[25] We tested the MLA's ability to detect KDIGO stage 2 or stage 3 AKI at 0, 12, and 24 hours before onset.

Machine Learning and Experimental Methods

All predictors trained in this work are boosted ensembles of decision trees produced using the XGBoost package for Python.[26] The boosting process improves predictions by successively adding new decision trees to a growing ensemble of trees, where new trees are trained to perform better on those patients who are misclassified by the current ensemble.

We used 3-fold cross-validation to assess the performance of the algorithm under the NHS gold standard separately on the BIDMC and Stanford data sets, in which we divided each data set into thirds, trained a predictor on two of the thirds, and tested the trained predictor on the remaining third. This process was repeated so that training occurred on all possible combinations of thirds; thus, all included patients were cycled through during the training process. The data sets were randomly divided, with randomization based on patient identification number. We measured area under the receiver operating characteristic (AUROC), accuracy, diagnostic odds ratio (DOR), and positive and negative likelihood ratios (LR+ and LR−) obtained by the MLA via this method. Our reported metrics are the average metrics of 30 independently trained models, each trained using 3-fold cross-validation. Randomization was performed before training of each model.

We compared these results with those same measures obtained by the SOFA organ dysfunction score[17] on both data sets. The SOFA score was calculated as in study of Pandharipande et al,[27] with SpO_2/FiO_2 ratios used in place of PaO_2/FiO_2 ratios due to data availability. Statistical comparisons were performed using pairwise, single-tailed *t* tests with significance set at $P < .01$.

To assess the algorithm under the KDIGO criteria, we used 10-fold cross-validation to separately train the algorithm on the BIDMC and Stanford data sets. We then assessed algorithm performance on test sets from the BIDMC data set.

Reported performance metrics are the average of each of the 10 models generated by the 10-fold cross-validation process.

Results

Participants

The final patient population used to train and test the algorithm included 48 582 patients from the BIDMC data set and 19 737 patients from the Stanford data set. Patient demographics differed in several important ways between the two data sets (Table 1). The BIDMC data set contains only patients admitted to the ICU, while the Stanford data set contains all inpatients; BIDMC patients therefore represent a more critically ill population. In addition, the data sets display differences in age and gender. The Stanford data set skewed younger than the BIDMC data set, with around 15% of Stanford patients in the 18- to 29-year-old group and only around 4.5% of BIDMC patients in this group. Around 41% of BIDMC patients were older than the age of 70, while only around 14% of Stanford patients fell into this age group. The BIDMC data set also skewed more heavily male than the Stanford data set, with more than 56% of BIDMC patients male. Around 49% of Stanford patients were male. Prevalence of AKI was higher in the BIDMC than in the Stanford data set.

Table 1. Patient Demographic Information for Complete BIDMC and Stanford Cohorts.

Characteristic	BIDMC (%)	Stanford (%)
Gender		
Female	43.66	51.19
Male	56.44	48.81
Age (years)		
18-29	4.51	15.23
30-39	5.26	11.22
40-49	10.64	11.22
50-59	17.50	13.20
60-69	20.98	12.69
70+	40.91	14.07
Severe AKI based on NHS England algorithm[a]		
Yes	2.7%	0.5%
No	97.3%	99.5%
In-hospital death		
Yes	9.2%	2.78%
No	90.8%	97.22%

Note. BIDMC = Beth Israel Deaconess Medical Center; AKI = acute kidney injury; NHS = National Health Service.
[a]Prevalence of stage 2 or stage 3 AKI before filtering patients according to inclusion criteria.

Main Results

For detecting severe AKI under the NHS gold standard, the MLA demonstrated higher AUROC, accuracy, and DOR than the SOFA score at all prediction windows and for each data set. When tested on data collected from BIDMC, the MLA demonstrated an AUROC of 0.841 at time of onset while the SOFA score achieved an AUROC of 0.762 at time of onset. The MLA AUROC improved upon that of SOFA for all prediction windows ($P < .01$ for all windows). In addition, MLA accuracy and DOR remained superior for all prediction windows (12, 24, 48, and 72 hours prior to onset) (Tables 2 and 3). The algorithm had higher or comparable positive likelihood ratios (LR+) and comparable negative likelihood ratios (LR−) at onset and for all prediction windows. Full performance metrics for the MLA and SOFA when tested on BIDMC data are presented in Table 2.

The algorithm also demonstrated superior performance when trained and tested for NHS gold standard AKI on patient data from Stanford Medical Center. At time of onset, the MLA demonstrated an AUROC of 0.872, while the SOFA score demonstrated an AUROC of 0.815. As on BIDMC data, the MLA AUROC exceeded that of the SOFA score for all prediction windows ($P < .01$ for all windows). MLA accuracy was higher than that of the SOFA score for all prediction windows (Table 3). The MLA also demonstrated improved DOR compared with the SOFA score. Full performance measures for the algorithm and SOFA when tested on Stanford data are presented in Table 3.

We note that the performance metrics in Tables 2 and 3 were measured for prediction sensitivities held near 0.80, to facilitate comparison of metrics across prediction times. The MLA performance across all such operating points (ie, choices of sensitivity) is summarized in Figures 2 and 3. Figure 2 provides a receiver operating characteristic (ROC) curve comparison across prediction times for the MLA trained and tested on BIDMC data, and algorithm performance on Stanford data is displayed in Figure 3. On both the BIDMC and Stanford data sets, MLA performance declined gradually as the prediction window was lengthened from 0 hours to 72 hours before AKI onset.

The algorithm also demonstrated high AUROC, sensitivity, and specificity for detecting stage 2 or stage 3 AKI under the KDIGO criteria (Table 4). When trained on patient data from Stanford and tested on data from BIDMC, the algorithm demonstrated AUROC above 0.75 for AKI detection up to 24 hours in advance of onset.

Discussion

The machine learning approach described here results in a prediction tool which demonstrates strong predictive performance, in terms of AUROC, up to 72 hours in advance of stage 2 or stage 3 AKI onset, under both the NHS and KDIGO criteria for AKI. Furthermore, for a given degree of sensitivity, the MLA outperforms the commonly used SOFA score in terms of specificity, accuracy, and other metrics. This performance was achieved using only 5 commonly collected patient measurements as inputs. By requiring the presence of only these 5

Table 2. Comparison of Performance Metrics for the MLA and for the SOFA Score Measured on Patient Data From Beth Israel Deaconess Medical Center.

Prediction time	Onset		12 hours		24 hours		48 hours		72 hours	
Predictor	MLA	SOFA	MLA	SOFA	MLA	SOFA	MLA	SOFA	MLA	SOFA
AUROC (95% CI)	0.841 (0.837-0.844)	0.762	0.749 (0.744-0.755)	0.734	0.758 (0.754-0.762)	0.716	0.707 (0.701-0.713)	0.675	0.674 (0.669-0.679)	0.653
Sensitivity	0.81	0.55	0.77	0.54	0.83	0.78	0.83	0.84	0.82	0.82
Specificity	0.75	0.79	0.62	0.78	0.56	0.57	0.48	0.41	0.45	0.39
Accuracy	0.81	0.57	0.76	0.55	0.82	0.76	0.82	0.81	0.80	0.79
DOR	13.1	4.8	5.5	4.2	6.2	4.7	4.5	3.6	3.7	3.0
LR+	3.3	2.7	2.0	2.5	1.9	1.8	1.6	1.4	1.5	1.3
LR−	0.25	0.56	0.37	0.59	0.30	0.39	0.35	0.39	0.40	0.46

Note. Predictions were made at 0, 12, 24, 48, and 72 hours before stage 2 or stage 3 AKI onset. Operating points for the MLA were chosen to keep sensitivities close to 0.80. 95% CIs were calculated only for the MLA. MLA = machine learning algorithm; SOFA = sequential organ failure assessment; AUROC = area under the receiver operating characteristic; CI = confidence interval; DOR = diagnostic odds ratio; LR = likelihood ratios; AKI = acute kidney injury.

Table 3. Comparison of Performance Metrics for the MLA and for the SOFA Score Measured on Patient Data From Stanford Medical Center.

Prediction time	Onset		12 hours		24 hours		48 hours		72 hours	
Predictor	MLA	SOFA	MLA	SOFA	MLA	SOFA	MLA	SOFA	MLA	SOFA
AUROC (95% CI)	0.872 (0.867-0.878)	0.815	0.800 (0.792-0.809)	0.781	0.795 (0.785-0.804)	0.764	0.761 (0.753-0.768)	0.732	0.728 (0.719-0.737)	0.720
Sensitivity	0.77	0.73	0.75	0.73	0.79	0.55	0.85	0.53	0.78	0.51
Specificity	0.82	0.78	0.73	0.74	0.64	0.83	0.51	0.79	0.53	0.81
Accuracy	0.78	0.73	0.75	0.73	0.79	0.56	0.84	0.54	0.79	0.53
DOR	15.5	9.7	8.0	7.3	6.9	5.9	5.8	4.3	4.4	4.3
LR+	4.3	3.4	2.7	2.7	2.2	3.2	1.7	2.6	1.7	2.7
LR−	0.28	0.35	0.34	0.37	0.32	0.55	0.30	0.60	0.38	0.61

Note. Predictions were made at 0, 12, 24, 48, and 72 hours before stage 2 or stage 3 AKI onset. Operating points were chosen to keep sensitivities close to 0.80. 95% CIs were calculated only for the MLA. MLA = machine learning algorithm; SOFA = sequential organ failure assessment; AUROC = area under the receiver operating characteristic; CI = confidence interval; DOR = diagnostic odds ratio; LR = likelihood ratio; AKI = acute kidney injury.

measurements for AKI predictions, this algorithm is designed to be able to predict AKI risk on a large portion of the hospital population in future clinical work. Based on these results, we believe this MLA could provide clinicians the opportunity to improve patient outcomes through earlier AKI detection and subsequent intervention, which may include volume resuscitation or avoidance of nephrotoxic medications to minimize further kidney injury.[28]

We emphasize that the MLA performed similarly well on the BIDMC and Stanford data sets, an observation which has important clinical implications. The BIDMC data included only patients admitted to the ICU, while the Stanford data set contained information about inpatient stays from all hospital wards. These two data sets thus represent hospital settings with different demographics, frequency of patient measurement collection, levels of care provision, and disease severity in patients. The predictive ability of the algorithm across these data sets suggests that the algorithm can identify patients at risk of AKI onset in a variety of hospital settings. Because AKI is a common complication of hospital stays of a diverse nature,[1] this ability is of central importance in an AKI prediction tool.

In previous studies, we assessed a MLA for sepsis detection in both retrospective[29-31] and prospective settings.[32,33] Our prospective work includes a randomized controlled trial (RCT) conducted in 2 adult ICUs at the University of California, San Francisco. In this study, there was a statistically significant decrease in the primary endpoint of mean hospital length of stay (13.0 days vs 10.3 days, $P = .0421$) and in the secondary endpoint of in-hospital mortality rate (21.3% vs 8.96%, $P = .0176$).[32] This study demonstrates the feasibility of bedside use of a machine learning–based prediction system, as well as the potential for such a system to improve patient outcomes through earlier and more accurate detection of patient conditions. Future clinical implementation of the system described in this work is intended not to disrupt clinical workflow or require additional work from care providers. As with past implementations of our sepsis detection system, more frequent collection of patient measurements is not required; the same forward-filling imputation method described in this paper can be used to make accurate predictions in clinical settings.

Machine learning methods have previously been applied to AKI detection. One such study used a random forest model

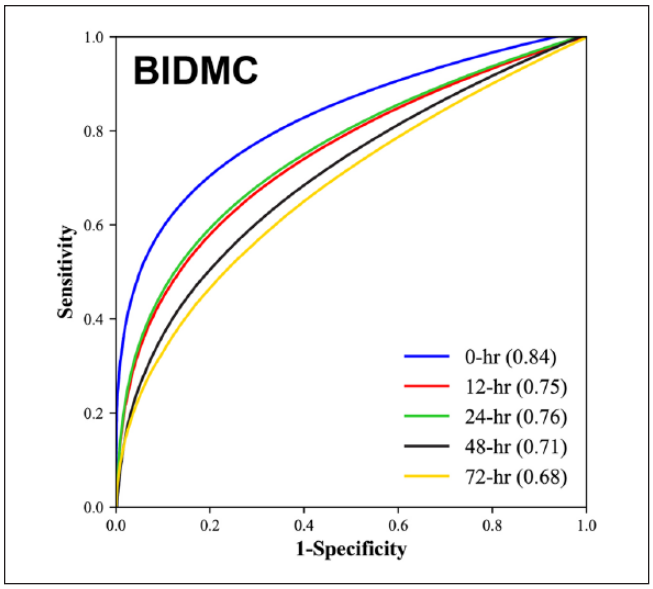

Figure 2. Comparison of the receiver operating characteristic and area under the receiver operating characteristic for machine learning algorithm 0-, 12-, 24-, 48-, and 72-hour advance prediction of stage 2 or stage 3 acute kidney injury development for BIDMC patient data.
Note. BIDMC = Beth Israel Deaconess Medical Center.

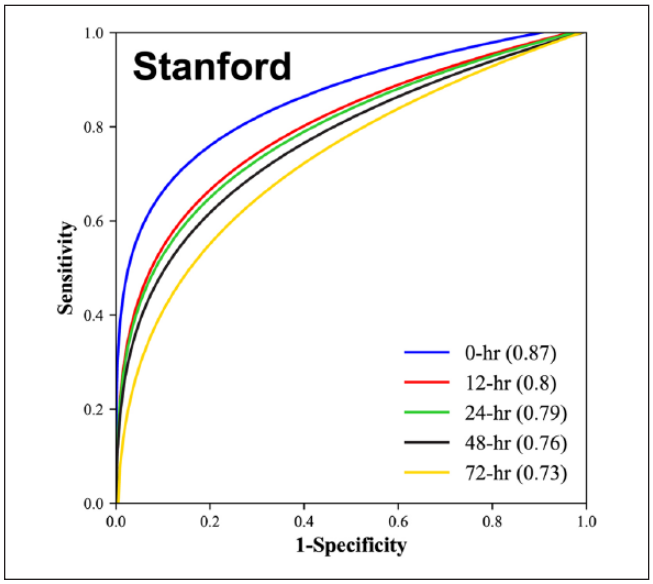

Figure 3. Comparison of the receiver operating characteristic and area under the receiver operating characteristic for machine learning algorithm 0-, 12-, 24-, 48-, and 72-hour advance prediction of stage 2 or stage 3 acute kidney injury development for Stanford Medical Center patient data.

to successfully predict AKI with AUROC up to 0.84.[14] However, this model was developed and validated using only ICU data, thus may have limited applicability to other settings. Another study demonstrated the successful application of a discrete-time survival model to predict AKI development on the ward.[13] Additional work has compared logistic regression, support vector machine, naive Bayes, and decision tree models for AKI detection on a population of patients 60 years or older.[15] The work presented here advances the field by evaluating a machine learning method trained and tested on a mixed-ward population that includes adults of all ages.

A machine learning approach provides advantages over currently used systems. Unlike the SOFA score, which is a generalized disease severity score, this method is specific to AKI. The measurable benefits (Tables 2 and 3) of focusing on AKI prediction could allow clinicians to more rapidly determine the cause of patient deterioration and, thus, administer appropriate treatments in a more timely manner. In addition, the ROC curves of Figures 2 and 3 provide a continuous range of sensitivity-specificity pairs at which the MLA can operate. If fewer alerts, greater specificity, and 72-hour notice were preferable over more alerts, greater sensitivity, and nearer-onset notice, the MLA could function accordingly. This flexibility is not available for a rules-based score like SOFA. The MLA also may provide advantages over manual AKI detection methods, which may not be implemented unless a physician already suspects AKI, and are subject to human error. Future work will investigate ensemble learning methods for further improved accuracy in AKI detection.

Limitations

Because this work presents a retrospective study, we cannot draw strong conclusions about this algorithm's performance in a live clinical setting. We cannot determine from the nature of this study what impact the algorithm might have on clinicians and the care which they provide. This study assesses algorithm performance only on US patients older than the age of 18, and results may not be generalizable to additional populations. In prospective settings, if the algorithm is implemented on patient populations which differ substantially from those used in this study, the predictive performance of the algorithm may differ. Indeed, our cross-validation analysis only allows us to conclude that the performance we report would generalize well to patient populations similar to the BIDMC and Stanford data sets. Because this study does not examine variables such as cause of admission or patient comorbidities, we cannot determine from this study whether there is a subpopulation of patients for whom this algorithm may be most useful.

Because there have been several proposed consensus definitions for AKI, our predictive algorithm may have different results when compared against various gold standard definitions, or in prospective clinical settings which utilize a different gold standard in their diagnostic procedures. Due to missing observations of urine output in the data sets used in these experiments, the training and testing sets for the

Table 4. Algorithm Performance for Detection of Stage 2 or Stage 3 Acute Kidney Injury Under the Kidney Disease: Improving Global Outcomes Criteria, Measured on Patient Data From BIDMC.

	Trained on BIDMC			Trained on Stanford		
	Onset	12 hours	24 hours	Onset	12 hours	24 hours
AUROC (95% CI)	0.924 (0. 872-0.975)	0.914 (0.814-0.999)	0.882 (0.669-0.999)	0.844 (0.716-0.972)	0.826 (0.716-0.935)	0.760 (0.591, 0.929)
Sensitivity	0.987	0.999	0.900	0.981	0.971	0.933
Specificity	0.912	0.907	0.879	0.715	0.719	0.602

Note. BIDMC = Beth Israel Deaconess Medical Center; AUROC = area under the receiver operating characteristic; CI = confidence interval.

KDIGO criteria[21] were limited in size. Performance against this criteria in a different retrospective or clinical setting may therefore differ from the results presented here.

Conclusion

The machine learning approach described in this study accurately predicts stage 2 or stage 3 AKI up to 72 hours in advance of onset on when trained and tested on two distinct data sets. This algorithm may improve detection of AKI in clinical settings, allowing for earlier intervention and improved patient outcomes.

Ethics Approval and Consent to Participate

Retrospective patient data was fully deidentified before access. Studies performed on deidentified data constitute nonhuman subject research, thus no institutional or ethical approvals were required for this study. Participant informed consent was not required for this study.

Consent for Publication

All authors consent for publication.

Availability of Data and Materials

Deidentified patient data from the Medical Information Mart for Intensive Care III (MIMIC-III) v1.3 database is publicly available at https://mimic.physionet.org/.

Acknowledgments

We gratefully acknowledge Jana Hoffman and Emily Huynh for their suggestions and assistance in editing this manuscript. We also thank Thomas Desautels for his feedback during this study.

Author Contributions

Study concept and design: Hamid Mohamadlou and Ritankar Das; Acquisition, analysis, or interpretation of data: Hamid Mohamadlou, Anna Lynn-Palevsky, Christopher Barton, Uli Chettipally, Lisa Shieh, Jacob Calvert, Nicholas R. Saber, and Ritankar Das; Drafting of the manuscript: Hamid Mohamadlou and Anna Lynn-Palevsky; Critical revision of the manuscript for intellectual content: Hamid Mohamadlou, Anna Lynn-Palevsky, Christopher Barton, Uli Chettipally, Lisa Shieh, Jacob Calvert, Nicholas R. Saber, and Ritankar Das; Statistical analysis: Hamid Mohamadlou and Nicholas R. Saber.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Mohamadlou, Barton, Calvert, Lynn-Palevsky, Saber, and Das are employees of Dascena, developers of the predictive algorithm. Shieh reports receiving grant funding from Dascena.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material for this article is available online.

References

1. Waikar SS, Liu KD, Chertow GM. Diagnosis, epidemiology, and outcomes of acute kidney injury. *Clin J Am Soc Nephrol*. 2008;3(3):844-861. doi:10.2215/CJN.05191107.
2. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J Am Soc Nephrol*. 2005;16(11):3365-3370. doi:10.1681/ASN.2004090740.
3. Silver SA, Long J, Zheng Y, Chertow GM. Cost of acute kidney injury in hospitalized patients. *J Hosp Med*. 2017;12(2): 70-76.
4. 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. doi:10.1186/cc5713.
5. Metnitz PG, Krenn CG, Steltzer H, et al. Effect of acute renal failure requiring renal replacement therapy on outcome in critically ill patients. *Crit Care Med*. 2002;30:2051-2058.
6. De Corte W, Dhondt A, Vanholder R, et al. Long-term outcome in ICU patients with acute kidney injury treated with renal replacement therapy: a prospective cohort study. *Crit Care*. 2016;20:256. doi:10.1186/s13054-016-1409-z.
7. Kolhe NV, Staples D, Reilly T, et al. Impact of compliance with a care bundle on acute kidney injury outcomes: a prospective observational study. *PLoS One*. 2015;10(7):e0132279. doi:10.1371/journal.pone.0132279.
8. Kellum JA, Lameire N, Aspelin P, et al. Kidney Disease: Improving Global Outcomes (KDIGO) acute kidney injury work group KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl*. 2012;2:1-138.
9. Star RA. Treatment of acute renal failure. *Kidney Int*. 1998; 54(6):1817-1831.

10. Praught ML, Shlipak MG. Are small changes in serum creatinine an important risk factor? *Curr Opin Nephrol Hypertens*. 2005;14:265-270.
11. Mehta RL, Pascual MT, Soroko S, et al. Spectrum of acute renal failure in the intensive care unit: the PICARD experience. *Kidney Int*. 2004;66(4):1613-1621. doi:10.1111/j.1523-1755.2004.00927.x.
12. Hoste EAJ, Clermont G, Kersten A, et al. RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis. *Crit Care*. 2006;10(3):R73. doi:10.1186/cc4915.
13. Koyner JL, Adhikari R, Edelson DP, Churpek MM. Development of a multicenter ward-based AKI prediction model. *Clin J Am Soc Nephrol*. 2016;11:1935-1943.
14. Fletchet M, Guiza F, Chetz M, et al. 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*. 2017;43(6):764-773. doi:10.1007/s00134-017-4678-3.
15. Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med Informat Decis Making*. 2016;16:39. doi:10.1186/s12911-016-0277-4.
16. Porter CJ, Juurlink I, Bisset LH, Bavakunji R, Mehta RL, Devonald MA. A real-time electronic alert to improve detection of acute kidney injury in a large teaching hospital. *Nephrol Dial Transplant*. 2014;29(10):1888-1893. doi:10.1093/ndt/gfu082.
17. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine. *Intensive Care Med*. 1996;22(7):707-710. doi:10.1007/BF01709751.
18. De Mendonca A, Vincent JL, Suter PM, et al. Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score. *Intensive Care Med*. 2000;26:915-921.
19. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3: 160035. doi:10.1038/sdata.2016.35.
20. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. 2017. *BMJ Open*. 2018;8(1):e017833.
21. Selby NM, Hill R, Fluck RJ. Standardizing the early identification of acute kidney injury: the NHS England national patient safety alert. *Nephron*. 2015; 131:113-117. doi:10.1159/000439146.
22. Kellum JA, Lameire N. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). On behalf of the KDIGO guideline AKI group. *Crit Care*. 2013;17(1):204. doi:10.1186/cc11454.
23. Ricci Z, Cruz D, Ronco C. The RIFLE criteria and mortality in acute kidney injury: a systematic review. *Kidney Int*. 2008; 73(5):538-546.
24. Schuur JD, Chambers JG, Hou PC. Urinary catheter use and appropriateness in U.S. emergency departments, 1995-2010. *Acad Emerg Med*. 2014;21:292-300.
25. Závada J, Hoste E, Cartin-Ceba R, et al. A comparison of three methods to estimate baseline creatinine for RIFLE classification. *Nephrol Dial Transplant*. 2010;25(12):3911-3918.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2016; San Francisco, CA, USA.
27. Pandharipande PP, Shintani AK, Hagerman HE, et al. Derivation and validation of SpO2/FiO2 ratio to impute for PaO2/FiO2 ratio in the respiratory component of the Sequential Organ Failure Assessment (SOFA) Score. *Crit Care Med*. 2009;37(4):1317-1321. doi:10.1097/CCM.0b013e31819cefa9.
28. Palevsky PM, Liu KD, Brophy PD, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for acute kidney injury. *Am J Kidney Dis*. 2013;61(5):649-672.
29. Calvert JS, Price DA, Chettipally UK, et al. A computational approach to early sepsis detection. *Comp Biol Med*. 2016;74: 69-73.
30. Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform*. 2016; 4(3):e28.
31. Calvert J, Desautels T, Chettipally U, et al. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg*. 2016;8:50-55.
32. Shimabukuro D, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomized clinical trial. *BMJ Open Respir Res*. 2017; 4:e000234.
33. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual*. 2017;6:e000158. doi:10.1136/bmjoq-2017-000158.