KI REPORTS
KIReports.org

Convolutional Neural Network Model for Intensive Care Unit Acute Kidney Injury Prediction

Sidney Le[1,5], Angier Allen[1,5], Jacob Calvert[1], Paul M. Palevsky[2], Gregory Braden[3], Sharad Patel[4], Emily Pellegrini[1], Abigail Green-Saxena[1], Jana Hoffman[1] and Ritankar Das[1]

[1]Dascena, Inc., Houston, Texas, USA; [2]VA Pittsburgh Healthcare System and University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [3]Baystate Medical Center, Springfield, Massachusetts, USA; and [4]Department of Critical Care Medicine, Cooper University Health Care, Camden, New Jersey, USA

Introduction: Acute kidney injury (AKI) is common among hospitalized patients and has a significant impact on morbidity and mortality. Although early prediction of AKI has the potential to reduce adverse patient outcomes, it remains a difficult condition to predict and diagnose. The purpose of this study was to evaluate the ability of a machine learning algorithm to predict for AKI as defined by Kidney Disease: Improving Global Outcomes (KDIGO) stage 2 or 3 up to 48 hours in advance of onset using convolutional neural networks (CNNs) and patient electronic health record (EHR) data.

Methods: A CNN prediction system was developed to use EHR data gathered during patients' stays to predict AKI up to 48 hours before onset. A total of 12,347 patient encounters were retrospectively analyzed from the Medical Information Mart for Intensive Care III (MIMIC-III) database. An XGBoost AKI prediction model and the sequential organ failure assessment (SOFA) scoring system were used as comparators. The outcome was AKI onset. The model was trained on routinely collected patient EHR data. Measurements included area under the receiver operating characteristic (AUROC) curve, positive predictive value (PPV), and a battery of additional performance metrics for advance prediction of AKI onset.

Results: On a hold-out test set, the algorithm attained an AUROC of 0.86 and PPV of 0.24, relative to a cohort AKI prevalence of 7.62%, for long-horizon AKI prediction at a 48-hour window before onset.

Conclusion: A CNN machine learning-based AKI prediction model outperforms XGBoost and the SOFA scoring system, revealing superior performance in predicting AKI 48 hours before onset, without reliance on serum creatinine (SCr) measurements.

Kidney Int Rep (2021) **6,** 1289–1298; https://doi.org/10.1016/j.ekir.2021.02.031

KEYWORDS: acute kidney injury; convolutional neural net; electronic health record data; machine learning; prediction; serum creatinine

Acute kidney injury (AKI) is a complex syndrome associated with large clinical and financial burdens.[1–12] Despite its prevalence in hospitalized patients[2,13] and reported incidence as high as 70% in the critically ill,[13,14] no treatment has been developed to effectively reverse injury to the kidney and restore kidney function.[1] The reasons for this failure have been attributed to delays in diagnosis and intervention,[2,15–23] the complex nature of the AKI syndrome and the staging of its severity,[3,21] and its multiple etiologies.[15,16]

Until recently, studies of incidence and outcomes of AKI have produced inconsistent results owing to varying definitions of AKI.[24–26] The Risk, Injury, Failure, Loss, End-stage kidney disease criteria,[27] followed by the AKI Network,[28] and most recently the Kidney Disease: Improving Global Outcomes (KDIGO) criteria[29,30] have provided consensus on AKI definition. KDIGO guidelines define AKI as an absolute increase of serum creatinine (SCr) of >0.3 mg/dl within 48 hours or a relative increase of >50% in no more than 7 days.[21,29] Doubling of SCr at steady state reflects an approximate 50% decrease in kidney function as evaluated by glomerular filtration rate.[31] Some studies have suggested that changes in SCr even smaller than 0.3 mg/dl within 48 hours are associated with

Correspondence: Abigail Green-Saxena, Dascena, Inc., 12333 Sowden Road, Suite B, PMB 65148, Houston, Texas 77080-2059, USA. E-mail: abigail@dascena.com

[5]SL and AA contributed equally to this work.

Received 11 August 2020; revised 4 February 2021; accepted 15 February 2021; published online 26 February 2021

significant increases in the risk of death, dialysis, and other morbidities,[6,21,32–38] and other studies are consistent with worsening outcomes with increasing AKI stage.[5,14,24,39–42] However, increases of SCr are known to lag kidney injury by hours to days after the initial kidney insult, and therefore recognition of AKI is delayed owing to reliance on SCr measurements.[43,44]

Early AKI detection is critical to improving patient outcomes.[45–48] Given that the components necessary for defining and staging AKI are routinely available in EHR,[3] a number of automated alerts have been developed to predict AKI events before onset. However, these alerts are generally triggered by detecting changes in SCr and urine output alone or in combination.[17] Because a range of kidney injuries can exist before a loss of kidney function can be estimated with these standard laboratory tests,[44,49] there is great interest in developing methods that can be used to detect AKI in patients at an earlier stage.[50–56] In this article, we describe our methodology for the development of a convolutional neural net (CNN) prediction system that predicts AKI up to 48 hours before onset using patient data extracted from the EHR. The CNN model does not require SCr or urine output values.

METHODS

Description of Data

This study uses data from the MIMIC-III version 1.3 data set,[57] collected at Beth Israel Deaconess Medical Center in Boston, Massachusetts, from 2001 to 2012. The MIMIC data set offers a variety of encounter information of more than 40,000 unique patients and includes both structured (e.g., laboratory results) and unstructured (e.g., clinician notes) data. Owing to differences in the storage of patient procedure information, we restrict our study to data collected from 2008 to 2012 using the iMDsoft MetaVision ICU (iMDsoft, Needham, MA) EHR system and do not include data collected from 2001 to 2008 using the Philips CareVue Clinical Information System (Philips Health-care, Andover, MA).[58] Because the collection of the MIMIC data did not affect patient safety and because all data were anonymized in accordance with the Health Insurance Portability and Accountability Act Privacy Rule, the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology waived the requirement for patient consent.

From the MetaVision EHR MIMIC encounters, we selected for inclusion stays involving adult patients (i.e., age 18 years or older) with at least one measurement of diastolic blood pressure, systolic blood pressure, temperature, respiratory rate, heart rate, oxygen saturation, and Glasgow Coma Scale. These measurements were selected because they were frequently available and easily collected at the patient bedside, even before clinical suspicion of AKI was present. These were the only direct variables used during the training and testing of the algorithm; clinical notes vectorized with the Doc2Vec algorithm were also used as inputs to the CNN model. Serum creatinine was used as part of the KDIGO criteria, which served as the gold standard of patients with true-positive AKI, but it was not used as an input in testing. To facilitate the analysis of the 48-hour advance prediction of AKI onset with a 5-hour window of measurements upon which to base such a prediction, we required the patient stay duration to be at least 53 hours long. For convenience and with minimal restrictions, we required that patient encounters lasted no more than 1000 hours. To train and test the algorithm on the broadest possible patient sample, no further inclusion or exclusion criteria were applied. Patients with prevalent AKI, those with chronic kidney disease, and who received dialysis were therefore included. Inclusion criteria are listed in Figure 1 for 24- and 48-hour prediction windows, and the demographic characteristics of encounters meeting the inclusion criteria are reported in Table 1.

Overview of Preprocessing, Training, and Testing

MIMIC-III intensive care unit (ICU) encounter data were gathered in the following ways: encounters from the MetaVision database in MIMIC-III were required to be at least 18 years of age and had to include at least 1 measurement for at least 1 of the required input features. For each prediction offset T, the encounters were filtered such that each encounter was between $5 + T$ hours and 1000 hours. A total of $5 + T$ hours were required to account for the offset and give the model the required 5 hours of measurements used for prediction. For each prediction offset T, positive examples measurements were taken between $5 + T$ and T hours before onset for prediction, whereas negative example measurements were taken during random 5-hour windows of the patient stays. Onset was defined as the first time that the relevant KDIGO criteria were met during the patient stay. Patient encounters satisfying the inclusion criteria were immediately allocated to training and testing sets. Approximately 90% and 10% of all encounters were randomly allocated to the training and testing sets, respectively, stratified by positive and negative classes to ensure equal representation of classes in both sets. We binned the data by the hour, imputed missing measurements, and standardized measurements on a variable-by-variable basis. AKI was defined according to KDIGO stage 2 or KDIGO stage 3 criteria, and positive cases were identified as those

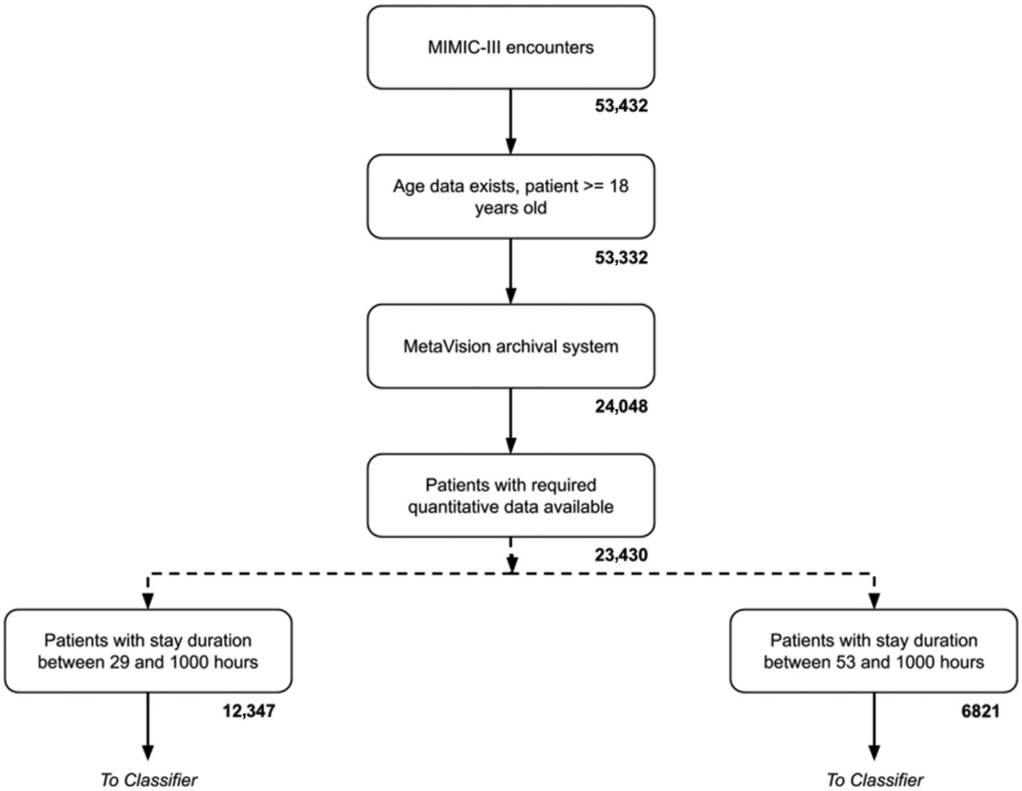
MIMIC-III encounters
53,432
Age data exists, patient >= 18 years old
53,332
MetaVision archival system
24,048
Patients with required quantitative data available
23,430
Patients with stay duration between 29 and 1000 hours
12,347
To Classifier
Patients with stay duration between 53 and 1000 hours
6821
To Classifier

Figure 1. Inclusion diagram. Patients were required to be at least 18 years of age and must have at least 1 measurement of at least 1 of the input features. MIMIC-III, Medical Information Mart for Intensive Care III.

patients reaching KDIGO stage 2 or stage 3 during the encounter. KDIGO stage 2 or stage 3 classifications were determined for each encounter, along with the corresponding times of KDIGO onset where appropriate. Stage 2 AKI is defined in the KDIGO staging system as an increase in SCr to more than 200% to 300% (>2- to 3-fold) from baseline or urine output <0.5 ml/kg per hour for more than 12 hours.[29] Stage 3 AKI is defined as an increase in SCr to more than 300% (>3-fold) from baseline, or ≥4.0 mg/dl (≥354 mmol/l), or kidney replacement therapy, or a decrease in estimated glomerular filtration rate to <35 ml/min per 1.73 m^2 (if <18 years of age), or urine output < 0.5 ml/kg per hour for ≥24 hours or anuria for ≥12 hours.[29] In both cases, the smaller of either the Modification of Diet in Renal Disease[27] SCr estimate based on KDIGO 2012 guidelines or the 20th percentile of observed creatinine measurements was used for the baseline creatinine measurement in each patient encounter. Any missing features required for measurement, including missing urine or SCr measures, made a contribution of 0 to the total KDIGO score.

A Doc2Vec embedding network was created to vectorize clinical text data. The Doc2Vec algorithm works by creating vectors for the most common words in all the documents and separate vectors for each document. These vectors are trained by selecting a window of words in each document; the corresponding vectors for these words, in addition to the vector for the document that the text came from, predict the next word in the sequence. The resulting document vectors are used as inputs, whereas the word vectors are

Table 1. Demographic characteristics of MIMIC-III ICU encounters found in the 48 hour data set and meeting the inclusion criteria of Figure 1

Characteristic		Count	%
Gender	Female	3186	46.71
	Male	3635	53.29
Age (d): Median 65, IQR (53–77)	18–29	317	4.65
	30–39	307	4.50
	40–49	665	9.75
	50–59	1246	18.27
	60–69	1599	23.44
	70+	2687	39.39
Length of stay (d): Median 5, IQR (4–9)	<3	43	0.63
	3–5	4282	62.78
	6–8	1200	17.59
	9–11	528	7.74
	≥12	768	11.26
Inhospital death	Yes	1747	25.61
	No	5074	74.39
KDIGO stage 2 or 3	Positive	520	7.62
	Negative	6301	92.38
KDIGO stage 1, 2, or 3	Positive	1410	20.67
	Negative	5411	79.33

ICU, intensive care unit; IQR, interquartile range; KDIGO, Kidney Disease: Improving Global Outcomes; MIMIC-III, Medical Information Mart for Intensive Care III.
We note that the determination of KDIGO positive or negative was made after the data preprocessing steps described in the Methods section.

discarded. The embedding network was prepared on a large collection of midstay clinical notes, ranging from the primary complaint to radiology notes, including everything up to, but not including, the discharge summary, from encounters allocated to the training set. The network embedded texts into 250-dimensional numeric vectors, which served as inputs to the classifiers, alongside the structured data associated with the stays. Any notes dated after the onset of AKI were not used as inputs for the model to ensure that the model used only data found at or before prediction time.

Training data were passed to a CNN structure, with hyperparameters optimized on the training set using the Python-based optimization package Talos (Autonomio Talos [Computer software]). Tuned hyperparameters include learning rate, batch size, optimization loss, L1 and L2 regularization coefficients, and the size of dense layers in the model. CNN was chosen instead of a recurrent neural network as it is faster to train and has fewer parameters (M. Blohm *et al.,* unpublished data, 2018). In addition, the window of time from which the structured data were gathered for prediction was relatively short (5 hours). CNN modeling techniques have been found to outperform recurrent neural network modeling techniques with improved generalizability when applied to speech recognition tasks (A.V. Oord *et al.*, unpublished data, 2016). After the end of the training on each fold, network performance was evaluated using the hold-out test set. Results were reported as the average test set performance across cross-validation folds.

Structured Data Preprocessing

Structured data were binned by the hour, with multiple intrahour measurements of the same variable replaced by its average. Missing measurements were handled separately for training and testing sets using the last observation that carried forward the imputation. Any remaining missing values were filled in using the measurement median in the training data. Quantitative data and document vectors were then standardized using the training data such that each feature had a mean of 0 and a variance of 1.

Document Vector Encoding Network and Unstructured Data Preprocessing

To facilitate the use of unstructured text data alongside the structured inputs, we trained a Doc2Vec (Q.V.Le., unpublished data, 2014) embedding network with 250 nodes on 238,468 midstay clinical notes. Document vectors were produced for the text data available from each encounter, using 125 epochs of the Doc2Vec algorithm—to better ensure the stability of the inferred document vectors—and an initial learning rate of 0.01. The choice of the number of epochs and learning rate was found through experimentation. Clinical notes dated after AKI onset were excluded from the input when training and testing CNN.

Training of Neural Network Classifier

We constructed a classifier to predict the probability of the presence of AKI at a given offset time from prediction using the Python deep learning library, Keras, that uses variants of multichannel, multiheaded attention together with convolutions to extract information from quantitative time series data. A separate network for handling the document vector produced by the Doc2Vec network was combined downstream through concatenation in a fully connected output layer. This allowed the model to incorporate information from both the time series data in the EHRs and the qualitative information found in the clinical notes. Model parameters were optimized using the Nadam optimizer[59] as implemented in the Keras library with a learning rate of 0.0009 and binary cross-entropy loss. A diagram of this neural network architecture is available as Supplementary Figure S1. Owing to the low prevalence of AKI in the data, random oversampling was performed to artificially inflate the positive population. This was performed by picking examples from the positive class at random with replacement until the number of positive examples matched the number of negative examples.

To fit the weights of the network with 10-fold cross-validation, we split the training data into 10 subsets of roughly equal size and iteratively used 9 subsets for intrafold training and the final subset for intrafold testing. Model parameters were fit over the course of 50 epochs on the 9 intrafold training subsets, with evaluation on the final subset. For each iterate, we obtained an ROC curve and a battery of performance metrics. We then randomly reset the model parameters before performing another iterate. From cross-validation, we obtained an average ROC curve and average performance metrics, along with standard deviation for the performance metrics. These results are presented in comparison with an XGBoost[60] classifier and the SOFA score,[61] which has been found to independently predict AKI outcomes[62–64] and therefore serves as a validated comparison measure for AKI prediction. SOFA was computed using all organ systems; any missing inputs required for computation contributed zero points to the total SOFA score. The XGBoost classifier was trained on the same processed training sets—5-hour windows of quantitative, clinical EHR data—and evaluated on the same testing set. The time series data were turned into a list of the binned measurements at the different hours and given to XGBoost as input, requiring no additional feature engineering. Document

Table 2. Results from 10-fold cross-validation of predictions 48 hours before onset on the MIMIC-III data set

Performance metric	CNN	XGBoost	SOFA	No Doc2Vec	Stage 1 included	Stage 3 only
AUROC mean (SD)	0.856 (0.034)	0.654 (0.011)	0.701	0.763 (0.035)	0.778 (0.037)	0.819 (0.036)
Sensitivity mean (SD)	0.804 (0.000)	0.798 (0.000)	0.798	0.805 (0.006)	0.806 (0.008)	0.806 (0.000)
Specificity mean (SD)	0.763 (0.057)	0.380 (0.006)	0.441	0.623 (0.064)	0.649 (0.074)	0.679 (0.079)
PPV mean (SD)	0.236 (0.039)	0.095 (0.001)	0.127	0.163 (0.022)	0.310 (0.044)	0.105 (0.023)
NPV mean (SD)	0.975 (0.002)	0.956 (0.001)	0.960	0.970 (0.003)	0.940 (0.006)	0.985 (0.002)
Accuracy mean (SD)	0.765 (0.052)	0.411 (0.005)	0.612	0.638 (0.056)	0.672 (0.062)	0.683 (0.076)
DOR mean (SD)	14.076 (3.779)	2.421 (0.059)	3.123	7.123 (1.899)	8.167 (2.425)	9.566 (3.410)
LR+ mean (SD)	3.558 (0.739)	1.287 (0.012)	1.429	2.191 (0.362)	2.389 (0.478)	2.658 (0.660)
LR– mean (SD)	0.258 (0.021)	0.532 (0.008)	0.458	0.316 (0.035)	0.301 (0.035)	0.288 (0.035)
F1 mean (SD)	0.361 (0.047)	0.169 (0.001)	0.214	0.270 (0.030)	0.444 (0.045)	0.184 (0.036)

AUROC, area under the receiver operating characteristic curve; CNN, convolutional neural network; DOR, diagnostic odds ratio; KDIGO, Kidney Disease: Improving Global Outcomes; LR+, positive likelihood ratio; LR–, negative likelihood ratio; MIMIC-III, Medical Information Mart for Intensive Care III; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation; SOFA, sequential organ failure assessment.

The CNN model is compared with an XGBoost classifier and the SOFA score. SOFA required no training and thus could be applied to the entire test set at once; hence, no SD is reported. Additional comparison is made to the CNN model without the use of the Doc2Vec network (i.e., without unstructured text data) and for the prediction of KDIGO criteria of any stage.

vectors were not given as input for XGBoost. XGBoost hyperparameters were tuned using a cross-validated grid search on the training data. Hyperparameters were optimized using grid search in the hyperparameters "gamma," which controls how often the trees are split, and "colsample_bytree," which controls the number of features randomly selected for inputs when constructing each tree.

RESULTS

The demographic characteristics associated with MIMIC-III ICU encounters meeting the inclusion criteria of Figure 1 are provided in Table 1. The study population consisted of 53.29% men, with a few patients younger than 30 years of age (4.65%) and a substantial percentage of patients aged 70 years or more (39.39%). More than half the patients had stays lasting between 3 and 5 days (62.78%), with a substantial percentage of patients experiencing stays of 12 days or longer (11.26%). The overall mortality rate was 39.39%, with 7.62% of encounters meeting the criteria for KDIGO stage 2 or stage 3 at some point during the stay, and 20.6% of stays meeting some stage of the KDIGO criteria at any point during the stay.

Performance was evaluated by predicting once in each encounter using 5 hours of data. These data were taken either from a random portion of the stay for negative examples, or from the specified model offset for positive examples. The results from 10-fold cross-validation on the 90% training set are reported in Tables 2 and 3 for 48 and 24 hour predictions, respectively. Test performance is reported for the best performing model, selected by cross-validation of the training data. The CNN model, with the use of the Doc2Vec embeddings of encounter text data, outperformed the XGBoost comparator model and the SOFA score for advance prediction of KDIGO stage 2 or stage 3 onset. We note that, to provide nonsummative performance metrics (i.e., the metrics other than AUROC), we selected an operating point for each model or score that provided a sensitivity nearest to 0.80. The CNN model performed better (AUROC of 0.86 for 24 and 48 hour predictions) when text data were made available through Doc2-Vec than when these data were unavailable (AUROC of 0.77 and 0.76 for 24 and 48 hour predictions, respectively). In addition, the quality of prediction was higher for KDIGO stage 2 or stage 3 onset, as compared with the prediction of onset for any of KDIGO stages 1–3. For corresponding CNN and XGBoost results without oversampling of the minority class, see Supplementary Table S1. Permutation feature importance methods were implemented to provide information on the relative importance of each input variable. A precision-recall curve comparison between the CNN model, the XGBoost model, and the SOFA score is presented in Supplementary Figure S2.

The CNN model averaged a PPV of 0.24 over cross-validation folds for the 48-hour prediction of KDIGO stages 2 and 3, compared with average PPVs of 0.09 and 0.13 for XGBoost and the SOFA score, respectively (Table 2). CNN had almost no advantage (PPV of 0.16) in the absence of text data through Doc2Vec input. The average PPV was highest when the CNN classifier was given access to Doc2Vec input and tasked with 48-hour prediction of KDIGO stages 1–3 (PPV of 0.31). Relative to the 7.62% prevalence of KDIGO stages 2 and 3, positive predictions made by the CNN model enriched KDIGO stage 2 or 3 encounters by a factor of 4.80, whereas XGBoost and the SOFA scores enriched these encounters by factors of 2.50 and 2.11, respectively.

The ROC curve comparison of the 48-hour prediction on the 10% hold-out test set is found in Figure 2. The CNN model, which was provided text data through the Doc2Vec input, performed substantially

Table 3. Results from 10-fold cross-validation of predictions 24 hours before onset on the MIMIC-III data set

Performance metric	CNN	XGBoost	SOFA	No Doc2Vec	Stage 1 included	Stage 3 only
AUROC mean (SD)	0.863 (0.009)	0.729 (0.009)	0.727	0.769 (0.028)	0.834 (0.004)	0.867 (0.009)
Sensitivity mean (SD)	0.803 (0.000)	0.801 (0.000)	0.784	0.801 (0.003)	0.798 (0.005)	0.795 (0.000)
Specificity mean (SD)	0.772 (0.021)	0.463 (0.026)	0.537	0.585 (0.066)	0.716 (0.018)	0.785 (0.024)
PPV mean (SD)	0.221 (0.016)	0.111 (0.005)	0.151	0.153 (0.019)	0.359 (0.014)	0.131 (0.014)
NPV mean (SD)	0.978 (0.001)	0.964 (0.002)	0.961	0.968 (0.003)	0.944 (0.001)	0.988 (0.000)
Accuracy mean (SD)	0.773 (0.020)	0.489 (0.024)	0.684	0.602 (0.060)	0.728 (0.014)	0.784 (0.023)
DOR mean (SD)	13.905 (1.617)	3.484 (0.367)	4.200	5.861 (1.440)	10.030 (0.822)	14.396 (2.212)
LR+ mean (SD)	3.545 (0.319)	1.494 (0.073)	1.692	1.970 (0.292)	2.821 (0.178)	3.740 (0.452)
LR− mean (SD)	0.256 (0.007)	0.431 (0.024)	0.403	0.344 (0.038)	0.282 (0.007)	0.261 (0.008)
F1 mean (SD)	0.345 (0.019)	0.194 (0.007)	0.247	0.256 (0.027)	0.494 (0.013)	0.224 (0.020)

AUROC, area under the receiver operating characteristic curve; CNN, convolutional neural network; DOR, diagnostic odds ratio; KDIGO, Kidney Disease: Improving Global Outcomes; LR+, positive likelihood ratio; LR−, negative likelihood ratio; MIMIC-III, Medical Information Mart for Intensive Care III; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation; SOFA, sequential organ failure assessment.

The CNN model is compared with an XGBoost classifier and the SOFA score. SOFA required no training and thus could be applied to the entire test set at once; hence, no SD is reported. Additional comparison is made to the CNN model without the use of the Doc2Vec network (i.e., without unstructured text data) and for the prediction of KDIGO criteria of any stage.

better than the XGBoost model and SOFA. The XGBoost model and SOFA had similar performance on the test set.

DISCUSSION

These experiments reveal that a CNN can predict AKI up to 48 hours in advance of KDIGO stage 2 or stage 3 AKI onset, with AUROC performance superior to that of an XGBoost classifier and the SOFA scoring system (Table 2, Figure 2). Unlike other diseases for which multiple severity scores exist, AKI represents a group of syndromes that are loosely connected by the characteristic rapid drop in estimated glomerular filtration rate found in patients with AKI.[65] With more than 30 definitions of AKI,[66] attempts at a uniform definition for AKI have included the Risk, Injury, Failure, Loss, End-stage kidney disease classification,[27] followed by the AKI Network,[28] and, most recently, the KDIGO criteria.[29,30] The absence of a consistent, uniform definition may explain the current lack of an AKI-specific risk score that serves as a standard-of-care. To provide context for the performance of their AKI prediction models, previous studies have used the biomarker serum neutrophil gelatinase–associated lipocalin as a comparator,[67] compared their model to other machine learning models,[68] or not included a standard-of-care comparator.[69,70] In the current study, we compare 2 machine learning models and provide the SOFA score as a comparator. Although the SOFA score was not developed for the purpose of long-horizon AKI prediction, because of the ubiquity of the SOFA score and its previous use in AKI outcome prediction, it serves as a validated comparator for our current approach.[62–64] The XGBoost comparator is similarly important, primarily owing to its broad and successful use in applications for other clinical prediction tasks (e.g., the 2019 Physionet Computing in Cardiology Challenge[71]).

The superiority of the CNN classifier over the XGBoost classifier and the commonly used SOFA score is evidenced by key performance metrics, such as AUROC and PPV (Table 2). The PPV performance improvement is of particular importance. Romero-Brufau *et al*. have argued that AUROC performance may be misleading for clinicians interested in evaluating the clinical impact of a diagnostic tool, as AUROC does not incorporate information on the prevalence of a condition.[72] In fact, for the same reason, AUROC is useful for comparing the performance of tools retrospectively validated on different data sets. This concern regarding PPV and prevalence is

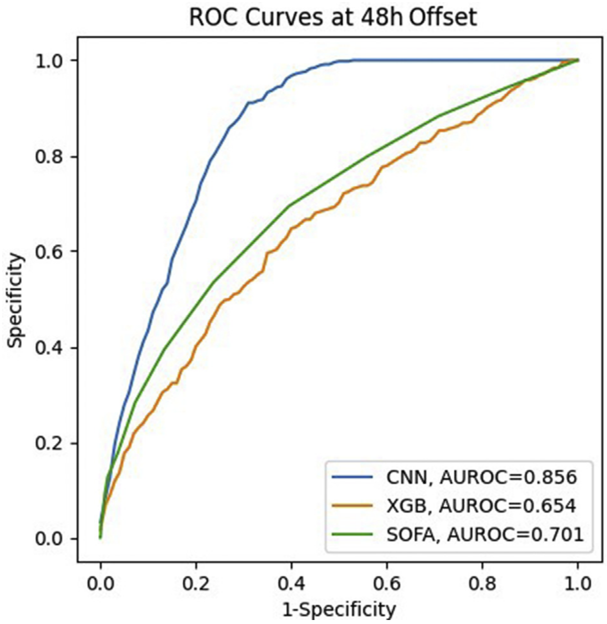

Figure 2. ROC curve comparison of prediction performance using a CNN classifier, an XGB classifier, and the SOFA score, 48 hours before AKI onset on the MIMIC-III ICU hold-out data set. AKI, acute kidney injury; AUROC, area under the receiver operating characteristic curve; CNN, convolutional neural network; ICU, intensive care unit; MIMIC-III, Medical Information Mart for Intensive Care III; ROC, receiver operating characteristic; SOFA, Sequential Organ Failure Assessment score; XGB, XGBoost.

relevant to our study, as we found that the prevalence of KDIGO stages 2 or 3 is roughly 7.6% in the cohort, an estimate consistent with previous epidemiologic studies.[73] The AUROC is a summative metric that may include ranges of operating points that are irrelevant to a given task, whereas PPV can be focused on a clinically relevant operating point. To produce the metrics in Table 2, we chose the operating points for the CNN and the comparators such that their sensitivities were fixed near 0.80.

Beyond the text data input through Doc2Vec, CNN predictions were made using only age and 7 routinely collected patient measurements (diastolic blood pressure, systolic blood pressure, temperature, respiratory rate, heart rate, oxygen saturation, and Glasgow Coma Scale) as inputs. Although this study was restricted to the MetaVision (iMDSoft) EHR system for technical reasons, the use of these widely available inputs supports the generalizability of the model to broad clinical practice. Importantly, the CNN model did not rely on SCr to make predictions, distinguishing it from other AKI prediction tools. Creatinine levels can take hours or days to rise to AKI thresholds as defined in the KDIGO staging system[74]; therefore, changes in SCr may reflect preexisting kidney damage. An AKI prediction tool that does not depend on SCr measurements may better afford clinicians the opportunity to intervene early, to prevent AKI development or progression, or to limit further kidney damage. Furthermore, using only often collected variables in the EHR for AKI prediction allows automatic screening of a general patient population for impending AKI without requiring specialized evaluation.

This study contributes to the growing body of retrospective machine learning literature for the prediction of AKI.[75] Chiofolo *et al*.[69] developed a model for AKI prediction and surveillance in patients in the ICU for a 6-hour prediction window with an AUROC of 0.88. Flechet *et al*.[67] developed the AKIpredictor, a prognostic calculator for prediction of AKI in patients in the ICU during the first week of stay. Their KDIGO stage 2 and 3 models produced AUROCs between 0.77 and 0.84. The AUROC of 0.84 corresponds to a prediction of KDIGO stage 2 and 3 after gathering 24 hours of data. As a point of comparison, the CNN model used only 5 hours of data before making a prediction. Recent work by Tomašev *et al*.[68] pursued a deep learning approach for continuous risk prediction of deterioration in patients with AKI and evaluated their tool on a Veteran's Health Administration data set of 703,782 adult patients. Algorithm performance for a 48-hour prediction window corresponded to a sensitivity of 55.8% and a specificity of 82.7%.[68] This performance is reported to be in the range required for regulatory approval.[76] Although these studies make important contributions to the domain of AKI research, they depend on the use of SCr to make predictions, which is a lagging marker of kidney function. In contrast, the CNN described in this work does not rely on SCr to make predictions of AKI onset, allowing for both longer lead times and improved predictive performance and for making predictions for patients who may not yet be clinically suspected of having AKI and who have not yet had their SCr measures drawn. CNN also offers improvement in performance as compared with our previous work,[77] which used the machine learning method of gradient-boosted trees to predict AKI before onset and included SCr as a model input. In comparison, results from our current work suggest that AKI predictions can be made with a more robust machine learning architecture, without reliance on SCr, while achieving stronger predictive performance.

Although the CNN described in this study offers substantial lead time in AKI identification (up to 48 hours) and offers improved predictive performance over our previous work,[77] it still requires prospective validation. Furthermore, we cannot determine from this retrospective study what impact the algorithm might have on clinicians and their provision of care in clinical settings, nor provide an analysis of model evaluation and its prediction performance in time. Although the CNN model performance was superior to that of SOFA and XGBoost, improvements in PPV achieved by CNN compared with XGBoost or SOFA are less pronounced without the use of clinical notes. Algorithm performance was evaluated only on patients in the United States older than 18 years with stays in the ICU, which limits the generalizability of our results to other patient populations and levels of care. Although most of the patients in the negative class had a SCr measurement at some point in the ICU stay, it is possible that inclusion of patients missing urine measures in the negative class led to the misclassification of some patients in our data set. It is also possible that misclassifications could have occurred for some patients in the data set owing to inclusion of patients with a previous diagnosis of chronic kidney disease or who received dialysis. Owing to the lack of a standard-of-care AKI score, we used the SOFA score and the XGBoost model to provide context for our model performance. Although the SOFA score has been used in AKI outcome prediction studies,[62–64] it was not developed for the purpose of long-horizon AKI prediction. Furthermore, although the XGBoost comparator was included owing to its use in other clinical prediction tasks,[71] it does not serve as a standard-of-care for AKI predictions. Last, because there have been several proposed consensus definitions for AKI, the algorithm we described may produce different results when compared against non–KDIGO definitions, or in settings that use a different standard in their diagnostic procedures.

CONCLUSION

A CNN for AKI prediction outperforms XGBoost and the traditional SOFA scoring system, revealing superior performance in predicting AKI up to 48 hours before onset without reliance on measurements of changes in SCr. Although the use of clinical text data through a Doc2Vec network substantially strengthened CNN prediction performance, CNN was found to have superior performance over both XGBoost and SOFA even when clinical notes were not included as model inputs, supporting the use of CNN models for the task of AKI prediction. Such a tool may improve prediction and early detection of AKI in clinical settings, thereby allowing for earlier intervention.

DISCLOSURE

SL, AA, JC, EP, AS, JH, and RD are or were employees or contractors of Dascena (Houston, Texas, USA) at the time the work was performed.

ACKNOWLEDGMENT

This work was supported by the National Institute on Alcohol Abuse and Alcoholism (grant ID: 1R43AA02767401).

AUTHOR CONTRIBUTIONS

RD, SL, JC, and AA conceived and designed this study; SL and AA performed the modeling and statistical analysis; all authors contributed to acquisition, analysis, or interpretation of data; SL, JH, AS, and EP drafted the article; all authors revised the article for important intellectual content; and RD obtained funding.

DATA SHARING PLAN

The data that support the findings of this study are publicly available from http://www.nature.com/articles/sdata201635.

SUPPLEMENTARY MATERIAL

Supplementary file (PDF)

Figure S1. Schematic diagram of the neural network architecture.

Table S1. Model performance metric results from 10-fold cross-validation.

Figure S2. Precision-recall curve comparison of model prediction performance.

REFERENCES

1. Kashani K, Ronco C. Acute kidney injury electronic alert for nephrologist: reactive versus proactive? *Blood Purif*. 2016;42:323–328.
2. Al-Jaghbeer M, Dealmeida D, Bilderback A, et al. Clinical decision support for in-hospital AKI. *J Am Soc Nephrol*. 2018;29:654–660.
3. Hoste EA, Bagshaw SM, Bellomo R, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med*. 2015;41:1411–1423.
4. Wang HE, Muntner P, Chertow GM, Warnock DG. Acute kidney injury and mortality in hospitalized patients. *Am J Nephrol*. 2012;35:349–355.
5. Uchino S, Bellomo R, Goldsmith D, et al. An assessment of the RIFLE criteria for acute renal failure in hospitalized patients. *Crit Care Med*. 2006;34:1913–1917.
6. Chertow GM, Burdick E, Honour M, et al. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J Am Soc Nephrol*. 2005;16:3365–3370.
7. Kellum JA, Sileanu FE, Murugan R, et al. Classifying AKI by urine output versus serum creatinine level. *J Am Soc Nephrol*. 2015;26:2231–2238.
8. Hoste EAJ, Kellum JA, Selby NM, et al. Global epidemiology and outcomes of acute kidney injury. *Nat Rev Nephrol*. 2018;14:607–625.
9. Freda BJ, Knee AB, Braden GL, et al. Effect of transient and sustained acute kidney injury on readmissions in acute decompensated heart failure. *Am J Cardiol*. 2017;119:1809–1814.
10. VA/NIH Acute Renal Failure Trial Network, Palevsky PM, Zhang JH, et al. Intensity of renal support critically ill patients with acute kidney injury [published correction appears in *N Engl J Med*. 2009;361:2391] *N Engl J Med*. 2008;359:7–20.
11. Kellum JA, Chawla LS, Keener C, et al. The effects of alternative resuscitation strategies in acute kidney injury patients with septic shock. *Am J Respir Crit Care Med*. 2016;193:281–287.
12. Khalil P, Murty P, Palevsky PM. The patient with acute kidney injury. *Prim Care*. 2008;35:239–264. vi.
13. Forni LG, Dawes T, Sinclair H, et al. Identifying the patient at risk of acute kidney injury a predictive scoring system for the development of acute kidney injury in acute medical patients. *Nephron Clin Pract*. 2013;123:143–150.
14. Uchino S, Kellum JA, Bellomo R, et al. Acute renal failure in critically ill patients: a multinational, multicenter study. *JAMA*. 2005;294:813–818.
15. Endre JH, Pickering JW. Acute kidney injury clinical trial design: old problems, new strategies. *Pediatr Nephrol*. 2013;28:207–217.
16. Pickering JW, Ralib AM, Nejat M, Endre ZH. New considerations in the design of clinical trials of acute kidney injury. *Clin Investig*. 2011;1:637–650.
17. Lachance P, Villeneuve PM, Rewa OG, et al. Association between e-alert implementation for detection of acute kidney injury and outcomes: a systematic review. *Nephrol Dial Transplant*. 2017;32:265–272.
18. 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, e0132279.

19. Terrell KM, Perkins AJ, Hui SL, et al. Computerized decision support for medication dosing in renal insufficiency: a randomized, controlled trial. *Ann Emerg Med*. 2010;56:623–629.

20. Colpaert K, Hoste EA, Steurbaut K, et al. Impact of real-time electronic alerting of acute kidney injury on therapeutic intervention and progression of RIFLE class. *Crit Care Med*. 2012;40:1164–1170.

21. Wilson FP, Shashaty M, Testani J, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. *Lancet*. 2015;385:1966–1974.

22. Thomas ME, Sitch A, Baharani J, Dowswell G. Earlier intervention for acute kidney injury: evaluation of an outreach service and a long-term follow-up. *Nephrol Dial Transplant*. 2015;30:239–244.

23. Jo SK, Rosner MH, Okusa MD. Pharmacologic treatment of acute kidney injury: why drugs haven't worked and what is on the horizon. *Clin J Am Soc Nephrol*. 2007;2:356–365.

24. Porter CJ, Juurlink I, Bisset LH, et al. A real-time electronic alert to improve detection of acute kidney injury in a large teaching hospital. *Nephrol Dial Transplant*. 2014;29:1888–1893.

25. Waikar SS, Curhan GC, Wald R, et al. Declining mortality in patients with acute renal failure, 1988 to 2002. *J Am Soc Nephrol*. 2006;17:1143–1150.

26. Xue JL, Daniels F, Star RA, et al. Incidence and mortality of acute renal failure in Medicare beneficiaries, 1992 to 2001. *J Am Soc Nephrol*. 2006;17:1135–1142.

27. Bellomo R, Ronco C, Kellum JA, et al. 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–R212.

28. 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.

29. 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.

30. 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;61:649–672.

31. Weisenthal 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.

32. Joannidis M, Metnitz B, Bauer P, et al. Acute kidney injury in critically ill patients classified by AKIN versus RIFLE using the SAPS 3 database. *Intensive Care Med*. 2009;35:1692–1702.

33. Kolli H, Rajagopalam S, Patel N, et al. Mild acute kidney injury is associated with increased mortality after cardiac surgery in patients with eGFR $<$ 60 mL/min/1·73 m (2). *Ren Fail*. 2010;32:1066–1072.

34. Wilson FP, Yang W, Feldman HI. Predictors of death and dialysis in severe AKI: the UPHS-AKI cohort. *Clin J Am Soc Nephrol*. 2013;8:527–537.

35. Bihorac A, Delano MJ, Schold JD, et al. Incidence, clinical predictors, genomics, and outcome of acute kidney injury among trauma patients. *Ann Surg*. 2010;252:158–165.

36. Bihorac A, Yavas S, Subbiah S, et al. Long-term risk of mortality and acute kidney injury during hospitalization after major surgery. *Ann Surg*. 2009;249:851–858.

37. Garzotto F, Piccinni P, Cruz D, et al. RIFLE-based data collection/management system applied to a prospective cohort multicenter Italian study on the epidemiology of acute kidney injury in the intensive care unit. *Blood Purif*. 2011;31:159–171.

38. Newsome BB, Warnock DG, McClellan WM, et al. Long-term risk of mortality and end-stage renal disease among the elderly after small increases in serum creatinine level during hospitalization for acute myocardial infarction. *Arch Intern Med*. 2008;168:609–616.

39. Coca SG, Yusuf B, Shlipak MG, et al. Long-term risk of mortality and other adverse outcomes after acute kidney injury: a systematic review and meta-analysis. *Am J Kidney Dis*. 2009;53:961–973.

40. Lafrance JP, Miller DR. Acute kidney injury associates with increased long-term mortality. *J Am Soc Nephrol*. 2010;21:345–352.

41. Ricci Z, Cruz D, Ronco C. The RIFLE criteria and mortality in acute kidney injury: a systematic review. *Kidney Int*. 2008;73:538–546.

42. Ali T, Khan I, Simpson W, et al. Incidence and outcomes in acute kidney injury: a comprehensive population-based study. *J Am Soc Nephrol*. 2007;18:1292–1298.

43. Ostermann M, Joannidis M. Acute kidney injury 2016: diagnosis and diagnostic workup. *Crit Care*. 2016;20:299.

44. Makris K. The role of the clinical laboratory in the detection and monitoring of acute kidney injury. *J Lab Precis Med*. 2018;3:69.

45. 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;24:1052–1061.

46. Park S, Baek SH, Ahn S, et al. Impact of electronic acute kidney injury (AKI) alerts with automated nephrologist consultation on detection and severity of AKI: a quality improvement study. *Am J Kidney Dis*. 2018;71:9–19.

47. de Virgilio C, Kim DY. Transient acute kidney injury in the postoperative period: it is time to pay closer attention. *JAMA Surg*. 2016;151:450–451.

48. Soares DM, Pessanha JF, Sharma A, et al. Delayed nephrology consultation and high mortality on acute kidney injury: a meta-analysis. *Blood Purif*. 2017;43:57–67.

49. Thomas ME, Blaine C, Dawnay A, et al. The definition of acute kidney injury and its use in practice. *Kidney Int*. 2015;87:62–73.

50. Hodgson LE, Dimitrov BD, Roderick PJ, et al. Predicting AKI in emergency admissions: an external validation study of the acute kidney injury prediction score (APS). *BMJ Open*. 2017;7, e013511.

51. Braitman LE, Davidoff F. Predicting clinical states in individual patients. *Ann Intern Med*. 1996;125:406–412.

52. Feinstein AR. "Clinical Judgment" revisited: the distraction of quantitative models. *Ann Intern Med*. 1994;120:799–805.

53. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med*. 1993;118:201–210.

54. Christensen E. Prognostic models including the Child-Pugh, MELD and Mayo risk scores–where are we and where should we go? *J Hepatol*. 2004;41:344–350.

55. Kashani K, Rosner MH, Haase M, et al. Quality improvement goals for acute kidney injury. *Clin J Am Soc Nephrol*. 2019;14: 941–953.

56. Garcia S, Bhatt DL, Gallagher M, et al. Strategies to reduce acute kidney injury and improve clinical outcomes following percutaneous coronary intervention: a subgroup analysis of the PRESERVE Trial. *JACC Cardiovasc Interv*. 2018;11:2254–2261.

57. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.

58. 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:e28.

59. Dozat T Incorporating Nesterov Momentum Into Adam. Available at: https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ. Accessed February 2, 2021.

60. Chen T, Guestrin C. A scalable tree boosting system. In Proceedings of: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13–17, 2016; San Francisco, CA.

61. Vincent JL, Moreno R, Takala J, et al. 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: 707–710.

62. Hoste EA, Clermont G, Kersten A, et al. RIFLE criteria for acute kidney injury is associated with hospital mortality in critically ill patients: a cohort analysis. *Crit Care*. 2006;10: R73.

63. de Mendonça A, Vincent JL, Suter PM, et al. Acute renal failure in the ICU: risk factors and outcomes evaluated by the SOFA score. *Intensive Care Med*. 2000;26:915–921.

64. Chang CH, Fan PC, Chang MY, et al. Acute kidney injury enhances outcome prediction ability of sequential organ failure assessment score in critically ill patients. *PLoS One*. 2014;9: e109649.

65. Kellum JA, Prowle JR. Paradigms of acute kidney injury in the intensive care setting. *Nat Rev Nephrol*. 2018;14:217–230.

66. Kellum JA, Levin N, Bouman C, Lameire N. Developing a consensus classification system for acute renal failure. *Curr Opin Crit Care*. 2002;8:509–514.

67. Flechet M, Güiza F, Schetz 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;43:764–773.

68. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572:116–119.

69. Chiofolo C, Chbat N, Ghosh E, et al. Automated continuous acute kidney injury prediction and surveillance: a random forest model. *Mayo Clin Proc*. 2019;94:783–792.

70. Simonov M, Ugwuowo U, Moreira E, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: a descriptive modeling study. *PLoS Med*. 2019;16:e1002861.

71. Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: the Physionet/Computing in Cardiology Challenge 2019. *Crit Care Med*. 2020;48:210–217.

72. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit Care*. 2015;19: 285.

73. Case J, Khan S, Khalid R, Khan A. Epidemiology of acute kidney injury in the intensive care unit. *Crit Care Res Pract*. 2013;2013:479730.

74. Waikar SS, Bonventre JV. Creatinine kinetics and the definition of acute kidney injury. *J Am Soc Nephrol*. 2009;20:672–679.

75. Palevsky PM. Electronic alerts for acute kidney injury. *Am J Kidney Dis*. 2018;71:1–2.

76. Kellum J, Bihorac A. Artificial intelligence to predict AKI: is it a breakthrough? *Nat Rev Nephrol*. 2019;15:663–664.

77. 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.