

RESEARCH

Open Access



ORAKLE: Optimal Risk prediction for mAke30 in patients with sepsis associated AKI using deep LEarning

Wonsuk Oh^{1,2}, Marinela Veshtaj^{3,4}, Ashwin Sawant^{1,2,5}, Pulkit Agrawal⁶, Hernando Gomez⁷, Mayte Suarez-Farinas⁸, John Oropello⁹, Roopa Kohli-Seth⁹, Kianoush Kashani^{10,11}, John A. Kellum⁷, Girish Nadkarni^{1,2,12†} and Ankit Sakhuja^{1,2,9*†}

Abstract

Background Major Adverse Kidney Events within 30 days (MAKE30) is an important patient-centered outcome for assessing the impact of acute kidney injury (AKI). Existing prediction models for MAKE30 are static and overlook dynamic changes in clinical status. We introduce ORAKLE, a novel deep-learning model that utilizes evolving time-series data to predict MAKE30, enabling personalized, patient-centered approaches to AKI management and outcome improvement.

Methods We conducted a retrospective study using three publicly available critical care databases: MIMIC-IV as the development cohort, and SiCdb and eICU-CRD as external validation cohorts. Patients with sepsis-3 criteria who developed AKI within 48 h of intensive care unit admission were identified. Our primary outcome was MAKE30, defined as a composite of death, new dialysis or persistent kidney dysfunction within 30 days of ICU admission. We developed ORAKLE using Dynamic DeepHit framework for time-series survival analysis and its performance against Cox and XGBoost models. We further assessed model calibration using Brier score.

Results We analyzed 16,671 patients from MIMIC-IV, 2665 from SiCdb, and 11,447 from eICU-CRD. ORAKLE outperformed the XGBoost and Cox models in predicting MAKE30, achieving AUROCs of 0.84 (95% CI: 0.83–0.86) vs. 0.81 (95% CI: 0.79–0.83) vs. 0.80 (95% CI: 0.78–0.82) in MIMIC-IV internal test set, 0.83 (95% CI: 0.81–0.85) vs. 0.80 (95% CI: 0.78–0.83) vs. 0.79 (95% CI: 0.77–0.81) in SiCdb, and 0.85 (95% CI: 0.84–0.85) vs. 0.83 (95% CI: 0.83–0.84) vs. 0.81 (95% CI: 0.80–0.82) in eICU-CRD. The AUPRC values for ORAKLE were also significantly better than that of XGBoost and Cox models. The Brier score for ORAKLE was 0.21 across the internal test set, SiCdb, and eICU-CRD, suggesting good calibration.

Conclusions ORAKLE is a robust deep-learning model for predicting MAKE30 in critically ill patients with AKI that utilizes evolving time series data. By incorporating dynamically changing time series features, the model captures the evolving nature of kidney injury, treatment effects, and patient trajectories more accurately. This innovation facilitates tailored risk assessments and identifies varying treatment responses, laying the groundwork for more personalized and effective management approaches.

Keywords Acute kidney injury, Major acute kidney events, Deep learning, Survival analysis, Time series

[†]Girish Nadkarni, Ankit Sakhuja have equal contribution as senior author.

*Correspondence:

Ankit Sakhuja

ankit.sakhuja@mssm.edu

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Introduction

Acute kidney injury (AKI) is common among critically ill patients with sepsis. It is associated with high mortality, long-term kidney dysfunction, and a substantial burden on the healthcare system [1–3]. To address the challenges in AKI research and direct the focus towards patient-centered outcomes, the National Institute of Diabetes and Digestive and Kidney Diseases workgroup on Clinical Trials in Acute Kidney Injury recommended a composite outcome measure in 2012 [4]. This measure, encompassing death, dialysis initiation, or a sustained increase in serum creatinine, is widely recognized as Major Adverse Kidney Events (MAKE). The MAKE metric provides a valuable benchmark for assessing the impact of AKI on patients.

Despite considerable progress in understanding the pathophysiology of AKI, clinical trials for its management have largely failed to demonstrate efficacy. Advancing both novel and existing treatment strategies requires the ability to accurately identify patients at high risk for meaningful endpoints, such as MAKE. Current management approaches rely on optimizing fluid status, avoiding nephrotoxins, controlling underlying etiologies (e.g., sepsis), and maintaining hemodynamic stability. Our prior work has demonstrated that these approaches can be personalized to better address patient-specific needs by leveraging advanced data-driven methods to refine clinical decision-making [5]. By accurately predicting outcomes like MAKE, we can not only guide personalization of care but also create enriched cohorts for clinical trials, improving the evaluation of both new and existing interventions.

However, current predictive models for MAKE rely on static, predefined time points [6, 7], limiting their utility in dynamic and complex clinical environments. These models fail to capture the fluctuating nature of kidney function [8] and critical illness, where patient conditions can evolve rapidly. This study seeks to bridge this gap by introducing ORAKLE (Optimal Risk prediction for mAKE30 in patients with acute Kidney injury using deep LEarning), a novel deep learning model to predict the development of MAKE within 30 days (MAKE30). ORAKLE leverages time-series data and event-time modeling to capture the temporal dynamics of patient characteristics and provide robust, individualized risk predictions. By incorporating these advanced techniques, ORAKLE has the potential to revolutionize AKI outcome prediction, enabling dynamic, patient-centered approaches to clinical management and improving outcomes in this high-risk population.

Materials and methods

Study design and data sources

We conducted a retrospective observational study utilizing three publicly available critical care databases—Medical Information Mart for Intensive Care (MIMIC-IV) version 3.1, Salzburg Intensive Care Database (SICdb), and eICU Collaborative Research Database (eICU-CRD) version 1.0.8. The MIMIC-IV [9] database comprises electronic health records of critically ill patients admitted to the intensive care units (ICUs) at Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA, between 2008 and 2022. The SICdb [10] includes ICU patient records from University Hospital Salzburg, Austria, spanning 2013 to 2021. The eICU-CRD [11] encompasses de-identified electronic health records of patients admitted to 208 ICUs across the United States during 2014–2015. The dataset provides insights into ICU practices and outcomes across diverse healthcare settings. All datasets were harmonized to the MIMIC-IV framework by mapping corresponding variables and standardizing definitions, and measurement units. Discrepancies were resolved through conversion to MIMIC-IV standards, ensuring clinical consistency across datasets.

We divided the MIMIC-IV dataset into two subsets—training set (80%) and internal test set (20%). We utilized SICdb and eICU-CRD databases for independent external validation to evaluate the generalizability of the model.

Study population

We included all adult (age ≥ 18 years), critically ill patients with sepsis, who developed AKI within the first 48 h of ICU admission. We defined sepsis based on the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) [12], and AKI based on the Kidney Disease Improving Global Outcomes (KDIGO) [13] guidelines. We identified patients with sepsis based on a combination of suspected infection, and increase in Sequential Organ Failure Assessment (SOFA) [14] score by ≥ 2 within a 24-h period [8]. Consistent with prior studies [12, 15], we assumed a SOFA score of zero prior to ICU admission. We defined suspicion of infection by concurrent administration of intravenous antibiotics and collection of blood cultures. Specifically, blood cultures must have been collected within 24 h after intravenous antibiotics were initiated, or intravenous antibiotics must have been administered within 72 h after the collection of blood cultures [16]. We determined sepsis as the earlier of the suspicion of infection time or SOFA time, provided that the SOFA time occurred no more than 24 h before or 12 h after the suspicion of infection [17]. As SICdb and eICU-CRD databases provide limited data on fluid samplings, we relied on the occurrence of antibiotic

administrations to define suspicion of infection, a method previously validated in literature [18].

As per KDIGO guidelines, we defined AKI as an increase in serum creatinine of 0.3 mg/dL or more within 48 h, or a rise of at least 1.5 times the baseline serum creatinine within 7 days [13]. To estimate baseline serum creatinine, we first determined the median serum creatinine value measured during the 12 months preceding hospital admission [8, 19]. For patients without measurements during this period, we estimated serum creatinine using the Modification of Diet in Renal Disease formula by assuming an estimated glomerular filtration rate of 75 mL/min per 1.73 m² [13]. Consistent with earlier research, the baseline serum creatinine was defined as the lower of either the measured or back-calculated value and the first serum creatinine obtained upon admission [8, 19]. We excluded patients with a history of end-stage kidney disease or prior kidney transplantation.

Outcomes

The primary outcome of interest was major adverse kidney events by 30 days (MAKE30) [20, 21], a composite measure defined by (1) death, (2) initiation of new dialysis, or (3) final inpatient serum creatinine $\geq 200\%$ of the baseline serum creatinine censored at the earlier of hospital discharge or 30 days after ICU admission.

Features

We extracted a comprehensive set of predictor variables from the electronic health records, spanning multiple domains of patient information (Table 1 and Supplementary Table S1). These included demographic

characteristics (e.g., age, sex, race/ethnicity), vital signs (e.g., heart rate, respiratory rate), laboratory test results (e.g., serum creatinine, blood urea nitrogen [BUN]), severity scores (e.g., AKI stage, Sequential Organ Failure Assessment [SOFA] [14] score), and treatment variables (e.g., mechanical ventilation, vasopressor administration).

Only features present in more than 70% of the cohorts were included. Clinical features were summed or averaged within each 8-h time interval as appropriate. Predictors were constructed for time windows extending from AKI onset (baseline) up to 7 days post-onset. Physiologically implausible values were excluded based on clinical expertise, as detailed in Supplementary Table S2, which outlines the thresholds used and the frequencies of values observed during the study periods. In accordance with standard practices for handling missing data in these datasets, forward fill imputation was applied to all features, except for medications, fluid intake, and urine output which were assigned a value of zero when missing. To address any remaining missing values, we applied multivariate imputation by chained equations (MICE) [22].

ORACLE: model development and validation

We used the framework of Dynamic DeepHit [23] to capture the temporal dynamics of patient characteristics. ORACLE, thus, uses a deep learning survival analysis framework designed to predict time-to-event outcomes while accommodating time-dependent covariates. It has two key subnetworks: (1) a time-series subnetwork, implemented as a Long Short-Term Memory (LSTM)

Table 1 Baseline characteristics of patients

	MIMIC-IV: training set (N = 13,336)	MIMIC-IV: test set (N = 3335)	SICdb (N = 2665)	eICU-CRD (N = 11,447)	Total (N = 30,783)	p value
Age, year (mean \pm SD)	67.51 \pm 15.61	67.66 \pm 15.72	66.52 \pm 15.34	66.22 \pm 15.25	66.96 \pm 15.48	< 0.001
Sex—Male, n (%)	7782 (58.4%)	1939 (58.1%)	1656 (62.1%)	6402 (55.9%)	17,779 (57.8%)	< 0.001
Race/ethnicity, n (%)						
White	8893 (66.7%)	2198 (65.9%)	2665 (100.0%)	9302 (81.3%)	23,058 (74.9%)	< 0.001
Black	1248 (9.4%)	295 (8.8%)	0 (0.0%)	1200 (10.5%)	2743 (8.9%)	
Hispanic	423 (3.2%)	107 (3.2%)	0 (0.0%)	172 (1.5%)	702 (2.3%)	
Other	2772 (20.8%)	735 (22.0%)	0 (0.0%)	773 (6.8%)	4280 (13.9%)	
Baseline serum creatinine, mg/dL (mean \pm SD)	1.26 \pm 1.21	1.24 \pm 1.18	0.87 \pm 0.18	0.89 \pm 0.17	1.09 \pm 0.91	< 0.001
First AKI stage, n (%)						
	11,043 (82.8%)	2769 (83.0%)	2321 (87.1%)	6770 (59.1%)	22,903 (74.4%)	< 0.001
	1292 (9.7%)	306 (9.2%)	134 (5.0%)	2063 (18.0%)	3795 (12.3%)	
	1001 (7.5%)	260 (7.8%)	210 (7.9%)	2614 (22.8%)	4085 (13.3%)	
SOFA score, (mean \pm SD)	5.36 \pm 2.84	5.34 \pm 2.87	6.06 \pm 2.58	5.69 \pm 3.12	5.54 \pm 2.94	< 0.001

Legend: The development cohort, MIMIC-IV, was divided into 80% training and 20% testing sets. The external validation cohorts include SICdb and eICU-CRD

network to model and project multivariate time-series observations effectively; (2) a cause-specific subnetwork, consisting of a fully connected network that outputs the probability of each cause-specific outcome at a given estimation time. The model parameters are trained by maximizing a composite likelihood function, that includes three key components: (1) the likelihood for cause-specific cumulative incidence function, defined as the cumulative probability of observing an event at a specific time divided by the marginal probability of observing an event across all times; (2) the likelihood for ranking cause-specific outcomes, based on the assumption that patients experiencing an event at a particular time should exhibit a higher predicted risk than those who do not; and (3) the likelihood for the time-series model, derived from the product of differences between actual and predicted time-varying covariates. This dual-subnetwork structure enables ORAKLE to capture time-varying patterns

in multivariate time-series data and model changes in risk over time simultaneously. Unlike traditional survival models, such as cox regression, which assume proportional hazards or static predictors, ORAKLE captures time-dependent risks dynamically, enabling precise and adaptable risk predictions at any given time point.

We trained ORAKLE using time-series features available at specific time horizons relative to AKI onset. At each time horizon, the time-series features were aggregated and used as input to the model. Predictions were made at 8-h intervals, starting from AKI onset, and were censored at the earliest of the following events: 7 days post-AKI, transfer out of the ICU, or the development of MAKE30 (Fig. 1a). Our neural network architecture consisted of three Long Short-Term Memory (LSTM) layers where each comprising 100 hidden nodes, followed by three fully connected layers, also containing 100 hidden nodes each. To prevent overfitting, we trained the model

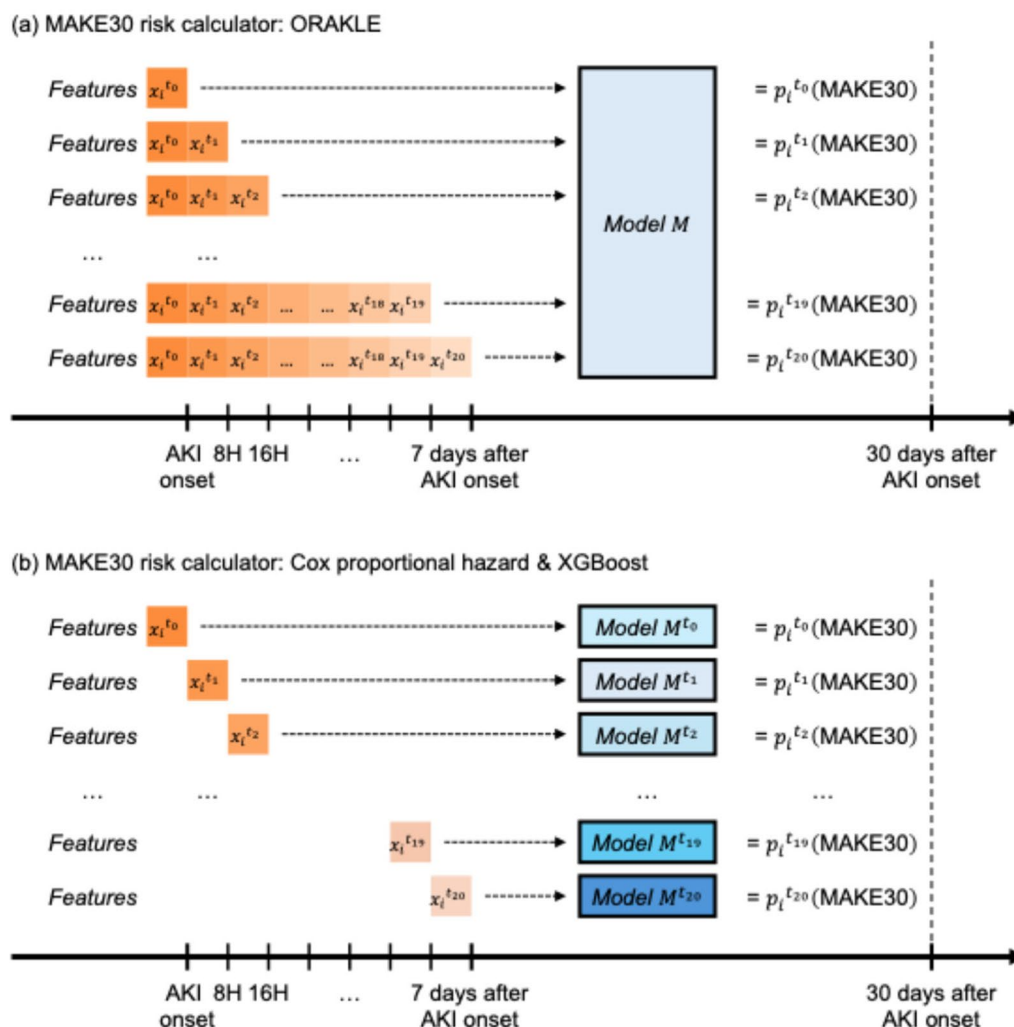


Fig. 1 Overview of study design

on a training dataset derived from MIMIC-IV and subsequently evaluated performance on an independent test dataset, also from MIMIC-IV. The model was iteratively trained until the likelihood function converged.

To evaluate the performance of our MAKE30 prediction model, we compared it against the Cox proportional hazard model and XGBoost model as benchmarks. Since the Cox and XGBoost models do not inherently support time-series data, we built separate Cox and XGBoost models for each time of prediction. At each specified time point, the models used the available features to estimate the risk of MAKE30 (Fig. 1b).

Statistical analysis

We have reported mean values with standard deviations for continuous features and proportions for categorical variables. We examined the difference across cohorts (MIMIC-IV, SICdb, and eICU-CRD) using analysis of variance (ANOVA) for continuous variables and chi-square tests for categorical variables, with a P value of <0.05 considered statistically significant across all statistical tests.

We used the C-index to compare ORAKLE and Cox proportional hazards model for survival analysis across each prediction time point. We further compared the performance of ORAKLE against Cox proportional hazards model and XGBoost at each prediction time point using two key metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). AUROC measures the overall discrimination ability by plotting the true positive rate against the false positive rate, offering a comprehensive evaluation of model performance across different thresholds. AUPRC measures the trade-off between precision (the proportion of true positive predictions out of all positive predictions) and recall (the proportion of true positives correctly identified out of all actual positives). We have also reported the macro-averaged AUROC and AUPRC for MAKE30 outcomes, aggregating performance across all prediction time points [24]. We used 10,000 bootstrap resampling to estimate 95% confidence intervals and determine the significance of differences between outcomes.

Additionally, we assessed the calibration performance of the proposed ORAKLE model using the Brier Score, a metric that evaluates the accuracy of probabilistic predictions. To estimate 95% confidence intervals and assess the significance of differences between outcomes, we employed 10,000 bootstrap resampling iterations.

To allow for explainability of predictions of the model, we assessed for feature importance using SHapley Additive exPlanations (SHAP) values. Mean absolute SHAP values quantify the influence of individual predictors on

the outcome (MAKE30), with higher values indicating greater importance in model predictions and lower values reflecting lesser influence.

All statistical analyses were performed using Python, version 3.6.13, with TensorFlow, version 1.13, and R software, version 4.3.0. We followed the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) AI guideline [25] for reporting clinical prediction models study, as detailed in Supplementary Table S3.

Results

Study population

This study employed three distinct databases to develop and validate, ORAKLE, a deep learning model to predict major adverse kidney events by 30 days (MAKE30) in a continuous fashion. From the MIMIC-IV database, 16,671 patients satisfied the inclusion criteria, with 13,336 (80%) used for model development and 3335 (20%) as internal test cohort (Supplementary Fig. S1). External validation was conducted using two independent cohorts: 2665 patients from the SICdb database and 11,447 patients from the eICU-CRD database. The mean age of patients in the MIMIC-IV cohort was 67.5 years, with 58.3% male. The SICdb cohort had a mean age of 66.5 years, with 62.1% males. In comparison, the eICU-CRD cohort had a mean age of 66.2 years, with 55.9% males. A detailed summary of baseline characteristics for all cohorts is presented in Table 1 and Supplementary Table S1. The overall incidence of MAKE30 was 29% in MIMIC-IV, 26% in SiCdb and 39% in eICU-CRD (Table 2).

Performance of ORAKLE

ORAKLE consistently demonstrated superior performance compared to Cox proportional hazard models across all cohorts. Figure 2 illustrates C-index at different prediction horizons, with detailed values provided in Supplementary Table S4. When predictions were made at the time of AKI onset, ORAKLE achieved C-index of 0.78 (95% CI: 0.77–0.80) for MIMIC-IV, 0.77 (95% CI: 0.79–0.81) for SICdb, and 0.78 (95% CI: 0.79, 0.80) for eICU-CRD, whereas Cox models achieved C-index of 0.75 (95% CI: 0.73–0.77) for MIMIC-IV, 0.76 (95% CI: 0.74–0.78) for SICdb, and 0.77 (95% CI: 0.76–0.78) for eICU-CRD, respectively. This superior performance was sustained across the entire prediction horizon; notably, at a prediction horizon of 7 days post-AKI onset, ORAKLE maintained concordance of 0.87 (95% CI: 0.86–0.89) for MIMIC-IV, 0.86 (95% CI: 0.84–0.88) for SICdb, 0.88 (95% CI: 0.87–0.88) for eICU-CRD, compared to Cox models of 0.83 (95% CI: 0.80–0.85) for MIMIC-IV, 0.83 (95% CI: 0.81–0.85) for

Table 2 Detailed distribution of individual components of MAKE30 outcomes

	MIMIC-IV: training set (N=13,336)	MIMIC-IV: test set (N=3335)	SICdb (N=2665)	eICU-CRD (N=11,447)	Total (N=30,783)	p value
MAKE30	3737 (28.0%)	938 (28.1%)	761 (28.6%)	4402 (38.5%)	9838 (32.0%)	< 0.001
Death	2487 (18.6%)	614 (18.4%)	303 (11.4%)	2151 (18.8%)	5555 (18.0%)	< 0.001
Initiation of new dialysis	452 (3.4%)	129 (3.9%)	112 (4.2%)	334 (2.9%)	1027 (3.3%)	< 0.001
Final inpatient serum creatinine ≥ 200% of the baseline serum creatinine	2025 (15.2%)	527 (15.8%)	467 (17.5%)	1904 (16.6%)	4923 (16.0%)	< 0.001

Legend 1: The development cohort, MIMIC-IV, was divided into 80% training and 20% testing sets. The external validation cohorts include SICdb and eICU-CRD

Legend 2: The individual components of the MAKE30 outcomes are not mutually exclusive. For instance, a patient may have died within 30 days following ICU transfer and simultaneously had a last serum creatinine value greater than 200% of their baseline level

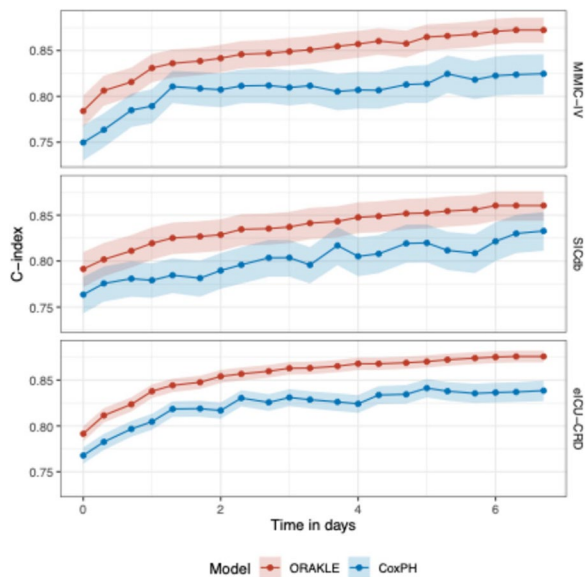


Fig. 2 Comparative C-index for predicting MAKE30 across time horizons

SICdb, 0.84 (95% CI: 0.83–0.85) for eICU-CRD. Survival curves for MAKE30 events from ORAKLE across MIMIC-IV, SICdb, and eICU-CRD cohorts are presented in Fig. 3.

In addition, ORAKLE consistently outperformed Cox proportional hazard and XGBoost models in prediction of MAKE30 when evaluated using AUROC and AUPRC (Figs. 4 and 5, and Supplementary Tables S5 and S6). For the internal test set, ORAKLE achieved a macro-averaged AUROC of 0.84 (95% CI: 0.83–0.86), which was significantly higher than the Cox model's macro-averaged AUROC of 0.80 (95% CI: 0.78–0.82) and XGBoost model's macro-averaged AUROC of 0.81 (95% CI: 0.79–0.83). This performance advantage was consistent across external validation cohorts. In the

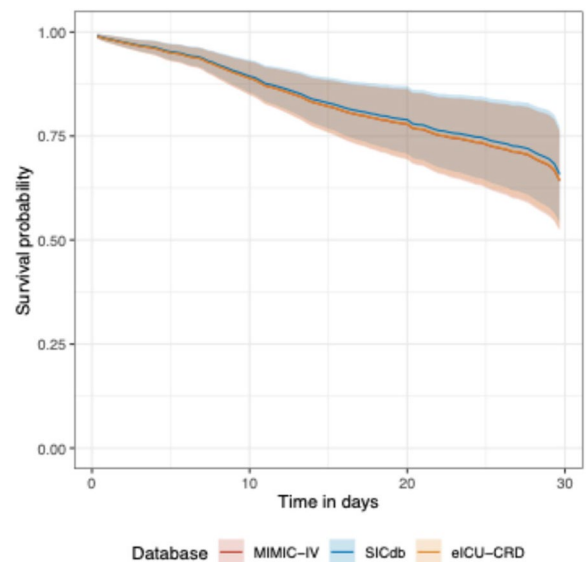


Fig. 3 Survival curves for MAKE30 across cohorts

SICdb cohort, ORAKLE achieved an AUROC of 0.83 (95% CI: 0.81–0.85) compared to 0.79 (95% CI: 0.77–0.81) for the Cox model and to 0.80 (95% CI: 0.78–0.83) for XGBoost model, while in the eICU-CRD cohort, ORAKLE achieved an AUROC of 0.85 (95% CI: 0.84–0.85) compared to 0.81 (95% CI: 0.80–0.82) for the Cox model and to 0.83 (95% CI: 0.83–0.84) for the XGBoost model. All differences were statistically significant.

At the onset of AKI, the ORAKLE demonstrated clear advantages in predictive performance compared to the Cox proportional hazard model. In the internal test set, ORAKLE achieved an AUROC of 0.79 (95% CI: 0.77–0.80), significantly outperforming the Cox model's AUROC of 0.75 (95% CI: 0.73–0.77), and the XGBoost model's AUROC of 0.75 (0.73, 0.77). This trend persisted in external validation cohorts, where ORAKLE achieved an AUROC of 0.79 (95% CI: 0.77–0.81) in the SICdb cohort compared to 0.76 (95% CI: 0.74–0.78) for the Cox

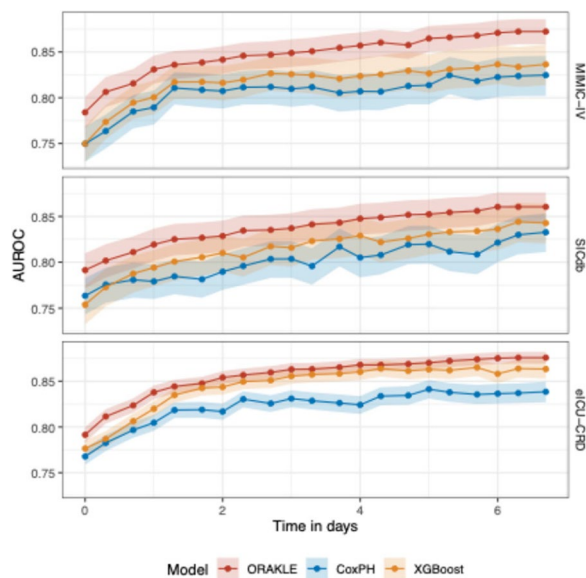


Fig. 4 Comparative AUROC for predicting MAKE30 across time horizons

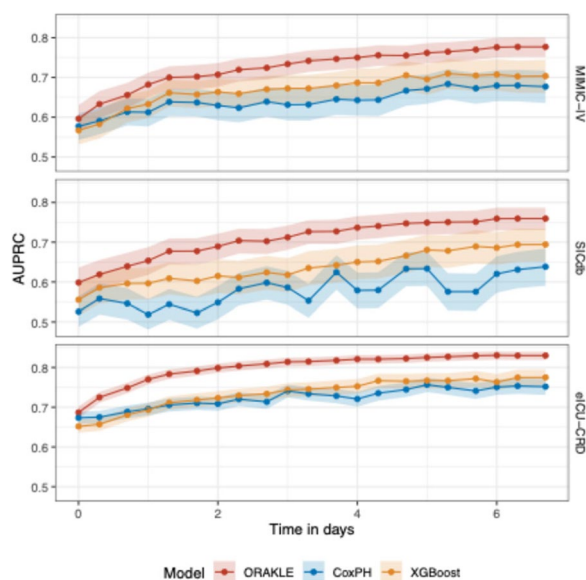


Fig. 5 Comparative AUPRC for predicting MAKE30 across time horizons

model and to 0.76 (0.74, 0.78) for the XGBoost model, and an AUROC of 0.79 (95% CI: 0.78–0.80) in eICU-CRD cohort compared to 0.77 (95% CI: 0.76–0.78) for the Cox model and to 0.78 (0.77, 0.79) for the XGBoost model.

The model performance steadily improved predictive performance as the prediction time point moved further from the onset of AKI. Both ORAKLE and the Cox proportional hazards model showed increasing AUROC values as the prediction window progressed from AKI onset

to 7 days post-onset across both the internal test set and external validation cohorts. However, the performance gap between ORAKLE and the two models persisted throughout this period, highlighting ORAKLE's superior ability to utilize extended time-series data for more accurate and dynamic risk prediction.

ORAKLE also consistently outperformed the Cox proportional hazards model in terms of AUPRC. The macro-averaged AUPRC for ORAKLE was 0.74 (95% CI: 0.71–0.76), 0.71 (95% CI: 0.68–0.74), and 0.80 (95% CI: 0.79–0.81) in the MIMIC-IV, SICdb, and eICU-CRD cohorts, respectively, compared to 0.64 (95% CI: 0.60–0.68), 0.58 (95% CI: 0.54–0.62), and 0.72 (95% CI: 0.71–0.74) for the Cox model and compared to 0.65 (95% CI: 0.62–0.69), 0.62 (95% CI: 0.58–0.66), and 0.72 (95% CI: 0.70–0.74) for the XGBoost model. At the onset of AKI, the AUPRC for ORAKLE was 0.61 (95% CI: 0.58–0.65) in the internal test set, compared to 0.58 (95% CI: 0.54–0.61) for the Cox model ($p < 0.01$) and to 0.57 (95% CI: 0.53–0.60) for the XGBoost model ($p < 0.01$). While all models demonstrated improved performance as the prediction time point moved further from AKI onset, ORAKLE exhibited more substantial gains, reflecting its enhanced ability to utilize time-series data. By day 7, the ORAKLE achieved an AUPRC of 0.79 (95% CI: 0.77–0.81) vs. 0.68 (95% CI: 0.64–0.72) for the Cox model and 0.70 (95% CI: 0.66–0.74) for the XGBoost model in internal test set, with p values < 0.01 for all comparisons (Fig. 5 and Supplementary Table S6).

Similar trends were observed in external cohorts. At AKI onset, ORAKLE's AUPRC was 0.60 (95% CI: 0.56–0.64) in the SICdb cohort and 0.69 (95% CI: 0.67–0.70) in the eICU-CRD cohort, compared to 0.53 (95% CI: 0.49–0.56) and 0.67 (95% CI: 0.66–0.69) for the Cox model and compared to 0.56 (95% CI: 0.52–0.59) and 0.65 (95% CI: 0.64–0.67) for the XGBoost model, respectively. By day 7, ORAKLE achieved an AUPRC of 0.76 (95% CI: 0.73–0.79) in the SICdb cohort and 0.83 (95% CI: 0.82–0.84) in the eICU-CRD cohort, outperforming the Cox model's AUPRC of 0.64 (95% CI: 0.59–0.68) and 0.75 (95% CI: 0.73–0.77) and the XGBoost model's AUPRC of 0.69 (95% CI: 0.65–0.73) and 0.77 (95% CI: 0.75–0.80), respectively.

Finally, we evaluated the calibration of the ORAKLE using the Brier Score. The Brier score for ORAKLE was 0.21 (95% CI: 0.21–0.21) in internal test set, 0.21 (95% CI: 0.20–0.21) in SICdb and 0.21 (95% CI: 0.20–0.21) in eICU-CRD. Figure 6a shows the model's Calibration Plot and Fig. 6b illustrates the Brier Scores for ORAKLE across three cohorts over time.

Feature importance

At the time of prediction, AKI stage emerged as the most influential variable for predicting MAKE30, followed

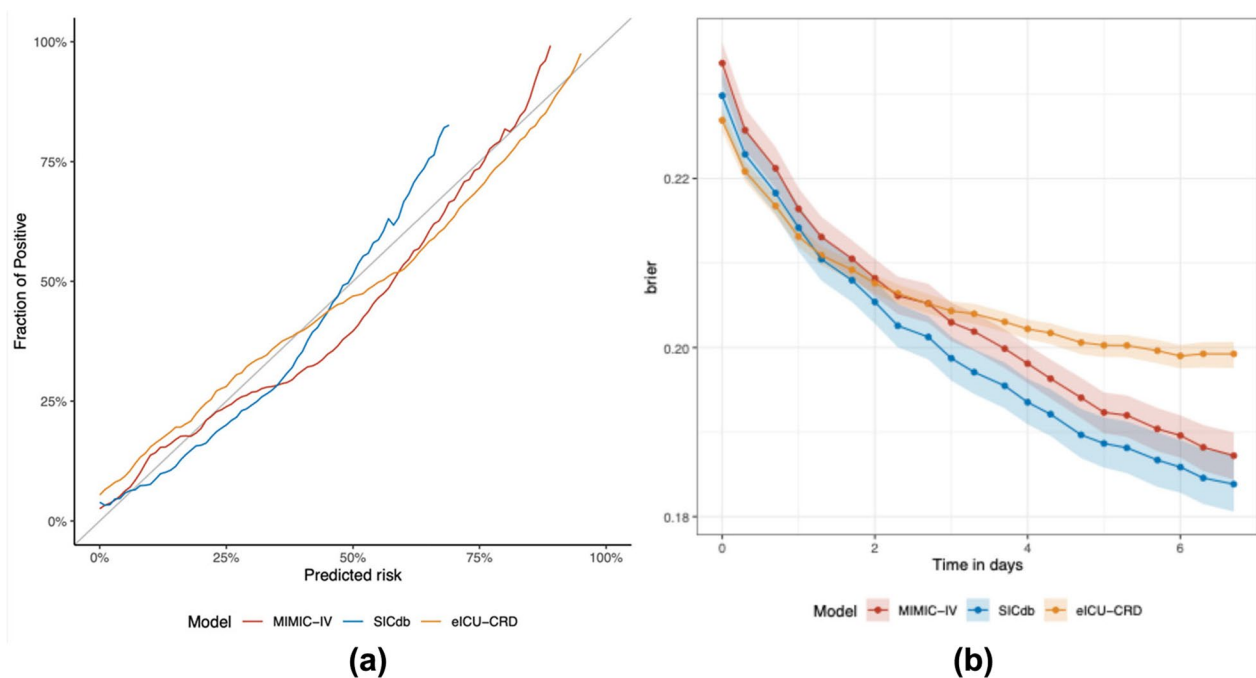


Fig. 6 ORAKLE's **a** Calibration plot, **b** Brier scores over time

by age, patient weight, serum creatinine, and the SOFA score (Fig. 7).

Discussion

In this study, we developed ORAKLE, a deep learning model designed for dynamic prediction of MAKE30, leveraging the evolving clinical characteristics of patients over time. The model generates predictions every 8 h, starting at AKI onset and continuing up to the first 7 days post-AKI. This 8-h interval balances the need for timely updates with sufficient time for clinicians to make informed decisions, ensuring clinical relevance. ORAKLE demonstrated excellent performance, with high AUROC and AUPRC values, and has been validated across diverse datasets—a heterogeneous U.S. cohort and a European ICU cohort. This robust validation underscores the model's generalizability and potential to support dynamic, actionable decision-making in varied clinical settings.

Most predictive models in the field of AKI have focused primarily on predicting the development of AKI itself [26]. However, among patients who already have AKI, it is crucial to identify those at high risk for adverse outcomes such as MAKE. Understanding this risk is vital for tailoring existing therapies to individual patient needs and for selecting enriched patient samples for clinical trials of new or existing interventions. The KDIGO guidelines provide a framework for AKI management, and their implementation has been shown to improve outcomes in randomized controlled trial setting [27]. However,

compliance with these guidelines is notably low in routine clinical practice [28, 29], largely due to their non-specific nature and the challenge of determining which patients would benefit most from specific components of the recommendations. Our prior work has demonstrated that specific consensus-based recommendations for management of AKI can be developed for subgroup of patients [30], demonstrating that this limitation in guidelines can be addressed. Nonetheless, the critical next step lies in identifying the differential effects of treatments for individual patients. Achieving this requires dynamic, time-sensitive risk assessment for adverse outcomes such as MAKE30. By enabling this evolving risk assessment, we can not only refine personalized treatment strategies but also optimize the evaluation of therapies, paving the way for truly individualized AKI management.

There is a paucity of studies focused on predicting MAKE, and none of the available models incorporate evolving time-series data into their predictions. For instance, Neyra and colleagues compared the performance of various machine learning algorithms such as logistic regression, random forest, support vector machine, and extreme gradient boosting to predict MAKE at 120 days [7]. While their model demonstrated good performance with an AUC of 0.73 in an external validation cohort, it did not account for time-varying covariates and was limited to critically ill adult patients during the first 3 days of hospitalization, restricting its generalizability. Similarly, Xiaoyuan and colleagues developed

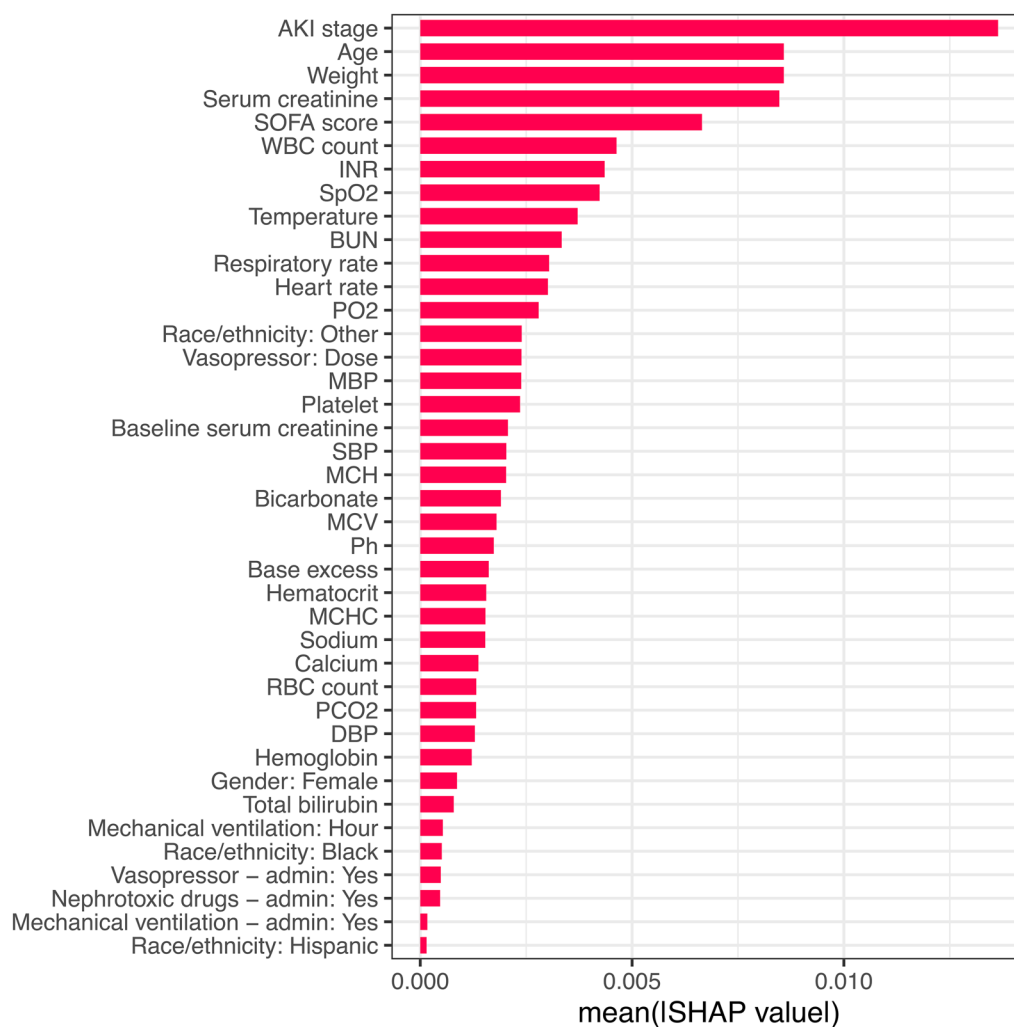


Fig. 7 Feature importance for MAKE30 in the ORAKLE. Legend: Features: Age, Race/ethnicity: Black, Race/ethnicity: Hispanic, Race/ethnicity: Other, Gender: Female, Heart rate, SBP, DBP, MBP, Respiratory rate, Temperature, SpO2, Weight, Hematocrit, Hemoglobin, MCH, MCHC, MCV, Platelet, RBC count, WBC count, Bicarbonate, BUN, Calcium, Serum creatinine, Sodium, INR, Total bilirubin, PO2, PCO2, Ph, Base excess, Baseline serum creatinine, AKI stage, SOFA score, Vasopressor—admin: Yes, Vasopressor: Dose, Mechanical ventilation—admin: Yes, Mechanical ventilation: Hour, Nephrotoxic drugs—admin: Yes

an early warning model for MAKE30 using a nomogram based on variables collected within the first 24 h of ICU admission [31]. Their model exhibited good predictive performance, with AUC of 0.82 in external validation cohort but was limited to predictions within first day of ICU admission and did not incorporate temporally evolving data. This is particularly critical in the ICU, where a patient's clinical condition is highly dynamic. Our prior work has demonstrated that changes in creatinine over time are a significant predictor of both acute kidney disease and mortality [8]. Therefore, capturing this dynamic evolution in clinical characteristics is essential for accurately predicting outcomes in AKI patients. Unlike static models that rely on fixed data points captured at specific

intervals, ORAKLE employs a double subnetwork architecture designed for dynamic prediction. It utilizes an LSTM network to process multivariate time-series data, coupled with a fully connected cause-specific network that estimates probabilities for each outcome at a given time. This innovative architecture provides a deeper and more nuanced understanding of AKI progression, effectively addressing the limitations inherent in static prediction models.

LSTMs, are a type of recurrent neural network (RNN) designed to handle time-series data by addressing the vanishing and exploding gradient problems of traditional RNNs [32, 33]. Unlike feedforward networks, which process inputs independently, RNNs maintain “memory”

through hidden states, enabling them to model temporal and sequential dependencies. However, standard RNNs often struggle to retain long-term information. LSTMs resolve these limitations through their unique architecture, which includes memory cells and gating mechanisms. These components enable the selective retention, updating, and removal of information, allowing LSTMs to maintain relevant context over extended sequences. This capability makes them particularly effective for tasks like time-series prediction. In clinical applications, LSTMs excel at modeling time-dependent health data such as electronic health records (EHRs) and continuous monitoring outputs, integrating both abstract and temporally distant variables [34]. Unlike traditional models like Cox regression and XGBoost, which struggle to model dynamic relationships over time, LSTMs are specifically designed to handle sequential data with changing patterns. This is especially true when the changes are nonlinear or when there are intricate, time-dependent dependencies that are not well captured by the Cox and XGBoost models. This makes LSTMs a superior choice for modeling outcomes in diseases like AKI, where kidney function can vary rapidly and treatment decisions need to be adjusted continuously based on the evolving patient data.

While ORAKLE demonstrates strong performance, it is important to acknowledge its limitations. First, as a retrospective analysis, it is inherently susceptible to biases, such as selection bias and potential confounding factors, which could influence the interpretation of the results. However, the study demonstrated good generalizability, as evidenced by its strong performance in both internal and two distinct external validation datasets. Second, we did not specifically adjust baseline serum creatinine for patients with chronic kidney disease (CKD), consistent with previous literature [8, 35–37]. This may lead our model to slightly overestimate MAKE30 risk in these patients. An alternative approach would be to use admission creatinine as the baseline for those with a documented history of CKD when historical values of creatinine are unavailable [38]. In our data, this adjustment would lower AKI rates by 1.2% in MIMIC-IV, 2.2% in SICdb, and 0.7% in eICU, and decrease MAKE30 rates by 1.6%, 3.7%, and 1.1%, respectively. However, using admission creatinine as baseline could underestimate MAKE30 risk in these patients. As our tool is intended for screening, we favor a conservative approach that slightly overestimates risk rather than underestimates it. Third, while the model shows strong performance, it has yet to be deployed prospectively in real-world clinical environments. While the model provides an easily deployable

framework, the next step is to implement it in clinical practice to evaluate its effectiveness in a dynamic, clinical context. Fourth, our sample consists of patients who developed AKI within the first 48 h of ICU admission, which may limit the generalizability of the findings to all patients with AKI or those who develop AKI outside of the ICU setting. Finally, our AKI definition did not include patients with isolated oliguria, and as a result, the findings may not be applicable to this subgroup of patients.

In conclusion, we have developed a robust model that marks a significant advancement in MAKE30 prediction and AKI management. By integrating dynamically evolving time-series features, the model offers a more accurate reflection of the dynamic nature of kidney injury, treatment responses, and patient progression. This approach enables individualized risk assessments and the identification of differential responses to treatments, paving the way for more personalized and effective management strategies.

Abbreviations

MAKE	Major adverse kidney events
MAKE30	Major adverse kidney events within 30 days
AKI	Acute kidney injury
ORAKLE	Optimal risk prediction for mAKE30 in patients with acute Kidney injury using deep LEarning
MIMIC-IV	Medical information mart for intensive care
SICdb	Salzburg intensive care database
eICU-CRD	EICU collaborative research database
ICU	Intensive care unit
BIDMC	Beth Israel deaconess medical center
SOFA	Sequential organ failure assessment
Sepsis-3	Third international consensus definitions for sepsis and septic shock
MICE	Multivariate imputation by chained equations
LSTM	Long short-term memory
ANOVA	ANalysis of VAriance
AUROC	Area under the receiver operating characteristic curve
AUPRC	Area under the precision-recall curve
SHAP	SHapley additive exPlanations
EHR	Electronic health records
RNN	Recurrent neural network
KDIGO	Kidney disease: improving global outcomes

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13054-025-05457-w>.

Supplementary Material 1.

Acknowledgements

Not applicable.

Author contributions

WO and AS designed and conceptualized the study. WO performed data collection and analysis. WO and AS interpreted the data analyses. WO, MV, and AS wrote the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content and read and approved the final version. All authors gave final approval for the version to be published. WO is first author, GNN and AS and co-senior authors.

Funding

This study was supported by NIH grant K08DK131286 (AS). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Publicly available datasets were analyzed in this study. The MIMIC-IV dataset is available at <https://physionet.org/content/mimiciv/>, the eICU-CRD dataset is available at <https://physionet.org/content/eicu-crd/>, and the SICdb dataset is available at <https://physionet.org/content/sicdb/>. Source code available upon request.

Declarations

Ethics approval and consent to participate

The study received approval from the Institutional Review Board at the Icahn School of Medicine at Mount Sinai (approval no. 19–00951).

Consent for publication

Not applicable.

Competing interests

GNN is a founder of Renalytix, Pensieve, Verici and provides consultancy services to AstraZeneca, Reata, Renalytix, Siemens Healthineer and Variant Bio, serves a scientific advisory board member for Renalytix and Pensieve. He also has equity in Renalytix, Pensieve and Verici. JAK reports receiving consulting fees from Astute Medical/bioMerieux, Astellas, Alexion, Chugai Pharma, Novartis, Mitsubishi Tenabe and GE Healthcare and is a Full-time employee of Spectral Medical. All remaining authors have declared no conflicts of interest.

Author details

¹Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ²Division of Data-Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ³Touro College of Osteopathic Medicine, New York, NY, USA. ⁴Department of Cardiovascular Surgery, Mount Sinai Morningside, New York, NY, USA. ⁵Division of Hospital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁶Improbable AI Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁷Department of Critical Care Medicine, Center for Critical Care Nephrology, University of Pittsburgh, Pittsburgh, PA, USA. ⁸Department of Population Health Science and Policy, Center for Biostatistics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁹Institute for Critical Care Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ¹⁰Division of Nephrology and Hypertension, Mayo Clinic, Rochester, MN, USA. ¹¹Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, USA. ¹²Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Received: 18 January 2025 Accepted: 10 May 2025

Published online: 26 May 2025

References

- Coca SG, Singanamala S, Parikh CR. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. *Kidney Int.* 2012;81:442–8.
- Hoste EA, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med.* 2015;41:1411–23.
- Dasta JF, Kane-Gill SL, Durtschi AJ, Pathak DS, Kellum JA. Costs and outcomes of acute kidney injury (AKI) following cardiac surgery. *Nephrol Dial Transplant.* 2008;23:1970–4.
- Palevsky PM, Molitoris BA, Okusa MD, Levin A, Waikar SS, Wald R, et al. Design of clinical trials in acute kidney injury: report from an Niddk workshop on trial methodology. *Clin J Am Soc Nephrol.* 2012;7:844–50.
- Oh W, Takkavatakarn K, Kittrell H, Shawwa K, Gomez H, Sawant AS, et al. Development and validation of a policy tree approach for optimizing intravenous fluids in critically ill patients with sepsis and acute kidney injury. *Intensive Care Crit Care Med.* 2024. <https://doi.org/10.1101/2024.08.06.24311556>.
- McKown AC, Wang L, Wanderer JP, Ehrenfeld J, Rice TW, Bernard GR, et al. Predicting major adverse kidney events among critically ill adults using the electronic health record. *J Med Syst.* 2017;41:156.
- Neyra JA, Ortiz-Soriano V, Liu LJ, Smith TD, Li X, Xie D, et al. Prediction of mortality and major adverse kidney events in critically ill patients with acute kidney injury. *Am J Kidney Dis.* 2023;81:36–47.
- Takkavatakarn K, Oh W, Chan L, Hofer I, Shawwa K, Kraft M, et al. Machine learning derived serum creatinine trajectories in acute kidney injury in critically ill patients with sepsis. *Crit Care.* 2024;28:156.
- Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10:1.
- Rodemund N, Wernly B, Jung C, Cozowicz C, Koköfer A. Harnessing big data in critical care: exploring a new European dataset. *Sci Data.* 2024;11:320.
- Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data.* 2018;5:180178.
- Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA.* 2016;315:801.
- Kellum JA, Lameire N, Aspelin P, Barsoum RS, Burdmann EA, Goldstein SL, et al. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl.* 2012;2:1–138.
- Vincent J, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Med.* 1996;22:707–10.
- Raith EP, Udy AA, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *JAMA.* 2017;317:290.
- Kamaleswaran R, Lian J, Lin D-L, Molakapuri H, Nunna S, Shah P, et al. Predicting volume responsiveness among sepsis patients using clinical data and continuous physiological waveforms. In: AMIA 2021 annual symposium. 2021.
- Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Crit Care Med.* 2020;48:210–7.
- Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinhäuser N, et al. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *eClinicalMedicine.* 2023;62:102124.
- Peerapornratana S, Priyanka P, Wang S, Smith A, Singbartl K, Palevsky PM, et al. Sepsis-associated acute kidney disease. *Kidney Int Rep.* 2020;5:839–50.
- Semler MW, Self WH, Wanderer JP, Ehrenfeld JM, Wang L, Byrne DW, et al. Balanced crystalloids versus saline in critically ill adults. *N Engl J Med.* 2018;378:829–39.
- Semler MW, Rice TW, Shaw AD, Siew ED, Self WH, Kumar AB, et al. Identification of major adverse kidney events within the electronic health record. *J Med Syst.* 2016;40:167.
- Buuren SV, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Soft.* 2011;45:1–67.
- Lee C, Yoon J, Schaar MVD. Dynamic-Deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans Biomed Eng.* 2020;67:122–33.
- Poulin-Moore A, Tuzcuoglu K. Forecasting recessions in Canada: an autoregressive probit model approach. Ottawa: Bank of Canada; 2024.
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378.
- Vagliano I, Chesnaye NC, Leopold JH, Jager KJ, Abu-Hanna A, Schut MC. Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal. *Clin Kidney J.* 2022;15:2266–80.
- Zarbock A, Küllmar M, Ostermann M, Lucchese G, Baig K, Cennamo A, et al. Prevention of cardiac surgery-associated acute kidney injury by implementing the Kidgo guidelines in high-risk patients identified by

- biomarkers: the PrevAKI-multicenter randomized controlled trial. *Anesth Analg*. 2021. <https://doi.org/10.1213/ANE.0000000000005458>.
28. Küllmar M, Weiß R, Ostermann M, Campos S, Grau Novellas N, Thomson G, et al. A multinational observational study exploring adherence with the kidney disease: improving global outcomes recommendations for prevention of acute kidney injury after cardiac surgery. *Anesth Analg*. 2020;130:910–6.
 29. Kolhe NV, Staples D, Reilly T, Merrison D, McIntyre CW, Fluck RJ, et al. Impact of compliance with a care bundle on acute kidney injury outcomes: a prospective observational study. *PLoS ONE*. 2015;10:e0132279.
 30. Justice CM, Nevin C, Neely RL, Dilcher B, Kovacic-Scherrer N, Carter-Templeton H, et al. Effect of tiered implementation of clinical decision support system for acute kidney injury and Nephrotoxin exposure in cardiac surgery patients. *Appl Clin Inform*. 2025;16:001–10.
 31. Yu X, Xin Q, Hao Y, Zhang J, Ma T. An early warning model for predicting major adverse kidney events within 30 days in sepsis patients. *Front Med*. 2024;10:1327036.
 32. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9:1735–80.
 33. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput*. 2019;31:1235–70.
 34. Lee JM, Hauskrecht M. Recent context-aware LSTM for clinical event time-series prediction. In: Riaño D, Wilk S, Ten Teije A, editors. *Artificial intelligence in medicine*. Cham: Springer; 2019. p. 13–23. https://doi.org/10.1007/978-3-030-21642-9_3.
 35. Shawwa K, Ghosh E, Lanius S, Schwager E, Eshelman L, Kashani KB. Predicting acute kidney injury in critically ill patients using comorbid conditions utilizing machine learning. *Clin Kidney J*. 2021;14:1428–35.
 36. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med*. 2018;46:1070–7.
 37. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572:116–9.
 38. Al-Jaghbeer M, Dealmeida D, Bilderback A, Ambrosino R, Kellum JA. Clinical decision support for in-hospital AKI. *JASN*. 2018;29:654–60.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.