Causal deep learning for real-time detection of cardiac surgery-associated acute kidney injury: derivation and validation in seven time-series cohorts

Qin Zhong, Yuxiao Cheng*, Zongren Li*, Dongjin Wang*, Chongyou Rao, Yi Jiang, Lianglong Li, Ziqian Wang, Pan Liu, Hebin Che, Pei Li, Xin Lu, Jinli Suo, Kunlun He*

Summary

Background Cardiac surgery-associated acute kidney injury (CSA-AKI) is a complex complication substantially contributing to an increased risk of mortality. Effective CSA-AKI management relies on timely diagnosis and interventions. However, many cases are detected too late. Despite the advancements in novel biomarkers and data-driven predictive models, existing practices are primarily constrained due to the limited discriminative and generalisation capabilities and stringent application requirements, presenting major challenges to the timely and effective diagnosis and interventions in CSA-AKI management. This study aimed to develop a causal deep learning architecture, named REACT, to achieve precise and dynamic predictions of CSA-AKI within the subsequent 48 h.

Methods In this retrospective model development and prospective validation study, we included adult patients (aged ≥18 years) from seven distinct cohorts undergoing major open-heart surgery for model training and validation. Data for model development and internal validation were sourced from electronic health records of two large centres in Beijing, China, between Jan 1, 2000, and Dec 31, 2022. External validation was conducted on three independent centres in China between Jan 1, 2000, and Dec 31, 2022, along with cross-national data from the public databases MIMIC-IV and eICU in the USA. To facilitate implementation, we also developed a publicly accessible web calculator and applet. The model's prospective application was validated from June 1, to Oct 31, 2023, at two centres in Beijing and Nanjing, China.

Findings The final derivation cohort included 14 513 eligible patients with a median age of 56 years (IQR 45–65), 5515 (38·0%) patients were female, and 3047 (21·0%) developed CSA-AKI. The external validation dataset included 20 813 patients from China and 28 023 from the USA. REACT reduced 1328 input variables to six essential causal factors for CSA-AKI prediction. In internal validation, REACT achieved an average area under the receiver operating characteristic curve (AUROC) of 0·930 (SD 0·032), outperforming state-of-the-art deep learning architectures, specifically transformer-based and long short-term memory-based models, which rely on more complex variables. The model consistently outperformed in external validation across different centres (average AUROC 0·920 [SD 0·036]) and regions (0·867 [0·073]), as well as in prospective validation (0·896 [0·023]). Compared with guideline-recommended pathways, REACT detected CSA-AKI on average 16·35 h (SD 2·01) earlier in external validation.

Interpretation We proposed a causal deep learning approach to predict CSA-AKI risk within 48 h, distilling the complex temporal interactions between variables into only a few universal, relatively cost-effective inputs. The approach shows great potential for deployment across hospitals with minimum data requirements and provides a general framework for causal deep learning and early detection of other conditions.

Funding The Construction Project and the National Natural Science Foundation of China.

Introduction

More than 2 million patients undergo cardiac surgery annually for valvular or coronary heart diseases globally.[1] Cardiac surgery-associated acute kidney injury (CSA-AKI) affects one third of these patients and increases mortality risk by two to eight fold.[2] Although the prognosis for patients with CSA-AKI can be substantially improved with timely interventions, observational data from one study involving 12 hospitals reveal that recommended acute kidney injury (AKI) prevention strategies are implemented in less than 10% of patients.[3] This low adherence is likely due to the absence of accurate and actionable intervention alerts in the guidelines.[4]

In clinical practice, AKI diagnosis often relies on creatinine levels. However, kidney damage can manifest before a substantial rise in creatinine occurs. For early AKI

Lancet Digit Health 2025; 7: 100901

Published Online September 25, 2025
https://doi.org/10.1016/j.landig.2025.100901

*Contributed equally

Medical Big Data Research Center, Chinese The People's Liberation Army General Hospital, Beijing, China (Q Zhong PhD, C Rao PhD, P Liu PhD, H Che MSc, P Li PhD, Prof K He PhD); **National Engineering Research Center of Medical Big Data Application Technology, The People's Liberation Army General Hospital, Beijing, China** (Q Zhong, C Rao, P Liu, H Che, P Li, Prof K He); **Medical Artificial Intelligence Research Center, Chinese The People's Liberation Army General Hospital, Beijing, China** (Z Li PhD); **Department of Automation, Tsinghua University, Beijing, China** (Y Cheng BEng, L Li BEng, Z Wang BEng, J Suo PhD); **Department of Cardiovascular Surgery, Nanjing Drum Tower Hospital, Chinese Academy of Medical Science & Peking Union Medical College, Nanjing, China** (Prof D Wang MD, Y Jiang MD); **College of Systems Engineering, National University of Defense Technology, Changsha, China** (Prof X Lu PhD)

Correspondence to:
Dr Jinli Suo, Department of Automation, Tsinghua University, Beijing, 100084, China
jlsuo@tsinghua.edu.cn

or

Prof Kunlun He, Medical Big Data Research Center, The People's Liberation Army General Hospital, Beijing, 100853, China
kunlunhe@plagh.org

Research in context

Evidence before this study

We searched PubMed on June 30, 2024, without language or date restrictions for publications on the development and validation of deep learning-based models for cardiac surgery-associated acute kidney injury. The following terms and related terms were used when searching: ("causal machine learning", "deep learning", "machine learning", "real-time prediction" OR "artificial intelligence") AND ("cardiac surgery associated", "cardiac surgery", "intensive care" OR "critical care") AND ("kidney injury" or "kidney failure" OR "nephropathy" OR "renal failure" OR "renal impairment" OR "renal injury"). We systematically reviewed 232 search results and identified 22 original studies on adult patients undergoing cardiac surgery. Of these, five used deep learning, and three made dynamic predictions; however, none used causal deep learning methods. Most models had limitations such as poor discrimination, reliance on static conditions, or requiring too many input variables, limiting their clinical applicability. The quality assessment indicated that these studies were at a high or unclear risk of bias. Most studies had sample sizes ranging from 500 to 5000 and lacked adequate validation, with only three being externally validated. Of these three, one focused on the cardiac surgery subgroup of the acute aortic syndrome and predicted only severe acute kidney injury (AKI), achieving an area under the receiver operating characteristic curve of 0·81 in an external validation with 319 patients. Another also predicted only severe AKI using publicly available datasets but required input of 52 variables, making it impractical in resource-limited health systems due to its high data input demands. The third predicted moderate-to-severe AKI in 4912 patients but was based on static conditions and could not adapt to dynamic patient changes. Our literature search highlights a major gap in high-performance, reliable, and practical AI models for whole-stage AKI prediction validated on large, multicentre datasets. Real-time prediction of AKI remains the topic with the most substantial evidence.

Added value of this study

The Real-time Evaluation and Anticipation with Causal Distillation (REACT) is a novel temporal causal deep learning architecture designed to predict the risk of cardiac surgery-associated acute kidney injury within the subsequent 48 h. Derivated and validated using the largest cardiac surgery patient time-series database, which includes 63 349 patients with more than 21·5 billion data entries from five independent hospitals and two commonly used public datasets, REACT distilled the complex temporal dynamics among variables into six minimal causal inputs. It achieved an average area under the receiver operating characteristic curve of 0·930 (SD 0·032) across 12 sub-tasks, outperforming models that rely on more complex variables. It efficiently predicts 97% (685/706) of events in internal validation, detecting cardiac surgery-associated acute kidney injury an average of 14·65 h (SD 3·17) earlier than guideline-recommended pathways. We have further validated this approach prospectively on 754 patients across two centres and made REACT publicly accessible via a user-friendly website and applet.

Implications of all the available evidence

With an innovative causal deep learning framework and large, diverse datasets, our model show high performance in both internal and external validation cohorts. REACT is a bedside assessment tool that can be easily implemented in routine clinical practice with minimal input variables during the application, shifting computational intensity and complex variable input to the training phase, thereby increasing the model's generalisability. This tool has the potential to support perioperative management and clinical decision making for cardiac surgery, addressing existing needs for personalised care following open-heart procedures.

detection, researchers have proposed various statistical risk tools,[5–8] such as the Cleveland[9,10] and Mehta scores.[11] Although being widely tested in different cardiac surgery cohorts and acknowledged due to their inherent simplicity, methodological transparency, and ease of implementation, these tools are often of constrained predictive accuracy, particularly for patients who have undergone cardiac surgery with prolonged hospital stays and rapidly evolving clinical conditions.[12] The limited accuracy is mainly attributed to the oversimplification of the intricate, non-linear dynamics of patients' physiological states, resulting in many AKI cases being detected only at the severe stages. Therefore, none of the statistical tools have been expressly recommended by the guidelines.

In the past decade—particularly since 2018—advances in AI techniques, especially deep neural networks, have shown great promise in fields such as internal medicine, ophthalmology, and radiology.[13,14] Several deep learning-based models have been developed to predict AKI, exhibiting technical feasibility with area under the receiver operating characteristic curve (AUROC) values ranging from 0·69 to 0·83.[15–19] However, many of these models suffer from low external generalisability attributed to potential data mismatches between model derivation and diverse clinical environments. Moreover, although neural networks excel at identifying underlying relationships from complex data, they often confuse correlation with causation, raising reliability concerns.[20] In addition, existing neural network-based models require extensive inputs for accurate prediction, limiting their clinical applicability. For example, Tomašev and colleagues[21] built a model using 620 000 features (input variables) and achieved real-time AKI prediction with hundreds of variables. Despite its impressive performance, the model requires extensive data collection, and the absence of any single input can compromise its accuracy.[22] The above attributes render

AI powerful for medical data analysis but simultaneously vulnerable from an application viewpoint.[20–23]

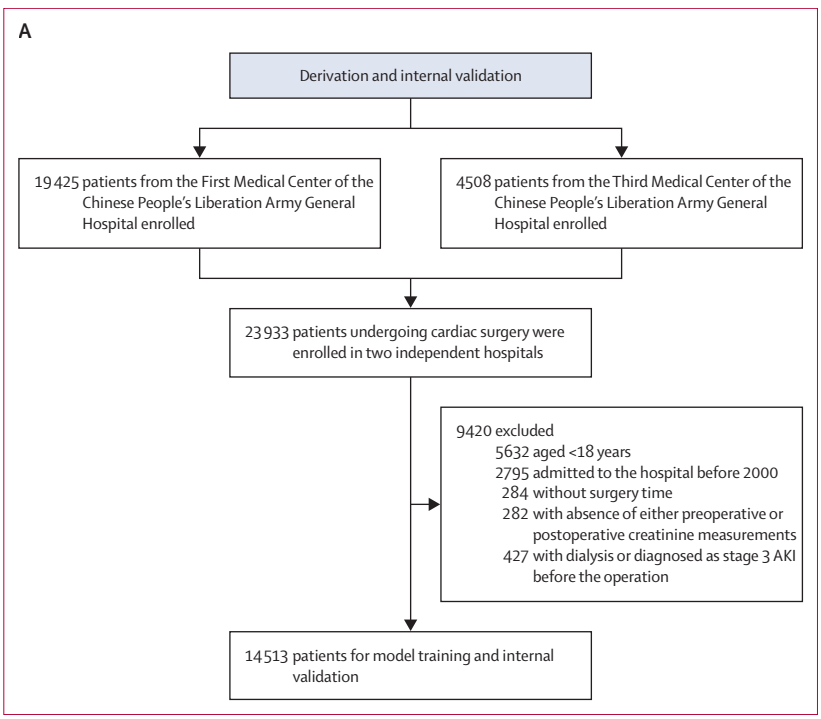
At the convergence of statistics and AI, causal deep learning combines the explainability and simplicity of statistical tools as well as the advanced predictive capabilities of deep neural networks, marking a shift from mere prediction to actionable insights.[24,25] Under this change, we developed a causal deep learning architecture named Real-time Evaluation and Anticipation with Causal Distillation (REACT) to dynamically predict the risk of patients developing CSA-AKI within the following 48 h from a small number of input variables and at arbitrary postoperative timepoints, thus providing physicians with timely intervention opportunities.

Methods

Study design and participants

In this retrospective model development and prospective validation study, we collected data from seven cohorts across different centres and conducted prospective trials on data from two of these centres but in a later time range. Adults who were 18 years or older and were admitted before the year 2000 were included. Patients who underwent major open-heart surgery, including coronary artery bypass grafting, valve replacement or repair, combined valve and artery surgery, aortic surgery, pericardiectomy, or other major cardiac surgical procedures, were included in the study. Patients who were receiving long-term dialysis and required preoperative dialysis, or those diagnosed with AKI stage 3 (ie, severe AKI) at the first test before surgery were excluded. Patients without baseline (preoperative) serum creatinine measurements or postoperative serum creatinine measurements were excluded. Patients with no operative time were excluded. Details of the data preprocessing are provided in figure 1 and the appendix (pp 1, 2).

For model development and internal validation, we consecutively enrolled eligible patients from two large centres: the First and Third Medical Centers of the People's Liberation Army General Hospital in Beijing, China, spanning from Jan 1, 2000, to Dec 31, 2022. For external validation, we consecutively enrolled patients from three independent centres (the Sixth and Seventh Medical Centers of the People's Liberation Army General Hospital and Nanjing Drum Tower Hospital in Nanjing, China) from Jan 1, 2000, to Dec 31, 2022. These data were extracted from an electronic medical system encompassing comprehensive data on demographics, encoded diagnoses, laboratory values, and treatment details. Additionally, to validate the model on a more diverse range of races, we conducted external validation using data from the MIMIC-IV[26] and eICU[27] databases from the USA for international validation. For prospective and application validation, we developed a web-based platform and an applet for REACT. These tools were deployed and tested in patients undergoing cardiac surgery at the First Medical Center and Nanjing Drum Tower Hospital from June to October, 2023.



(Figure 1 continues on next page)

Data from the patients' entire hospitalisation (ie, preoperative and postoperative) were included in the analysis. Input variables encompass static terms such as demographics and comorbidities, and dynamic ones that can undergo substantial changes over time (eg, heart rate and laboratory tests). Laboratory tests were taken at clinically determined timepoints. Input variables were time-aligned for real-time prediction. All the variables documented for at least 50% of the patients were included in the model. We did not use any kind of imputation to keep fidelity to the true observations. Missingness might carry information. For example, certain laboratory tests might be absent because the treating clinician judged that the patient's condition did not warrant the test, meaning that whether the test was performed or not can reflect illness severity, so our model used a missingness indicator to retain that information (figure 2; appendix pp 1, 2).

See Online for appendix

The primary outcome was the prediction of severe (stage 3) CSA-AKI 6–48 h before onset. Secondary outcomes included the prediction of any AKI occurrence (stage 1, 2, or 3) and the prediction of moderate or severe AKI (stage 2 or 3). CSA-AKI was defined according to the modified Kidney Disease: Improving Global Outcomes criteria,[28] which uses serum creatinine trajectories and is further complemented by clinician-diagnosed AKI via ICD-9 or ICD-10 codes, depending on the coding system in use at each site. The most recent preoperative creatinine value was used as baseline. Overall, we set 12 predictive goals, focusing on the prediction of three CSA-AKI categories (any AKI, moderate or severe AKI, and severe AKI) each at 6 h, 12 h, 24 h, and 48 h in the future.

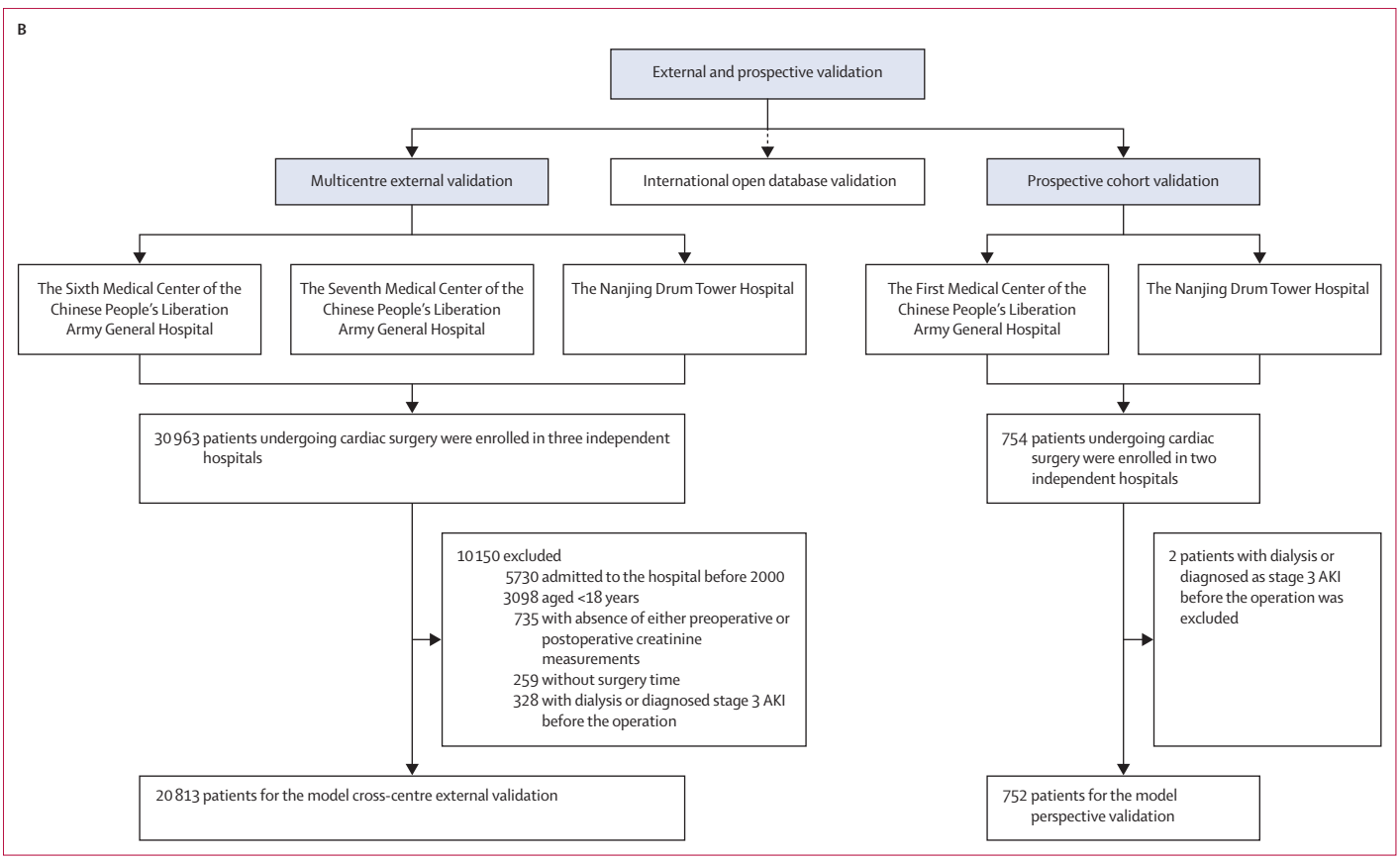


Figure 1: **Study flowchart**
Flowchart illustrating patient selection for model derivation, internal validation, multicentre external validation, and prospective validation. (A) Derivation and internal validation. (B) External and prospective validation. An additional external validation was conducted using publicly available international datasets (MIMIC-IV and eICU). AKI=acute kidney injury.

The retrospective analysis was approved by the institutional review board of the Chinese People's Liberation Army General Hospital (S2021-305-01) and Nanjing Drum Tower Hospital (S2020-281-01), with a waiver of informed consent. The prospective application phase of the website and decision support system received approval from the ethics committee (S2022-281-01 and S2021-305-01). Written informed consent was obtained from all prospective patients. We adhered to the TRIPOD guideline and the Ensuring Fairness in Machine Learning to Advance Health Equity checklist to report prediction models (appendix pp 14–17).

The casual deep learning model: REACT

The literature has widely discussed that deep learning, when based on associations rather than causations, could lead to unstable predictions that struggle to be generalised effectively (appendix p 4).[29] To address this issue, we proposed REACT, which imposes an explicit causal graph tailored for time-series data[30,31] onto the deep neural network to extract the Granger causal variables for CSA-AKI among a high-dimensional dataset (appendix pp 7–9). Specifically, we designed a two-phase learning strategy to reveal stable underlying causal structures and achieve more generalisable predictive outcomes. During the prediction phase, the model takes current and historical patient data as input. Dynamic time-series data are processed with long short-term memory (LSTM) networks to capture temporal features, whereas static data are processed through several fully connected layers. The features generated from these models are then aggregated and passed through additional fully connected layers and a SoftMax layer to output the risk score for future CSA-AKI events. In the causal discovery phase, the model simulates intervention by performing counterfactual inference on all variables to evaluate their respective Granger causal effects (a fundamental concept in causal learning, to determine whether historical data from one time series can predict future data in another, indicating a directional influence between variables; appendix pp 4–5), and progressively distilling the causal variables to eliminate the spurious ones. The causal discovery phase iteratively removes variables with negligible influence on prediction. We simulate interventions by randomly masking inputs (appendix pp 7–9) and testing the resulting performance degradation—the variables are likely non-causal and dropped from the model if their removal hardly degrades the prediction accuracy. In this way, the model distills 1328 initial variables down to the six most impactful causal factors. These two phases are iteratively

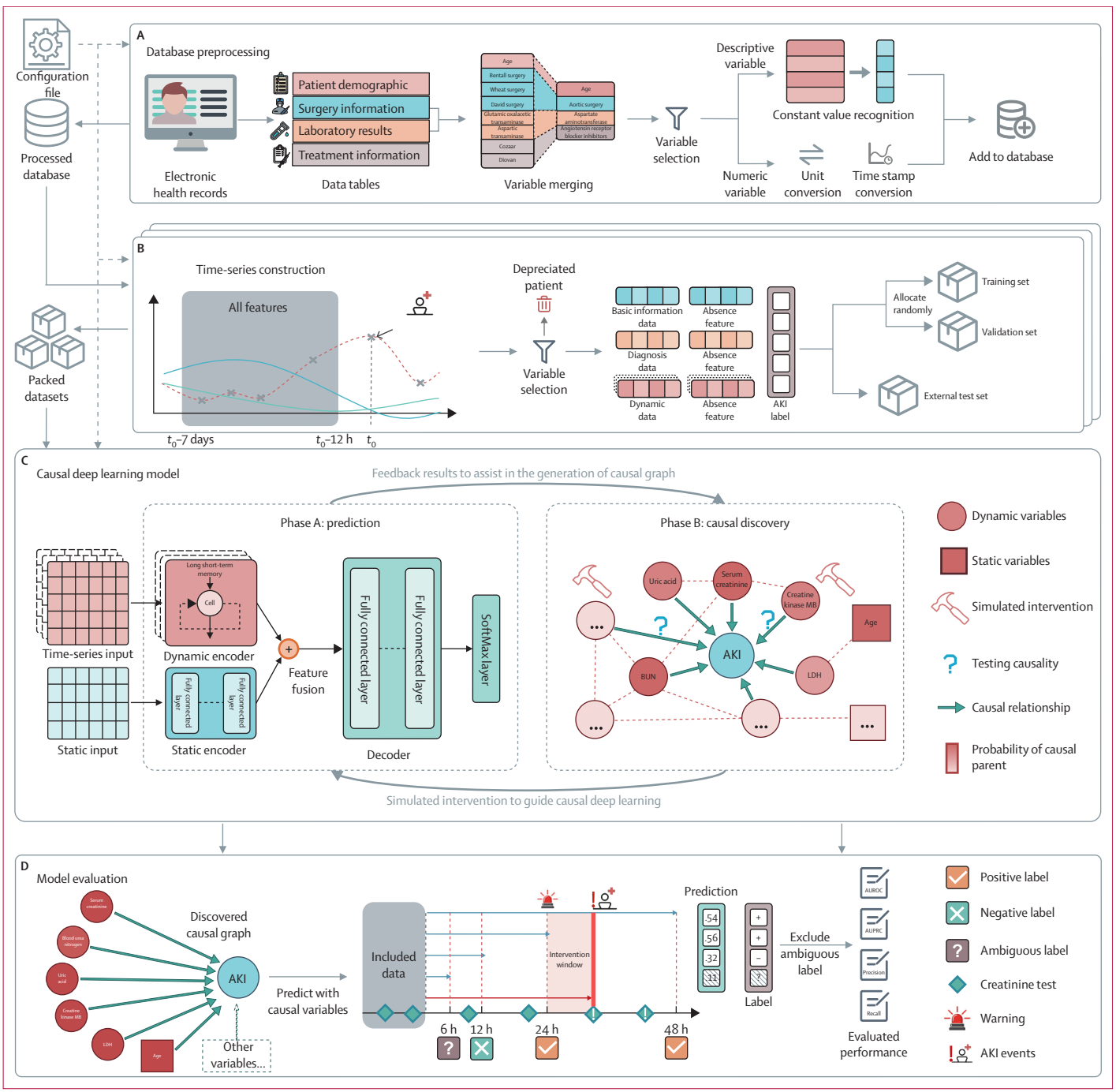


***Figure 2*: Overview of REACT**

Panel A shows the database preprocessing. The raw database was collected from electronic health records of patients who had undergone cardiac surgery and was then processed by merging the same items with different names, excluding the rarely present variables, converting units and timestamps, excluding artifacts, and mapping textual test results to real-valued numbers. Panel B shows the time-series construction for each patient. The dynamic variables of each patient were extracted from the preprocessed database and then sampled to structured time series with a 2 h time interval. The presence and absence of each variable were indicated with an additional feature. The model input at prediction time t_0 comprises all features from the preceding 7 days up to 12 h before t_0, with an alert generated at t_0. CSA-AKI labels were obtained at each timepoint of assessment with multiple prediction windows (ie, CSA-AKI within 6, 12, 24, or 48 h). All data-label pairs were divided into a training set, an internal (or in-distribution) validation set, and an external (or out-of-distribution) testing set. Panel C shows the causal deep learning model. REACT consists of two iterative phases: the prediction phase predicts risks for CSA-AKI at each timepoint; the causal discovery phase learns causal graphs with a fixed prediction model, performing simulated intervention to mimic randomised controlled trials and progressively distilling the causal variables to eliminate the spurious ones. Panel D shows the model evaluation. Our method took only six causal variables as input, and was evaluated in terms of AUROC and AUPRC along with other criteria. Here we excluded ambiguous labels, ie, no serum creatinine tests are within the corresponding prediction window. AKI=acute kidney injury. AUPRC=area under the precision-recall curve. AUROC=area under the receiver operating characteristic curve. BUN=blood urea nitrogen. CSA-AKI=cardiac surgery-associated acute kidney injury. LDH=lactate dehydrogenase. MB=myocardial band. REACT=Real-time Evaluation and Anticipation with Causal Distillation.

trained together (figure 2). The loss function of the model comprises both prediction error and causal regularisation error, and it is iteratively optimised to achieve minimal loss. When, during the causal discovery phase, a variable's change no longer effects model performance (its Granger causal effect falls below the threshold λ), this variable is excluded from the input in the prediction phase. However, at the neural network parameter level, the model retains the influence of all variables (including excluded ones) on prediction accuracy (mathematical details and assumptions are provided in appendix p 10).

Model evaluation

We evaluated performance using the AUROC and the Area Under the Precision-Recall Curve (AUPRC). The AUPRC is especially valuable in imbalanced datasets because it focuses on the performance related to the positive class, providing a more informative measure when dealing with rare events. The 95% CIs for AUROC and AUPRC were evaluated using bootstrap methods. We also reported sensitivity, specificity, positive predictive value, and negative predictive value. To compare model discrimination, we used DeLong's test for correlated receiver operating characteristic curves, considering differences significant if 95% CIs did not overlap. Calibration was assessed by the Brier score and calibration plots. We compared REACT with four algorithms: XGBoost, a multilayer perceptron, an LSTM network, and a Transformer model. All models were fine-tuned and optimised. Hyperparameter search details are in the appendix (pp 11, 19). Subgroup analyses were also conducted based on age, sex, surgical type, and admission year to explore the model's performance across different patient demographics.

Model application

For the **website** see, https://www.causal-cardiac.com

To facilitate clinical application, we developed a website and an applet of REACT (appendix p 54) for physician use. These tools were deployed and validated at the First Medical Center of People's Liberation Army General Hospital and Nanjing Drum Tower Hospital. Physicians could input patient demographics, surgical details, and laboratory results either manually or by uploading laboratory report images processed via optical character recognition. All data are standardised, and timestamps are recorded automatically. Users can request CSA-AKI risk predictions for the next 6, 12, 24, or 48 h, typically triggered by new test results or during handovers. The system allows adjustment of alert thresholds to reduce false positives for mild AKI and false negatives for moderate or severe AKI.

Statistical analysis

Categorical variables were assessed using the χ^2 test or Fisher's exact test. Continuous variables were compared using Student's *t* test or Welch's *t* test, as appropriate, based on Levene's test for variance homogeneity. All tests were two-sided, and a p value of less than 0·05 was considered statistically significant. Furthermore, we compared in-hospital mortality between cohorts with CSA-AKI first observed at stage 1 and stage 2 or 3, using risk ratios with 95% CIs. Statistical analyses were performed using R software (version 4.1.2).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

14 513 admissions (14 403 patients) from an initial pool of 4·1 million records were included in the derivation cohort, for whom a corresponding 23 933 cardiac surgeries were done, with a median age of 56 years (IQR 45–65); 5515 (38·0%) of 14 513 were female and 8998 (62·0%) were male (table). The incidence of CSA-AKI was 3047 (21·0%) of 14 513, with 512 (3·5%) experiencing severe CSA-AKI. Among these patients, 1715 (56·3%) developed CSA-AKI within 48 h post surgery (table). Among the 2710 patients who initially developed mild CSA-AKI, 848 (31·3%) experienced progression to more severe stages. The external validation dataset comprised 20 552 patients (20 813 admissions) from three medical institutions: Nanjing Drum Tower Hospital and the Sixth and Seventh Medical Centers of the Chinese People's Liberation Army General Hospital, with a median age of 57 years (IQR 45–66); and 8608 (41·4%) of 20 813 were female and 12 205 (58·6%) were male. Additionally, we included 14 229 patients and 13 794 patients from the MIMIC-IV dataset and the eICU dataset for international validation.

REACT, after being trained on the entire dataset, had reduced 1328 input variables (appendix p 18) to only six essential causal variables (ie, age, serum creatinine, urea nitrogen, uric acid, lactate dehydrogenase, and creatine kinase enzyme) for CSA-AKI prediction. In the internal validation set, REACT achieved an average AUROC of 0·930 (SD 0·032) across all prediction windows (6 h, 12 h, 24 h, and 48 h) for predicting severe CSA-AKI; REACT's AUROC varied from 0·949 (95% CI 0·945–0·953, at 48 h) to 0·971 (0·967–0·974, at 12 h), with an AUPRC spanning 0·663 (0·647–0·677, at 48 h) to 0·739 (0·712–0·761, at 6 h; figure 3; appendix pp 23, 25). Predicting severe CSA-AKI within 24 h, REACT achieved an AUROC of 0·969 (0·966–0·972) and an AUPRC of 0·725 (0·709–0·741). At a probability threshold chosen to approximate a 0·671 false discovery rate, REACT achieved a sensitivity of 0·864, a specificity of 0·955, and an F1 score of 0·476 for severe AKI prediction in the 24 h post surgery (appendix p 27). For the shorter prediction windows of 6 h and 12 h, REACT's performance remained strong. For instance, in internal validation, the AUROC for predicting severe CSA-AKI within 6 h was 0·970 (0·965–0·975) and within 12 h was 0·971 (0·967–0·974; appendix p 23). A similar trend was seen in external cohorts, with slightly lower discrimination for the 48 h window (AUROC 0·904 [95% CI 0·902–0·905],

	The Derivation Dataset		The External Testing Dataset		
	The First Medical Center of the Chinese People's Liberation Army General Hospital (n=12 685)	The Third Medical Center of the Chinese People's Liberation Army General Hospital (n=1828)	The Sixth Medical Center of the Chinese People's Liberation Army General Hospital (n=2261)	The Seventh Medical Center of the Chinese People's Liberation Army General Hospital (n=1570)	The Nanjing Drum Tower Hospital (n=16 982)
Patient demographics					
Age, years	58·0 (47·0–66·0)	45·0 (29·0–58·0)	59·0 (48·0–66·0)	56·0 (40·0–68·0)	57·0 (47·0–67·0)
Male	8101 (63·9%)	897 (49·1%)	1426 (63·1%)	955 (60·8%)	9824 (57·8%)
Female	4584 (36·1%)	931 (50·9%)	835 (36·9%)	615 (39·2%)	7158 (42·2%)
Height, cm	165 (159·0–171·0)	164 (160·5–170·0)	168 (160·0–172·0)	165 (158·5–170·5)	165 (160·0–171·0)
Weight, kg	67·0 (59·0–75·5)	58·0 (50·0–68·0)	66·0 (58·0–75·0)	66·0 (59·0–74·2)	65 (57·3–75·0)
Ethnicity					
Chinese	12 656 (99·8%)	1828 (100·0%)	2259 (99·9%)	1570 (100·0%)	16 968 (99·9%)
Other	29 (0·2%)	0	2 (0·1%)	0	14 (0·1%)
Comorbidities					
Hypertension	6798 (53·6%)	888 (48·6%)	1328 (58·7%)	609 (38·8%)	10 763 (63·4%)
Diabetes	3890 (30·7%)	178 (9·7%)	424 (18·8%)	240 (15·3%)	3826 (22·5%)
Congestive heart failure	3970 (31·3%)	594 (32·5%)	465 (20·6%)	288 (18·3%)	8480 (49·9%)
Pulmonary disease	897 (7·1%)	70 (3·8%)	922 (40·8%)	102 (6·5%)	5437 (32·0%)
Chronic kidney disease	398 (3·1%)	36 (2·0%)	54 (2·4%)	17 (1·1%)	999 (5·9%)
Type of surgery					
Coronary artery bypass grafting alone	4877 (38·4%)	324 (17·7%)	1147 (50·7%)	690 (43·9%)	2031 (12·0%)
Valve surgery alone	4263 (33·6%)	712 (38·9%)	618 (27·3%)	351 (22·4%)	5272 (31·0%)
Coronary artery bypass grafting and valve surgery	451 (3·6%)	0	6 (0·3%)	0	1046 (6·2%)
Aortic surgery	776 (6·1%)	14 (0·8%)	135 (6·0%)	94 (6·0%)	3420 (20·1%)
Coronary heart disease corrective surgery	1057 (8·3%)	646 (35·3%)	213 (9·4%)	320 (20·4%)	1756 (10·3%)
Others	1261 (9·9%)	132 (7·2%)	142 (6·3%)	115 (7·3%)	3457 (20·4%)
Surgery characteristics					
Number of surgeries involving cardiopulmonary bypass	11 167 (88·0%)	1565 (85·6%)	1038 (45·9%)	197 (12·5%)	12 717 (74·9%)
Use of intra-aortic balloon pump	478 (3·8%)	29 (1·6%)	258 (11·4%)	19 (1·2%)	257 (1·5%)
Use of extracorporeal membrane oxygenation	22 (0·2%)	2 (0·1%)	19 (0·8%)	0	166 (1·0%)
Preoperative laboratories					
Serum platelet, 10*9/L	193·0 (156·0–234·0)	199·0 (163·0–241·0)	204·0 (166·0–246·0)	207·0 (168·5–256·5)	176·0 (138·0–220·0)
Mean corpuscular haemoglobin concentration, g/L	338·0 (330·0–346·0)	333·0 (324·0–340·0)	338·0 (330·0–34 5·0)	329·0 (322·0–336·0)	335·0 (327·0–342·0)
Serum albumin, g/L	41·0 (38·5–43·4)	41·9 (39·2–44·7)	40·3 (37·6–42·7)	40·6 (37·8–43·7)	39·8 (37·5–41·9)
Serum potassium, mmol/L	4·08 (3·85–4·33)	3·99 (3·74–4·25)	3·92 (3·60–4·20)	3·96 (3·76–4·19)	3·99 (3·75–4·24)
Blood urea nitrogen, mmol/L	5·64 (4·60–7·02)	5·73 (4·58–7·12)	5·60 (4·60–6·90)	5·54 (4·40–6·95)	6·37 (5·10–8·21)
Serum creatinine, µmol/L	75·2 (64·1–88·2)	63·0 (53·0–76·0)	84·6 (73·9–97·0)	69·0 (58·0–81·0)	67·0 (56·0–81·0)
Outcomes					
AKI	2601 (20·5%)	446 (24·4%)	468 (20·7%)	414 (26·4%)	2989 (17·6%)
AKI stage 1	2397 (18·9%)	313 (17·1%)	421 (18·6%)	338 (21·5%)	2513 (14·8%)
AKI stage 2	634 (5·0%)	194 (10·6%)	145 (6·4%)	96 (6·1%)	1087 (6·4%)
AKI stage 3	368 (2·9%)	144 (7·9%)	72 (3·2%)	52 (3·3%)	594 (3·5%)

Data are n (%) or median (IQR). AKI=acute kidney injury.

Table: **Patient characteristics**

AUPRC 0·606 [0·602–0·610]; appendix pp 24, 26) but overall high performance across all windows, including AUROCs of 0·964 (0·961–0·966) at 6 h, 0·964 (0·963–0·966) at 12 h, and 0·964 (0·963–0·965) at 24 h (appendix p 24).

REACT outperformed all other methods (pairwise DeLong test p<0·0001 for AUROC and AUPRC with other models) trained on all variables: multilayer perceptron (predicting severe CSA-AKI within 24 h, AUROC 0·913, 95% CI 0·908–0·918; AUPRC 0·370, 95% CI 0·353–0·390), LSTM (0·881, 0·875–0·888; 0·397, 0·379–0·415), and Transformer (0·943, 0·939–0·947; 0·591, 0·576–0·605; appendix pp 23, 25). Furthermore, REACT showed improved accuracy for all stages and moderate or severe stage CSA-AKI predictions within 24 h (figure 3 and appendix pp 23, 25), registering an AUROC of 0·892 (0·889–0·895) and 0·936 (0·932–0·940) and an AUPRC of 0·637 (0·629–0·646) and 0·671 (0·658–0·680), respectively (figure 3; appendix pp 23–25). Sensitivity ranged from

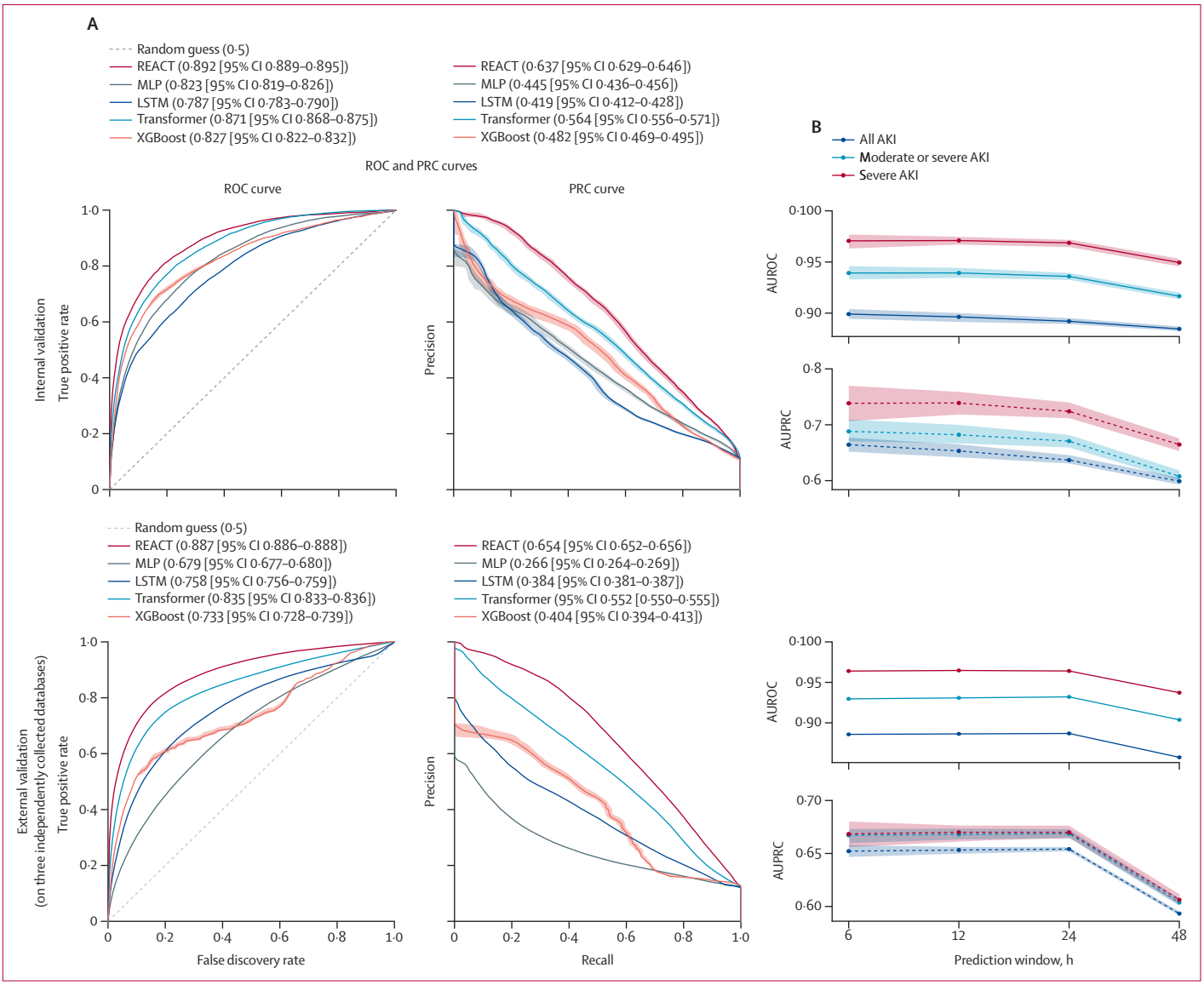


Figure 3: **Model performance**

Panel A shows the ROC and PRC curves for predicting any-stage AKI within 24 h post surgery, for each model. Panel B shows the prediction AUROC and AUPRC on three tasks (all AKI, moderate or severe AKI, and severe AKI) and at four different prediction windows (6, 12, 24, and 48 h). The shaded regions in the plots represent the 95% CIs of the respective performance curves. AKI=acute kidney injury. AUPRC=area under the precision-recall curve. AUROC=area under the receiver operating characteristic curve. LSTM=long short-term memory. MLP=multilayer perceptron. PRC=precision-recall curve. ROC=receiver operating characteristic curve.

0·716 to 0·876, and specificity ranged from 0·780 to 0·959 (appendix p27). Calibration was good for severe AKI prediction (Brier score=0·064; appendix p 47).

On average, REACT identified CSA-AKI 14·65 h (SD 3·17) earlier than guideline-based clinical recognition on the internal validation. Notably, for AKI events occurring within 24 h, REACT was approximately 45·4% quicker in detection (10·43 h *vs* 19·10 h). For AKI events within 48 h, the model achieved a 57·7% quicker recognition (13·57 h *vs* 32·09 h), providing critical lead time for early intervention (appendix p 22). In the external validation, REACT detected CSA-AKI on average 16·35 h (SD 2·01) earlier than guideline-based recognition (appendix p 22).

As the event approached, the model's predictive accuracy incrementally improved: 535 (75·8%) of 706 cases were predicted successfully 48 h in advance, and 600 (85·0%) were predicted 24 h earlier in the internal validation. For the total inaccurate predictions, a deeper analysis revealed that 495 (34·4%) of 1440 inaccuracies were due to delayed AKI onset within 24 h and 118 (8·2%) were attributed to 24–48 h after the time of assessment. 730 (50·7%) were actual false positives that did not reach the respective thresholds—a trade-off made to avoid alarm fatigue (appendix p 53).

Our model consistently outperformed comparison approaches on external validation datasets, achieving an average AUROC of 0·920 (SD 0·036; appendix pp 24, 26, 28).

Calibration was also good for severe AKI prediction, with a Brier score of 0·062 (appendix p 47). Notably, despite being based on entirely new data, REACT exhibited optimal performance and minimal performance fluctuation in the external validation set (figure 3; appendix pp 24, 26, 28), maintaining a median AUROC loss of just 0·008 (IQR 0·006–0·012). This performance markedly exceeded those of the comparator models (multilayer perceptron 0·152, IQR 0·143–0·164; LSTM 0·028, 0·022–0·031; and Transformer: 0·031, 0·023–0·041). Our model showed robust performance on MIMIC-IV and eICU datasets, with an average AUROC of 0·867 (SD 0·073). For the 24 h prediction window on the MIMIC-IV and eICU datasets, AUROCs were 0·752, 0·883, and 0·943 for all AKI, moderate or severe AKI, and severe AKI in MIMIC-IV, respectively, and 0·806, 0·898, and 0·950 in eICU, respectively (appendix p 52).

Experiments showed that separately training a regular neural network (eg, multilayer perceptron, LSTM, or Transformer) with only the six selected variables, instead of our causal deep learning strategy, resulted in lower AUROC and AUPRC scores (appendix pp 23–26). Taking the prediction of severe CSA-AKI within 24 h as an example, this separate training scheme underwent a performance degeneration of 11% compared with REACT, which achieved an AUPRC of 0·725 (95% CI 0·709–0·741) in internal validation. Thus, our causal deep learning approach enhances generalisability and reduces overfitting by integrating causal discovery.

REACT consistently showed robust performance across all demographic and surgery type subgroups except for moderate degradation in the pericardiectomy subgroup (appendix pp 30–38). In a prospective study conducted from June to October 2023, we included 754 patients who had major open-heart cardiac surgery (appendix pp 20, 21). During the study period, 129 patients developed CSA-AKI. REACT pre-emptively identified 121 (93·8%) of these episodes across all prediction windows, achieving a sensitivity of 0·825 and a specificity of 0·811. Among these patients, six developed severe CSA-AKI, and REACT successfully predicted five of these cases, with a positive predictive value of 0·608 and an negative predictive value of 0·998 (figure 4). When predictions were spot-on, the REACT system granted clinicians an average lead time of 16·64 (SD 0·78) h for intervention. Sensitivity analyses (eg, repeat admissions, creatinine-only definition, recent creatinine tests—defined as using only the most recent serum creatinine measurement at each prediction timepoint rather than cumulative values—and sampling intervals) confirmed the robustness of primary findings (appendix pp 40–43).

Discussion

REACT integrates the advances of deep learning techniques and causal discovery to provide a practical solution for CSA-AKI prediction. On the one hand, this approach leverages the powerful capabilities of neural networks to generate accurate real-time predictions based on a patient's evolving clinical state. On the other hand, with causal discovery, REACT evaluates the causal effects of variables, distilling the essential causal variables, which substantially reduces the number of required inputs for prediction and increases the generalisability of the neural network. Achieving reliable prediction across seven diverse cohorts using only six variables, REACT bridges the gap from the prediction model to clinical applications.

To show the effectiveness of this approach, we validated REACT's performance with five cohorts from different regions in China and with two international databases in the USA, detecting CSA-AKI 16·35 h earlier in external tests and 14·65 h earlier in internal validation. Additionally, we developed a user-friendly website and applet, which were updated with clinician feedback from an ongoing perspective implementation study.

Early detection of AKI in patients at high risk enables timely intervention before complications arise. For example, promptly applying a postoperative Kidney Disease: Improving Global Outcomes care bundle (eg, optimising volume and haemodynamics while avoiding nephrotoxins and hyperglycaemia) can reduce CSA-AKI in patients at high risk.[3] Traditional risk scores for CSA-AKI are often calculated from a small number of variables based on previous knowledge, discarding substantial data from electronic health records.[9,32] Many continuously changing vital signs, laboratory tests, and other clinical features are not used, so such conventional methods do not capture abrupt changes in the patients' conditions for in-time AKI prediction, which occur frequently in patients with CSA-AKI. These models reached AUROCs ranging from 0·69–0·83 in their respective internal validation cohorts.[5] A model by Demirjian and colleagues[10] used postoperative laboratory tests and timing to predict moderate-to-severe AKI within 72 h (AUROC 0·876). Despite the high AUROC, this model was applied at a median of 10 h after surgery, potentially delaying interventions. In contrast, REACT achieved higher accuracy (AUROC of 0·936 and 0·932 internally and externally for moderate or severe AKI within 24 h) from any postoperative timepoint, outperforming traditional models. It also maintains high positive and negative predictive values.

Deep learning methods, such as LSTM networks, excel in identifying complex relationships and handling high-dimensional time-series data from electronic health records, which is seldomly addressed in traditional risk scores. However, they often require very large input datasets, limiting their clinical use. For instance, Tomašev and colleagues[21] developed a model using 620 000 entries from hundreds of variables for real-time AKI prediction, showing the power of deep learning but also its strong dependency on extensive data inputs. In practical applications, the absence of a single feature or changes in data structure can potentially lead to a decline in the overall performance of the model. Furthermore, data distribution mismatches can affect model generalisability. In Tomašev's study, 94% of participants were male; subsequent validation on a sex-balanced cohort

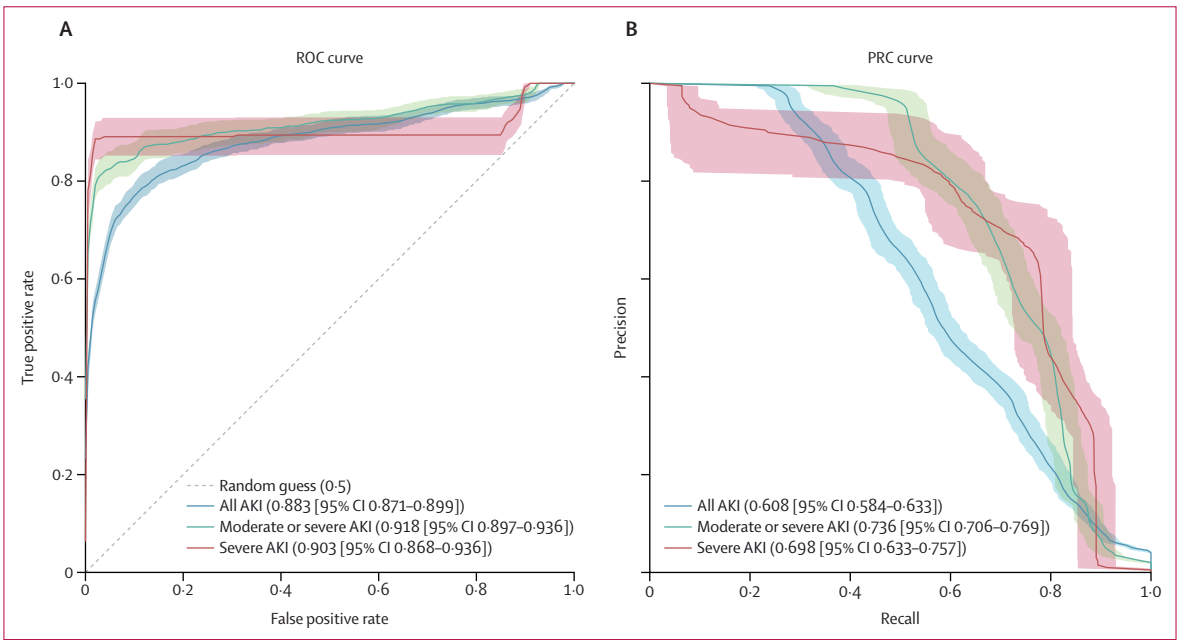


Figure 4: **Prospective validation**
The performance of REACT in prospective validation. This subgraph presents ROC and PRC curves for predicting AKI occurrences within 24 h, and the shaded regions represent the CIs for the performance curves. AKI=acute kidney injury. PRC=precision-recall curve. REACT=Real-time Evaluation and Anticipation with Causal Distillation. ROC=receiver operating characteristic curve.

showed reduced performance in females.[22] In contrast, REACT combines the strengths of LSTM with causal discovery techniques to distill the causal variables of CSA-AKI, thereby reducing dependency on extensive variable inputs. It showed robust performance in cross-centre and cross-international external validations, addressing one of the primary limitations of applying existing deep learning models in clinical practice.

Most previously developed models focused on severe AKI requiring kidney replacement therapy. However, even milder degrees of CSA-AKI (with minor serum creatinine changes) worsen outcomes for patients.[6,10,33] This study shows that 31·2% of mild CSA-AKI patients progressed to moderate or severe stages, highlighting the importance of early detection of mild CSA-AKI. Additionally, patients with mild and moderate CSA-AKI, whose clinical symptoms are subtle and overlooked by clinicians, might benefit more from deep learning models compared with those with severe CSA-AKI. As discussed by Rank and colleagues,[19] patients with mild and moderate CSA-AKI often have subtle clinical symptoms that tend to be overlooked by clinicians. Physicians can only mainly identify severe AKI or dialysis cases, and the overall sensitivity is as low as 0·594. In their experiments, deep learning models outperformed experienced physicians in AKI prediction, but were limited to moderate and severe AKI. In contrast, REACT accurately predicted AKI across all stages (including mild cases). This finding is particularly important given the scarcity of research in this field. In addition, although the current model focuses on postoperative prediction, the REACT framework could be adapted for preoperative risk assessment based on baseline characteristics. Future work could also explore automatic hyperparameter selection methods for λ to minimise the risk of overfitting associated with manual tuning.

Our causal discovery approach assumes that all major confounders are measured. This assumption is common in Granger causal analysis. In a clinical setting, this is an approximation—there might always be factors that were not captured (eg, genetic susceptibility and subtle aspects of surgical technique). However, we attempted to include a comprehensive set of perioperative variables to minimise unmeasured factors. Moreover, the success of the model in external validations suggests that unmeasured confounders did not crucially undermine its generalisability—if they had, the model likely would have performed poorly, showing substantially reduced AUROC and calibration when applied to new data.

This study has several notable strengths. First, to the best of our knowledge, this represents the largest dataset used for CSA-AKI prediction to date. The study was conducted on a large amount of consecutive patients' data from five medical centres of the People's Liberation Army General Hospital, which are tertiary grade A hospitals—the highest level in China's hospital classification system, providing the most comprehensive and specialised medical care. These medical centres admit patients from all over the country, aligning the distribution of patients' regions with China's population density. This extensive dataset facilitated robust internal and external validation, enhancing the generalisability and reliability of our findings. Second, REACT is the first model to integrate deep learning with causal

discovery for CSA-AKI prediction, establishing a new benchmark in the domain. Furthermore, our model exhibited stable performance across various subgroups, confirming its robustness and adaptability in diverse clinical settings.

Nevertheless, our study has limitations. First, as a retrospective study with little prospective validation, our findings could be subject to inherent biases. Despite favourable external validation results, prospective randomised controlled trials are needed to confirm clinical efficacy. Furthermore, although we enhance model practicality and stability for external validation using a causal deep learning method and identifying causal variables through counterfactual inference, this causality emphasises more on predictive power, rather than fundamental physiological mechanisms. If some confounders were unmeasured, causal attribution could be biased, and results should be interpreted with appropriate caution. Future exploration of the mechanisms of CSA-AKI will require more experiments related to physiological mechanisms and additional randomised controlled trials. Additionally, some intraoperative variables were not included in the model training. For instance, although we considered the reception of blood transfusions during surgery, we did not account for the number of red blood cell units transfused, which could introduce bias in our results. Similarly, emerging biomarkers such as NGAL and TIMP-2·IGFBP7 were not included due to scarce availability in routine data, but our framework readily accommodates their future incorporation. The pericardiectomy subgroup showed lower predictive accuracy for any-stage AKI due to small sample size and unique haemodynamics; expanding this subgroup in future studies could enhance stability and generalisability. Lastly, our dataset was predominantly Asian. Although we validated REACT on more diverse US datasets (MIMIC-IV and eICU), further validation in other populations would be beneficial to assess model generalisability.

In conclusion, our study presents a pioneering model that integrates deep learning with causal discovery, facilitating dynamic and accurate prediction of CSA-AKI up to 48 h in advance. This model reduces the number of required input variables to six and generalises well across different cohorts, which are both crucially important for clinical applications of AI models. REACT's success in early CSA-AKI prediction also highlights the potential of causal deep learning for broader applications to enable earlier interventions.

Contributors

KH and JS supervised and reviewed the study. KH also provided the study materials, data, computing resources, and secured financial support for the project. QZ, YC, and ZL were responsible for the study design, execution of all project components, and manuscript preparation. YC performed algorithm design, tuning, and validation. DW provided a dataset from Nanjing Drum Tower Hospital for external tests and offered guidance on data curation as well as in writing the clinical discussions. CR and HC contributed expert clinical opinions in case selection and data collection. XL assisted in manuscript review. LL and ZW participated in the development, debugging, and optimisation of the model. PLi conducted the final formatting and grammatical checks. PLiu was responsible for developing and maintaining the website. YJ organised the prospective validation of the study at two centres, coordinating and managing parts of the research process. QZ, YC, ZL, CR, YJ, and HC verified the raw data. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The demonstration simulated data are available at our GitHub. Researchers from non-commercial institutions interested in accessing more data can submit a request by emailing either the corresponding author or the first author (mikozhong@outlook.com). We have opensourced all codes, models, and a Google Colab example on Github; the latest version of the model is also accessible via a user-friendly website.

For the **demonstration simulated data** see https://github.com/jarrycyx/UNN/tree/main/REACT

For the **latest version of the model** see http://www.causal-cardiac.com/

Acknowledgments

This research was supported by the Construction Project (AHQ140010000X) and the National Natural Science Foundation of China (72025405 and 72421002).

References

1 Hu J, Chen R, Liu S, Yu X, Zou J, Ding X. Global incidence and outcomes of adult patients with acute kidney injury after cardiac surgery: a systematic review and meta-analysis. *J Cardiothorac Vasc Anesth* 2016; **30:** 82–89.
2 Bove T, Monaco F, Covello RD, Zangrillo A. Acute renal failure and cardiac surgery. *HSR Proc Intensive Care Cardiovasc Anesth* 2009; **1:** 13–21.
3 Zarbock A, Küllmar M, Ostermann M, et al. Prevention of cardiac surgery–associated acute kidney injury by implementing the KDIGO guidelines in high-risk patients identified by biomarkers: the PrevAKI-multicenter randomized controlled trial. *Anesth Analg* 2021; **133:** 292–302.
4 Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin Pract* 2012; **120:** c179–84.
5 Huen SC, Parikh CR. Predicting acute kidney injury after cardiac surgery: a systematic review. *Ann Thorac Surg* 2012; **93:** 337–47.
6 Hobson CE, Yavas S, Segal MS, et al. Acute kidney injury is associated with increased long-term mortality after cardiothoracic surgery. *Circulation* 2009; **119:** 2444–53.
7 Aronson S, Fontes ML, Miao Y, et al. Risk index for perioperative renal dysfunction/failure: critical dependence on pulse pressure hypertension. *Circulation* 2007; **115:** 733–42.
8 Wijeysundera DN, Karkouti K, Dupuis J-Y, et al. Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. *JAMA* 2007; **297:** 1801–09.
9 Thakar CV, Arrigain S, Worley S, Yared J-P, Paganini EP. A clinical score to predict acute renal failure after cardiac surgery. *J Am Soc Nephrol* 2005; **16:** 162–68.
10 Demirjian S, Bashour CA, Shaw A, et al. Predictive accuracy of a perioperative laboratory test-based prediction model for moderate to severe acute kidney injury after cardiac surgery. *JAMA* 2022; **327:** 956–64.
11 Mehta RH, Grab JD, O'Brien SM, et al. Bedside tool for predicting the risk of postoperative dialysis in patients undergoing cardiac surgery. *Circulation* 2006; **114:** 2208–16.
12 Birnie K, Verheyden V, Pagano D, et al. Predictive models for kidney disease: improving global outcomes (KDIGO) defined acute kidney injury in UK cardiac surgery. *Crit Care* 2014; **18:** 606.
13 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; **380:** 1347–58.
14 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022; **28:** 31–38.
15 Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann Surg* 2019; **269:** 652–62.
16 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:** 1070–77.

17 Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* 2018; **6:** 905–14.

18 Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One* 2016; **11:** e0155705.

19 Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *NPJ Digit Med* 2020; **3:** 139.

20 Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun* 2020; **11:** 3923.

21 Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; **572:** 116–19.

22 Cao J, Zhang X, Shahinian V, et al. Generalizability of an acute kidney injury prediction model across health systems. *Nat Mach Intell* 2022; **4:** 1121–29.

23 Hunter DJ, Holmes C. Where medical statistics meets artificial intelligence. *N Engl J Med* 2023; **389:** 1211–19.

24 Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.

25 Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. *Int J Epidemiol* 2021; **49:** 2058–64.

26 Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023; **10:** 1.

27 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.

28 Kellum JA, Lameire N, Aspelin P, et al. Kidney Disease: Improving Global Outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl* 2012; **2:** 1–138.

29 Zhang X, Cui P, Xu R, Zhou L, He Y, Shen Z. Deep stable learning for out-of-distribution generalization. *arXiv* 2021; published online April 16, 2021. https://doi.org/10.48550/arXiv.2104.07876.

30 Cheng Y, Yang R, Xiao T, et al. CUTS: neural causal discovery from irregular time-series data. *arXiv* 2023; published online Feb 15, 2023. https://doi.org/10.48550/arXiv.2302.07458.

31 Cheng Y, Li L, Xiao T, et al. CUTS+: high-dimensional causal discovery from irregular time-series. *Proc AAAI Conf AI* 2024; **38:** 11525–33.

32 Kristovic D, Horvatic I, Husedzinovic I, et al. Cardiac surgery-associated acute kidney injury: risk factors analysis and comparison of prediction models. *Interact Cardiovasc Thorac Surg* 2015; **21:** 366–73.

33 Cho JS, Shim J-K, Lee S, et al. Chronic progression of cardiac surgery associated acute kidney injury: intermediary role of acute kidney disease. *J Thorac Cardiovasc Surg* 2021; **161:** 681–88.