

RESEARCH

Open Access



Beyond binary AKI classification: development and external validation of a distributional model predicting serum creatinine and urine output trajectories in ICU patients

Jef Jonkers^{1*†}, Anahita Rouzé^{2*†}, Jarne Verhaeghe³, Wim Van Biesen⁴, Jan De Waele⁵, Luc Duchateau⁶, Glenn Van Wallendael³, Eric Hoste⁵ and Sofie Van Hoecke³

Abstract

Background Acute kidney injury (AKI) is a frequent, severe complication in the intensive care units (ICU). Existing machine learning models are typically inflexible, classification-based (i.e., predicting AKI occurrence as yes/no), and of limited clinical utility. This study proposes and externally validates the first multi-step, multivariate distributional regression model that directly predicts future distributions of serum creatinine (sCr) and urine output across multiple time horizons, thereby enhancing AKI risk stratification and personalized clinical decision support.

Methods The model was developed using a training cohort of 4,118 adult ICU stays from the MIMIC-IV dataset and externally validated on four independent, diverse cohorts: MIMIC-IV (N=3,838), UZGent (N=4,442), eICU (N=10,760), and AmsterdamUMC (N=6,129). The model used clinical data to generate multivariate predictive distributions hourly for urine output and sCr (up to 48 hours ahead). Predictors included demographics, vital signs, laboratory results, medications, and recent urine output, with time-varying variables summarized over the preceding 72 hours (recent value, slope, minimum, maximum, variability). Performance was evaluated by comparing our predictive distributions with state-of-the-art tree-based classifiers for 24-hour ahead prediction of KDIGO stages 1-3 AKI and persistent stage 3 AKI.

Results Across all external cohorts, the distributional regression model demonstrated high discrimination (mean AUC-PR 0.774 for all stages) and excellent calibration, consistently outperforming the benchmark classifiers. By jointly predicting sCr and urine output distributions, a single model successfully enables flexible risk stratification across all stages, capturing AKI onset and persistence, and allowing changes to stage definitions.

[†]Jef Jonkers and Anahita Rouzé contributed equally to this work.

*Correspondence:

Jef Jonkers

jef.jonkers@ugent.be

Anahita Rouzé

anahita.rouze@chu-lille.fr

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



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

Conclusion This multi-step, multivariate distributional regression model is a reliable, more flexible, transparent, and clinically interpretable approach for AKI prediction compared to traditional classification methods. It represents a necessary step toward bedside implementation of predictive models for personalized AKI management in the ICU.

Keywords Acute kidney injury, Intensive care unit, Multivariate distributional regression, External validation, Machine learning

Introduction

Acute kidney injury (AKI) is a common complication in critically ill patients, affecting more than 50% of ICU admissions [1]. It contributes not only to increased acute morbidity and mortality, but also to long-term renal sequelae that can compromise overall prognosis [2]. Prediction of AKI onset and progression could better inform preventive and management strategies, such as haemodynamic optimization and avoidance of nephrotoxins, and facilitate planning of kidney replacement therapy (KRT) in patients at risk of persistent severe injury [3].

Current approaches to AKI prediction rely on clinical risk scores, typically derived from linear logistic regression models, as well as functional, damage, and stress biomarkers. However, clinical scores are largely static and provide limited information on timing, severity, or trajectory of AKI [4]. Alternatively, while biomarkers can identify high-risk patients for AKI onset before KDIGO (kidney disease improving global outcomes) [5] criteria are met, predict disease progression or recovery, and provide physiopathological insights, they are not routinely used due to their cost, limited availability, and variable specificity [6].

To overcome these limitations, the past decade has seen an abundance of machine learning-based risk prediction models for AKI, particularly in critically ill patients [4, 7]. Yet, only a small subset has undergone external validation [8–11], and prospective evaluation remains even more exceptional [12].

Strikingly, nearly all existing AKI prediction models are classification-based: they estimate whether a patient will meet a predefined AKI outcome at a future horizon, such as KDIGO stage 1, 2, or 3 AKI or persistent stage 3 AKI. This categorical approach is inherently inflexible, reduces interpretability, and limits integration into routine decision-making. Only one study, to our knowledge, formulates renal function prediction as a regression task [13]; however, it provides point forecasts only (i.e., single-value predictions) and does not model predictive uncertainty.

In contrast, recent advances in machine learning, originally pioneered in fields such as weather forecasting, have introduced probabilistic approaches that predict complete distributions of continuous variables at future-defined horizons, rather than binary outputs [14]. Such models can flexibly address multiple clinical questions, such as onset, severity, and progression within a single

framework, while explicitly quantifying the uncertainty around predictions. Importantly, this approach could also generate patient-specific dashboards, offering an intuitive way to visualize model outputs (see Fig. 1).

Given this background, we aimed to develop and externally validate the first multi-step, multi-target, distributional regression model, which predicts future multi-horizon probability distributions of serum creatinine and urine output in critically ill patients rather than a single fixed AKI label or point estimate. Based on routinely collected patient data, this model predicts both serum creatinine (sCr) and urine output trajectories (i.e., multi-target) over multiple time horizons (i.e., multi-steps), providing clinically actionable forecasts to support personalized AKI management in the ICU.

Methods

Study data sources & population

We used four large-scale ICU datasets spanning diverse healthcare systems and geographic regions. The training cohort consisted of adult ICU patients from the MIMIC-IV (v3.1) database [15], a publicly available critical care resource containing de-identified records from a single center in Boston, USA. We chose MIMIC-IV for training due to its public accessibility (supporting reproducibility), high completeness and consistency of key time-varying variables (notably sCr and urine output), and to enable a realistic temporal evaluation by training on earlier admissions. Model validation was performed using four independent cohorts. First, temporal validation was conducted on a held-out subset of MIMIC-IV, comprising ICU admissions between 2017 and 2019. For this purpose, MIMIC-IV ICU stays were partitioned strictly by admission year into three disjoint sets: 2011–2013 for model training, 2014–2016 for early stopping and hyperparameter tuning (validation), and 2017–2019 for temporal generalization testing. Second, external validation was carried out on three additional datasets: two publicly available databases, the multi-center U.S. eICU Collaborative Research Database [16], containing ICU stays from 2014–2015, and the AmsterdamUMC database from the Netherlands [17], containing ICU stays from 2003–2016; and one proprietary dataset from UZGent (University Hospital Ghent, Belgium), comprising ICU stays from 2013–2017.

Eligible patients were adults (≥ 18 years) admitted to the ICU with available sCr and urine output measurements.

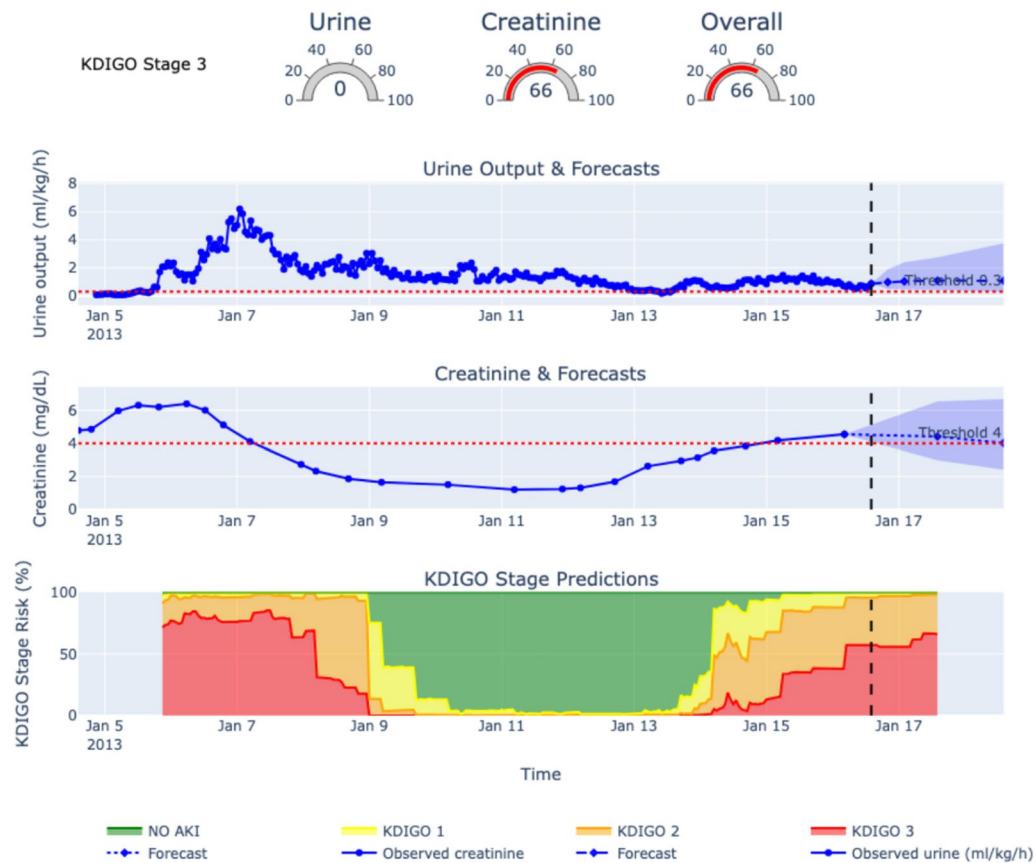


Fig. 1 Patient-level dashboard showing predicted trajectories for stage 3 AKI predominantly driven by sCr elevation. The continuous blue lines represent observed values of urine output and sCr over the past, while the blue dashed lines show the median predicted trajectory for each variable at future time points (urine output at 6, 12, 24, 48 hours; sCr at 24 and 48 hours). The shaded blue areas around the dashed lines indicate the full predicted distributions. Predictions are updated hourly, allowing dynamic assessment over time. Users can select the AKI stage of interest (in this example, stage 3) and enter patient-specific baseline sCr, which personalizes the thresholds for AKI classification. The model provides the predicted risk of reaching the selected stage within the next 24 hours, along with the relative contributions of sCr and urine output to that risk

Exclusion criteria comprised both static exclusions (entire ICU stays removed) and dynamic exclusions (time periods within ICU sessions removed). These criteria, together with the sequential filtering steps and the number of ICU stays remaining after each step, are summarized in Table 1. Key exclusions included patients with pre-existing chronic kidney disease (CKD, defined based on available clinical information in each dataset), as CKD would invalidate our sCr baseline derivation approach for this cohort; patients receiving KRT before ICU admission (acute or chronic); ICU stays shorter than 48 hours (which is only applied to allow evaluation of 48 hours-ahead predictions); periods without sufficient follow-up data (e.g., insufficient sCr or urine measurements available); and censoring once KRT was initiated. Across most databases, approximately 5–27% of the original ICU stays were retained after filtering. The characteristics of the final cohorts are presented in Appendix Table E1.

Preprocessing of outcomes

Urine output

Urine output was collected as discrete measurements at variable intervals and converted to hourly rates. Values were normalized by admission body weight (kg) to obtain the result in ml/kg/h. The first urine output measurement after ICU admission was excluded from analysis, as the volume collected does not correspond to a well-defined time interval. To calculate hourly urine flow, we implemented a retrospective windowing approach: first examining the preceding 6-hour period, and if fewer than three measurements were available, extending the look-back window to 12 hours. Periods with no measurements for 12 consecutive hours were classified as anuria. ICU sessions with urine volume outliers exceeding 2000 mL in a single measurement were excluded, including their corresponding session.

Serum Creatinine (sCr)

Serum Creatinine (sCr) values were collected from laboratory results. To generate outcome labels, we

Table 1 Exclusion criteria and number of ICU sessions remaining after sequential filtering steps across datasets. Filters are applied sequentially from the initial dataset to the final usable ICU sessions

Type	Criterion (applied at each step)	MIMIC-IV	UZ Gent	eICU	AmsterdamUMC
Initial ICU sessions					
	Initial ICU sessions	52,074 (100%)	16,433 (100%)	200,859 (100%)	23,106 (100%)
Data filters					
Data	Unknown ICU admission or discharge time	52,071 (100%)	16,433 (100%)	200,859 (100%)	23,106 (100%)
Data	Weight ≤ 0	52,071 (100%)	16,111 (98.0%)	184,125 (91.7%)	22,160 (95.9%)
Model filters					
Model	ICU stay <48 h	25,845 (49.6%)	7,345 (44.7%)	75,461 (37.6%)	9,236 (40.0%)
Clinical filters					
Clinical	Age on admission <18 y	25,845 (49.6%)	7,312 (44.5%)	75,365 (37.5%)	9,236 (40.0%)
Clinical	Acute or chronic RRT before ICU admission	25,674 (49.3%)	7,137 (43.4%)	71,922 (35.8%)	9,032 (39.1%)
Clinical	Chronic kidney disease	20,137 (38.7%)	5,762 (35.1%)	64,554 (32.1%)	9,032 (39.1%)
Target filter					
Data	No urine and creatinine measurements	18,958 (36.4%)	5,738 (34.9%)	56,095 (27.9%)	8,932 (38.7%)
Urine output filter					
Data	Negative urine output or single value >2000 mL	18,958 (36.4%)	5,684 (34.6%)	44,849 (22.3%)	8,875 (38.4%)
Data	No urine measurement for 12h*	18,958 (36.4%)	5,684 (34.6%)	23,155 (11.5%)	8,875 (38.4%)
Dynamic filters (entire sessions or periods excluded)					
Dynamic	– After initiation of RRT during ICU stay	12,084 (23.2%)	4,442 (27.0%)	10,760 (5.36%)	6,129 (26.5%)
	– ICU stay > 14 days				
	– No creatinine or urine output available ≥ 48 h ahead				

* The no urine measurement for 12 hours exclusion criterion is only applied for the eICU dataset

interpolated linearly between measurements, as we aim to train a real-time prediction model. sCr is typically measured daily, except in specific circumstances, where it is measured more frequently. This is a common practice in the literature [18].

AKI KDIGO stages

AKI stages were defined according to KDIGO 2012 guidelines [5] using patient-specific baseline sCr values. Because pre-ICU baseline sCr was generally unavailable, we estimated a back-calculated baseline for all patients by assuming an estimated glomerular filtration rate (eGFR) of 75 mL/min/1.73m² and applying the chronic kidney disease epidemiology collaboration (CKD-EPI) equation [19]; patients with CKD were excluded from evaluation because this assumption may be invalid. Persistent stage 3 AKI was defined as uninterrupted stage 3 lasting ≥ 48 hours. As a sensitivity analysis, we additionally evaluated an alternative baseline definition from Alfieri et al. [10], using the minimum recorded sCr over the patient's available history; this alternative baseline was used only during evaluation.

Feature engineering

To predict future patient outcomes, we constructed a set of predictors, referred to as features, from routinely collected clinical data. These features were organized into five groups: demographics, vital signs, laboratory results, medications, and urine output.

For time-series data like vital signs and lab values, we analyzed the 72-hour period preceding a prediction to capture dynamic trends. This was done by calculating summary statistics for each variable within this window, such as its mean, minimum, maximum, range, and slope. To ensure data quality, values falling outside of clinically plausible ranges were excluded (see Appendix Table E2).

Starting from 265 candidate covariates, PowerShap [20] was used to identify the most informative features, yielding a compact set of 124 covariates while maintaining coverage across all key features. Table C2, in the Appendix, summarizes the resulting feature space by clinical category and contrasts the number of candidate predictors with the subset retained after feature selection. Additionally, the complete list of selected predictors is provided in Table C3, for transparency and reproducibility.

Predictive models

The distributional regression model was implemented using NGBoost [21], a gradient boosting algorithm that learns the parameters of specified parametric distributions conditional on the input features. Creatinine was modeled on the log-scale for 24- and 48-hour horizons, while urine output was modeled as $\log(x+1)$ to accommodate zero values for 6-, 12-, 24-, and 48-hour horizons. The quality of the predicted distributions was rigorously evaluated using appropriate probabilistic metrics, with detailed technical results provided in Appendix D.

Additionally, for benchmarking, separate gradient boosting tree classifiers were trained for each KDIGO AKI stage to forecast the risk of being in that stage 24 hours ahead, as well as for 48 hours of persistent stage 3 AKI.

Both distributional regression and classification models were trained on MIMIC-IV ICU sessions from 2011–2013 (N=4,118) and early-stopped based on a validation set from 2014–2016 (N=4,128), where training was halted once performance stopped improving, to prevent overfitting. Default hyperparameters were used for both approaches.

From predictive distributions to AKI stage probabilities

The predictive distributions enable estimation of the probability of meeting stage-specific KDIGO criteria at any horizon, including persistent stage 3 AKI. This allows calculation of clinically relevant outcomes such as: *probability of stage 2 at 24 hours* or *probability of persistent stage 3 over 48 hours*. Detailed formulas and illustrations are provided in Appendix C.2.

Results

The distributional regression model forecasts full future trajectories of sCr and urine output, expressed as predictive distributions rather than single values. This allows estimation of both the most likely course of renal function and the uncertainty around those forecasts, which is clinically important for decision making. To facilitate bedside use, we generated patient-specific dashboards, providing clinicians with an intuitive and potentially actionable tool (Fig. 1). Prediction variance reflected clinical complexity: higher variance indicated greater uncertainty in patient trajectories, while narrow intervals corresponded to more stable and predictable courses, typically associated with longer ICU stays, high or low sCr levels, and frequent measurements. Overall, the model demonstrated robust predictive performance and generalization across validation cohorts (see Appendix D for technical details of distributional performance).

Feature importance

The most influential predictors, based on tree-based feature importance, for sCr predictions included recent sCr trends and urine output; however, phosphate, lactate, magnesium, FiO₂ trends, as well as age on admission and diuretic administration (furosemide), were also influential. For urine output predictions, current urine output, diuretic administration (furosemide), mean and diastolic arterial pressure, as well as weight and age on admission, emerged as key predictors. SHAP (SHapley Additive exPlanations) analysis revealed similar insights (see Appendix Figs. E5 and E68).

AKI stage risk prediction

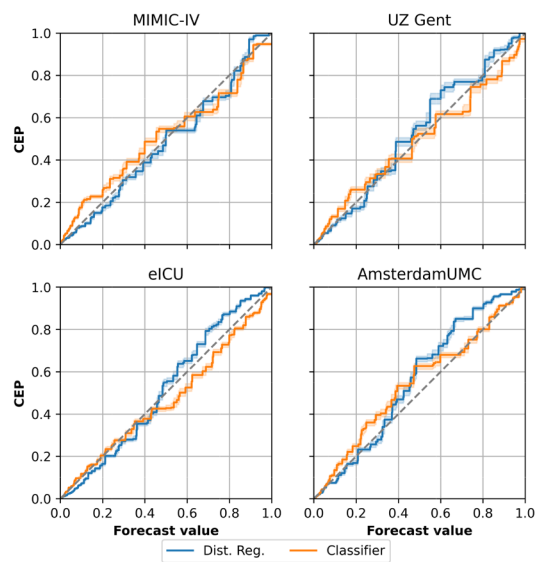
When converted to classification tasks for comparison with existing approaches, the distributional model achieved strong discrimination across all KDIGO stages (Fig. 2 and Table 2; additionally, in the supplementary material, you can find the results for stages 1 and 2: Appendix Figs. E2, E33, and Tables E4, E6). For 24-hour ahead stage 3 AKI prediction, AUC-ROC values exceeded 0.95 across all external cohorts, with AUC-PR ranging from 0.672–0.877.672.877, which is equivalent to the area under the precision-recall (PR) curves in Fig. 2. The model consistently outperformed gradient boosting classifiers trained on the same features, with improvements in F1-scores ranging from 1 to 5 % across different AKI stages and datasets. For rare outcomes, the AUC-PR should be interpreted in the context of event prevalence, since a no-skill model would already achieve the baseline event rate. When deploying the model, the choice of threshold matters: the reported sensitivity and specificity correspond to a single threshold, but the model produces continuous probabilities that can be adjusted to meet local priorities. Good calibration is key to reducing alarm fatigue. Well-calibrated probabilities allow hospitals to implement risk-based alerting strategies, such as using probability thresholds, requiring sustained high risk over time, or triggering alerts only for the highest-risk patients. This approach helps control the number of alerts, reduce unnecessary alarms, and still maintain the ability to increase sensitivity when needed.

Calibration performance was excellent (Fig. 2a and Table 2; and in supplementary material: Figs. E1a, E2a and Tables E4, E6), with expected calibration error (ECE) values below 0.02 for most cohorts and AKI stages, assuming a bin size of 10 to calculate the ECE. The ECE can be seen as the distance between the ideal calibration curve and the empirical curve plotted in Fig. 2a.

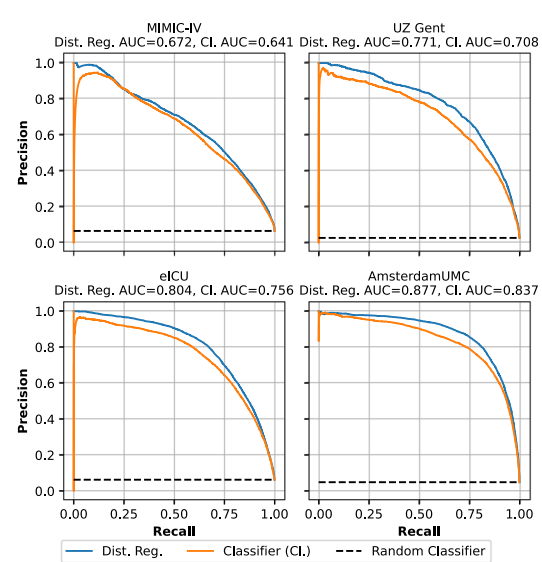
Subgroup analysis

To further assess robustness and potential heterogeneity in performance, we report subgroup-stratified results for 24h-ahead KDIGO stage 3 prediction across clinically relevant cohorts. We stratify by common admission descriptors, when available (sepsis vs. non-sepsis, medical vs. surgical). For each subgroup, we report the cohort size and event prevalence, along with discrimination, diagnostic accuracy, and calibration metrics. Results are depicted in Appendix Tables E8 and E9.

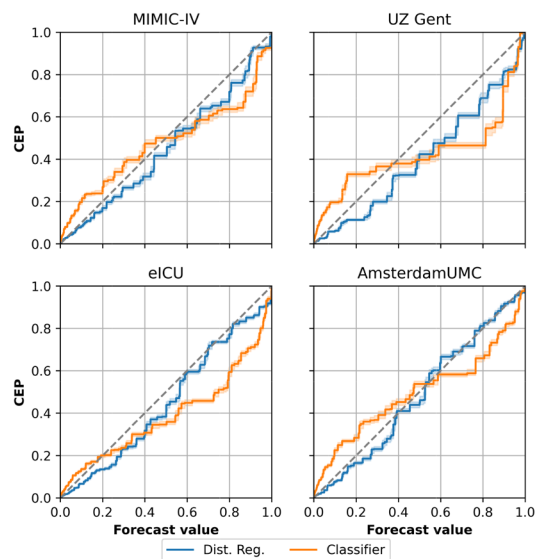
Across cohorts, subgroup performance showed clinically plausible heterogeneity, while calibration remained overall good (low ECE). Sepsis versus non-sepsis performance was broadly comparable across cohorts. When medical and surgical admission strata were available, surgical admissions were consistently less predictable than medical admissions, suggesting that postoperative renal



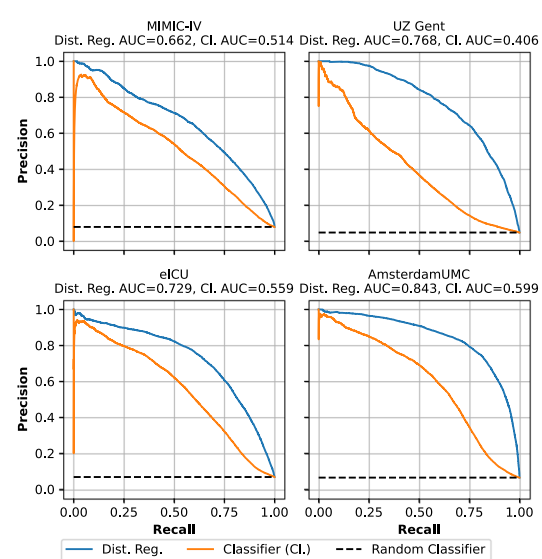
(a) CORP (consistent, optimal reproducible, PAV (pool-adjacent violators) algorithm-based) reliability diagrams [22], and the y-axis shows the observed frequency of the outcome. A perfectly calibrated model lies along the diagonal: predictions match observed outcomes. Curves above the diagonal indicate underestimation, while curves below indicate overestimation.



(b) PR (precision-recall) curve comparison. Higher curves indicate better discrimination across decision thresholds, particularly in the high-precision region.



(c) CORP reliability diagrams under alternative definitions of baseline sCr used in the KDIGO sCr criteria (Table A1).



(d) PR curve comparison under alternative definitions of baseline sCr used in the KDIGO sCr criteria (Table A1).

Fig. 2 Comparison of a tree-based distributional regressor and a classifier for predicting KDIGO stage 3. **(a, b)** Calibration and PR under the back-calculated baseline sCr definition. **(c, d)** Evaluation under an alternative baseline sCr definition used to derive the KDIGO stage 3 labels, with no retraining

Table 2 Comparison of distributional regression vs. baseline classifier across datasets for 24h-ahead KDIGO stage 3 prediction

Dataset	Model	Discrimination			Diagnostic accuracy				Calibration		
		F1	AUC-PR	AUC-ROC	Sens	Spec	PPV	NPV	Brier	NLL	ECE
Temporal validation											
MIMIC-IV (N=364,449; stage 3=6.7%)	Dist. Reg.	0.599	0.672	0.948	0.519	0.985	0.708	0.966	0.035	0.123	0.005
	Classifier	0.546	0.641	0.942	0.437	0.988	0.727	0.961	0.037	0.131	0.013
External validation											
UZ Gent (N=466,901; stage 3=2.8%)	Dist. Reg.	0.678	0.771	0.984	0.584	0.996	0.809	0.988	0.012	0.042	0.002
	Classifier	0.646	0.708	0.980	0.568	0.995	0.748	0.988	0.013	0.049	0.003
eICU (N=991,810; stage 3=6.3%)	Dist. Reg.	0.721	0.804	0.969	0.637	0.991	0.831	0.976	0.025	0.090	0.008
	Classifier	0.693	0.756	0.964	0.634	0.987	0.764	0.976	0.027	0.098	0.003
AmsterdamUMC (N=833,277; stage 3=4.9%)	Dist. Reg.	0.785	0.877	0.988	0.707	0.995	0.882	0.985	0.015	0.053	0.005
	Classifier	0.745	0.837	0.985	0.675	0.993	0.830	0.983	0.017	0.060	0.005

A patient is classified as KDIGO stage 3 if sCr exceeds at least 3×baseline or 4 mg/dL, or urine output is below 0.3 mL/kg/h for at least 24 hours. Reported are discrimination, calibration, and diagnostic accuracy metrics. Best values per dataset and metric are in **bold**

trajectories and care pathways may introduce additional heterogeneity. Overall, these results support the robustness of the proposed framework while highlighting that subgroup-specific evaluation is important and that performance may vary with case-mix and documentation practices.

Temporal patterns & early warning capabilities

The model distinguished between transient and persistent (> 48 hours) KDIGO stage 3 episodes predicted at KDIGO stage 3 onset, achieving an AUC-ROC of 0.825–0.889, 0.825.889 and an AUC-PR of 0.699–0.869, 0.699.869 (Appendix Fig. E46a and Table E10).

Effect of sCr definition changes on risk predictions

KDIGO staging based on sCr depends critically on the definition of the patient-specific baseline, which is frequently unavailable in ICU datasets and may vary across institutions and studies. Because our approach forecasts the underlying physiological variables (future sCr and urine output distributions) rather than directly learning a single fixed AKI label, it should be more robust to such definition changes: only the final mapping from predicted trajectories to KDIGO criteria changes, while the model outputs remain unchanged.

To assess this robustness, we repeated the evaluation after redefining KDIGO labels using an alternative baseline sCr definition (minimum historical/available sCr, as used in prior work), without any retraining. Figure 2 (c,d) shows that the distributional regressor maintained strong discrimination and stable probability estimates under this label shift, whereas the benchmark classifier exhibited a larger degradation, consistent with its dependence on the label definition used during training. Quantitatively, across all external cohorts, the distributional model retained high performance for 24-hour stage 3 prediction (AUC-PR 0.662–0.843, 0.662.843; AUC-ROC 0.936–0.981, 0.936.981) and remained well calibrated (ECE 0.004–0.011, 0.004.011; Table E3). Similar robustness was

observed for stage 1 and stage 2 prediction (Tables E5 and E10), where the distributional regressor consistently achieved higher F1 and AUC-PR and lower calibration error than the classifier.

These findings highlight a practical advantage of distributional forecasting for clinical deployment: the same trained model can support alternative baseline assumptions (and, more broadly, evolving diagnostic definitions) through a transparent post-processing step, rather than requiring task-specific retraining whenever the outcome definition is modified.

Discussion

This study developed and externally validated a probabilistic forecasting model that predicts full distributions of sCr and urine output in critically ill patients over multiple horizons (24h, 48h for sCr; 6h, 12h, 24h, 48h for urine output). The model enables early detection of AKI, assessment of its persistence, and stratification across all KDIGO stages, while explicitly quantifying uncertainty. Predictions are summarized in a patient-specific dashboard that flexibly displays the desired outcome together with its predicted probability, providing a potentially actionable tool for bedside decision-making. Across validation cohorts, it achieved high discriminative performance, with mean AUC-PR values of 0.78, 0.76, and 0.78, and maintained good calibration (for 24-hour stage 1, 2, and 3 AKI prediction). When benchmarked against state-of-the-art classifiers, distributional regression demonstrated superior discrimination while providing more robust estimates under definition changes. The most influential predictors were dominated by recent sCr and urine output trends, while clinically coherent features, including age, blood pressure, lactate, electrolytes, and diuretic use, also emerged as important contributors.

To our knowledge, this is the first application of a distributional, multi-step, multivariate forecasting framework for AKI risk stratification in critical care. Its predictive performance was comparable to established

classification-based benchmarks [9–11, 23]. Beyond performance, the approach offers considerably greater flexibility: a single trained model can generate forecasts for all AKI stages without retraining, while accommodating patient-specific sCr baselines, as we demonstrate by evaluating on different baseline definitions. Classification-based approaches, by contrast, are inherently rigid: Alfieri et al. [10] published separate models for distinct prediction tasks, one for stage 2–3 AKI, and another for persistent AKI at 72h [11], which led them to bring to market a CE-certified predictive software solely for stage 2–3 AKI within 24h¹. Such fixed-category tools risk becoming obsolete if diagnostic thresholds are redefined, whereas forecasting the underlying physiological variables of renal function preserves relevance across evolving AKI definitions. This adaptability is particularly timely, as recent evidence suggests that the conventional oliguria cut-off of <0.5 ml/kg/h over 6h may be overly conservative [24]. The model also produces separate predictions for sCr and urine output, each with distinct pathophysiological, prognostic, and therapeutic implications, preserving clinical interpretability. In addition, it provides multivariate probability distributions, which allow intuitive visualization of forecast confidence over time (see Fig. 1).

From a clinical perspective, the model provides richer clinical utility by delivering full predictive distributions of both sCr and urine output rather than a single binary “AKI: yes/no” output. First, it enables stratification across all AKI stages, thereby supporting tailored prevention strategies and the identification of high-risk patients for interventional trials. Additionally, the model may help anticipate the need for KRT in patients most likely to benefit, allowing better planning of both material and personnel resources. Evidence converges toward a delayed strategy for most patients, since waiting can avoid unnecessary treatment when renal recovery occurs without worsening outcomes [25]. Importantly, the diagnostic criterion driving AKI matters: in patients with severe sCr elevation but preserved urine output, early initiation has been associated with increased mortality; conversely, in patients with sustained oliguria, prolonged deferral is unlikely to be beneficial [26]. By jointly forecasting urine output and sCr trajectories, and combining explicit uncertainty quantification, the model could support more personalized decisions on the timing of KRT initiation, tailored to both AKI persistence prediction and its underlying drivers (see Fig. 1 and Fig. E3 for stage 3 AKI driven predominantly by urine output vs. creatinine).

Limitations & future directions

Several limitations should be acknowledged. First, the choice and preprocessing of input features may have influenced performance; alternative feature sets and representations could yield different results. In addition, patient selection may have affected performance, as patients with pre-existing CKD were excluded. This was a deliberate choice to allow a consistent KDIGO-based benchmark for comparison with classification-based models, which rely on baseline creatinine values that are often unavailable or not estimable in CKD. Importantly, the distributional forecasting framework itself does not require such a baseline and can readily be applied to CKD populations in future studies. Second, urine output was normalized by actual body weight, as patient length was unavailable in some datasets. However, recent evidence suggests that normalization by ideal rather than actual body weight leads to a stable incidence of oliguria across weight categories, and provides stronger associations between oliguria and outcomes [27]. In addition, urine output is prone to heterogeneity in documentation and missingness in routine EHR data, introducing unavoidable label noise shared by most urine output-based AKI studies. Third, while more flexible, non-parametric approaches, such as normalizing flows or diffusion models, could also be used to model predictive distributions, we chose tree-based distributional models as challengers of state-of-the-art tree-based classification models, since our primary goal was to demonstrate the benefit of modeling the full predictive distribution rather than training separate classifiers for each stage. Additionally, tree-based models are relatively more transparent and interpretable, which is an essential consideration in this context. Fourth, the model jointly forecasts sCr and urine output, but visualizing and interpreting multivariate uncertainty remains challenging, requiring usability research on visualizations in this context. Fifth, although the model was validated on diverse ICU datasets, all validations were retrospective; prospective, real-time evaluation will be required to confirm robustness and assess clinical impact on patient outcomes. In addition, our study evaluates predictive performance rather than causal effects of treatments: because interventions are time-varying and confounded by indication, quantifying how treatments change predicted trajectories (or using the model to recommend interventions) will require dedicated causal inference designs and prospective studies. Finally, although the input variables are routinely available in most ICUs, integration into electronic health records and clinical workflows will require additional technical development, interface design, and user training.

¹<https://u-caremedical.com/>

Conclusion

In summary, this study developed and externally validated the first multi-step, multivariate, distributional regression model for renal function in critically ill patients. By providing future distributions for both sCr and urine output, rather than relying on categorical AKI stage outputs, the approach improves flexibility, reliability, transparency, and clinical interpretability, key attributes for trustworthy artificial intelligence. Such a model has the potential to guide preventive and early intervention strategies, as well as support planning for KRT, and ultimately foster more personalized AKI management. Prospective studies are warranted to evaluate its clinical utility and impact on patient outcomes.

Abbreviations

AKI	Acute kidney injury
AUC-PR	Area under the precision-recall curve
AUC-ROC	Area under the receiver operating characteristic curve
CE	Conformité Européenne (CE marking/certification)
CDF	Cumulative distribution function
CKD	Chronic kidney disease
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration (equation)
CHU	Centre Hospitalier Universitaire
CORP	Consistent, Optimal, Reproducible, PAV (pool-adjacent violators) algorithm-based
CRPS	Continuous ranked probability score
ECE	Expected calibration error
eGFR	Estimated glomerular filtration rate
eICU	eICU Collaborative Research Database
ES	Energy score
FiO ₂	Fraction of inspired oxygen
F1	F1 score
FN	False negatives
FP	False positives
ICU	Intensive care unit
KDIGO	Kidney Disease: Improving Global Outcomes
KRT	Kidney replacement therapy
MIMIC-IV	Medical Information Mart for Intensive Care IV
mL/kg/h	Milliliters per kilogram per hour
NGBoost	Natural Gradient Boosting
NIBP	Non-invasive blood pressure
NLL	Negative log-likelihood
NPV	Negative predictive value
P/F ratio	PaO ₂ /FiO ₂ ratio
PDF	Probability density function
PPV	Positive predictive value
PR	Precision-recall
sCr	Serum creatinine
Sens	Sensitivity
SHAP	SHapley Additive exPlanations
Spec	Specificity
TN	True negatives
TP	True positives
UO	Urine output
UZGent	University Hospital Ghent (UZ Gent)
AmsterdamUMC (AmsterdamUMCdb)	Amsterdam University Medical Center database

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13054-026-06017-6>.

Supplementary Material 1: See Appendix A, B, C, D, and E

Author contributions

JJ, AR, JV, EH, and SVH contributed to the study conception and design. Data preprocessing, model development and performance analyses were performed by JJ, AR, and JV. JJ and AR drafted the manuscript. All authors revised the manuscript critically for important intellectual content and approved the final version.

Funding

Jef Jonkers is funded by the Research Foundation Flanders (FWO, Ref. 1S11525N). Anahita Rouzé received an international mobility grant from the Interregional Healthcare Cooperation Group of the University Hospitals of Amiens, Caen, Lille, and Rouen (GCS G4). Jarne Verhaeghe is funded by the Research Foundation Flanders (FWO, Ref. 1S59522N). Part of the research was funded through the Research Foundation Flanders senior research project on Trustworthy Time-to-Event Predictions (FWO, Ref. G0AH525N). The funders had no role in the study design; collection, analysis, or interpretation of data; writing of the manuscript; or the decision to submit for publication.

Data availability

This study used de-identified data from four intensive care cohorts. The MIMIC-IV database, the eICU Collaborative Research Database, and AmsterdamUMCdb are publicly available to qualified researchers subject to completion of the respective data access and use agreements and any required training/certification. The UZGent cohort consists of patient data from Ghent University Hospital and is not publicly available due to privacy regulations (including GDPR), institutional policies, and ethics approval restrictions. Access to the UZGent data may be considered upon reasonable request to the corresponding author, subject to approval by Ghent University Hospital and completion of the necessary legal and ethical data-sharing agreements. The code supporting the analyses and findings of this study is available here: <https://github.com/predict-idlab/beyond-binary-aki>.

Declarations

Ethics approval

The study was conducted using de-identified data from publicly available critical care databases (MIMIC-IV, AmsterdamUMCdb, and the eICU Collaborative Research Database). According to the policies governing these databases, their use for research purposes does not require additional institutional review board approval. In addition, the study included data from a local database at Ghent University Hospital, for which approval was obtained from the Ethics Committee of Ghent University Hospital (EC nr 2019/0705). The study was carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments.

Consent to participate

For the publicly available databases, individual informed consent was waived, because the databases contain only de-identified health information that is made available for research in compliance with applicable regulations. For the local database at Ghent, informed consent was obtained according to applicable regulations.

Code availability

The code supporting the analyses and findings of this study is available here: <https://github.com/predict-idlab/beyond-binary-aki>.

Competing Interests

The authors declare no competing interests.

Author details

¹IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

²Univ. Lille, CNRS, Inserm, CHU Lille, U1285 - UMR 8576 - UGSF -Unité de Glycobiologie Structurale et Fonctionnelle, Service de Médecine Intensive Réanimation, 59000 Lille, France

³IDLab, Ghent University - imec, Ghent, Belgium

⁴Department of Nephrology, Ghent University Hospital, Ghent University, Ghent, Belgium

⁵ICU, Ghent University Hospital, Department of Internal Medicine and Pediatrics, Ghent University, Ghent, Belgium

⁶Biometrics Research Group, Department of Morphology, Imaging, Orthopedics, Rehabilitation and Nutrition, Ghent University, Ghent, Belgium

Received: 31 January 2026 / Accepted: 7 April 2026

Published online: 10 May 2026

References

- Hoste EAJ, 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(8):1411–23. <https://doi.org/10.1007/s00134-015-3934-7>.
- Pickkers P, Darmon M, Hoste E, Joannidis M, Legrand M, Ostermann M, et al. Acute kidney injury in the critically ill: an updated review on pathophysiology and management. *Intensive Care Med.* 2021;47(8):835–50. <https://doi.org/10.1007/s00134-021-06454-7>.
- Kashani KB, Awdishu L, Bagshaw SM, Barreto EF, Claire-Del Granado R, Evans BJ, et al. Digital health and acute kidney injury: consensus report of the 27th Acute Disease Quality Initiative workgroup. *Nat Rev Nephrol.* 2023;19(12):807–18. <https://doi.org/10.1038/s41581-023-00744-7>.
- Feng Y, Wang AY, Jun M, Pu L, Weisbord SD, Bellomo R, et al. Characterization of risk prediction models for acute kidney injury: a systematic review and meta-analysis. *JAMA Netw Open.* 2023;6(5):2313359. <https://doi.org/10.1001/jamanetworkopen.2023.13359>.
- Kdigo clinical practice guideline for acute kidney injury. section 2: AKI Definition. *Kidney International Supplements.* 2012;2(1):19–36. <https://doi.org/10.1038/kisup.2011.32>.
- Ostermann M, Legrand M, Meersch M, Srisawat N, Zarbock A, Kellum JA. Biomarkers in acute kidney injury. *Ann Intensive Care.* 2024;14(1):145. <https://doi.org/10.1186/s13613-024-01360-9>.
- De Vlieger G, Kashani K, Meyfroidt G. Artificial intelligence to guide management of acute kidney injury in the ICU: a narrative review. *Curr Opin Crit Care.* 2020;26(6):563–73. <https://doi.org/10.1097/MCC.0000000000000775>.
- Alfieri F, Ancona A, Tripepi G, Randazzo V, Paviglianiti A, Pasero E, et al. External validation of a deep-learning model to predict severe acute kidney injury based on urine output changes in critically ill patients. *J Nephrol.* 2022;35(8):2047–56. <https://doi.org/10.1007/s40620-022-01335-8>.
- Huang C-Y, Güiza F, De Vlieger G, Meyfroidt G. External validation of the AKI predictor in critically ill adults. *Intensive Care Med.* 2022;47(8):952–3. <https://doi.org/10.1007/s00134-022-06746-6>.
- Alfieri F, Ancona A, Tripepi G, Rubéis A, Arjoldi N, Finazzi S, et al. Continuous and early prediction of future moderate and severe acute kidney injury in critically ill patients: Development and multi-centric, multi-national external validation of a machine-learning model. *PLoS One.* 2023;18(7):0287398. <https://doi.org/10.1371/journal.pone.0287398>.
- Zappalà S, Alfieri F, Ancona A, Taccone FS, Maviglia R, Cauda V, et al. Development and external validation of a machine learning model for the prediction of persistent acute kidney injury stage 3 in multi-centric, multi-national intensive care cohorts. *Crit Care.* 2024;28(1):189. <https://doi.org/10.1186/s13054-024-04954-8>.
- Flechet M, Falini S, Bonetti C, Güiza F, Schetz M, Berghe G, et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: A prospective evaluation of the AKI predictor. *Critical Care (London England).* 2019;23(1):282. <https://doi.org/10.1186/s13054-019-2563-x>.
- Huang C-Y, Güiza F, Wouters P, Mebis L, Carra G, Gunst J, et al. Development and validation of the creatinine clearance predictor machine learning models in critically ill adults. *Crit Care.* 2023;27(1):1–9. <https://doi.org/10.1186/s13054-023-04553-z>.
- Gneiting T, Katzfuss M. Probabilistic forecasting. *Annu Rev Stat Appl.* 2014;1(1):125–51.
- Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data.* 2023;10(1):1. <https://doi.org/10.1038/s41597-022-01899-x>.
- 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(1):180178. <https://doi.org/10.1038/sdata.2018.178>.
- Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force: Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/Euro-pean Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. *Crit Care Med.* 2021;49(6):563–77. <https://doi.org/10.1097/CCM.0000000000004916>.
- Zappalà S, Alfieri F, Ancona A, Dell'Anna AM, Kashani KB. External validation of persistent severe acute kidney injury prediction with machine learning model. *Mayo Clin Proc Digit Health.* 2025;3(2):100200. <https://doi.org/10.1016/j.mcpdig.2025.100200>.
- Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New creatinine-and cystatin c-based equations to estimate gfr without race. *N Engl J Med.* 2021;385(19):1737–49.
- Verhaeghe J, Van Der Donck J, Ongenaes F, Van Hoeck S. Powershap: a power-full shapley feature selection method. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 71–87 (2022). Springer
- Duan T, Avati A, Ding DY, Thai KK, Basu S, Ng AY, et al. NGBoost: natural gradient boosting for probabilistic prediction. *arXiv.* 2020. <https://doi.org/10.48550/arXiv.1910.03225>.
- Dimitriadis T, Gneiting T, Jordan AI. Stable reliability diagrams for probabilistic classifiers. *Proc Natl Acad Sci U S A.* 2021;118(8):201619118.
- Flechet M, Güiza F, Schetz M, Wouters P, Vanhorebeek I, Derese I, et al. AKI predictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. *Intensive Care Med.* 2017;43(6):764–73. <https://doi.org/10.1007/s00134-017-4678-3>.
- Bianchi NA, Altarelli M, Monard C, Kelevina T, Chaouch A, Schneider AG. Identification of an optimal threshold to define oliguria in critically ill patients: an observational study. *Crit Care.* 2023;27(1):1–10. <https://doi.org/10.1186/s13054-023-04505-7>.
- Barbar SD, Wald R, Quenot J-P. Acute kidney injury: when and how to start renal replacement therapy. *Intensive Care Med.* 2025;51(6):1172–5. <https://doi.org/10.1007/s00134-025-07933-x>.
- Barbar SD, Bourredjem A, Trusson R, Dargent A, Binquet C, Quenot J-P. Differential effect on mortality of the timing of initiation of renal replacement therapy according to the criteria used to diagnose acute kidney injury: an IDEAL-ICU substudy. *Crit Care.* 2023;27(1):1–9. <https://doi.org/10.1186/s13054-023-04602-7>.
- Monard C, Tebib N, Trächsel B, Kelevina T, Schneider AG. Comparison of methods to normalize urine output in critically ill patients: A multicenter cohort study. *Crit Care.* 2024;28(1):1–10. <https://doi.org/10.1186/s13054-024-05200-x>.
- Henzi A, Ziegel JF, Gneiting T. Isotonic distributional regression. *J R Stat Soc Ser B Stat Methodol.* 2021;83(5):963–93.
- Allen S, Gavrilopoulos G, Henzi A, Kleger G-R, Ziegel J. In-sample calibration yields conformal calibration guarantees. *arXiv preprint arXiv:2503.03841* (2025)
- Vovk V, Petej I. Venn-abers predictors. *arXiv preprint arXiv:1211.0025* (2012)

Publisher's Note

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