

Six-hour landmark prediction of poor subsequent urine output in critically ill adults with heart failure: model development and external validation

Xiaohuan Zhuang

Guangdong Provincial Second Hospital of Traditional Chinese Medicine

Xinyuan Zhang

Nanjing University of Chinese Medicine

Tongnian Fang

Ftn111@163.com

Guangdong Provincial Second Hospital of Traditional Chinese Medicine

Research Article

Keywords: heart failure, intensive care, urine output, landmark prediction, external validation, machine learning, calibration, clinical decision support

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10958487/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Early response to loop diuretics may help identify critically ill patients with heart failure who will have low subsequent urine output. We assessed whether information recorded during the first six hours improved prediction across intensive care databases.

Methods

We developed models in MIMIC-IV and externally evaluated them in eICU-CRD. The prediction landmark was six hours after the first qualifying intravenous loop-diuretic record. Poor subsequent response was defined as mean urine output during hours 6–24 at or below the development-cohort lower quartile (approximately 0.718 mL/kg/h). Four candidate algorithms underwent subject-grouped, candidate-specific nested cross-validation. A development-selected XGBoost model was applied unchanged externally. Evaluation included area under the receiver operating characteristic curve (AUROC), calibration, threshold performance, decision curves and inverse-probability-of-inclusion weighting.

Results

The development cohort included 1,450 stays and 363 events. The external cohort included 545 stays and 147 events across 69 hospitals. Development out-of-fold AUROC was 0.760 (95% confidence interval 0.732–0.787), conditional on algorithm-family selection and potentially optimistic. External AUROC was 0.729 (0.690–0.789). Compared with baseline elastic-net regression, external AUROC increased by 0.080 (0.028–0.126). However, early-response elastic-net regression achieved a higher external AUROC of 0.756; the paired XGBoost-minus-elastic-net difference was – 0.027 (– 0.057 to – 0.004). External calibration slope was 0.682. At the locked threshold, sensitivity was 75.5%, specificity 56.0% and positive predictive value 38.8%; incremental net benefit over reviewing all patients was uncertain. Inclusion-weighted external AUROC was 0.690 (0.632–0.750).

Conclusions

Six-hour early-response information improved prediction of subsequent low urine output, but the more complex algorithm showed no advantage. Calibration limitations, selection for observable outcomes and measurement differences between databases limit immediate clinical use. Prospective evaluation should compare parsimonious models and establish whether risk-guided review improves care.

Background

Intravenous loop diuretics are central to treating congestion in hospitalized patients with heart failure, but the response to an initial dose varies ^[1, 2]. Randomized trials have evaluated several approaches to

intensifying diuresis. DOSE compared loop-diuretic dosing strategies, whereas ADVOR and CLOROTIC tested adjunctive acetazolamide and hydrochlorothiazide, respectively [3, 4, 5]. These trials support attention to decongestion and safety rather than treating urine volume alone as the therapeutic objective. In intensive care, interpreting a low output additionally requires assessment of perfusion, kidney function, administered therapy and measurement completeness. Diuretic reviews and the Heart Failure Association position statement emphasize that decongestion must be assessed alongside renal function and electrolyte safety [6, 7]. The 2024 ACC inpatient pathway likewise recommends assessing the clinical trajectory rather than relying on an isolated response measure [8].

Early treatment response may provide information beyond that available at admission. Physiological work showed that urine measurements shortly after a loop-diuretic dose can anticipate subsequent sodium excretion [9]. PUSH-AHF subsequently demonstrated greater natriuresis with a natriuresis-guided strategy, without establishing a reduction in its longer-term clinical composite outcome [10]. ENACT-HF, a two-phase pragmatic implementation study, and the randomized EASY-HF study further examined structured urine-sodium-guided care [11, 12]. The 2025 European Journal of Heart Failure expert consensus reviews these findings and identifies unresolved questions about early response assessment [13]. Urine volume and sodium excretion are related, but routinely recorded output does not directly measure natriuresis. Physiological studies have identified renal tubular response as an important determinant of diuretic resistance, which cannot be inferred from urine volume alone [14]. Residual congestion is also associated with adverse outcomes, reinforcing the distinction between a recorded output endpoint and clinically successful decongestion [15].

Several studies have developed and evaluated models for predicting diuretic response. The BAN-ADHF score was developed using pooled acute-heart-failure trial data and validated in additional cohorts [16]. Recent studies have extended its evaluation to ICU patients and other acute-heart-failure populations, including a 2026 ICU external validation, RESIST, and a post hoc analysis of PUSH-AHF [17, 18, 19]. An unresolved question is whether an early observation window provides transferable predictive information about response during a subsequent, non-overlapping interval. Evaluation should also compare simpler models and assess the effects of incomplete observation.

We therefore developed a six-hour landmark model for poor urine-output response during the following 18 hours in critically ill adults with heart failure. We quantified the predictive value of early-response information beyond baseline assessment. We then evaluated the development-selected model without modification in a separate multicentre database. Additional analyses compared algorithms and assessed calibration and performance at a clinical-review threshold. Sensitivity analyses examined missing-data handling, predictor composition and inclusion in the external cohort. The intended use was prioritization for clinical review, not automated selection or escalation of diuretic treatment.

Methods

Study design, data sources and reporting framework

We developed prediction models retrospectively in the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. External validation used the eICU Collaborative Research Database (eICU-CRD) [20, 21]. MIMIC-IV provides single-centre hospital and ICU records, whereas eICU-CRD includes multicentre US ICU records. The database releases were MIMIC-IV version 3.1 and eICU-CRD version 2.0 [22, 23]. The source databases cover 2008–2022 and 2014–2015, respectively. These periods do not imply that eligible study records were available in every year. Reporting followed the structure of TRIPOD + AI, with STROBE and RECORD guiding the description of routinely collected data [24, 25, 26]. TRIPOD-Cluster informed reporting of between-hospital heterogeneity in the external database [27]. Both databases are distributed through PhysioNet, a controlled research-data platform [28].

Population and landmark construction

Adults were eligible when a qualifying intravenous loop-diuretic record occurred within six hours of ICU admission. The first qualifying dose record defined time zero; the prediction landmark occurred six hours later. We used landmarking to separate the predictor and outcome windows. The analysis did not involve fitting a competing-risks survival model [29]. We retained the first qualifying ICU record per hospitalization. Repeat hospitalizations could contribute to development, with all stays from the same individual assigned to the same validation fold. MIMIC-IV heart failure was identified using hospital diagnostic codes whose descriptions indicated acute heart failure. eICU-CRD used a broader admission/diagnosis-based heart-failure phenotype; the external population therefore did not have an identical acute-heart-failure case definition.

Eligibility required a body weight of 20–350 kg, survival and ICU residence through the landmark, and observable outcomes through hour 24. Records also required at least three charted hourly urine-output bins in hours 0–6 and at least four in hours 6–24. MIMIC-IV exclusions addressed end-stage/chronic dialysis-dependent kidney disease, kidney transplantation, urinary obstruction, renal replacement therapy before or during the predictor window, and relevant genitourinary irrigation records. External ascertainment was more limited and included the available admission dialysis indicator; equivalent ascertainment of every MIMIC-IV exclusion was not available. Figure 1 summarizes cohort construction. Restrictions based on subsequent outcome observability could introduce selection after the landmark.

Outcome definition

The primary outcome was poor future urine-output response, defined by total recorded urine during hours 6–24 divided by body weight and 18 hours. The lower-quartile threshold was estimated once in the full development cohort and transported unchanged externally. A value at or below that threshold constituted an event. The outcome measured relative urine-output response within the cohort. It did not diagnose diuretic resistance, quantify sodium excretion or estimate a dose-normalized treatment effect.

Urine entries were aggregated within prespecified time bins. After minimum charting-density criteria were applied, unrecorded component intervals contributed zero recorded volume to the aggregate. This

convention does not establish physiological anuria during those intervals. Subsequent treatment was left as delivered under usual care, and positive future diuretic exposure was not required for the primary endpoint. Alternative analyses examined mean output below 0.5 mL/kg/h, dose-normalized response among those with recorded subsequent exposure, and selected clinical outcomes. A mean output threshold across 18 hours should not be equated with a fully adjudicated acute-kidney-injury diagnosis. Previous studies defined diuretic efficiency using fluid output relative to loop-diuretic exposure. Our primary endpoint did not normalize urine output to diuretic dose ^[30].

Predictors and database harmonization

Baseline predictors included age, sex, body weight, pre-index six-hour urine output, mean heart rate, and systolic and mean arterial pressure. Additional predictors were creatinine, blood urea nitrogen, sodium, potassium, initial loop-diuretic dose, vasoactive exposure and mechanical ventilation. Laboratory values used the nearest eligible pre-index measurement within the specified 24-hour look-back window. The early-response model added cumulative weight-normalized urine output during hours 0–6 and the number of charted hourly urine bins. It also included integrated mean-arterial-pressure deficit below 65 mmHg, cumulative loop-diuretic dose during hours 0–6, and vasoactive escalation. All predictor calculations ended by the landmark. Definitions and missingness are detailed in Supplementary Tables S1 and S2.

For analysis, doses were converted to intravenous furosemide equivalents using multipliers of 1 for furosemide, 40 for bumetanide and 2 for torsemide. The 0–6-hour cumulative exposure included the index exposure; it was not an additional-dose-only variable. MIMIC-IV infusion exposure was allocated according to overlap with the observation window. Medication orders and start records in eICU-CRD do not identify administration times in the same way, limiting harmonization. Blood-pressure deficit integrated the depth and duration below the threshold, with bounded carry-forward of observations in the MIMIC-IV extraction. Vasoactive escalation represented a new start or increased number of agents. SOFA and Charlson scores were described in development data but excluded from the primary harmonized predictor set because comparable external values were unavailable.

Missing data and model development

The primary pipeline used median imputation with missingness indicators, fitted only within the corresponding training partition. Missing-data handling was incorporated into the prediction pipeline to reflect information available at application. External performance was not used to select the missing-data strategy ^[31]. Elastic-net regression additionally used robust scaling. Four early-response candidate families were evaluated: elastic-net logistic regression, random forest, LightGBM and XGBoost ^[32, 33, 34, 35]. A baseline-only elastic-net model served as the clinical-information comparator. Five outer folds and five inner folds used stratified, subject-grouped splitting. Within each candidate family, hyperparameters were chosen by inner-fold log loss; preprocessing was repeated inside training partitions. The random seed was 20260903, with fold-specific offsets for inner splitting. Fold-level results and selected hyperparameters are supplied in Supplementary Table S3.

Among candidates within 0.01 of the best outer-fold AUROC, selection prioritized the lowest Brier score, with a 0.002 tolerance favouring the predefined simpler family. Final XGBoost settings were 350 trees, learning rate 0.03, maximum depth 3 and minimum child weight 5, selected using grouped development-data tuning. The model and operating threshold were locked before primary external performance evaluation; feasibility and cohort-level descriptive information had already been examined. No external result was used to replace the locked model. The random-forest comparator used balanced class weights recalculated within each tree bootstrap sample; the other candidates did not use class weighting. No synthetic oversampling or undersampling was used. The random-forest probabilities were evaluated without a separate recalibration step.

Hyperparameter tuning was nested within each candidate, but family selection subsequently used the candidate outer-fold summaries. Internal estimates for the selected model are conditional on family selection and may remain optimistic. They do not provide an unbiased evaluation of the complete selection procedure. Independent evaluation of model selection is necessary to avoid this source of bias^[36]. Post-lock analyses were kept separate from the primary locked estimates, consistent with transparent development and external-validation practice^[37, 38].

Performance and uncertainty

We assessed discrimination using AUROC and area under the precision–recall curve (PR-AUC). Brier score and log loss measured overall probabilistic performance. PR-AUC supplements, rather than replaces, AUROC and depends on event prevalence^[39]. Calibration assessment included an intercept and slope estimated jointly by logistic regression on predicted log odds, flexible calibration curves and the observed-to-expected event ratio^[40]. Primary external predictions were not recalibrated. Confidence intervals for performance and paired differences used 1,000 bootstrap samples, clustering by individual in MIMIC-IV and hospital in eICU-CRD. These analyses resampled fixed predictions; they did not repeat imputation fitting, training, family selection or threshold selection. Post-lock algorithm comparisons were exploratory and not multiplicity-adjusted.

We selected the operating threshold using pooled out-of-fold predictions from the selected model. A fixed 0.001 probability grid maximized development specificity subject to sensitivity of at least 0.80. The same threshold was applied externally. Sensitivity, specificity and predictive-value intervals used the Wilson method; these conditional intervals do not account for clustering or threshold-selection uncertainty. Decision-curve analysis compared risk-guided review with review-all and review-none strategies^[41]. The proposed action was structured senior review of congestion, haemodynamics, renal function, treatment exposure and data completeness, not automatic dose escalation. Thresholds represented hypothetical relative costs of unnecessary review and missed events. These costs were not derived from patient or clinician preferences^[42].

Sensitivity analyses and model interpretation

Development-only missing-data benchmarks compared the selected strategy with native missing-value handling, iterative Bayesian-ridge and extra-trees imputation, and median imputation without indicators. These were alternative prediction pipelines, not multiple-imputation inference combined by Rubin's rules [43]. A tolerance of 0.005 in log loss and 0.002 in Brier score supported a parsimonious choice. Predictor-ablation and expanded-panel analyses used fixed XGBoost settings and the original outer partitions without new hyperparameter searches. Complete-case performance used original out-of-fold predictions in the complete subset rather than refitting a complete-case model. Other analyses restricted recording completeness, diagnostic definitions or repeated admissions and explored hospital-grouped cross-fitted recalibration. For the composite of low output, death or new renal replacement therapy before hour 24, we reconstructed the five primary-outcome outer-fold models using their original selected parameters. Original participants retained their held-out fold. Additional participants were assigned to the fold containing any other admission from that individual. Those without a development admission were assigned deterministically by subject identifier. Each composite prediction therefore came from a model trained without that individual. This was a post-lock assessment of the primary score, not development of a composite-outcome model. Bootstrap intervals resampled individuals with predictions fixed.

To assess measured selection in eICU-CRD, a five-fold hospital-grouped penalized logistic model estimated inclusion probability among 2,267 stays clinically eligible through the landmark. Predictors included available demographic, physiological, renal, treatment and hospital-volume characteristics. Training-fold preprocessing handled numeric and categorical missingness. Probabilities were bounded at 0.02 and 0.98. Stabilized inverse-probability-of-inclusion weights were truncated at the first and 99th percentiles among included stays and normalized to mean one. In this study, IPCW denotes inverse-probability-of-inclusion weighting; it does not refer to a time-to-event censoring model. Weighting requires assumptions about measured selection and overlap [44]. Hospital bootstrap intervals held selection probabilities and truncation limits fixed, so selection-model estimation uncertainty was not propagated.

A post hoc external-validation precision assessment used the development risk distribution and precision targets rather than observed external model performance [45, 46]. It indicated that approximately 2,609 observations would be needed for a calibration-slope interval width of 0.20 under its planning assumptions, compared with 545 available observations. This was not an a priori sample-size justification (Supplementary Table S4). TreeSHAP contributions were obtained from development outer-fold models and expressed on the log-odds scale [47]. Reconstructed predictions were checked against saved out-of-fold predictions. Global displays and illustrative low-, middle- and high-risk cases describe model behaviour, not causal mechanisms. Analyses used Python 3.11.3, scikit-learn 1.6.1, XGBoost 3.0.2 and LightGBM 4.6.0.

ChatGPT/Codex (OpenAI) provided language editing and coding assistance (see Declarations).

Results

Cohort selection and characteristics

The development cohort comprised 1,450 ICU stays from 1,331 individuals in MIMIC-IV. The external cohort comprised 545 stays from 69 eICU-CRD hospitals (Fig. 1). Poor future urine-output response occurred in 363 development stays (25.0%) and 147 external stays (27.0%). The development-derived threshold was 0.7176404163 mL/kg/h, approximately 0.718 mL/kg/h, and was transported unchanged to the external cohort.

Median age was 72 years in MIMIC-IV and 74 years in eICU-CRD; 49.7% and 53.4% were male, respectively (Table 1). Median creatinine was 1.50 versus 1.26 mg/dL, and median blood urea nitrogen was 35 versus 27 mg/dL. Vasoactive exposure was more frequent in development than external data (20.8% versus 11.4%). Ventilation was recorded in 49.6% versus 19.4% of stays with observed values. The development cohort had a median SOFA score of 2 and Charlson comorbidity index of 7; these scores were not harmonized externally. Missingness and distributional differences are reported in Supplementary Table S2 and Supplementary Figure S1.

Table 1
Baseline characteristics at the index loop-diuretic dose

Characteristic	Development cohort (MIMIC-IV; n = 1,450)	External validation cohort (eICU-CRD; n = 545)
Demographic Characteristics		
Age, years	72.0 (62.0–81.0)	74.0 (64.0–83.0)
Male sex, n (%)	720 (49.7)	291 (53.4)
Body weight, kg	82.6 (68.8–99.2)	83.9 (67.6–102.5)
Pre-Index Clinical Measurements		
Pre-index urine output, mL	250.0 (120.0–463.8)	430.0 (200.0–761.5)
Mean heart rate, beats/min	87.2 (74.4–103.0)	86.4 (72.6–100.8)
Mean systolic blood pressure, mm Hg	118.5 (105.3–132.3)	125.8 (110.4–143.2)
Minimum systolic blood pressure, mm Hg	108.0 (94.0–124.0)	110.0 (97.0–130.0)
Mean arterial pressure, mm Hg	79.4 (71.0–88.6)	84.0 (73.6–96.3)
Minimum arterial pressure, mm Hg	71.0 (62.0–81.0)	74.0 (64.0–86.0)
Minimum peripheral oxygen saturation, %	94.0 (91.0–97.0)	94.0 (90.0–97.0)
Pre-Index Laboratory Measurements		
Serum creatinine, mg/dL	1.50 (1.10–2.20)	1.26 (0.97–1.80)
Blood urea nitrogen, mg/dL	35.0 (22.0–55.0)	27.0 (19.0–43.0)
Serum sodium, mmol/L	138.0 (135.0–141.0)	137.0 (134.0–141.0)
Serum potassium, mmol/L	4.30 (3.90–4.80)	4.30 (3.90–4.80)
Hemoglobin, g/dL	10.4 (8.8–12.3)	11.3 (9.6–13.0)
White blood cell count, 10 ⁹ /L	10.4 (7.7–13.7)	10.4 (7.8–14.0)
Platelet count, 10 ⁹ /L	205.0 (156.0–271.8)	210.5 (167.8–278.0)
Serum albumin, g/dL	3.30 (3.00–3.60)	3.40 (2.92–3.70)
Serum lactate, mmol/L	1.60 (1.20–2.40)	1.70 (1.10–2.60)
Note. Values are median (interquartile range) or n (%), unless otherwise indicated. Percentages for categorical variables use ICU stays with observed values as the denominator. Missingness and cross-cohort distributional differences are reported in Supplementary Table S2. SOFA score and Charlson comorbidity index were not harmonized in eICU-CRD. No between-cohort significance testing was performed.		

Characteristic	Development cohort (MIMIC-IV; n = 1,450)	External validation cohort (eICU-CRD; n = 545)
Demographic Characteristics		
Severity And Comorbidity		
SOFA score, points	2.0 (1.0–4.0)	Not harmonized
Charlson comorbidity index, points	7.0 (5.0–8.0)	Not harmonized
Pre-Index Support And Treatment		
Initial loop-diuretic dose, mg furosemide equivalent	40.0 (20.0–80.0)	40.0 (40.0–40.0)
Vasoactive support at index, n (%)	301 (20.8)	62 (11.4)
Mechanical ventilation at index, n (%)	719 (49.6)	97 (19.4)
Outcome during 6–24 h		
Poor future urine-output response, n (%)	363 (25.0)	147 (27.0)
Note. Values are median (interquartile range) or n (%), unless otherwise indicated. Percentages for categorical variables use ICU stays with observed values as the denominator. Missingness and cross-cohort distributional differences are reported in Supplementary Table S2. SOFA score and Charlson comorbidity index were not harmonized in eICU-CRD. No between-cohort significance testing was performed.		

Table 2
A. Candidate-specific internal validation

Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Brier score (95% CI)	Log loss (95% CI)	Selected
Baseline elastic-net logistic	0.640 (0.610–0.673)	0.375 (0.333–0.423)	0.179 (0.168–0.189)	0.540 (0.515–0.564)	No
Early-response elastic-net logistic	0.757 (0.728–0.785)	0.518 (0.468–0.572)	0.158 (0.148–0.169)	0.489 (0.460–0.517)	No
Early-response LightGBM	0.752 (0.724–0.780)	0.534 (0.479–0.589)	0.161 (0.147–0.174)	0.513 (0.470–0.555)	No
Early-response random forest	0.763 (0.733–0.790)	0.552 (0.499–0.610)	0.166 (0.156–0.175)	0.504 (0.480–0.528)	No
Early-response XGBoost	0.760 (0.732–0.787)	0.554 (0.505–0.607)	0.154 (0.142–0.165)	0.477 (0.445–0.506)	Yes
Note. Estimates and 95% confidence intervals use 1,000 subject-cluster bootstrap samples. Within each candidate algorithm, hyperparameters were tuned inside the outer training partition using grouped inner cross-validation. The algorithm family was then selected using the outer-fold summaries; reported selected-model out-of-fold performance is therefore conditional on that selection and may retain selection optimism. Bootstrap intervals resample fixed predictions and do not repeat training or family selection.					

Table 2B. External comparison with early-response elastic-net regression

Metric	XGBoost	Early-response elastic-net	Difference XGBoost minus elastic-net (95% CI)
ROC-AUC	0.729	0.756	-0.027 (-0.057 to -0.004)
PR-AUC	0.500	0.528	-0.028 (-0.069 to 0.016)
Brier score	0.174	0.166	0.00813 (0.00007 to 0.01604)
Log loss	0.530	0.507	0.02357 (0.00047 to 0.04837)

Note. Differences are XGBoost minus elastic-net with 95% confidence intervals from 1,000 paired hospital-cluster bootstrap samples. Higher ROC-AUC and PR-AUC, and lower Brier score and log loss, are better. These post-lock comparisons are exploratory and not multiplicity-adjusted. The comparator did not replace the locked primary model. Full paired results for both cohorts are in Supplementary Table S5.

Internal validation and incremental information

The baseline elastic-net model achieved an AUROC of 0.640 (95% CI 0.610–0.673), compared with 0.757 (0.728–0.785) for elastic-net regression incorporating early-response information. Early-response random

forest, LightGBM and XGBoost achieved AUROCs of 0.763, 0.752 and 0.760, respectively (Table 2A; Fig. 2). XGBoost was selected by the development-only discrimination-and-Brier-score rule, rather than by the highest AUROC alone. Its PR-AUC was 0.554 (0.505–0.607), Brier score 0.154 (0.142–0.165) and log loss 0.477 (0.445–0.506).

Compared with the baseline elastic-net model, the selected early-response model increased development AUROC by 0.120 (95% CI 0.087–0.154) and PR-AUC by 0.180 (0.128–0.230). The Brier score also decreased by 0.025 (0.017–0.034) in development. In the external cohort, the corresponding AUROC difference was 0.080 (0.028–0.126), PR-AUC difference 0.092 (0.026–0.146), and Brier-score difference – 0.013 (– 0.023 to – 0.002). These comparisons combined changes in predictor information and model family. Elastic-net comparisons and ablation analyses separately assessed the contribution of early-response information.

The post-lock comparison did not support an external advantage of XGBoost over early-response elastic-net regression. External elastic-net AUROC was 0.756, compared with 0.729 for XGBoost; the paired difference was – 0.027 (95% CI – 0.057 to – 0.004). Brier scores were 0.166 and 0.174, respectively, with a difference of 0.00813 (0.00007–0.01604) when expressed as XGBoost minus elastic-net. Log loss also favoured elastic-net, whereas the PR-AUC difference was uncertain (Table 2B; Supplementary Table S5). Internal paired intervals crossed zero. These exploratory comparisons did not alter the locked primary model.

External discrimination and calibration

The unchanged XGBoost model had an external AUROC of 0.729 (95% CI 0.690–0.789) and PR-AUC of 0.500 (0.432–0.570). Brier score was 0.174 (0.143–0.196), and log loss was 0.530 (Table 3; Fig. 3). Calibration slopes were 0.804 (0.701–0.921) internally and 0.682 (0.545–0.950) externally. The jointly estimated external calibration intercept was – 0.115 (– 0.352 to 0.170). This differs from calibration-in-the-large, which fixes the slope at one. The external observed-to-expected event ratio was 1.144, indicating that average predicted risk was lower than the observed event proportion (Supplementary Table S6; Supplementary Figure S2 and Supplementary Figure S3).

Table 3
Performance of the final early-response prediction model

Cohort	n	Events	ROC-AUC (95% CI)	PR-AUC (95% CI)	Brier score (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)
MIMIC-IV	1,450	363	0.760 (0.732–0.787)	0.554 (0.505–0.607)	0.154 (0.142–0.165)	-0.169 (-0.329–-0.012)	0.804 (0.701–0.921)
eICU-CRD	545	147	0.729 (0.690–0.789)	0.500 (0.432–0.570)	0.174 (0.143–0.196)	-0.115 (-0.352–0.170)	0.682 (0.545–0.950)

Note. Values are estimates with 95% confidence intervals. Confidence intervals were estimated using subject-clustered bootstrap for MIMIC-IV and hospital-clustered bootstrap for eICU-CRD. The model was developed in MIMIC-IV and evaluated unchanged in eICU-CRD. Internal estimates are conditional on algorithm-family selection; these intervals do not capture full model-development uncertainty.

Operating threshold and decision-curve analysis

At the development-selected threshold of 0.161, development sensitivity was 80.2% and specificity was 56.5%. External sensitivity was 75.5% (95% CI 68.0–81.8%), specificity 56.0% (51.1–60.8%), positive predictive value 38.8% (33.3–44.6%) and negative predictive value 86.1% (81.4–89.8%) (Table 4). The rule flagged 286 of 545 external stays (52.5%), including 111 true positives and 175 false positives; 36 events were missed. Thus, the development sensitivity target was not reproduced externally.

Table 4
Classification performance at the development-selected operating threshold

Cohort	n (events)	Threshold	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	TP/FP/TN/FN
MIMIC-IV	1,450 (363)	0.161	0.802 (0.758–0.839)	0.565 (0.535–0.594)	0.381 (0.347–0.416)	0.895 (0.870–0.916)	291/473/614/72
eICU-CRD	545 (147)	0.161	0.755 (0.680–0.818)	0.560 (0.511–0.608)	0.388 (0.333–0.446)	0.861 (0.814–0.898)	111/175/223/36

Note. The threshold of 0.161 was selected in MIMIC-IV on a fixed 0.001 grid to maximize specificity subject to sensitivity of at least 0.80 and was then applied unchanged to eICU-CRD. Confidence intervals are Wilson 95% intervals. PPV, positive predictive value; NPV, negative predictive value; TP, true positive; FP, false positive; TN, true negative; FN, false negative. The development estimates are conditional on threshold selection and should not be described as independent validation of threshold performance. Wilson intervals do not account for clustering or threshold-selection uncertainty.

At a risk threshold of 0.161, external net benefit was 0.142 for model-guided review and 0.130 for reviewing all eligible patients. The difference was 0.012 (95% CI – 0.015 to 0.046), leaving the advantage over reviewing everyone uncertain at this operating point. Differences were more favourable at selected higher

thresholds, including 0.065 (0.029–0.114) at 0.25 and 0.112 (0.073–0.163) at 0.30 (Supplementary Table S7). These analyses quantify a hypothetical review policy, not observed improvements in patient outcomes.

Interpretation and sensitivity analyses

Early cumulative urine output contributed most strongly to the model's predictions, followed by cumulative loop-diuretic exposure during hours 0–6, body weight and the initial dose (Fig. 4). Individual TreeSHAP explanations illustrated how combinations of predictors shifted estimated log odds; they did not identify treatment effects. Removing early dose information reduced development discrimination, but the external difference remained uncertain (AUROC difference – 0.013, 95% CI – 0.032 to 0.011). Removing urine-output charting density had little effect externally (difference – 0.002, – 0.007 to 0.006). A model containing baseline variables and early urine output achieved an external AUROC of 0.723, close to the full model's 0.729 (Supplementary Table S8; Supplementary Figure S4).

Development-only comparisons showed little difference between median imputation with indicators, native missing-value handling and iterative strategies. AUROCs ranged from 0.759 to 0.762. Median imputation with missingness indicators was retained under the prespecified parsimony rule. This choice was specific to the benchmark (Supplementary Table S9; Supplementary Figure S5). Adding eight common clinical variables yielded only a small development change and lower external discrimination (0.711 versus 0.729; difference – 0.018, 95% CI – 0.038 to – 0.004; Supplementary Table S10). Other analyses examined MIMIC-IV-specific severity variables, complete cases, stricter recording completeness and narrower heart-failure definitions (Supplementary Tables S10, S11, S12 and S13; Supplementary Table S14). Development subgroup estimates were descriptive (Supplementary Table S15).

Outcome observability substantially restricted external inclusion. Only 545 of 2,267 stays meeting clinical eligibility through the landmark entered the primary analysis (24.0%). Cross-fitted inclusion weighting produced an effective sample size of 386.2. Weighted AUROC was 0.690 (95% CI 0.632–0.750), Brier score 0.198 (0.169–0.218), and calibration slope 0.557 (0.413–0.810) (Supplementary Tables S16 and S17; Supplementary Table S18; Supplementary Figure S6). These estimates were less favourable than the primary unweighted results. Excluding the largest contributing hospital left 418 stays in 68 hospitals and an AUROC of 0.744. Hospital and subgroup results were descriptive and often imprecise (Supplementary Tables S19 and S20; Supplementary Table S21; Supplementary Figure S7 and Supplementary Figure S8).

For the alternative endpoint of mean subsequent urine output below 0.5 mL/kg/h, AUROC was 0.786 internally and 0.743 externally. The primary score had an AUROC of 0.595 for mortality during the 28 days after the 24-hour response-window endpoint in MIMIC-IV. This result does not support its use as a general mortality model. The composite analysis included 1,470 stays and 384 events, comprising the 1,450 development stays and 20 additional stays. Subject-held-out predictions yielded an AUROC of 0.762 (95% CI 0.735–0.790), PR-AUC of 0.574 and Brier score of 0.156. These estimates assess transport of the primary score to a different endpoint and do not establish a validated composite-outcome probability model. Dose-normalized response analyses were limited by subsequent treatment exposure and different medication-record semantics (Supplementary Table S22 and Supplementary Table S23; Supplementary Figure S9). Exploratory hospital-grouped cross-fitted recalibration yielded an external Brier score of 0.172

and calibration slope of 0.944, but did not replace the primary unrecalibrated evaluation (Supplementary Table S24).

Discussion

Principal findings

Information recorded during the first six hours after a qualifying loop-diuretic record improved prediction in critically ill patients with heart failure. That information improved performance relative to baseline assessment and retained moderate discrimination in a separate multicentre database. A simpler early-response elastic-net model nevertheless outperformed the locked XGBoost model externally. External predictions were overly extreme, and performance was less favourable after weighting for measured selection. These findings support the predictive value of early-response information while identifying limitations to its use across databases. They do not establish readiness for clinical deployment.

Relationship to response-guided care and existing prediction research

Our findings align with the clinical rationale for assessing response soon after treatment begins, but should not be conflated with evidence for a therapeutic strategy. PUSH-AHF, ENACT-HF and EASY-HF evaluated protocols using urinary sodium, whereas our model predicted a later recorded urine-volume endpoint ^[10, 11, 12]. Greater diuresis does not necessarily indicate adequate sodium removal. The benefits and harms of intensifying treatment depend on congestion and perfusion, not urine volume alone. The 2025 expert consensus highlights urine-sodium assessment as an evolving approach, not a justification for equating low output with a single mechanism ^[13]. Analyses from PUSH-AHF also indicate that renal-function changes during guided decongestion need clinical context, supporting clinical interpretation of predicted low output rather than its use as an automatic dose instruction ^[48].

Our study extends existing research on prediction of diuretic response, including BAN-ADHF. The original BAN-ADHF study developed a practical score in acute-heart-failure trial populations. Subsequent studies evaluated ICU applicability, modified-score performance and associations with response to natriuresis-guided care ^[16, 17, 18, 19]. Our design differs in the six-hour landmark, subsequent non-overlapping 18-hour outcome window, routine-record predictors and explicit examination of inclusion-related transportability. The published ICU BAN-ADHF validation reported discrimination for dose-efficiency outcomes, which are not identical to ours ^[17]. These endpoint differences preclude a meaningful ranking of AUROCs across studies. A head-to-head comparison with BAN-ADHF using harmonized endpoints remains necessary to assess the practical value of this approach.

Information content versus algorithmic complexity

The early-response elastic-net comparison and ablation analyses suggest that the information observed during treatment may matter more than nonlinear model complexity. Early urine output was the dominant predictor, and a baseline-plus-early-output model approached the full model externally. This is clinically

plausible because the predictor and outcome measure related physiology in consecutive periods. Persistence of urine output across consecutive periods is expected and does not, by itself, establish a new mechanism. Its clinical usefulness depends on whether it supports reliable decisions after accounting for documentation and treatment context.

The external advantage of elastic-net regression argues against describing XGBoost as the best-performing or optimal clinical model. Broader comparative literature has likewise cautioned that machine learning does not consistently outperform logistic regression in clinical prediction, especially when comparisons are methodologically sound ^[49]. Our post-lock comparison was exploratory and cannot establish universal superiority of elastic-net either. In this study, internal selection of a more complex model did not translate into better external performance. Prospective evaluation should therefore include the simpler early-response model alongside XGBoost.

Calibration, selection and potential clinical use

Discrimination alone was insufficient to support implementation. An external calibration slope below one indicated that risk differences were too extreme, while the observed-to-expected ratio showed underprediction of average event frequency. Both findings can coexist because calibration slope and average risk assess different properties ^[40]. Cross-fitted recalibration suggested some room for improvement, but estimating an updating procedure and evaluating its future performance require additional data. The precision assessment indicated that the available sample was insufficient to estimate calibration slope within the specified interval-width target ^[45, 46].

Outcome observability imposed a stronger restriction than the final sample size alone suggests. Only about one quarter of clinically eligible external stays entered the primary analysis, and weighting for measured determinants of inclusion reduced discrimination and worsened probabilistic performance. Weighted estimates may still differ from performance in all eligible patients. Unmeasured determinants of missing output, unavailable outcomes and limited overlap remain unresolved. The unweighted cohort may therefore be unrepresentative of patients encountered in routine ICU use. Variation in documentation practices directly affects which patients can be assessed. Changes in populations, clinical practice and measurement processes can produce dataset shift, so retrospective transportability does not ensure stable performance after deployment ^[50].

At the locked threshold, more than half of external stays would have triggered review and positive predictive value was below 40%. Whether that workload is acceptable depends on the review's costs, the consequences of missed events and existing monitoring intensity. At the chosen threshold, decision-curve analysis did not establish an advantage over reviewing everyone ^[41]. More favourable estimates at higher thresholds do not retrospectively validate a new operating policy. Prospective silent evaluation should assess data availability, calibration and alert burden. An impact study would then be needed to determine whether acting on predictions improves decongestion, safety or patient-centred outcomes. DECIDE-AI provides a framework for assessing clinical performance, safety and human factors during early clinical

evaluation^[51]. FUTURE-AI further emphasizes lifecycle monitoring, robustness and governance; these requirements remain unmet by retrospective performance assessment alone^[52].

Strengths and limitations

The study separated predictor and outcome windows and used subject-grouped tuning within each candidate algorithm. Further strengths included unchanged external application, paired algorithm comparisons, cluster-aware intervals and sensitivity analyses of predictor information and measured inclusion. Several limitations affect interpretation. First, the external heart-failure definition was broader, and disease exclusions, medication records and clinical severity differed between databases. The external evaluation therefore assessed transportability across different populations, rather than validation in an identically defined acute-decompensated-heart-failure cohort.

Second, the outcome was a development-defined quartile of recorded urine output, not a prospectively validated clinical target. Minimum charting density did not guarantee complete collection; absent component volumes could be treated as zero recorded output. Weight normalization and shared measurement processes across consecutive windows may strengthen associations without representing a new biological signal. Conditioning on survival, ICU residence and later observability also narrows the population to which performance applies. Subsequent treatment, fluid administration and unmeasured congestion could influence outcomes; this design could not separate their causal effects.

Third, selection among algorithm families reused outer-fold summaries, and uncertainty intervals did not repeat the complete development process. The quartile definition and operating threshold were also development-derived. Internal estimates and confidence intervals therefore do not fully account for selection. Post-lock imputation, ablation, subgroup and recalibration analyses were sensitivity analyses, not independent confirmatory experiments. Contemporary reporting and risk-of-bias guidance emphasizes these distinctions and the need to evaluate a model's intended use rather than only technical performance^[24, 53].

Fourth, the external sample contained 147 events and substantial hospital imbalance, limiting precise calibration, subgroup and hospital-specific inference. Weighting addressed measured inclusion only. The study lacked a direct comparison with BAN-ADHF under identical endpoint definitions, and it did not test a live clinical workflow or patient outcomes. Finally, SHAP rankings describe fitted-model contributions; correlations, treatment-by-indication and documentation patterns can all influence them. Post hoc explanations should not be interpreted as causal evidence or as assurance that a model is clinically trustworthy^[54].

Conclusions

Information recorded during the first six hours after an initial qualifying loop-diuretic record improved prediction of subsequent low urine output in critically ill adults with heart failure. External evaluation supported moderate discrimination but did not support an advantage of XGBoost over a simpler early-response model. Calibration limitations, selection for observable outcomes and differences in data capture

preclude immediate use as an automated treatment rule. Prospective studies should compare simpler models, assess data completeness and calibration, and determine the clinical consequences of risk-guided review.

Abbreviations

AUROC

area under the receiver operating characteristic curve

BUN

blood urea nitrogen

CI

confidence interval

eICU-CRD

eICU Collaborative Research Database

ICU

intensive care unit

IPCW

inverse-probability-of-inclusion weighting in this study

MAP

mean arterial pressure

MIMIC-IV

Medical Information Mart for Intensive Care IV

PR-AUC

area under the precision–recall curve

SHAP

Shapley additive explanations

SOFA

Sequential Organ Failure Assessment.

Declarations

Ethics approval and consent to participate

This study was a secondary analysis of deidentified MIMIC-IV and eICU-CRD records, with no participant recruitment or intervention. The collection of patient information and creation of MIMIC-IV were reviewed by the Institutional Review Board of Beth Israel Deaconess Medical Center, which waived informed consent and approved data sharing. The investigator responsible for data access completed the required human-subject research training and obtained authorized access to both databases under the applicable PhysioNet data-use agreements. No additional local institutional ethics review was required for this secondary analysis under the authors' institutional requirements.

Consent for publication

Not applicable. No identifiable individual information is included in this manuscript.

Availability of data and materials

MIMIC-IV version 3.1 and eICU-CRD version 2.0 are available through PhysioNet to researchers who meet its credentialing, training and data-use requirements. Patient-level records are not redistributed. Code related to this study is available at <https://github.com/zxinyuan20-hue/ADHF-UO-Landmark-Framework>. Code-related enquiries can be directed to Xinyuan Zhang (zhangxyuan1@163.com).

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

Xiaohuan Zhuang: conceptualization, clinical interpretation, writing – original draft, and writing – review and editing. Tongnian Fang: conceptualization, supervision, clinical interpretation, and writing – review and editing. Xinyuan Zhang: methodology, data curation, software, formal analysis, visualization, and writing – review and editing. Xiaohuan Zhuang and Tongnian Fang contributed equally to this work. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the teams responsible for developing and maintaining MIMIC-IV, eICU-CRD and PhysioNet.

Use of generative artificial intelligence

ChatGPT/Codex (OpenAI) was used for language editing and coding assistance. All AI-assisted work was reviewed and verified by the authors, who take full responsibility for the final content.

Protocol, registration and patient involvement

The study was not registered, and no study protocol was made publicly available. Patients and members of the public were not involved in this retrospective database study.

References

1. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. 2021;42(36):3599–726. <https://doi.org/10.1093/eurheartj/ehab368>.
2. Writing Committee Members, ACC/AHA Joint Committee Members. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure. *J Card Fail*. 2022;28(5):e1–167. <https://doi.org/10.1016/j.cardfail.2022.02.010>.
3. Felker GM, Lee KL, Bull DA, Redfield MM, Stevenson LW, Goldsmith SR, et al. Diuretic strategies in patients with acute decompensated heart failure. *N Engl J Med*. 2011;364(9):797–805. <https://doi.org/10.1056/nejmoa1005419>.
4. Mullens W, Dauw J, Martens P, Verbrugge FH, Nijst P, Meekers E, et al. Acetazolamide in Acute Decompensated Heart Failure with Volume Overload. *N Engl J Med*. 2022;387(13):1185–95. <https://doi.org/10.1056/nejmoa2203094>.
5. Trullàs JC, Morales-Rull JL, Casado J, Carrera-Izquierdo M, Sánchez-Martel M, Conde-Martel A, et al. Combining loop with thiazide diuretics for decompensated heart failure: the CLOROTIC trial. *Eur Heart J*. 2023;44(5):411–21. <https://doi.org/10.1093/eurheartj/ehac689>.
6. Felker GM, Ellison DH, Mullens W, Cox ZL, Testani JM. Diuretic Therapy for Patients With Heart Failure: JACC State-of-the-Art Review. *J Am Coll Cardiol*. 2020;75(10):1178–95. <https://doi.org/10.1016/j.jacc.2019.12.059>.
7. Mullens W, Damman K, Harjola VP, Mebazaa A, Brunner-La Rocca HP, Martens P, et al. The use of diuretics in heart failure with congestion - a position statement from the Heart Failure Association of the European Society of Cardiology. *Eur J Heart Fail*. 2019;21(2):137–55. <https://doi.org/10.1002/ejhf.1369>.
8. Writing Committee, Hollenberg SM, Stevenson LW, Ahmad T, Bozkurt B, Butler J, et al. 2024 ACC Expert Consensus Decision Pathway on Clinical Assessment, Management, and Trajectory of Patients Hospitalized With Heart Failure Focused Update: A Report of the American College of Cardiology Solution Set Oversight Committee. *J Am Coll Cardiol*. 2024;84(13):1241–67. <https://doi.org/10.1016/j.jacc.2024.06.002>.
9. Testani JM, Hanberg JS, Cheng S, Rao V, Onyebeke C, Laur O, et al. Rapid and Highly Accurate Prediction of Poor Loop Diuretic Natriuretic Response in Patients With Heart Failure. *Circ Heart Fail*. 2016;9(1):e002370. <https://doi.org/10.1161/circheartfailure.115.002370>.
10. Ter Maaten JM, Beldhuis IE, van der Meer P, Krikken JA, Postmus D, Coster JE, et al. Natriuresis-guided diuretic therapy in acute heart failure: a pragmatic randomized trial. *Nat Med*. 2023;29(10):2625–32. <https://doi.org/10.1038/s41591-023-02532-z>.

11. Dauw J, Charaya K, Lelonek M, Zegri-Reiriz I, Nasr S, Paredes-Paucar CP, et al. Protocolized Natriuresis-Guided Decongestion Improves Diuretic Response: The Multicenter ENACT-HF Study. *Circ Heart Fail*. 2024;17(1):e011105. <https://doi.org/10.1161/circheartfailure.123.011105>.
12. Meekers E, Martens P, Dauw J, Gruwez H, Dhont S, Nijst P, et al. Nurse-led diuretic titration via a point-of-care urinary sodium sensor in patients with acute decompensated heart failure (EASY-HF): A single-centre, randomized, open-label study. *Eur J Heart Fail*. 2024;26(10):2129–39. <https://doi.org/10.1002/ejhf.3429>.
13. Meekers E, Dauw J, Ter Maaten JM, Martens P, Nijst P, Verbrugge FH, et al. Urinary sodium analysis: The key to effective diuretic titration? European Journal of Heart Failure expert consensus document. *Eur J Heart Fail*. 2025;27(6):940–9. <https://doi.org/10.1002/ejhf.3632>.
14. Ter Maaten JM, Rao VS, Hanberg JS, Perry Wilson F, Bellumkonda L, Assefa M, et al. Renal tubular resistance is the primary driver for loop diuretic resistance in acute heart failure. *Eur J Heart Fail*. 2017;19(8):1014–22. <https://doi.org/10.1002/ejhf.757>.
15. Rubio-Gracia J, Demissei BG, Ter Maaten JM, Cleland JG, O'Connor CM, Metra M, et al. Prevalence, predictors and clinical outcome of residual congestion in acute decompensated heart failure. *Int J Cardiol*. 2018;258:185–91. <https://doi.org/10.1016/j.ijcard.2018.01.067>.
16. Segar MW, Khan MS, Patel KV, Butler J, Ravichandran AK, Walsh MN, et al. A Phenomapping Tool and Clinical Score to Identify Low Diuretic Efficiency in Acute Decompensated Heart Failure. *JACC Heart Fail*. 2024;12(3):508–20. <https://doi.org/10.1016/j.jchf.2023.09.029>.
17. Otabor E, Lam J, Alomari L, Afolabi-Brown O, Lo KB. Predicting diuretic efficiency in critically ill patients with acute decompensated heart failure: External validation of the BAN-ADHF score. *Am Heart J Plus*. 2026;69:100841. <https://doi.org/10.1016/j.ahjo.2026.100841>.
18. Alfonso DM, Docherty KF, Lee MMY, Mark PB, McConnachie A, Osmanska J, et al. External validation of the BAN-ADHF diuretic-resistance score in the DAPA-RESIST clinical trial. *Br J Clin Pharmacol*. 2026;92(8):2960–5. <https://doi.org/10.1002/bcp.70619>.
19. Zonneveld LEEC, Pandey A, Beldhuis IE, van der Meer P, Coster JE, Nieuwland W, et al. Diuretic Resistance Risk and the Efficacy of Natriuresis-Guided Diuretic Therapy in Acute Heart Failure: Post Hoc Analysis From the PUSH-AHF Trial. *JACC Heart Fail*. 2026;14(7):103017. <https://doi.org/10.1016/j.jchf.2026.103017>.
20. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. <https://doi.org/10.1038/s41597-022-01899-x>.
21. 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. <https://doi.org/10.1038/sdata.2018.178>.
22. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet. 2024. <https://doi.org/10.13026/kpb9-mt58>
23. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. <https://doi.org/10.13026/C2WM1R>

24. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):e296. <https://doi.org/10.1371/journal.pmed.0040296>.
26. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*. 2015;12(10):e1001885. <https://doi.org/10.1371/journal.pmed.1001885>.
27. Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ*. 2023;380:e071018. <https://doi.org/10.1136/bmj-2022-071018>.
28. Pollard T, Moody BE, Lehman LW, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. *Nat Health*. 2026;1(8):792–5. <https://doi.org/10.1038/s44360-026-00096-z>.
29. Nicolaie MA, van Houwelingen JC, de Witte TM, Putter H. Dynamic prediction by landmarking in competing risks. *Stat Med*. 2013;32(12):2031–47. <https://doi.org/10.1002/sim.5665>.
30. Testani JM, Brisco MA, Turner JM, Spatz ES, Bellumkonda L, Parikh CR, et al. Loop diuretic efficiency: a metric of diuretic responsiveness with prognostic importance in acute decompensated heart failure. *Circ Heart Fail*. 2014;7(2):261–70. <https://doi.org/10.1161/circheartfailure.113.000895>.
31. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. 2020;125:183–7. <https://doi.org/10.1016/j.jclinepi.2020.03.028>.
32. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
33. Breiman L, Random, Forests. *Mach Learn*. 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.
34. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*. 2017;30. <https://proceedings.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
35. Chen T, Guestrin C, XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. <https://doi.org/10.1145/2939672.2939785>
36. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>.
37. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e078276. <https://doi.org/10.1136/bmj-2023-078276>.
38. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820. <https://doi.org/10.1136/bmj-2023-074820>.

39. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
40. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. <https://doi.org/10.1186/s12916-019-1466-7>.
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–74. <https://doi.org/10.1177/0272989x06295361>.
42. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*. 2019;3:18. <https://doi.org/10.1186/s41512-019-0064-7>.
43. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. <https://doi.org/10.1136/bmj.b2393>.
44. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–64. <https://doi.org/10.1093/aje/kwn164>.
45. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. 2021;40(19):4230–51. <https://doi.org/10.1002/sim.9025>.
46. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ*. 2024;384:e074821. <https://doi.org/10.1136/bmj-2023-074821>.
47. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell*. 2020;2(1):56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
48. Damman K, Beldhuis IE, van der Meer P, Krikken JA, Coster JE, Nieuwland W, et al. Renal function and natriuresis-guided diuretic therapy - a pre-specified analysis from the PUSH-AHF trial. *Eur J Heart Fail*. 2024;26(6):1347–57. <https://doi.org/10.1002/ejhf.3228>.
49. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>.
50. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*. 2021;385(3):283–6. <https://doi.org/10.1056/nejmc2104626>.
51. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924–33. <https://doi.org/10.1038/s41591-022-01772-9>.
52. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. <https://doi.org/10.1136/bmj-2024-081554>.

53. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. <https://doi.org/10.1136/bmj-2024-082505>.
54. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745–50. [https://doi.org/10.1016/s2589-7500\(21\)00208-9](https://doi.org/10.1016/s2589-7500(21)00208-9).