Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Haemodynamic intolerance at the de-resuscitation decision in critically ill adults: development, temporal and external validation of a prediction model across two intensive care databases

**Lujun Shao**
The Affiliated Taizhou People's Hospital of Nanjing Medical University
**Yue Li**
The Affiliated Taizhou People's Hospital of Nanjing Medical University
**Cheng Xu**
The Affiliated Taizhou People's Hospital of Nanjing Medical University
**Huiyu Tai**
The Affiliated Taizhou People's Hospital of Nanjing Medical University
**Li Wang**
thdzydm@163.com
The Affiliated Taizhou People's Hospital of Nanjing Medical University
**Jinyu Shi**
The Affiliated Taizhou People's Hospital of Nanjing Medical University

**Research Article**

**Keywords:** Clinical prediction model, De-resuscitation, Fluid overload, Haemodynamic instability, Furosemide, External validation, Calibration, Intensive care unit, MIMIC-IV, eICU-CRD

**Posted Date:** July 21st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10050782/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.** Active fluid removal (de-resuscitation) is frequently required after the resuscitation of critically ill patients, but it can precipitate haemodynamic instability, and no validated tool estimates this risk at the moment fluid removal is initiated. We developed and externally validated a model to predict haemodynamic intolerance within 24 hours of the first intravenous furosemide dose in haemodynamically stable intensive care patients, and examined whether its discrimination and calibration transport across time and across populations.

**Methods.** In a retrospective cohort study reported in accordance with the TRIPOD + AI statement, the decision point (t0) was the first intravenous furosemide administration in adults with no vasopressor in the preceding 6 hours. The primary outcome was haemodynamic intolerance within 24 hours, defined as new vasopressor initiation or a mean arterial pressure below 60 mmHg recorded on at least three occasions. Twenty routinely collected predictors available in both databases were used. Penalised (ridge) logistic regression and gradient-boosted trees were developed in MIMIC-IV and evaluated by internal cross-validation, internal temporal validation, and external geographic validation in the eICU-CRD without refitting. Discrimination, calibration, and decision-curve net benefit were assessed, and miscalibration was addressed by out-of-sample logistic recalibration.

**Results.** The development cohort comprised 14,294 admissions (27.9% with intolerance) and the external cohort 21,677 (31.4%). Discrimination was moderate and stable across all settings (gradient-boosting area under the receiver-operating-characteristic curve 0.764 [95% CI 0.755–0.773] for internal cross-validation, 0.733 [0.710–0.754] for temporal validation, and 0.767 [0.760–0.774] for external validation). In contrast, calibration shifted predictably with event prevalence — over-prediction when prevalence fell across the temporal split and under-prediction when it rose in the external cohort — while the calibration slope remained close to, or slightly below, one; a single logistic recalibration largely corrected the shift in both directions. Decision-curve analysis showed net benefit across a clinically plausible range of thresholds. The gradient-boosting and ridge models performed similarly.

**Conclusions.** A model predicting haemodynamic intolerance at the de-resuscitation decision showed portable discrimination but prevalence-dependent calibration that was largely correctable by simple recalibration. The model is positioned as decision support and should be recalibrated locally before use; prospective impact evaluation is the essential next step.


# Background

Intravenous fluids are a cornerstone of early resuscitation in critically ill patients, yet the volume administered during haemodynamic stabilisation is frequently not mobilised as the patient improves, so that a positive cumulative fluid balance accumulates [1]. Across diverse critically ill populations, fluid accumulation and overload have been consistently associated with prolonged organ support and increased mortality [1, 2]. Contemporary frameworks therefore describe intravenous fluid therapy as a sequence of phases — resuscitation, optimisation, stabilisation, and evacuation — in which the early

benefit of volume expansion gives way to the potential harm of accumulated fluid, and in which active fluid removal, or de-resuscitation, becomes a therapeutic goal [3, 4].

De-resuscitation denotes the deliberate removal of accumulated fluid once shock has resolved, most commonly achieved with loop diuretics such as furosemide [4, 5]. The rationale for limiting fluid accumulation is supported by randomised evidence: in acute lung injury, a conservative fluid-management strategy increased ventilator-free days without an excess of shock or mortality [6], and a systematic review and meta-analysis found that conservative or de-resuscitative fluid strategies were associated with more ventilator-free days and a tendency toward improved outcomes [7]. Subsequent trials in septic shock have compared restrictive with liberal fluid strategies; these did not demonstrate a mortality benefit but supported the safety of restricting fluids after initial resuscitation [8, 9].

Fluid removal is not without hazard, however. Diuresis lowers intravascular volume and can precipitate hypotension and a renewed requirement for vasopressors, particularly in patients who remain preload-dependent; goal-directed fluid removal with furosemide has accordingly been investigated in a dedicated randomised trial [5]. The decision to begin de-resuscitation in an apparently stable patient thus carries a real but unquantified risk of haemodynamic deterioration, and no validated tool currently estimates, at the moment fluid removal is initiated, which patients will fail to tolerate it.

Clinical prediction models derived from routinely collected intensive care data have proliferated, yet most newly developed models are never externally validated, and those that are tend to exhibit poorer discrimination than in their development setting [10]. Calibration — the agreement between predicted and observed risk — and transportability across populations are frequently neglected, even though they determine whether predicted probabilities can be trusted to guide decisions; current reporting guidance accordingly emphasises external validation, calibration, and transparent, reproducible methods [11].

In this study we developed and externally validated a model to predict haemodynamic intolerance within 24 hours of the de-resuscitation decision — operationalised as the first intravenous furosemide administration — in haemodynamically stable adult intensive care patients. The model was developed in the Medical Information Mart for Intensive Care IV (MIMIC-IV) [12] and externally validated in the eICU Collaborative Research Database (eICU-CRD) [13], two large, publicly available critical care databases that were collected in different eras and in geographically distinct intensive care units. Rather than pursuing maximal predictive performance, we addressed a deployment-relevant question: whether a model built on harmonised, routinely available predictors retains its discrimination and calibration when transported across time and across populations, and whether any miscalibration that emerges can be corrected by simple model updating.

## Methods

# Study design and data sources

We conducted a retrospective cohort study to develop and externally validate a clinical prediction model, reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis for artificial intelligence (TRIPOD + AI) statement [11]; the completed reporting checklist is provided in Additional file 1: Table S9. The model was developed in the Medical Information Mart for Intensive Care IV (MIMIC-IV) version 3.1, a single-centre database of intensive care unit (ICU) admissions at the Beth Israel Deaconess Medical Center (Boston, USA) [12], and externally validated in the eICU Collaborative Research Database (eICU-CRD) version 2.0, a multi-centre database drawn from intensive care units across the continental United States [13]. Both databases are de-identified and publicly accessible to credentialed investigators through PhysioNet. Because all data are de-identified, the requirement for individual informed consent was waived by the institutional review boards that approved the source databases; the present secondary analysis carried no additional patient contact.

## Participants and decision point

The unit of analysis was the ICU admission. We studied the clinical decision point of the first intravenous furosemide administration during the ICU stay, taken as an operational marker of the de-resuscitation decision (hereafter t0). To restrict the cohort to relatively haemodynamically stable patients at this decision point — the population in whom the question of haemodynamic tolerance is clinically meaningful — we applied the following criteria. We included adults who received a first intravenous furosemide dose and who had a haemodynamically stable baseline, defined as the absence of any vasopressor infusion during the 6 hours preceding t0. We required follow-up to extend to the 24-hour horizon — achieved either by at least 24 hours of subsequent in-unit observation or, for the few patients who died within the window, by death (treated as a competing event). We analysed the first qualifying furosemide administration per ICU stay. All candidate predictors were derived exclusively from data recorded before t0, precluding outcome leakage. Admissions in which the outcome-defining criterion was met before t0 — a small number attributable to charting-time granularity — were excluded so that the outcome was assessed strictly after the decision point. Further detail on cohort construction and database-specific considerations is provided in Additional file 1: Supplementary Methods.

Application of these criteria yielded a development cohort of 14,294 admissions from MIMIC-IV and an external validation cohort of 21,677 admissions from eICU-CRD. The flow of admissions through eligibility is shown in Fig. 1.

## Outcomes

The primary outcome was haemodynamic intolerance within 24 hours of t0, defined as the occurrence of either of the following: (i) initiation of a new vasopressor infusion, which is necessarily new given the stable-baseline criterion; or (ii) hypotension, defined as a mean arterial pressure below 60 mmHg recorded on at least three occasions within this 24-hour window (a threshold chosen to require repeated rather than isolated low readings). Death within 24 hours without prior intolerance was treated as a

competing event; a patient in whom intolerance occurred before death was counted as having experienced the outcome. Because follow-up extended to the 24-hour horizon or to an earlier competing death for every included patient, the 24-hour outcome was completely ascertained, with competing death the only alternative end point.

A secondary, descriptively reported outcome was acute kidney injury within 48 hours of the decision point, defined by the Kidney Disease: Improving Global Outcomes (KDIGO) serum-creatinine criterion (an absolute rise of at least 0.3 mg/dL, or a rise to at least 1.5 times the pre-decision baseline) [14]. This outcome was reported to characterise the adverse-event landscape of the cohorts and was not the subject of a separate prediction model.

## Predictors

Candidate predictors were routinely collected variables available at the decision point and present in both databases, to permit an external validation in which exactly the same predictors are used during development and validation. The final predictor set comprised 20 variables: age; sex; the first furosemide dose; the mean heart rate and mean peripheral oxygen saturation over the 6 hours before t0, together with the minimum, the mean, and the t0 value of mean arterial pressure over the same window; the most recent pre-decision serum creatinine, blood urea nitrogen, lactate, sodium, potassium, haemoglobin, platelet count, white cell count, and bicarbonate; and invasive mechanical ventilation status. Because serum albumin was unmeasured in the majority of admissions, its value was not used as a predictor; instead, a binary indicator of whether albumin had been measured was retained, and a corresponding measurement indicator was retained for lactate, to capture potentially informative measurement patterns. Variables that could not be harmonised across both databases — insurance status, primary language, the Charlson comorbidity index, cumulative fluid-balance and urine-output derivations, and categorical variables (race and admission type) whose category coding differed between the databases — were not included in the model used for external validation; their exploratory use within the development database is reported in Additional file 1: Table S7. A full data dictionary, including the source tables, time windows, and definitions for each predictor, is provided in Additional file 1: Table S1.

## Sample size

The analysis used all eligible admissions in each database rather than a pre-specified target sample. The number of outcome events available for development (3,982 events against 20 candidate predictors) comfortably exceeds the minimum sample size indicated by contemporary criteria for clinical prediction models [15] — equivalent to approximately 199 events per candidate predictor — supporting stable estimation without further variable reduction.

# Missing data

Missing predictor values were handled according to a pre-specified, deployment-realistic strategy. For the logistic regression model, continuous predictors were imputed to the median. Critically, the imputation values were estimated in the development data and then applied unchanged to the external validation data, so that no information from the external cohort entered preprocessing. The gradient-boosting model handled missing values natively through its split-finding mechanism and required no imputation. Informative-missingness indicators were included as described above. Missingness was more frequent in eICU-CRD than in MIMIC-IV for most laboratory variables — serum albumin being a notable exception, as it was measured more often in eICU-CRD — reflecting differences in measurement frequency between the databases; this difference was accommodated by the frozen imputation and the native missing-value mechanism rather than by re-estimation in the external cohort. The frequency of missingness for each predictor in each database is reported in Additional file 1: Table S2.

# Statistical analysis

*Model development.* Two models were fitted on the harmonised predictor set: penalised (ridge) logistic regression, as an interpretable model suited to transparent deployment, and gradient-boosted decision trees (the extreme gradient boosting, XGBoost, implementation), as a flexible non-linear comparator. Continuous predictors were standardised for the logistic regression model using means and standard deviations estimated in the development data and applied unchanged to the external data. The fitted ridge model, including its intercept and the per-predictor standardisation constants, is reported in full in Additional file 1: Table S10, allowing predicted risks to be computed for new patients. Model hyperparameters for the gradient-boosting model were pre-specified rather than tuned against any evaluation data (learning rate 0.05, maximum tree depth 4, 400 boosting rounds, L2 regularisation, and row and column subsampling fractions of 0.8). The ridge logistic regression used L2 regularisation with a fixed inverse-regularisation strength (C = 1.0) and was likewise not tuned against evaluation data; no data-driven predictor selection was performed for either model.

*Validation.* Three complementary evaluations were performed, spanning internal, temporal, and geographic validation. First, internal cross-validation: stratified five-fold cross-validation within the development cohort, with all preprocessing (imputation and standardisation) and model fitting performed inside each training fold to avoid optimism. Second, internal temporal validation: models were trained on admissions through 2016 (n = 11,026; outcome prevalence 29.9%) — with imputation and standardisation estimated on the training era and applied unchanged to the test era — and evaluated on admissions from 2017 onward (n = 3,268; outcome prevalence 21.0%), using the admission-era groupings recorded in MIMIC-IV to emulate prospective deployment under temporal drift (calendar years in MIMIC-IV are de-identified to multi-year anchor groups, so the split reflects de-identified admission era rather than exact calendar year). Third, geographic external validation: the final models were trained on the entire development cohort and applied to the external cohort without any refitting or re-estimation of preprocessing parameters.

*Performance measures.* Discrimination was summarised by the area under the receiver-operating-characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), the latter

interpreted against the outcome prevalence in each cohort. Calibration was assessed by the calibration slope and calibration intercept (calibration-in-the-large), by the integrated calibration index (ICI; the mean absolute difference between predicted and flexibly estimated observed risk, with the flexible curve fitted by locally weighted regression without robustifying iterations, which would otherwise distort a binary outcome) [16], by the Brier score, and by flexible calibration plots. Clinical utility was assessed by decision-curve analysis [17] applied to the original (un-recalibrated) predictions, reporting the net benefit of each model relative to the treat-all and treat-none strategies across a clinically plausible range of threshold probabilities.

*Model updating.* Where calibration was imperfect, the model was updated by logistic recalibration. To avoid optimism, recalibration was both fitted and evaluated out-of-sample: the relevant evaluation cohort was randomly split into two halves, the recalibration parameters were estimated on one half, and calibration before and after recalibration was assessed on the held-out half.

*Sensitivity analyses.* Two pre-specified sensitivity analyses were performed: a parsimonious five-variable model, to gauge how much discrimination derived from a few routine variables (Additional file 1: Table S6), and a re-estimation of internal cross-validation under patient-grouped folds, to assess robustness to the minority of patients contributing more than one admission (Additional file 1: Table S8).

*Competing risk.* Because follow-up was complete to the 24-hour horizon and competing mortality within 24 hours was negligible (17 of 14,294 development admissions, 0.1%; 1 of 21,677 external admissions, < 0.1%), a competing-risk (subdistribution) analysis yields results practically indistinguishable from the binary analysis. We therefore report the binary model as the primary analysis and treat the equivalence as a robustness consideration rather than fitting a separate subdistribution model; the basis for this equivalence is set out in Additional file 1: Supplementary Methods.

*Uncertainty and software.* Ninety-five percent confidence intervals for performance measures were obtained by bootstrap resampling of the relevant evaluation cohort (1,000 replications). Analyses were performed in Python using the scikit-learn [18], XGBoost [19], and statsmodels [20] libraries. The cohort-extraction queries and analysis code that reproduce all results, together with the random seeds used, are available (see Declarations); detailed environment and version information is provided in Additional file 1: Table S3.

## Results

# Participants

The development cohort comprised 14,294 ICU admissions from MIMIC-IV and the external validation cohort comprised 21,677 admissions from eICU-CRD. Baseline characteristics at the decision point are summarised in Table 1. The two cohorts were broadly similar in age (median 70 versus 68 years), sex distribution, and admission physiology, but differed in several respects relevant to model transfer. Haemodynamic intolerance within 24 hours occurred in 3,982 admissions (27.9%) in the development

cohort and in 6,809 admissions (31.4%) in the external cohort. Acute kidney injury within 48 hours occurred in 25.4% and 16.4%, respectively, and competing death within 24 hours was rare in both (17 admissions, 0.1%, and 1 admission, < 0.1%). The composite outcome was driven predominantly by new vasopressor initiation in both cohorts rather than by the blood-pressure criterion: a new vasopressor was involved in 81% of development events and 82% of external events, with the hypotension criterion alone accounting for the remaining 19% and 18%, respectively. Because vasopressor initiation is a discrete, charted clinical action rather than a monitoring-frequency-dependent threshold — and because the hypotension-defined fraction was, if anything, marginally smaller in the more densely sampled external database — the difference in blood-pressure sampling density between the databases is unlikely to have materially inflated the external event rate, which reflects predominantly more frequent vasopressor initiation. Laboratory measurement frequency differed markedly between databases — for example, lactate was recorded in 61.6% of development admissions but only 22.8% of external admissions — consistent with differing measurement practices and motivating the measurement-indicator and frozen-imputation strategy described in the Methods.

Table 1
Baseline characteristics at the decision point in the development and external validation cohorts.

| Characteristic | Development: MIMIC-IV (n = 14,294) | External: eICU-CRD (n = 21,677) |
|---|---|---|
| Age, years | 70 (59–80) | 68 (58–78) |
| Female sex | 6,444 (45.1) | 10,105 (46.6) |
| Invasive mechanical ventilation | 4,066 (28.4) | 7,259 (33.5) |
| Heart rate, beats/min | 87 (75–100) | 87 (75–100) |
| MAP at $t_0$, mmHg | 80 (71–91) | 83 (73–95) |
| Minimum MAP, mmHg | 70 (62–79) | 70 (61–81) |
| $SpO_2$, % | 96 (94–98) | 97 (95–98) |
| Serum creatinine, mg/dL | 1.0 (0.7–1.6) | 1.1 (0.8–1.6) |
| Blood urea nitrogen, mg/dL | 23 (15–38) | 23 (15–37) |
| Sodium, mmol/L | 139 (136–142) | 139 (136–142) |
| Potassium, mmol/L | 4.2 (3.8–4.6) | 4.1 (3.7–4.5) |
| Bicarbonate, mmol/L | 24 (21–27) | 25 (22–28) |
| Haemoglobin, g/dL | 9.7 (8.5–11.2) | 10.0 (8.6–11.7) |
| Platelet count, $\times 10^9$/L | 183 (126–256) | 183 (129–249) |
| White cell count, $\times 10^9$/L | 10.6 (7.8–14.1) | 10.9 (8.1–14.7) |
| Lactate, mmol/L | 1.5 (1.1–2.0) | 1.5 (1.1–2.3) |
| Lactate measured | 8,812 (61.6) | 4,936 (22.8) |
| Albumin measured | 3,844 (26.9) | 9,963 (46.0) |
| First furosemide dose, mg | 20 (20–40) | 40 (20–40) |
| Haemodynamic intolerance within 24 h | 3,982 (27.9) | 6,809 (31.4) |
| Acute kidney injury within 48 h | 3,625 (25.4) | 3,562 (16.4) |
| Death within 24 h | 17 (0.1) | 1 (< 0.1) |

*Values are median (interquartile range) for continuous variables and n (%) for categorical variables. t0 denotes the decision point (first intravenous furosemide administration). MAP, mean arterial pressure; SpO2, peripheral oxygen saturation; AKI, acute kidney injury (KDIGO serum-creatinine criterion). "Lactate measured" and "albumin measured" denote the proportion of admissions with a recorded pre-decision*

*value; serum albumin value was not used as a predictor owing to high missingness. Haemodynamic intolerance is the primary outcome; acute kidney injury within 48 hours is a secondary, descriptively reported outcome.*

# Model discrimination

Discrimination was consistent across all three validation settings and both models (Fig. 2; Table 2). For the gradient-boosting model, the area under the receiver-operating-characteristic curve (AUROC) was 0.764 (95% CI 0.755–0.773) under internal cross-validation, 0.733 (0.710–0.754) under internal temporal validation, and 0.767 (0.760–0.774) under external geographic validation. The penalised logistic regression model performed slightly lower in every setting (0.752, 0.721, and 0.764, respectively). The area under the precision–recall curve for the gradient-boosting model was 0.593, 0.478, and 0.634 across the three settings, in each case well above the corresponding outcome prevalence, indicating useful positive-predictive performance. Discrimination was therefore portable across both a within-centre temporal split (training through 2016, testing from 2017 onward) and a transfer to a different, multi-centre database. External discrimination was marginally higher than internal cross-validation (for example, 0.767 versus 0.764 for the gradient-boosting model); because AUROC is insensitive to outcome prevalence, this reflects a difference in case-mix rather than the higher external event rate. The correspondingly higher external AUPRC (0.634 versus 0.593) is driven largely by the higher external prevalence and should not be read as superior performance. The gradient-boosting model was marginally superior to the ridge model on both discrimination (Table 2) and calibration (for example, external integrated calibration index 0.036 versus 0.066), so non-linear structure contributed a small but non-zero gain. The choice between the two models is therefore a trade-off: the gradient-boosting model offers slightly better performance, whereas the ridge model offers fully interpretable coefficients (Additional file 1: Table S4) and a transparent functional form that may be preferred where auditability is paramount. The two models agreed closely on which predictors most influenced risk — the pre-decision mean arterial pressure summaries dominated both, followed by age, heart rate, lactate, and blood urea nitrogen, a pattern consistent with clinical expectation (Additional file 1: Table S4). Much of the discrimination was retained by a parsimonious five-variable model (Additional file 1: Table S6), and internal cross-validation discrimination was essentially unchanged when folds were grouped by patient to account for re-admissions (Additional file 1: Table S8). Both models exhibited the same prevalence-dependent calibration behaviour and both were corrected by a single recalibration (Additional file 1: Table S5).

Table 2
Discrimination, calibration, and overall accuracy of the prediction models under internal cross-validation, internal temporal validation, and external geographic validation.

| Validation setting | Model | AUROC (95% CI) | AUPRC (95% CI) | Brier | Calibration slope / intercept (95% CI) | ICI |
|---|---|---|---|---|---|---|
| Internal cross-validation | Ridge logistic regression | 0.752 (0.743–0.762) | 0.564 (0.547–0.581) | 0.166 | 0.98 (0.93–1.04) / −0.01 (− 0.07 to 0.04) | 0.030 |
| | XGBoost | 0.764 (0.755–0.773) | 0.593 (0.578–0.610) | 0.161 | 0.87 (0.83–0.90) / −0.10 (− 0.15 to − 0.05) | 0.017 |
| Internal temporal validation | Ridge logistic regression | 0.721 (0.697–0.742) | 0.443 (0.402–0.484) | 0.151 | 0.88 (0.77–1.00) / −0.46 (− 0.57 to − 0.33) | 0.063 |
| | XGBoost | 0.733 (0.710–0.754) | 0.478 (0.437–0.517) | 0.145 | 0.82 (0.73–0.91) / −0.47 (− 0.60 to − 0.36) | 0.047 |
| External (geographic) validation | Ridge logistic regression | 0.764 (0.758–0.772) | 0.626 (0.613–0.638) | 0.177 | 1.00 (0.96–1.04) / +0.39 (0.34 to 0.44) | 0.066 |
| | XGBoost | 0.767 (0.760–0.774) | 0.634 (0.621–0.646) | 0.172 | 0.93 (0.90–0.97) / +0.16 (0.12 to 0.21) | 0.036 |

*Ninety-five percent confidence intervals were obtained by bootstrap resampling (1,000 replications). AUROC, area under the receiver-operating-characteristic curve; AUPRC, area under the precision–recall curve; ICI, integrated calibration index. The calibration slope and intercept describe the original (un-recalibrated) model in each setting; an intercept below zero indicates over-prediction and above zero indicates under-prediction. Internal cross-validation reflects performance within the development cohort; temporal validation trains on admissions through 2016 and tests on 2017 onward; external validation applies the final models to eICU-CRD without refitting. The Brier score, like the AUPRC, is influenced by outcome prevalence and is not directly comparable across cohorts.*

# Calibration and the effect of changing event prevalence

In contrast to discrimination, calibration depended on the direction in which event prevalence shifted between development and validation, and did so predictably. Under temporal validation, where outcome prevalence fell from 29.9% in the training era to 21.0% in the test era, the gradient-boosting model over-predicted risk: the calibration intercept was − 0.47 and the calibration slope 0.82. Under external validation, where prevalence rose from 27.9% to 31.4%, the same model under-predicted risk, with a positive calibration intercept (+ 0.16) and a slope of 0.93. The ridge model under-predicted more strongly in the external cohort (intercept + 0.39); because a change in prevalence alone would predict an intercept

of approximately + 0.17, this larger value reflects additional sensitivity of the linear model's mean prediction to between-database differences in predictor distribution. The magnitude of miscalibration was modest in both settings (integrated calibration index 0.047 and 0.036, respectively), and the slope remained close to, or slightly below, one throughout (0.82 to 0.93 for the gradient-boosting model), indicating that the miscalibration was dominated by a shift in calibration-in-the-large that tracked the change in prevalence rather than by a distortion of the risk gradient. A single logistic recalibration substantially reduced this shift in both directions. Estimated on one randomly selected half of each validation cohort and evaluated on the held-out half, recalibration moved the calibration intercept from − 0.40 to + 0.21 (slope 0.86 to 1.11) under temporal validation and from + 0.20 to + 0.06 (slope 0.94 to 1.01) under external validation, markedly reducing the calibration-in-the-large (Fig. 3). Under temporal validation the recalibrated intercept slightly overshot zero (+ 0.21), reflecting the smaller held-out sample available for this comparison, so recalibration in that setting should be read as a proof of principle whose stability is not fully characterised. These half-sample values differ slightly from the full-cohort figures in Table 2 owing to sampling; full detail is provided in Additional file 1: Supplementary Methods.

# Clinical utility

Decision-curve analysis, applied to the original (un-recalibrated) predictions, indicated that using the model to identify patients at high risk of impending intolerance — for example, to prompt closer haemodynamic monitoring, a more cautious diuretic rate, or earlier escalation planning — provided greater net benefit than the default strategies of treating all or no patients across a clinically plausible range of threshold probabilities (approximately 0.10 to 0.40 under temporal validation and 0.10 to 0.50 under external validation; Fig. 3). Because the original model was miscalibrated in the large under both settings, net benefit at a specific threshold should be interpreted together with the calibration results, and local recalibration is advisable before such thresholds guide care.

# Competing risk

Because follow-up was complete to the 24-hour horizon and death within 24 hours occurred in only 17 development admissions (0.1%) and 1 external admission (< 0.1%), the competing-risk subdistribution analysis reduced to the binary analysis; the two were practically indistinguishable, and the binary model is reported throughout. The basis for this equivalence is set out in Additional file 1: Supplementary Methods.

# Secondary outcome

Acute kidney injury within 48 hours of the decision point, reported descriptively, occurred in 3,625 development admissions (25.4%) and 3,562 external admissions (16.4%). The lower external incidence is consistent with the lower frequency of serum-creatinine measurement in the external database and should be interpreted in that light; further detail is provided in Additional file 1: Supplementary Methods.

## Discussion

In a large, multi-database study, we developed and externally validated a model to predict haemodynamic intolerance within 24 hours of the first intravenous furosemide administration in haemodynamically stable intensive care patients. Discrimination was moderate and, importantly, stable across internal cross-validation, a within-centre temporal split, and geographic transfer to a different multi-centre database, with the area under the receiver-operating-characteristic curve lying between 0.73 and 0.77 in every setting. A flexible gradient-boosting model offered only a marginal advantage over penalised logistic regression in both discrimination and calibration, consistent with the broader observation that machine-learning algorithms rarely outperform well-specified regression for structured clinical data [21]. In contrast to discrimination, calibration depended on the direction in which event prevalence shifted between cohorts; the resulting miscalibration was nonetheless modest and was substantially reduced by a single logistic recalibration.

The dissociation between portable discrimination and prevalence-sensitive calibration is expected on theoretical grounds and has practical consequences. Because the area under the receiver-operating-characteristic curve is invariant to outcome prevalence, a model that ranks patients well continues to do so when transported to a population with a different event rate; measures that depend on the absolute risk level, including the area under the precision–recall curve and the Brier score, do not, and should not be compared directly across cohorts of differing prevalence [22]. Calibration — the agreement between predicted and observed risk — is the property that governs threshold-based decisions, yet it is the most frequently neglected aspect of prediction-model evaluation [23]. We summarised it with the calibration slope, the calibration intercept, and the integrated calibration index [16]; the slope remained close to, or slightly below, one in every setting, whereas the intercept tracked prevalence, falling (over-prediction) when the event rate declined across the temporal split and rising (under-prediction) when the event rate was higher in the external database.

This pattern is the signature of a shift in calibration-in-the-large rather than a distortion of the risk gradient, and it mirrors reports of calibration drift in intensive care models, in which predictions progressively over- or under-estimate risk as the underlying event rate changes over time [24]. A shift confined to the intercept is also the most readily corrected form of miscalibration: a single logistic recalibration, estimated on a sample from the target setting and evaluated out-of-sample, realigned predicted with observed risk in both directions, consistent with evidence that simple updating restores the performance of models transported to new patients [25]. The deployment implication is concrete: the model’s ranking can be transported as is, but its absolute probabilities should be recalibrated to local outcome frequencies before they are used to trigger action, and such recalibration requires only locally observed outcomes rather than redevelopment of the model.

Previous intensive care prediction research has largely addressed either the early-warning problem — continuously forecasting deterioration such as circulatory failure from streaming monitoring data, where reported discrimination is high but the prediction target, the time horizon, and the very low event prevalence differ fundamentally from ours [26] — or the prediction of acute kidney injury and mortality. To our knowledge, this is the first model to predict haemodynamic intolerance at the de-resuscitation

decision itself — the first intravenous furosemide dose in a haemodynamically stable patient — from routinely recorded data available at that moment, and to examine its transportability across time and across databases. Earlier work has addressed related but distinct questions, such as predicting diuretic responsiveness or relying on dedicated physiological challenge tests, rather than the bedside tolerance decision modelled here. Our more modest discrimination (area under the curve approximately 0.75) is therefore not directly comparable with continuous early-warning systems and reflects a harder, single-time-point judgement made from pre-decision data alone; it is nonetheless sufficient to separate risk groups and, by decision-curve analysis, to add net benefit over the default strategies across a plausible range of thresholds.

Several features of the prediction target warrant careful interpretation. First intravenous furosemide is an operational marker of the decision to begin fluid removal, not a direct measure of de-resuscitation intent: furosemide is administered for several overlapping indications — decongestion, management of oliguria, control of hypertension, and even as a physiological probe of renal tubular reserve [27] — so the cohort is necessarily heterogeneous in the clinical reasoning that prompted the diuretic. The model therefore predicts haemodynamic intolerance following a charted diuretic decision rather than the causal effect of de-resuscitation, and it should be read as a prognostic tool conditioned on that decision, not as evidence that furosemide causes the subsequent instability. The outcome is likewise shaped in part by local practice: the threshold at which a clinician initiates a vasopressor — the dominant component of the composite outcome — varies between providers and institutions, so the model partly reflects the prevailing response to early haemodynamic compromise rather than a purely physiological endpoint. These considerations bound the causal interpretation that the data can support and motivate the prognostic, decision-support framing we adopt.

This study has several methodological strengths. The model was developed in one large single-centre database and validated both across a within-centre temporal split and in a second, geographically distinct multi-centre database; because geographic and temporal validation probe partly different threats to generalisability, addressing both provides stronger evidence than either alone [28], and such external validation remains uncommon, since most published models are never tested outside their development data [10]. The development sample contained approximately 200 events per candidate predictor, comfortably exceeding contemporary minimum sample-size requirements for model development [15], while the validation databases were themselves large, in keeping with recommended minimums for precise estimation of external performance [29]. Both algorithms were pre-specified and left untuned, with no data-driven predictor selection, and the same 20 harmonised predictors were used identically in development and validation, reported in line with current guidance [11, 30]. Predictive performance was assessed across the levels of the calibration hierarchy [31] and summarised with several complementary measures and bootstrap confidence intervals rather than discrimination alone, following an established framework for evaluating prediction models [32].

Our study also has important limitations. Both databases comprise routinely collected data from United States intensive care units, and the analysis is therefore subject to the biases inherent in observational,

electronic-health-record research, including confounding by indication for the diuretic decision [33]. We were able to harmonise only variables that were comparably recorded in both databases; in particular, we could not include a direct measure of cumulative fluid balance or volume status — the very quantity that motivates de-resuscitation and is itself associated with mortality in critically ill patients [34] — so the model infers tolerance from indirect physiological and laboratory signals rather than from a measured fluid state. Measurement frequency also differed between databases for some variables, which we addressed with measurement-indicator features and a development-frozen imputation rule; nevertheless, heterogeneity in how and when predictors are measured between settings can itself degrade out-of-sample performance [35]. The external database aggregates intensive care units from many hospitals; we characterised transportability between databases rather than heterogeneity across individual sites within the external database, so site-level variation in calibration — which would inform the level at which local recalibration should be applied — remains to be quantified. Requiring at least 24 hours of in-unit observation to ascertain the outcome also excludes admissions discharged or transferred earlier, who tend to be at lower risk, which may have raised the observed event rate relative to the full population receiving intravenous furosemide, although the outcome was completely ascertained for included admissions. We also did not undertake a formal fairness or subgroup analysis of model performance across sociodemographic groups (for example by sex, age, or self-reported race or ethnicity); characterising any differential performance and the model's implications for health equity is an important direction for future work.

Several further constraints should be acknowledged. Although we followed practices intended to minimise bias, prediction models built on routinely collected data carry recognised methodological risks that structured appraisal tools are designed to detect, and machine-learning prediction studies in particular have frequently shown such shortcomings [36, 37]. Most importantly, we evaluated discrimination, calibration, and net benefit, but did not conduct a prospective impact study; showing that a model discriminates and is clinically informative does not establish that its use changes decisions or improves patient outcomes [38].

Within these limits, the model is best positioned as an adjunct to clinical judgement rather than a replacement for it, flagging patients at higher predicted risk of haemodynamic intolerance for closer monitoring or a more cautious approach to fluid removal, in keeping with the intended role of computer-based decision support [39]. Decision-curve analysis indicated that using the model to guide management would add net benefit over treat-all or treat-none strategies across a clinically plausible range of risk thresholds [17]. However, the threshold at which a predicted risk should alter management is a clinical and operational decision that depends on the local consequences of acting or not acting, and should not be set mechanically [40]; and because the model's absolute risks shift with local event prevalence, they should be recalibrated to the deploying unit before any threshold is applied [25].

Several directions for further work follow. Because patient populations and measurement procedures vary and change over time, a model cannot be regarded as definitively validated; the emphasis should therefore shift from developing new models towards extensive, well-reported validation, performance

monitoring, and updating of promising ones [41]. The present model requires prospective, preferably multi-centre, validation and ultimately an impact evaluation before clinical adoption [38]. Linking the prediction to the specific interventions tested in completed and ongoing de-resuscitation and fluid-strategy trials [5, 8, 9] would clarify how an at-risk prediction should change management at the bedside. Future iterations should incorporate a direct measure of fluid balance or volume status where it is reliably available, and should be deployed with a local recalibration step so that absolute risks remain accurate as case mix and event rates evolve [25].

# Conclusions

We developed and externally validated a model to predict haemodynamic intolerance within 24 hours of the first intravenous furosemide dose in haemodynamically stable intensive care patients — the point at which clinicians decide whether to begin de-resuscitation. Discrimination was moderate and consistent across a temporal split and an external multi-centre database, whereas calibration shifted predictably with the underlying event prevalence, a shift that a single logistic recalibration substantially reduced. A gradient-boosting model and penalised logistic regression performed similarly, supporting use of the simpler, more transparent model where auditability is valued. The model is intended as decision support, conditioned on the diuretic decision rather than its cause, and should be recalibrated locally before use; prospective evaluation of its effect on care remains the essential next step.

# Abbreviations

AKI: acute kidney injury; AUROC: area under the receiver-operating-characteristic curve; AUPRC: area under the precision–recall curve; CI: confidence interval; eICU-CRD: eICU Collaborative Research Database; ICI: integrated calibration index; ICU: intensive care unit; KDIGO: Kidney Disease: Improving Global Outcomes; MAP: mean arterial pressure; MIMIC-IV: Medical Information Mart for Intensive Care IV; SpO2: peripheral oxygen saturation; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence; XGBoost: extreme gradient boosting.

# Declarations

### Ethics approval and consent to participate

This study is a secondary analysis of two de-identified, publicly available critical care databases, MIMIC-IV and the eICU Collaborative Research Database, accessed through PhysioNet by credentialed investigators who completed the required human-subjects research training. The creation of these databases was approved by the responsible institutional review boards, and the requirement for individual informed consent was waived because all records are de-identified. No additional ethics approval was required for the present secondary analysis.

### Consent for publication

Not applicable.

**Clinical trial number**

Not applicable.

**Availability of data and materials**

The datasets analysed during the current study are available to credentialed users from PhysioNet: MIMIC-IV (https://physionet.org/content/mimiciv/) and the eICU Collaborative Research Database (https://physionet.org/content/eicu-crd/), subject to completion of the required training and a signed data use agreement. The cohort-extraction queries and analysis code that reproduce all results, together with the random seeds used, are openly available in the Zenodo repository (https://doi.org/10.5281/zenodo.20674156).

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Authors' contributions**

LW and JS conceived and designed the study, supervised the project, contributed to the methodology and to the interpretation of the results, and critically revised the manuscript for important intellectual content. LS performed the data extraction, statistical analysis, and model development, and drafted the manuscript. YL contributed to the study design and methodology, interpreted the results, and critically revised the manuscript for important intellectual content. CX contributed to data curation, statistical analysis, and validation of the results. HT contributed to interpretation of the results, visualisation, and critical revision of the manuscript. LW served as the first (primary) corresponding author and JS as the second corresponding author. All authors read and approved the final manuscript.

**Acknowledgements**

The authors thank the investigators who developed and maintain the MIMIC-IV and eICU-CRD databases and the PhysioNet platform for making these data publicly available.

# References

1. Boyd JH, Forbes J, Nakada TA, Walley KR, Russell JA. Fluid resuscitation in septic shock: a positive fluid balance and elevated central venous pressure are associated with increased mortality. Crit

Care Med. 2011;39(2):259–265.

2. Malbrain MLNG, Marik PE, Witters I, Cordemans C, Kirkpatrick AW, Roberts DJ, et al. Fluid overload, de-resuscitation, and outcomes in critically ill or injured patients: a systematic review with suggestions for clinical practice. Anaesthesiol Intensive Ther. 2014;46(5):361–380.
3. Hoste EA, Maitland K, Brudney CS, Mehta R, Vincent JL, Yates D, et al. Four phases of intravenous fluid therapy: a conceptual model. Br J Anaesth. 2014;113(5):740–747.
4. Malbrain MLNG, Van Regenmortel N, Saugel B, De Tavernier B, Van Gaal PJ, Joannes-Boyau O, et al. Principles of fluid management and stewardship in septic shock: it is time to consider the four D's and the four phases of fluid therapy. Ann Intensive Care. 2018;8(1):66.
5. Wichmann S, Schønemann-Lund M, Perner A, Itenov TS, Lange T, Gluud C, et al. Goal-directed fluid removal with furosemide versus placebo in intensive care patients with fluid overload: a randomised, blinded trial (GODIF trial—first version). Acta Anaesthesiol Scand. 2023;67(4):470–478.
6. National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network, Wiedemann HP, Wheeler AP, Bernard GR, Thompson BT, Hayden D, et al. Comparison of two fluid-management strategies in acute lung injury. N Engl J Med. 2006;354(24):2564–2575.
7. Silversides JA, Major E, Ferguson AJ, Mann EE, McAuley DF, Marshall JC, et al. Conservative fluid management or deresuscitation for patients with sepsis or acute respiratory distress syndrome following the resuscitation phase of critical illness: a systematic review and meta-analysis. Intensive Care Med. 2017;43(2):155–170.
8. Meyhoff TS, Hjortrup PB, Wetterslev J, Sivapalan P, Laake JH, Cronhjort M, et al. Restriction of intravenous fluid in ICU patients with septic shock. N Engl J Med. 2022;386(26):2459–2470.
9. National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury (PETAL) Clinical Trials Network, Shapiro NI, Douglas IS, Brower RG, Brown SM, Exline MC, et al. Early restrictive or liberal fluid management for sepsis-induced hypotension. N Engl J Med. 2023;388(6):499–510.
10. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25–34.
11. 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.
12. 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.
13. 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.
14. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney Int Suppl. 2012;2(1):1–138.

15. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
16. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019;38(21):4051–4065.
17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–574.
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–794.
20. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference (SciPy 2010). 2010. p. 57–61.
21. 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.
22. 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.
23. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
24. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052–1061.
25. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61(1):76–86.
26. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020;26(3):364–373.
27. Chawla LS, Davison DL, Brasha-Mitchell E, Koyner JL, Arthur JM, Shaw AD, et al. Development and standardization of a furosemide stress test to predict the severity of acute kidney injury. Crit Care. 2013;17(5):R207.
28. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol. 2016;79:76–85.
29. 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):4230–4251.
30. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ.

2015;350:g7594.
31. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–176.
32. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138.
33. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. JAMA. 2016;316(17):1818–1819.
34. Messmer AS, Zingg C, Müller M, Gerber JL, Schefold JC, Pfortmueller CA. Fluid overload and mortality in adult critical care patients—a systematic review and meta-analysis of observational studies. Crit Care Med. 2020;48(12):1862–1870.
35. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med. 2019;38(18):3444–3459.
36. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–58.
37. Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2023;154:8–22.
38. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–698.
39. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199–2200.
40. Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Van Calster B. Three myths about risk thresholds for prediction models. BMC Med. 2019;17(1):192.
41. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023;21(1):70.

# Figures

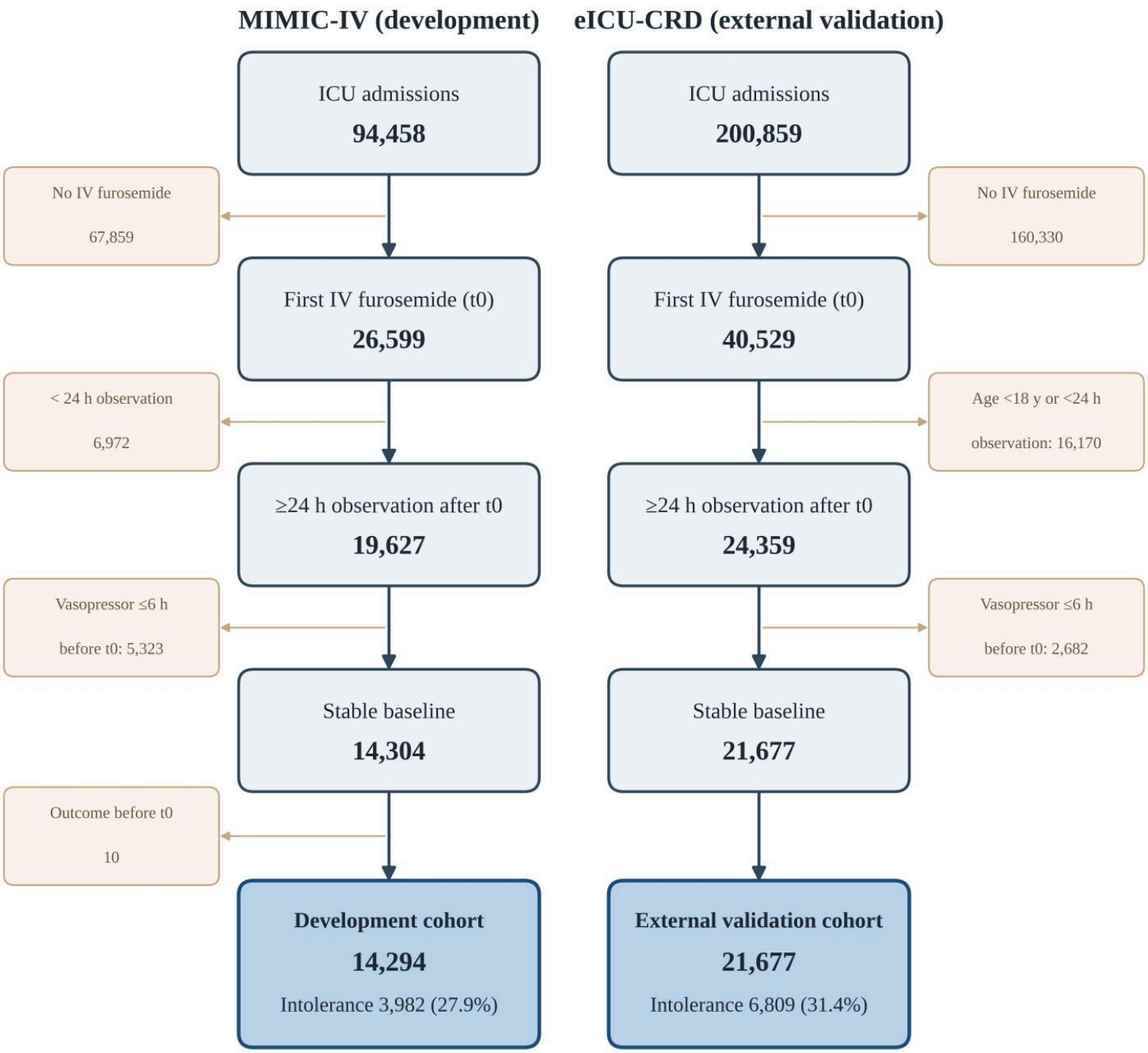
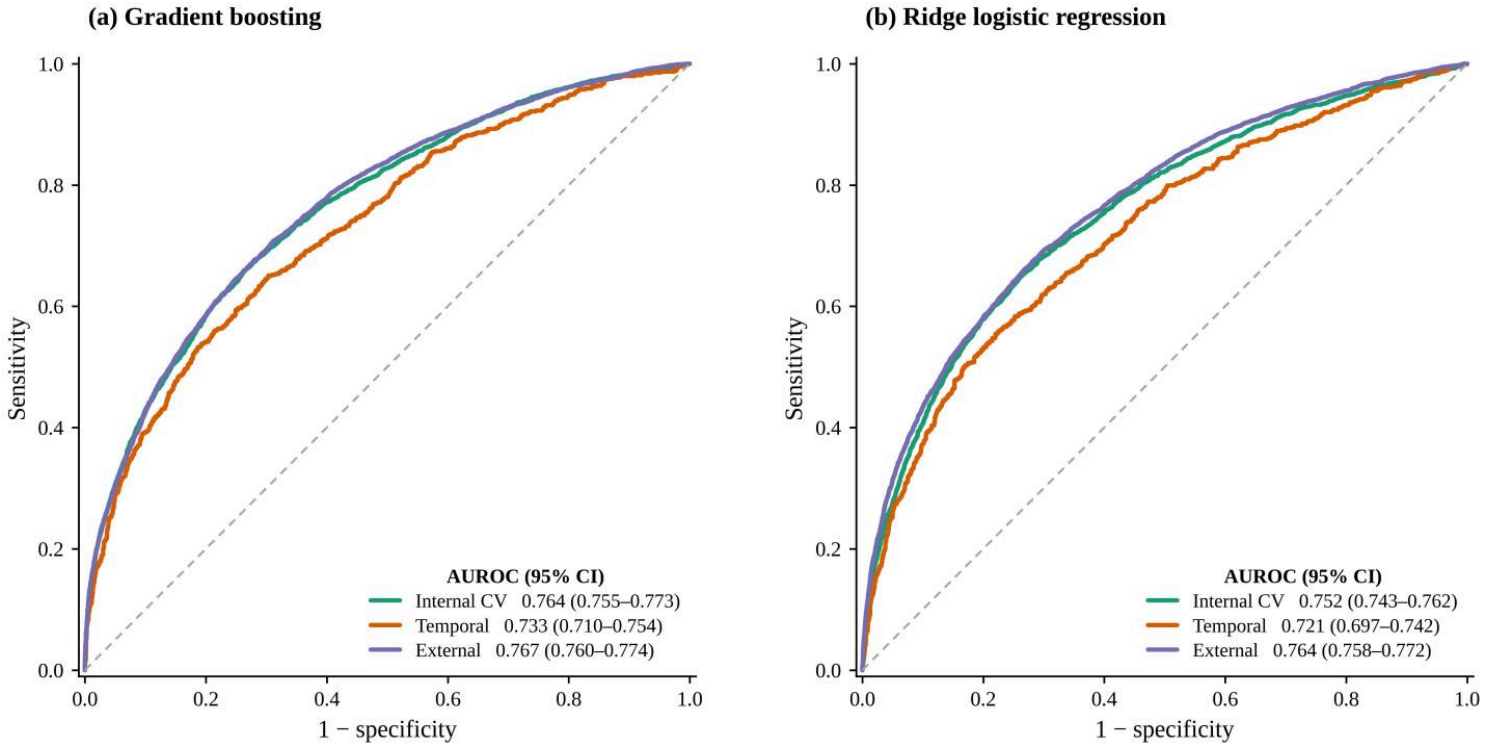



**Figure 1**

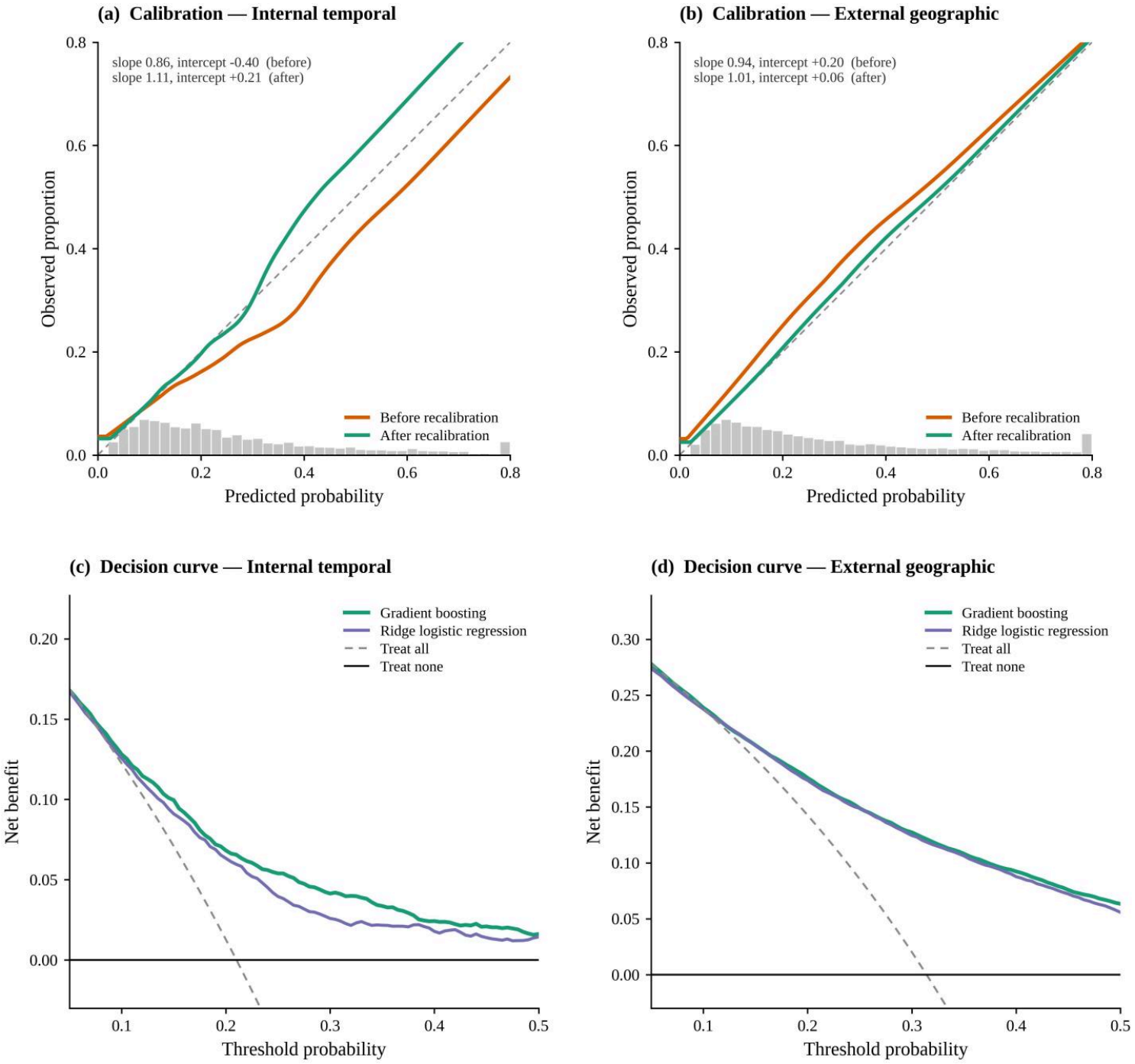
*Flow of adult intensive care unit admissions through the eligibility criteria for the development (MIMIC-IV v3.1) and external validation (eICU-CRD v2.0) cohorts. The decision point (t0) was the first intravenous furosemide dose. From all ICU admissions, the cohort comprised those receiving a first intravenous furosemide dose and was then restricted by excluding admissions with fewer than 24 hours of subsequent in-unit observation, those receiving a vasopressor in the 6 hours before t0 (non-stable baseline), and those for whom the outcome criterion was met before t0. Age below 18 years was an eligibility criterion in both*

*cohorts but was non-binding in MIMIC-IV, which contains only adults, and the exclusion of admissions meeting the outcome criterion before t0 applied to both cohorts and removed no eICU-CRD admissions.*




## Figure 2

*Receiver-operating-characteristic curves for the gradient-boosting model (a) and the ridge logistic-regression model (b) under internal cross-validation, internal temporal validation, and external geographic validation. The diagonal denotes chance performance. Areas under the curve with 95% confidence intervals (bootstrap, 1,000 replications) are shown in each legend. Discrimination was similar across the three settings, indicating portability across the temporal split and the transfer to the external database.*




## Figure 3

*Calibration and clinical utility of the gradient-boosting model under internal temporal validation (left column) and external geographic validation (right column). Top row: flexible calibration curves relating predicted to observed risk before and after a single logistic recalibration; the diagonal denotes perfect calibration. Under temporal validation the original model over-predicts risk (curve below the diagonal), reflecting the fall in event prevalence from the training to the test era; under external validation it under-predicts risk (curve above the diagonal), reflecting the higher prevalence in the external database; recalibration restores agreement in both settings. Bottom row: decision curves showing the net benefit of the model relative to*

*the treat-all and treat-none strategies across threshold probabilities. Recalibration parameters were estimated on one randomly selected half of each validation cohort and evaluated on the held-out half.*

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx