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.

Specification sensitivity of physiological clustering in intensive care: a retrospective methodological case study

Tianyi Yu

tianyi.1.yu@kcl.ac.uk

King's College London

Research Article

Keywords: cluster analysis, reproducibility, electronic health records, intensive care, measurement, renal function

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11261786/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

Physiological clusters are often evaluated by resampling stability and outcome differences. These checks answer different questions from robustness to measurement representation. We examined whether a reproducible intensive-care partition remained consistent under defensible changes in observation eligibility, renal representation and numerical estimation.

Methods

This retrospective methodological case study used MIMIC-IV, including 55,125 four-hour periods from 9,040 adults with coded diabetes, hypertension and early physiological stress. A three-component Gaussian mixture used arterial pressure, heart rate, recorded urine output and creatinine. Patient-disjoint analyses evaluated resampling, observation rules and absolute versus relative creatinine. A post-rejection amendment fitted 90 multistart seed models across nine specifications, selecting solutions by training likelihood within each specification. A planned quantile-spline prognostic comparator was rerun after identifying a uniform-knot implementation deviation. An eICU audit evaluated external measurement feasibility.

Results

Fixed-preprocessing bootstrap agreement had median adjusted Rand index (ARI) 0.944. After multiseed estimation, absolute and relative creatinine partitions in the same 9,232 test periods had ARI 0.377 (conditional patient-bootstrap 95% interval 0.351 to 0.404). Selecting higher-likelihood solutions also weakened the earlier apparent freshness robustness of absolute creatinine in the reference-available subgroup. In 1,441 twelve-hour test patients with 157 deaths, adding state to the corrected source-physiology model changed AUROC by -0.00068 (95% interval -0.00565 to 0.00438). Across 208 eICU hospitals, nine had no records meeting the urine-label screen; measurement equivalence remained unresolved.

Conclusions

High resampling stability did not establish specification invariance. The observed partitions were associated with risk, but did not demonstrate added discrimination in the tested physiological comparator or establish clinical response phenotypes. The contribution is a reproducible, bounded demonstration of the different evidence supplied by numerical stability, representation robustness and clinical validation.

Background

The research question was whether an intensive-care physiological partition that is reproducible under a fixed analysis remains consistent when reasonable measurement and modelling choices change. The distinction matters because a cluster can repeatedly reappear without identifying a discrete disease process. Published sepsis and acute respiratory distress syndrome studies have supported subphenotype interpretation through combinations of replication, biological characterization and randomized-treatment data [1, 2, 3]. Latent-class guidance likewise distinguishes model fit, numerical estimation and clinical interpretation [4]. Those forms of evidence cannot be replaced by a high agreement statistic.

Stability is conditional on the perturbation used to measure it. Resampling patients asks whether a fitted structure is sensitive to the sampled observations. Changing a variable from absolute concentration to relative change asks whether a different representation supports the same partition. Restricting the age of an eligible laboratory result changes which observations contribute to fitting. A convergence flag answers a computational question. Cluster-validation work treats these assessments as complementary, and shows why a stable algorithm need not recover a substantively meaningful grouping [5, 6]. Transparent specification analyses make the consequences of defensible analytical decisions visible, although no finite set can exhaust all reasonable analyses [7, 8].

Routine intensive-care records make these distinctions concrete. Measurements reflect physiology and the clinical process that generates observations. Completeness, plausibility and temporal availability are separate aspects of data quality [9, 10]. Testing intensity can be associated with illness even when the test value itself is not the principal source of information [11]. Models may exploit missingness patterns, but doing so does not establish that they have isolated an underlying physiological state [12]. A carried creatinine value and a newly recorded arterial pressure also have different temporal meanings.

Renal representation presents a particular interpretive problem. Absolute creatinine combines prior renal function, acute change and other determinants. A ratio to a reference value changes that mixture but introduces dependence on when and why the reference was measured [13]. Urine-output and creatinine criteria do not identify identical acute kidney injury populations, and renal recovery further complicates interpretation of a single value [14, 15, 16]. Restricting analysis to patients with a reference value changes the population as well as enabling normalization. A matched-population comparison is therefore necessary before attributing a different partition to relative renal change.

The clinical motivation was deliberately narrower than treatment selection. Static pressure, urine and creatinine measurements cannot establish fluid responsiveness, which concerns a change in flow after a defined preload challenge [17, 18, 19]. Trials of fluid composition, restriction and conservative management address different interventions in different clinical settings [20, 21, 22, 23, 24, 25]. None supplies treatment meaning to the labels of this four-variable mixture. Dynamic assessment itself requires a valid physiological measurement and a defensible decision context [26, 27].

We therefore used a selected MIMIC-IV cohort as an empirical case study of the limits of commonly reported clustering evidence [28, 29]. The primary objective was to compare conditional stability with sensitivity to eligibility, renal representation and numerical estimation. Secondary analyses asked whether frozen state labels improved a specified source-physiology prognostic comparator and whether existing records could support independent physiological or external validation. Reporting follows the relevant principles of STROBE and RECORD; the study is exploratory, and subsequent amendments are explicitly distinguished from the initial analysis [30, 31].

Methods

Design, cohort and analysis chronology

MIMIC-IV version 3.1 contains deidentified records from Beth Israel Deaconess Medical Center during 2008–2022 [28, 29]. The inherited cohort comprised adults with coded type 2 diabetes and hypertension, without coded chronic dialysis, and with an early physiological-stress screen. The screen required at least one of a positive creatinine-based kidney-injury screen, recorded urine output below 0.5 mL/kg/h, lactate at least 2 mmol/L, mean arterial pressure below 65 mmHg or recorded vasopressor exposure during ICU hours 0–6. The earliest eligible ICU stay per hospital admission was retained. Multiple admissions from one person could contribute to clustering, but always followed the same training/test assignment. Discharge diagnosis codes defined retrospective eligibility; this was not an admission-time screening tool.

State observation used complete four-hour periods starting at ICU hour 8 and ending by hour 48. Follow-up ended at ICU discharge, death or the first recorded comfort-measures time. The first possible state assignment was at hour 12. The sample was determined by available eligible records; no power-based sample size was asserted. Figure 1 separates people, ICU stays, periods and the subsequent clinical landmark cohort.

The original split and three-component specification preceded the reported diagnostic comparisons. Resampling, representation, clinical and physiological-reference amendments were subsequently recorded as exploratory extensions. The present post-rejection protocol specified the multiseed comparison, spline correction and timing audit before their new results were examined. All historical unfavorable results were retained. Because the test partition informed later questions, it is a patient-disjoint internal assessment set, not an untouched confirmatory sample. The supplementary amendment ledger records this chronology. Additional files 1, 2, 3 and 4 provide the historical state analyses, historical clinical analyses, physiological-reference audit and new methodological amendment, respectively.

Features and numerical model

Four features were required for each period: mean arterial pressure, heart rate, recorded urine output per kilogram per hour, and creatinine. Vital signs were within-period means of eligible positive values below 300, with chart and storage times before the period ended. Arterial-pressure identifiers were 220052, 220181 and 225312; heart rate used 220045. Urine entries were summed, required a nonnegative total and were divided by four hours and the upstream weight available before hour 6. Missing urine was not assigned zero. This calculation describes recorded output, not verified continuous collection.

Creatinine was the latest eligible chart-time value available before the period ended, with a primary maximum age of twelve hours. Alternatives required an observation in the current period or allowed maximum ages of eight or twenty-four hours. Source identifiers were 50912 and 52546, with positive values no greater than 150 mg/dL; these broad extraction limits were followed by training-derived winsorization. Concentrations are displayed in SI units where appropriate, using 88.4 micromol/L per mg/dL. The optional reference was the latest positive pre-ICU creatinine within seven days and stored before ICU hour 6. It was not a verified stable outpatient baseline, and no reference value was imputed.

Unique patients were sorted and shuffled with seed 20260922; 80% were assigned to training before period eligibility was applied. At most six eligible periods per training patient were sampled using the same recorded seed. The fitting sample supplied feature-specific first and ninety-ninth percentile winsorization limits. Urine and creatinine were log(1 + x)-transformed and all features standardized. Training transformations were applied unchanged to test periods.

The Gaussian mixture used three full-covariance components, covariance regularization 0.001, ten initializations, at most 300 iterations and convergence tolerance 0.001. K-means initialization was used. Assignment was the component with highest posterior probability; labels S1–S3 were ordered by the fitted transformed-creatinine mean. Maximum posterior probability quantified assignment uncertainty. Two-, four- and five-component alternatives were reported, without selecting the component count using outcomes. Three components constitute the worked example, not an estimated number of biological states. State profiles, transitions and clinical comparisons retain the original frozen model; the new likelihood-selected models are sensitivity fits and do not replace those labels. Code and aggregate frozen parameters support computational replay.

Stability, eligibility and representation comparisons

Adjusted Rand index (ARI) quantified agreement without requiring arbitrary labels to match [32]. Twenty original patient bootstraps refitted the mixture while holding preprocessing fixed, with two initializations per fit. Thirty later bootstraps re-estimated preprocessing, with five initializations. Both resampled whole groups of capped patient periods and evaluated the original test set. Their differing start counts prevent a clean causal attribution of distributional differences to preprocessing alone. Neither distribution includes uncertainty from a new patient split, cap or hospital.

Freshness rules filter a common physiological table: the feature values of periods eligible under both rules are unchanged. Frozen scoring should therefore agree on those periods. Independent refitting additionally changes the fitting observations, cap, transformations and numerical solution. We distinguish this eligibility-and-refitting experiment from a perturbation of a common measured value. Other historical stress tests removed creatinine or required at least three urine entries, or three entries for both vital signs. An observation-only random forest used recording counts, creatinine age and period index to predict state assignments, with a majority-class comparator. This diagnostic was not a test for absence of observation bias.

The new multiseed experiment covered nine specifications: full-cohort absolute creatinine at three freshness limits, and absolute and relative creatinine at those same limits in the reference-available population. Absolute and relative fits used identical eligible patients and the same capped rows within each freshness limit. Each specification used ten seeds, 20260922–20260931, with ten internal starts per seed. All ten multistart seed-fit outputs were retained. The converged seed fit with highest training mean log likelihood was selected separately within each specification; likelihoods were never ranked across representations or samples. This search does not establish a global optimum.

Within-rule seed agreement and all across-rule seed pairs in the full cohort were evaluated on the common eight-hour test periods. Representation comparisons used the common eligible test population at each freshness limit. Pairwise ARIs are dependent descriptive comparisons, not independent inferential replicates. For the selected twelve-hour representation contrast, 2,000 whole-patient bootstrap resamples of test assignments yielded a conditional percentile interval. The fitted models remained fixed. A cross-tabulation and maximum-overlap label alignment describe which assignments differ; the alignment is mathematical and does not equate clinical diagnoses.

Temporal observation and prognostic comparison

Transitions required consecutive four-hour indices within the same ICU stay, with both states observed; gaps were not bridged. Row probabilities used 500 patient-bootstrap samples. Unavailable next states were reported separately. One hundred within-stay permutations preserved observed positions and state counts while changing their order. Their central range is a descriptive reference, not a confidence interval for a biological mechanism. A further analysis stratified persistence by whether the selected creatinine specimen chart time changed, with 1,000 patient-bootstrap samples.

The exploratory clinical analysis retained the earliest upstream admission per person and required survival in ICU and hospital at hours 12 or 24. It excluded missing or inconsistent death records. The twelve-hour analysis required the current state; the twenty-four-hour analysis also required the twelve-hour state. Outcomes were subsequent in-hospital death and descriptively recorded new renal replacement therapy (RRT) in the indexed ICU stay after the landmark. RRT was not complete hospital-wide ascertainment, and death could preclude treatment. No treatment effect was estimated.

Logistic models with fixed L2 regularization (C = 1) compared age/sex alone, age/sex plus state, source physiology with age/sex, and that physiological comparator plus state. The twenty-four-hour extension added the earlier hard label and a change indicator to current physiology and state. This does not evaluate the full physiological history. Continuous predictors used cubic splines with four knots; winsorization, knots and scaling were fitted in training only. No outcome-guided tuning or comprehensive severity-score comparison was undertaken.

The original clinical plan specified quantile knots, but the executed code omitted the knot argument and therefore used uniform knots. The corrected analysis explicitly used quantile knots; explicit uniform-knot reruns and historical results were preserved. Both report AUROC, average precision, Brier score, calibration intercept and slope, and five-bin calibration displays. Grouped displays are coarse descriptive summaries that may conceal local miscalibration; no smooth calibration curve was estimated. Paired test-patient bootstraps, 2,000 in this amendment, estimated AUROC and Brier differences while holding both models fixed. These intervals omit training and cluster-estimation uncertainty. Prognostic reporting and interpretation draw on prediction-model evaluation guidance [33, 34, 35, 36]. We did not prespecify an equivalence margin or a clinical action threshold.

Timing and validation-feasibility audits

An audit identified that the early creatinine screen restricted the current result's storage time but not every prior reference result contributing to its rolling minima. We recomputed the screen after requiring all such references to be stored before hour 6 and compared ordinal stages, binary screens and final cohort membership. State-feature availability rules already included storage-time restrictions. This audit tests a specific extraction issue rather than certifying every temporal construct.

Existing passive-leg-raise (PLR) audits examined whether recorded manoeuvres and flow measurements could provide an independent physiological reference. An expanded audit removed the diabetes/hypertension restriction, used strictly prior four-hour features and retained the frozen model. It distinguished original training, original test and previously unseen patients from the same hospital. Charted CO/CI percentage changes were described at both 10% and 15% thresholds, without choosing by significance; they were not treated as adjudicated responses. Full methods and all negative feasibility results remain in Additional file 3.

The new eICU-CRD version 2.0 audit used inherited local extracts to describe hospital-level availability and candidate-cohort support [37, 38]. Its site screen counted stays with mapped vital signs, positive creatinine and urine-related labels over the recorded stay. Thus, coverage is not complete-period eligibility or validation of collection intervals. The supported early-stress subset used available MAP/lactate criteria and was not equated with the full MIMIC screen. We inspected source documentation on intake/output semantics and periodic vitals [39, 40]. No external state assignments or outcome validation were performed.

Research assistance

OpenAI Codex assisted with analysis-code development and execution, diagnostic review, figure code and manuscript drafting. Numerical results were generated by executing the supplied local scripts. No patient observations or experimental results were invented. The author retains responsibility for the analysis, interpretation and final submission. Computational checks and same-family review are not independent clinical adjudication or external peer review.

Results

Cohort and descriptive partition

The upstream cohort contained 12,175 ICU stays from 10,183 people. There were 90,862 candidate periods and 55,125 complete primary periods from 10,576 stays and 9,040 people. The eligible training set included 7,236 people; capping contributions yielded 33,436 fitting periods. The test set included 1,804 people and 10,978 periods. Table 1 describes patients at their first eligible stay.

Table 1
Patient characteristics at the first eligible stay.

Characteristic	Training	Test
People	7236	1804
Age, median (IQR), years	71 (63–79)	71 (63–79)
Female sex	2896 (40.0%)	733 (40.6%)
Chronic kidney disease code	2386 (33.0%)	587 (32.5%)
Heart-failure code	2845 (39.3%)	689 (38.2%)
Reference creatinine at first eligible stay	6032 (83.4%)	1489 (82.5%)

Diagnosis codes are retrospective admission codes. A reference may be available at a later admission even when unavailable at the first eligible stay.

State medians differed mainly in pressure, urine output and creatinine (Fig. 2). S1–S3 had mean arterial pressure medians of 82, 71 and 73 mmHg, recorded urine rates of 1.379, 0.474 and 0.521 mL/kg/h, and creatinine concentrations of 88.4, 97.2 and 238.7 micromol/L. Chronic kidney disease codes occurred in admissions contributing 23.3%, 29.4% and 69.9% of their periods. These are period-weighted descriptions; patients can contribute to multiple states. Maximum posterior probability was below 0.7 in 23.8%, 24.6% and 20.8% of periods, respectively.

All two- through five-component models converged, but test mean log likelihood improved from -5.550 with two components to -5.481, -5.434 and -5.426 with three, four and five. Three components were

therefore not the best-fitting choice among those examined. Neither fit nor convergence establishes a discrete clinical taxonomy.

Numerical stability and eligibility sensitivity

The fixed-preprocessing patient bootstraps had median ARI 0.944 and minimum 0.615. The historical preprocessing-refitted bootstraps had median 0.929 and range 0.287–0.976 (Fig. 3). These high medians coexist with substantially different recovered partitions. Removing creatinine gave ARI 0.348. The observation-only classifier had balanced accuracy 0.406 versus 0.333 for a majority classifier, but lower ordinary accuracy, 0.389 versus 0.471. Recording patterns contain some information about model labels; this result cannot quantify or exclude observation bias.

All 90 new multistart seed models converged. The likelihood-selected full-cohort twelve-hour model agreed with the eight-hour model at ARI 0.240 and with the twenty-four-hour model at ARI 0.767. Within the eight-hour specification, seed-pair ARIs ranged from 0.241 to 0.999, whereas the twelve-hour range was 0.992–1.000. The selected low-agreement eight-hour solution had better training likelihood within its own specification; it was not discarded because it differed from the primary labels. The common evaluation set contained 7,471 periods. Frozen scoring retained identical common-period assignments, as expected from identical feature values. The disagreement under refitting therefore concerns eligibility, estimation and representation, rather than an isolated effect of changing the age of a value for the same period. Full likelihoods, pairwise comparisons and selected seeds are supplied in Additional file 4.

Renal representation after multiseed estimation

The twelve-hour reference-available test population contained 9,232 periods from 1,519 people. After selecting the highest training-likelihood fit within each representation, absolute and relative creatinine still disagreed at ARI 0.377 (conditional 95% interval 0.351–0.404). Both selected twelve-hour fits used the original seed, so the point estimate reproduced the historical 0.377 result rather than replacing it with a more favorable contrast. The bootstrap interval is conditional on the selected fitted models and does not encompass another model search. Across the ten retained seed fits, the 100 twelve-hour cross-representation ARIs ranged from 0.376 to 0.384. Within-representation ARIs ranged from 0.980 to 1.000 for absolute and 0.981 to 1.000 for relative creatinine. However, the newly selected eight-hour absolute model agreed with the twelve-hour absolute model at only ARI 0.233 on 6,314 common periods; the historical single-seed comparison had been 0.923. The corresponding selected relative-model comparison was 0.939. Thus, the earlier claim that both representations were comparatively robust to freshness changes was not sustained for absolute creatinine. Figure 4 displays the full between-representation seed comparisons and the selected-model cross-tabulation. Maximum-overlap alignment gave cluster Jaccard indices of 0.706, 0.591 and 0.222. The limited third-cluster overlap (736 shared periods among 3,315 in the union) shows why a common numerical label cannot establish a shared clinical construct.

The result concerns two mathematically different representations of the same selected records. It does not identify either representation as the physiological truth. A pre-ICU reference can already reflect acute

illness; dividing by it can remove chronic variation while introducing reference error. The separately ordered state labels consequently should not be matched by their numerical names.

Persistence, missing observations and clinical associations

Among 7,338 adjacent observed pairs, conditional persistence was 62.4% for S1, 73.0% for S2 and 89.6% for S3. However, 35.0%, 27.5% and 23.1% of eligible origins lacked an observed next state (Fig. 5). Overall persistence was 74.5%, compared with a within-stay permutation mean of 67.7% and central range 67.1–68.3%. Persistence was 74.2% when the creatinine specimen time was unchanged and 75.3% when it changed. This near similarity supplies no evidence that specimen reuse alone explains persistence, while observational selection and other serial dependence remain unresolved.

At twelve hours, 7,448 of 9,770 eligible landmark survivors had a complete state. The test subset included 1,441 patients and 157 deaths. Mortality was 26/414 (6.3%) in S1, 67/788 (8.5%) in S2 and 64/239 (26.8%) in S3; recorded new indexed-stay RRT was 3/413, 12/787 and 32/233 (Fig. 6). The S3-versus-S1 mortality difference was 20.5 percentage points, with an interval of 12.0–29.8 after the historical twelve-comparison adjustment. These unadjusted gradients do not establish three distinct pathophysiological conditions.

Selection was clinically relevant: across both splits, 347/2,322 excluded landmark survivors died, compared with 773/7,448 included patients. At twenty-four hours, 4,654 patients had both required states, including 872 test patients and 89 deaths. These cohorts are conditional on surviving and remaining observable and are not independent replications.

Corrected prognostic comparison

With the corrected quantile knots, twelve-hour source-physiology AUROC was 0.7499, compared with 0.7492 after adding state. The paired difference was -0.00068 (95% interval -0.00565 to 0.00438). Brier scores were 0.08459 and 0.08445, a difference of -0.00014 (95% interval -0.00050 to 0.00024). Calibration intercept/slope were 0.115/0.992 and 0.112/0.986, respectively. Figure 7 presents the corrected and uniform-knot increments and corrected calibration. At twenty-four hours, adding the earlier state and change indicator to current physiology and current state changed AUROC by -0.00161 (95% interval -0.00985 to 0.00694). All comparisons, including the positive increments over demographics alone, are reported in Additional file 4. The explicit uniform-knot rerun reproduced all original patient predictions to a maximum absolute difference of 1.11e-16; its twelve-hour state increment was -0.00015 (2,000-bootstrap interval -0.00465 to 0.00451). The historical 1,000-bootstrap interval and this amended interval use different resampling realizations and are both retained. An interval spanning zero is not proof of equivalence; the estimates bound only the tested incremental comparison under the stated fixed-fit resampling procedure. No claim is made that every use of state labels or every temporal model lacks value.

Timing integrity and limits of external validation

Requiring early reference creatinine results to be available before hour 6 changed two of 19,711 candidate ordinal screens from stage zero to missing. No positive-screen indicator or overall stress indicator changed. All 12,175 retained ICU stays were unchanged, with no additions or removals. The extraction defect therefore had no measured cohort-membership effect in these data.

The inherited expanded PLR audit retained 497 patients with complete prior features and a charted CO/CI percentage change, including 380 previously unseen patients. This increased coverage within the same hospital. Raw-flow pairs were available for one of 1,408 recorded tests in the strict one-minute post-manoeuvre window and three with a five-minute window. Neither charted changes nor pairs establish valid manoeuvre execution, recovery or a clinical response standard (Additional file 3).

The external feasibility audit covered 200,859 eICU stays from 208 hospitals. It identified 18,737 chronic-disease candidate stays and 9,620 with the supported MAP/lactate stress definition. The median hospital fraction of stays with a urine-related record was 77.8% (interquartile range 52.8–89.9%); nine hospitals had none under this label screen (Supplementary Fig. 5). These all-stay summaries do not demonstrate complete four-hour urine collection. Unresolved interval semantics, weight availability, vital-sign processing and cohort equivalence prevented interpretation as external phenotype validation.

Discussion

This study provides a quantitative example of why different robustness checks should not be treated as interchangeable evidence. A twelve-hour physiological partition was often recovered under patient resampling, yet reasonable changes to eligibility and renal representation yielded different assignments. The principal twelve-hour representation disagreement persisted across seeds, whereas the apparent freshness robustness of absolute creatinine in the reference-available subgroup did not. This asymmetry is scientifically important: the additional experiment supports one part of the interpretation while overturning another. It prevents a selectively favorable narrative in which restricting to a renal-reference population appears to solve specification sensitivity. The conclusion is limited to this cohort, model family and set of specifications. The scientific contribution is not that clustering depends on its inputs, but that a reassuring resampling summary can coexist with materially different interpretations even after a more extensive numerical search.

The matched-population comparison is central to that argument. It separates a change in renal representation from the elementary confounding created by restricting analysis to patients with a reference value. It does not separate all consequences of normalization: ratios alter distributional shape and dependence on the other features, and a reference measured during illness introduces uncertainty. Clinical studies of creatinine trajectories motivate attention to relative change, but do not validate this particular ratio or partition [13, 41]. Neither high stability nor a more acute-looking profile can select the biologically preferable representation without independent evidence.

The numerical experiment also changes how the freshness analysis should be read. For periods eligible under both rules, the stored feature values are identical. The difference arises through the observations admitted to fitting and the model subsequently estimated. Reporting only the first eight-hour refit would mix that eligibility effect with initialization variability. Retaining all multistart seed fits and selecting within-specification training likelihood makes the comparison more defensible, but does not guarantee that the largest likelihood corresponds to a clinically useful grouping. Full covariance mixtures and three components remain modelling decisions [4, 5, 6].

Mortality separation is compatible with a severity-related partition rather than distinct pathophysiology. Creatinine is part of the state definition, and the high-creatinine group has substantial chronic kidney disease enrichment. A model can therefore separate outcome risk without isolating a treatment-responsive mechanism. Because frozen state assignment is a deterministic function of the source measurements, adding it to the physiological comparator adds a derived representation, not a new measured variable. The corrected comparison did not establish additional discrimination for the tested additive spline model. It also does not establish lossless compression, irrelevance to every prediction task, or equivalence within a clinically important margin.

The spline discrepancy illustrates the value of checking the executed analysis against its plan. Uniform and quantile knots are both possible modelling choices, but they are not interchangeable descriptions. Reporting the planned correction alongside the original implementation preserves the distinction between a repaired deviation and a search for a favorable comparator. The small observed increments require interpretation alongside calibration and the limited event count, rather than a binary conclusion from whether an interval includes zero [33, 36]. The twenty-four-hour result applies only to adding an earlier hard label and change indicator; it does not evaluate all longitudinal information.

Clinical validation remains a separate task. Physiological response requires an interpretable challenge and contemporaneous flow measurement, while fluid tolerance concerns whether additional fluid can be accommodated safely [18, 27, 42]. The expanded chart audit increased the number of assessable records without resolving the validity of the manoeuvre. Its negative feasibility finding illustrates why more records alone need not strengthen construct validation. Similarly, the eICU audit documents multicentre heterogeneity in measurement availability; it does not demonstrate transport of the original biological construct. Official source documentation cautions that cumulative intake/output totals can repeat and include non-urine output, and that periodic vital signs undergo monitor-based aggregation [39, 40]. Aligning column names would not resolve those differences.

The distinction between computation and interpretation should guide the next study. An independent investigator could reproduce frozen state assignments from the supplied parameters without thereby validating their clinical meaning. A transport study should first establish comparable collection intervals, renal references and availability times, then assess a frozen definition in patients who did not inform development. It should distinguish failure of a source definition to transport from the emergence of a new partition after refitting. Prior replicated subphenotype studies provide examples of stronger

evidential chains, not universal validation of retrospective clustering [1, 2, 3]. The published corrections to the fluid-strategy subphenotype study further show why treatment labels and supporting measurements must be verified before transferring an interpretation [43, 44].

Several limits remain. This is a single-centre physiological analysis in a selected, code-defined population. The early stress screen is broad, diagnosis timing is retrospective and missingness is informative. Repeated test-set use limits confirmation. The model omits validated congestion, respiratory support, complete fluid balance and vasopressor intensity. Recorded urine volume divided by four hours does not verify collection coverage; a fresh creatinine result is not necessarily an unbiased sample of renal physiology. Imputation could address specified missing-data mechanisms, but cannot make those mechanisms ignorable by assumption [45, 46, 47]. The transition matrix conditions on a subsequently observed state and should not be interpreted as a fully observed biological process.

Statistical scope is also restricted. Bootstrap intervals hold the split, cap and fitted comparator fixed as specified; seed-pair distributions are dependent; the historical preprocessing experiments use different initialization counts. No experiment tests whether a continuous biological distribution fits better than all discrete-state models. The fixed component count, feature set, regularization and four-hour grid are not globally optimal. The timing audit closes one measured cohort-impact concern, not every possible extraction limitation. Outcome follow-up depends on hospitalization, and indexed-stay RRT is a practice-dependent record with competing death. These limits are retained rather than converted into a claim that a larger number of analyses would make the experiment definitive.

For future clinical use, the appropriate design depends on the claim. A prognostic model needs separate validation, adequate calibration and a clearly defined decision; demographic descriptions alone cannot establish equitable performance [34, 35, 48]. A treatment comparison requires intervention definitions, aligned eligibility and time zero, confounding control and adequate treatment support [49, 50]. Prospective decision support requires evaluation of its actual use and consequences [51, 52]. None of these requirements is satisfied by an outcome gradient across retrospective clusters.

Table 2 summarizes the evidence supplied by each assessment and its limits.

Table 2
What each assessment can establish.

Assessment	Evidence in this study	Inference boundary
Patient resampling	Median fixed-preprocessing ARI 0.944	Conditional reproducibility, not biological validity
Matched renal representation	Selected 12-hour ARI 0.377	Different partitions; neither representation is a reference standard
Likelihood-selected refitting	Reference-population 8/12-hour absolute ARI 0.233	Earlier apparent freshness robustness is weakened
Outcome association	S3 mortality 26.8%; S1 6.3%	Unadjusted prognosis, not mechanism or treatment effect
Corrected physiological comparator	Added AUROC -0.00068	No demonstrated increment in this model; not equivalence
External availability	208 eICU hospitals audited	Feasibility evidence; no external phenotype validation
These are interpretation boundaries, not validated deployment criteria.		

Conclusions

In this selected intensive-care cohort, high within-specification resampling stability did not establish robustness to reasonable changes in renal representation and eligibility-dependent refitting. Multiseed estimation clarified the numerical contribution, and correction of a spline implementation deviation made the prognostic comparison consistent with its stated plan. The results support treating these labels as specification-dependent analytical partitions. They do not establish distinct biological states, fluid responsiveness, treatment benefit or an externally validated clinical tool.

Abbreviations

ARI, adjusted Rand index; AUROC, area under the receiver operating characteristic curve; CO/CI, cardiac output/cardiac index; GMM, Gaussian mixture model; ICU, intensive care unit; MAP, mean arterial pressure; PLR, passive leg raising; RRT, renal replacement therapy.

Declarations

Ethics approval and consent to participate

The study used deidentified secondary MIMIC-IV and eICU data, with no participant contact, new collection or intervention. The MIMIC data-sharing initiative was approved by the Beth Israel Deaconess Medical Center institutional review board with a waiver of informed consent. The eICU resource was released under its documented deidentification and source governance arrangements [29, 37, 38]. No

new study-specific approval was obtained for this secondary analysis; no institutional exemption certificate or additional approval number is claimed. Source access remains subject to PhysioNet credentialing and the applicable data-use agreements.

Consent for publication

Not applicable. No identifiable individual information is presented.

Clinical trial number

not applicable.

Competing interests

The author declares no competing interests.

Funding

This research received no funding.

Author Contribution

Tianyi Yu is the sole author and is responsible for the study conception, methodology, analysis, interpretation, manuscript revision and final approval.

Acknowledgements

The author thanks the patients, clinical teams and data custodians whose contributions made the MIMIC and eICU resources possible.

Data Availability

MIMIC-IV version 3.1 and eICU-CRD version 2.0 are available through PhysioNet to researchers who satisfy their access requirements [29, 38]. Aggregate results supporting this article, executable analysis code, amendment records, codebooks and frozen aggregate scoring parameters are supplied in the reproducibility files. Patient-level extracts, identifiers, individual predictions and serialized fitted patient-linked objects are excluded. The external data were used for feasibility auditing, not phenotype validation.

References

1. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA. 2019;321:2003–17. https://doi.org/10.1001/jama.2019.5791.
2. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA. NHLBI ARDS Network. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respiratory Med. 2014;2:611–20. https://doi.org/10.1016/s2213-2600(14)70097-9.
3. Famous KR, Delucchi K, Ware LB, Kangelaris KN, Liu KD, Thompson BT, et al. Acute Respiratory Distress Syndrome Subphenotypes Respond Differently to Randomized Fluid Management Strategy. Am J Respir Crit Care Med. 2017;195:331–8. https://doi.org/10.1164/rccm.201603-0645oc.
4. Sinha P, Calfee CS, Delucchi KL. Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. Crit Care Med. 2021;49:e63–79. https://doi.org/10.1097/ccm.0000000000004710.
5. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal. 2007;52:258–71. https://doi.org/10.1016/j.csda.2006.11.025.
6. Ullmann T, Hennig C, Boulesteix AL. Validation of cluster analysis results on validation data: A systematic framework. WIREs Data Min Knowl Discov. 2022;12:e1444. https://doi.org/10.1002/widm.1444.
7. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. Perspect Psychol Sci. 2016;11:702–12. https://doi.org/10.1177/1745691616658637.
8. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nat Hum Behav. 2020;4:1208–14. https://doi.org/10.1038/s41562-020-0912-z.
9. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inf Assoc. 2013;20:144–51. https://doi.org/10.1136/amiajnl-2011-000681.
10. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. EGEMS (Wash DC). 2016;4:1244. https://doi.org/10.13063/2327-9214.1244. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.
11. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479. https://doi.org/10.1136/bmj.k1479.
12. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci Rep. 2018;8:6085. https://doi.org/10.1038/s41598-018-24271-9.
13. Siew ED, Matheny ME. Choice of Reference Serum Creatinine in Defining Acute Kidney Injury. Nephron. 2015;131:107–12. https://doi.org/10.1159/000439144.
14. Kellum JA, Sileanu FE, Murugan R, Lucko N, Shaw AD, Clermont G. Classifying AKI by Urine Output versus Serum Creatinine Level. J Am Soc Nephrol. 2015;26:2231–8. https://doi.org/10.1681/asn.2014070724.

15. Chawla LS, Bellomo R, Bihorac A, Goldstein SL, Siew ED, Bagshaw SM, et al. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative (ADQI) 16 Workgroup. Nat Rev Nephrol. 2017;13:241–57. https://doi.org/10.1038/nrneph.2017.2.
16. Forni LG, Darmon M, Ostermann M, Oudemans-van Straaten HM, Pettilä V, Prowle JR, et al. Renal recovery after acute kidney injury. Intensive Care Med. 2017;43:855–66. https://doi.org/10.1007/s00134-017-4809-x.
17. Myburgh JA, Mythen MG, Resuscitation Fluids. N Engl J Med. 2013;369:1243–51. https://doi.org/10.1056/nejmra1208627.
18. Monnet X, Marik PE, Teboul JL. Prediction of fluid responsiveness: an update. Ann Intensive Care. 2016;6:111. https://doi.org/10.1186/s13613-016-0216-7.
19. Marik PE, Cavallazzi R. Does the Central Venous Pressure Predict Fluid Responsiveness? An Updated Meta-Analysis and a Plea for Some Common Sense. Crit Care Med. 2013;41:1774–81. https://doi.org/10.1097/ccm.0b013e31828a25fd.
20. Semler MW, Self WH, Wanderer JP, Ehrenfeld JM, Wang L, Byrne DW, et al. Balanced Crystalloids versus Saline in Critically Ill Adults. N Engl J Med. 2018;378:829–39. https://doi.org/10.1056/nejmoa1711584.
21. Finfer S, Micallef S, Hammond N, Navarra L, Bellomo R, Billot L, et al. Balanced Multielectrolyte Solution versus Saline in Critically Ill Adults. N Engl J Med. 2022;386:815–26. https://doi.org/10.1056/nejmoa2114464.
22. Zampieri FG, Machado FR, Biondi RS, Freitas FGR, Veiga VC, Figueiredo RC, et al. Effect of Intravenous Fluid Treatment With a Balanced Solution vs 0.9% Saline Solution on Mortality in Critically Ill Patients: The BaSICS Randomized Clinical Trial. JAMA. 2021;326:818–29. https://doi.org/10.1001/jama.2021.11684.
23. 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;386:2459–70. https://doi.org/10.1056/nejmoa2202707.
24. National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury Clinical Trials Network, Shapiro NI, Douglas IS, Brower RG, et al. Early Restrictive or Liberal Fluid Management for Sepsis-Induced Hypotension. N Engl J Med. 2023;388:499–510. https://doi.org/10.1056/nejmoa2212663.
25. National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network, Wiedemann HP, Wheeler AP, Bernard GR, et al. Comparison of Two Fluid-Management Strategies in Acute Lung Injury. N Engl J Med. 2006;354:2564–75. https://doi.org/10.1056/nejmoa062200.
26. Douglas IS, Alapat PM, Corl KA, Exline MC, Forni LG, Holder AL, et al. Fluid Response Evaluation in Sepsis Hypotension and Shock: A Randomized Clinical Trial. Chest. 2020;158:1431–45. https://doi.org/10.1016/j.chest.2020.04.025.

27. Monnet X, Teboul JL. Passive leg raising: five rules, not a drop of fluid! Crit Care. 2015;19:18. https://doi.org/10.1186/s13054-014-0708-5.
28. 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. https://doi.org/10.1038/s41597-022-01899-x.
29. 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
30. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335:806–8. https://doi.org/10.1136/bmj.39335.541782.ad.
31. Benchimol EI, Smeeth L, Guttmann 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:e1001885. https://doi.org/10.1371/journal.pmed.1001885.
32. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218. https://doi.org/10.1007/bf01908075.
33. 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.
34. 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.
35. 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;170:51–8. https://doi.org/10.7326/m18-1376.
36. 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;17:230. https://doi.org/10.1186/s12916-019-1466-7.
37. 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.
38. 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
39. MIT Laboratory for Computational Physiology. eICU Collaborative Research Database documentation: intakeOutput. https://eicu.mit.edu/eicutables/intakeoutput/. Accessed 1 October 2026.
40. MIT Laboratory for Computational Physiology. eICU Collaborative Research Database documentation: vitalPeriodic. https://eicu.mit.edu/eicutables/vitalperiodic/. Accessed 1 October

2026.

41. Bhatraju PK, Mukherjee P, Robinson-Cohen C, O'Keefe GE, Frank AJ, Christie JD, et al. Acute kidney injury subphenotypes based on creatinine trajectory identifies patients at increased risk of death. Crit Care. 2016;20:372. https://doi.org/10.1186/s13054-016-1546-4.
42. Kattan E, Castro R, Miralles-Aguiar F, Hernández G, Rola P. The emerging concept of fluid tolerance: A position paper. J Crit Care. 2022;71:154070. https://doi.org/10.1016/j.jcrc.2022.154070.
43. Erratum. Acute Respiratory Distress Syndrome Subphenotypes Respond Differently to Randomized Fluid Management Strategy. Am J Respir Crit Care Med. 2018;198:1590. https://doi.org/10.1164/rccm.v198erratum5.
44. Erratum. Acute Respiratory Distress Syndrome Subphenotypes Respond Differently to Randomized Fluid Management Strategy [Additional Corrections]. Am J Respir Crit Care Med. 2019;200:649. https://doi.org/10.1164/rccm.v200erratum3.
45. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66. https://doi.org/10.2147/clep.s129785.
46. Sterne JAC, 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.
47. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30:377–99. https://doi.org/10.1002/sim.4067.
48. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–53. https://doi.org/10.1126/science.aax2342.
49. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016;183:758–64. https://doi.org/10.1093/aje/kwv254.
50. Hernán MA, Dahabreh IJ, Dickerman BA, Swanson SA. Ann Intern Med. 2025;178:402–7. https://doi.org/10.7326/annals-24-01871. The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful?.
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;28:924–33. https://doi.org/10.1038/s41591-022-01772-9.
52. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2:e537–48. https://doi.org/10.1016/s2589-7500(20)30218-1.

Figures

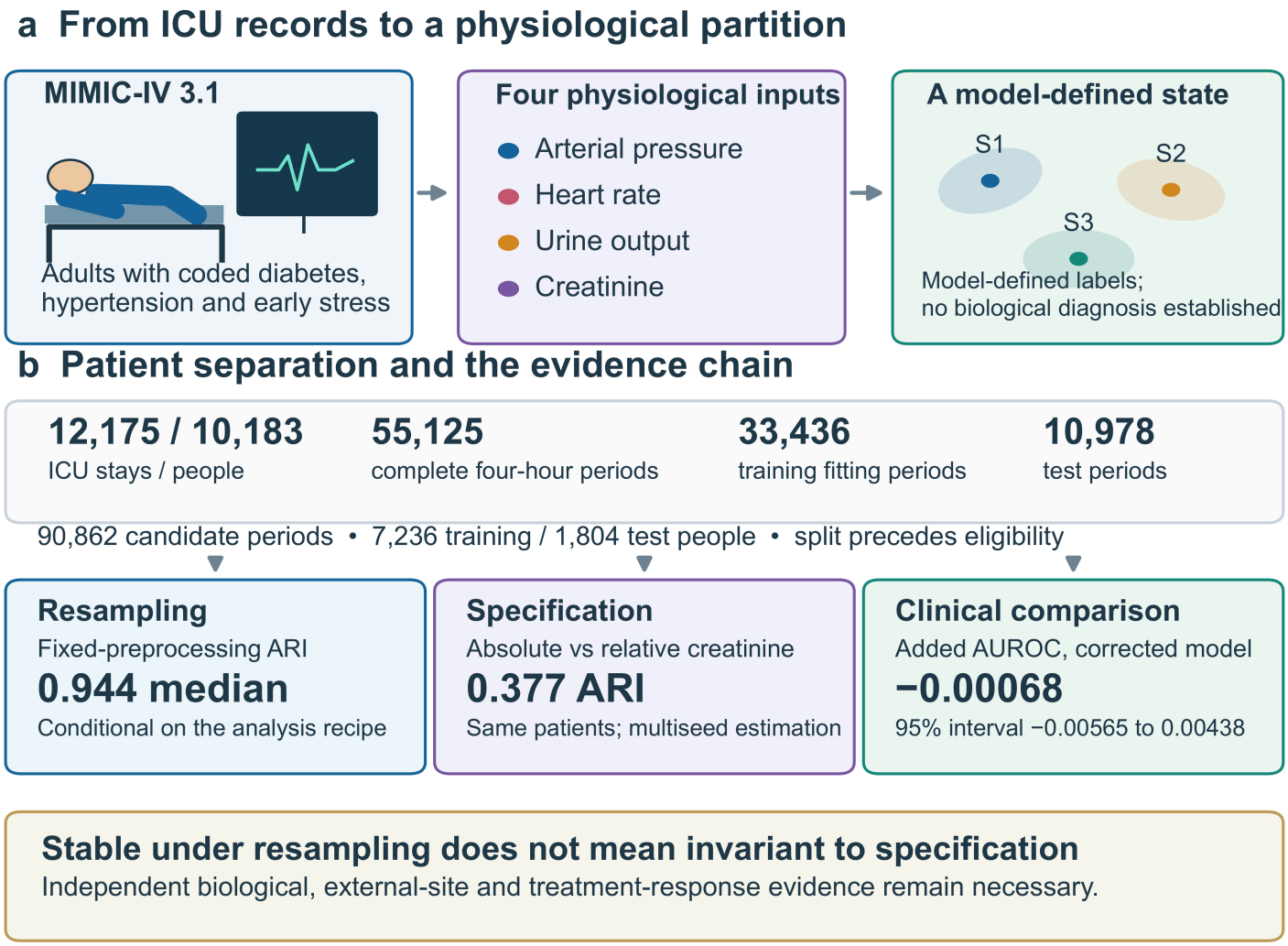


Figure 1

Study design and evidence chain. Original editable pictograms depict ICU records and model-defined partitions schematically; ellipses are not patient observations. Counts distinguish upstream stays/people, complete periods, capped fitting periods and test periods. The 80:20 patient split precedes feature eligibility. The numerical summaries refer to historical fixed-preprocessing bootstrap stability, the likelihood-selected twelve-hour matched-population representation comparison, and the corrected quantile-spline twelve-hour prognostic comparison. Later diagnostics reuse the internal test partition.

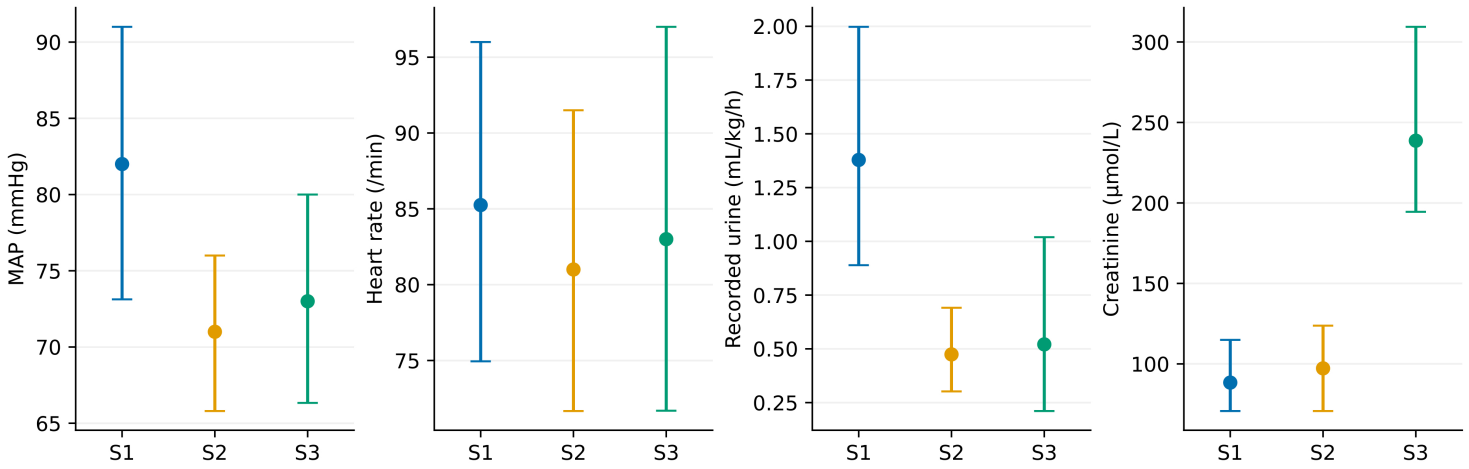


Figure 2

Primary state profiles. Points and bars indicate period-weighted medians and interquartile ranges, not confidence intervals. Patients can contribute periods to more than one state. Creatinine uses 88.4 micromol/L per mg/dL. Urine refers to recorded volume divided by weight and four hours. Labels are descriptive component identifiers.

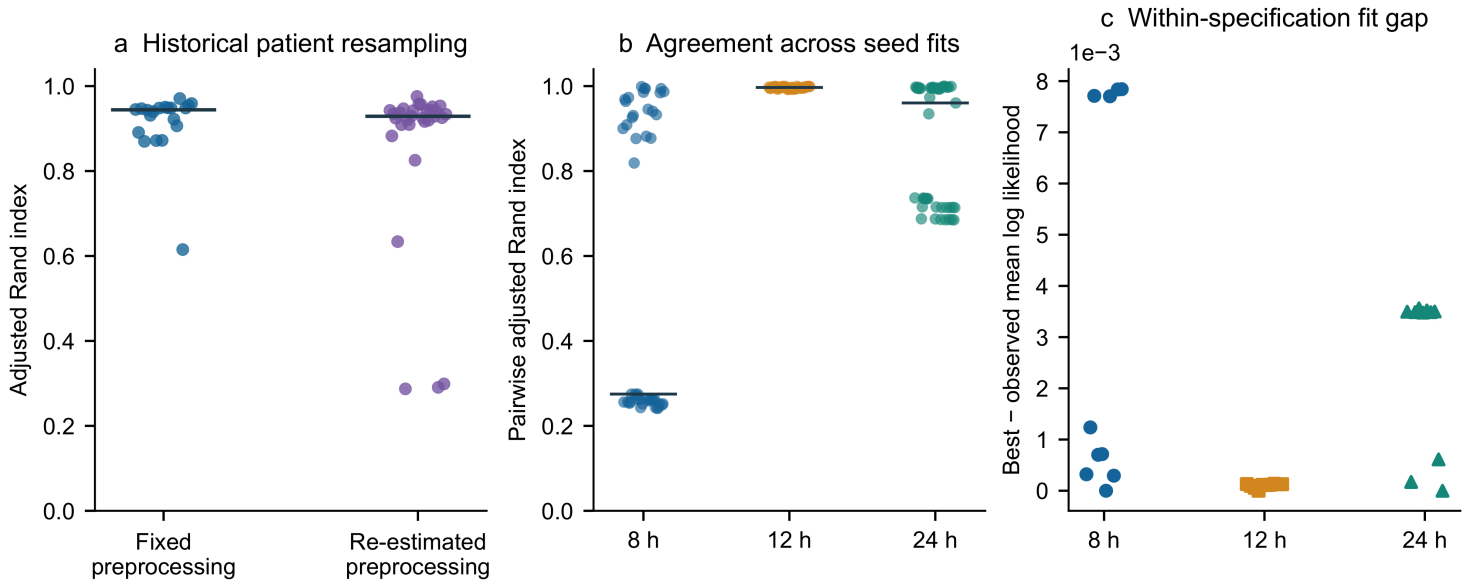


Figure 3

Conditional stability and numerical optimization. (a) Historical patient bootstraps with fixed preprocessing (20 fits, two starts each) and re-estimated preprocessing (30 fits, five starts each). Their difference does not isolate preprocessing. (b) All 45 pairwise comparisons among ten multistart seed fits per freshness specification, evaluated on the same 7,471 eight-hour-eligible full-cohort test periods. (c) Training mean log-likelihood deficit from the best seed fit within each specification; zero identifies the selected fit. Likelihoods are not ranked across specifications. Points are fits or dependent pairs, not independent patients; horizontal segments mark medians.

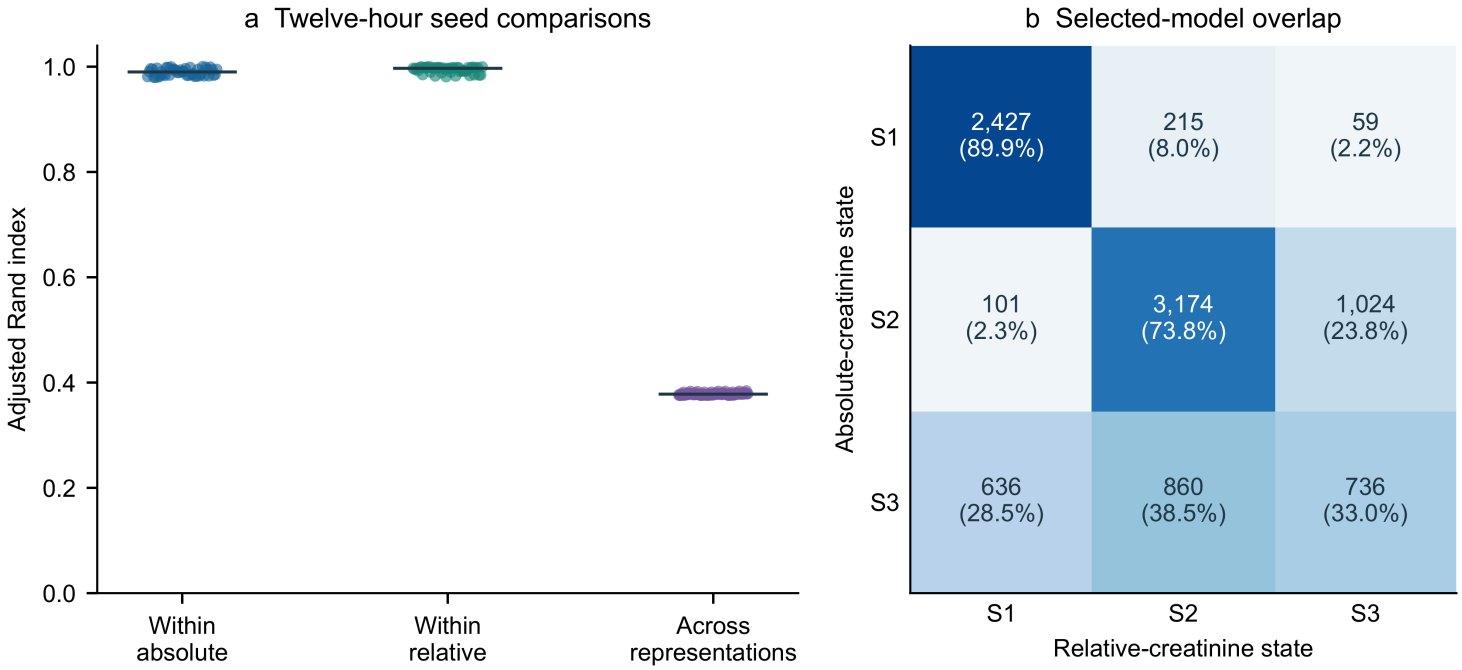


Figure 4

Renal representation and assignment overlap. (a) All seed-pair ARIs for twelve-hour absolute and relative representations in the same 9,232 test periods: 45 within each representation and 100 across representations. Horizontal segments mark medians; comparisons are dependent. (b) Counts and row percentages for the selected highest-training-likelihood models. Rows are absolute-creatinine labels and columns relative-creatinine labels. Separate ordering by renal coordinate does not imply matched diagnoses. The selected-model ARI is 0.377, with a conditional patient-bootstrap interval of 0.351–0.404.

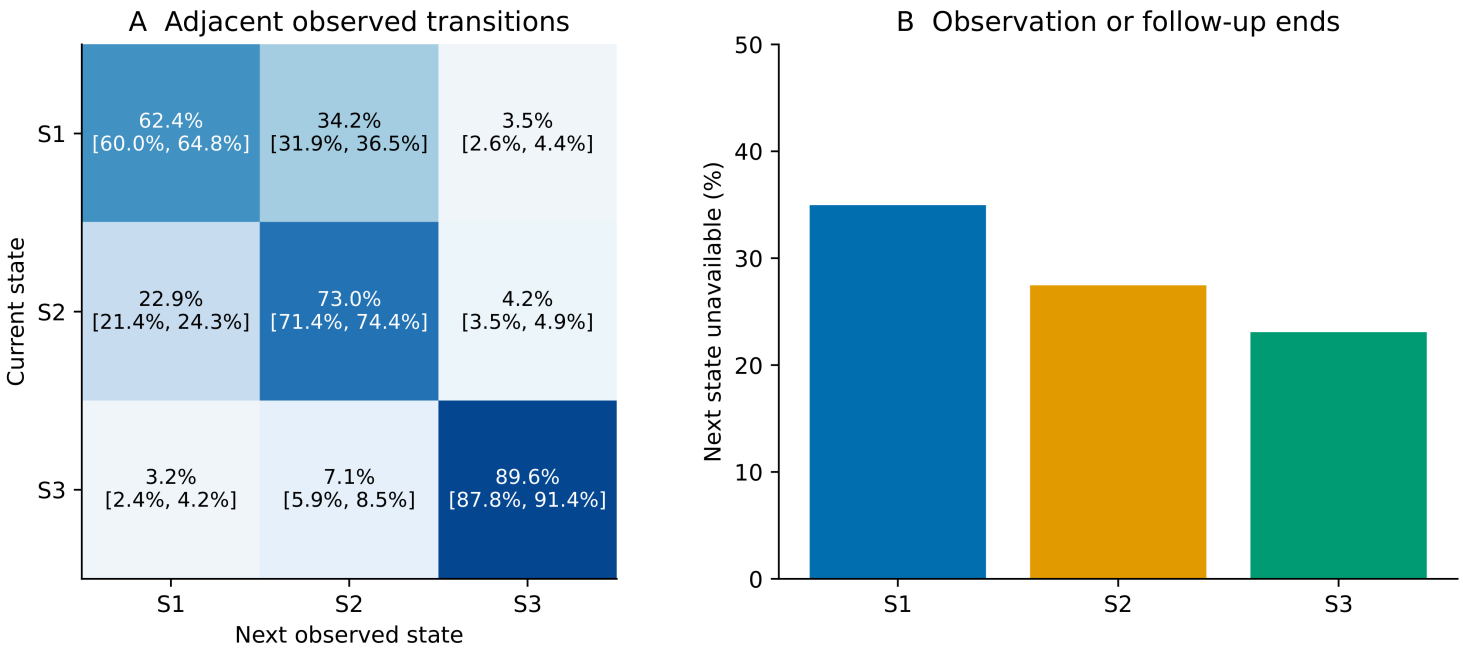


Figure 5

Conditional transitions and unavailable next states. (a) Row-normalized transitions among 7,338 adjacent observed test pairs, with 500 patient-bootstrap percentile intervals. (b) Fractions of eligible origins without a next observed state before the final scheduled period. Unavailability includes follow-up termination and insufficient measurements; missing states are not assigned persistence.

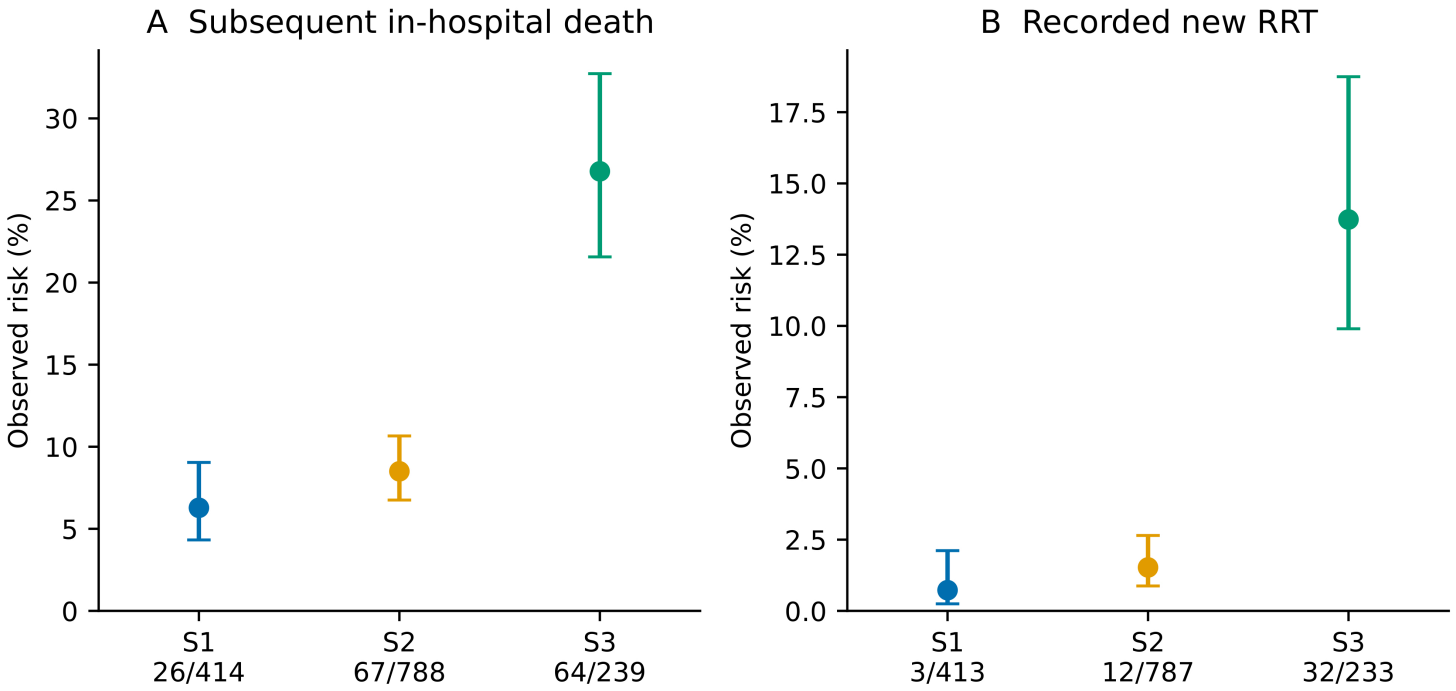


Figure 6

Outcome gradients in the twelve-hour test cohort. Points and Wilson 95% intervals describe subsequent in-hospital mortality and recorded new RRT in the indexed ICU stay. RRT denominators exclude treatment already initiated by the landmark. Outcome differences are unadjusted descriptive associations, not evidence of treatment response.

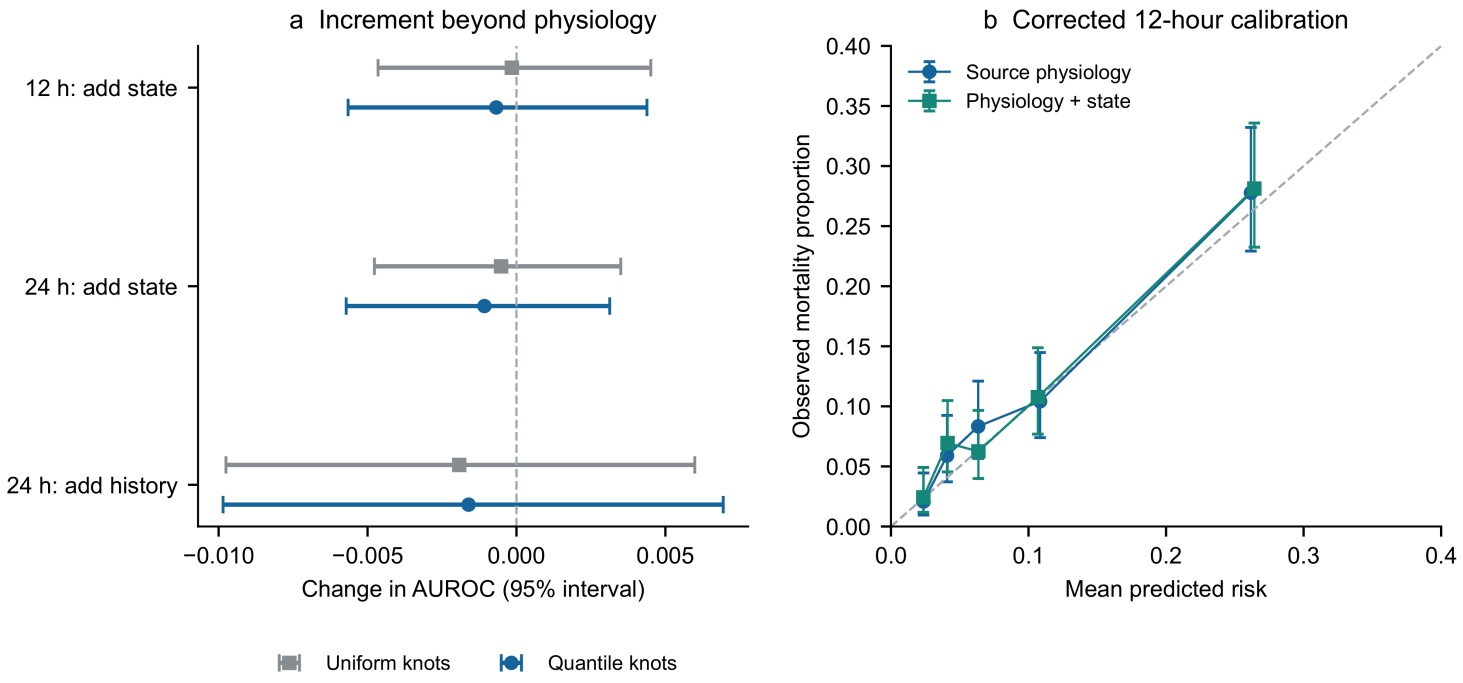


Figure 7

Corrected incremental prognosis and calibration. (a) AUROC differences from 2,000 paired bootstrap resamples of test patients for quantile knots (planned correction) and uniform knots (original implementation). The history comparison adds the earlier hard label and change indicator to current physiology and state. Fitted models remain fixed; intervals do not include development uncertainty. (b) Twelve-hour corrected quantile-knot calibration in five equal-frequency groups, with Wilson intervals for observed mortality. The diagonal denotes agreement. These displays do not establish equivalence or clinical utility.

Supplementary Files

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

- SFig1.jpeg
- SFig3.jpeg
- SFig4.jpeg
- SFig2.jpeg
- SFig5.jpeg
- supplement.pdf