Supplementary material

First-day SOFA changes and mortality in older adults with sepsis: a retrospective study of two ICU databases

Lei Wang; Ying-chao Liu; Yu Song

Contents

Page

Supplementary methods and results **S1**

S1. Source cohort rules S1
S2. Local score construction S2
S3. Time policies and remaining uncertainty S3
S4. Statistical implementation details S3
S5. Legacy analyses retained as context S3
S6 Version and source mapping S4
S7 Completion analysis of score construction S4
S8 Clinical provenance and follow-up audit S5
S9 Precision and independence in external prediction S5
S10 Reporting scope S6
S11 Completion results for scoring and selection S6
S12. Exact pairwise score comparisons for Figure 2a S7
S13 Revised descriptive and balance tables S9
S14 Rule crosswalk and criterion-validation status S9

Supplementary tables **S11**

Table S1 MIMIC-IV characteristics by restricted-policy SOFA change S11
Table S2 Flow from source selection to the 24-hour landmark S12
Table S3 Broad path associations conditional on attained SOFA S13
Table S4 Strict-path eligibility and matched comparisons across policies S14
Table S5 Matched mortality at 28 90 and 365 days S16
Table S6 Additional deaths after the day 28 landmark S17
Table S7 External prediction differences with hospital bootstrap intervals S18
Table S8 All prediction metrics by recording policy S19
Table S9 eICU score and component sensitivity across timestamp policies S22
Table S10 eICU input record timestamps S23
Table S11 MIMIC-IV score and component sensitivity to storage restrictions S24
Table S12 Recording and component source values S25

Contents continued

Page

Table S13 Strict-path associations before and after early care adjustment S26

Table S14 Prediction performance after available timestamp restrictions S27

Table S15 Full-source screening under inherited definitions S28

Table S16 Documented pre-index clinical history S29

Table S17 Mortality horizons and date-only uncertainty S30

Table S18 Score and category sensitivity S31

Table S19 Rematched 365-day recorded mortality S32

Table S20 Conditional associations across score policies S33

Table S21 Outcomes within the archived net-zero groups S34

Table S22a Absolute external prediction performance S35

Table S22b External calibration and patient sensitivity S36

Table S23 Paired external model differences S37

Table S24 Demographic coverage and attained-model precision S38

Table S25 Direct carried and missing inputs by window S39

Table S26 Recorded zeros and coherent GCS S40

Table S27 Entry-service selection within the archived risk set S41

Table S28 Exact pairwise SOFA score comparisons for Figure 2a S42

Table S29 Landmark cohorts and measurement coverage S43

Table S30 Organ component scores by change group in MIMIC-IV S44

Table S30 Organ component scores by change group in eICU S45

Table S31 Matching balance before matching S46

Table S31 Matching balance after matching S47

Table S32 Local scoring rules and interpretation S48

Supplementary figures **S49**

Figure S1 Score changes and class reassignments under timestamp restrictions S49

Figure S2 Early care records and adjusted mortality associations S50

Figure S3 Discrimination of locked models in eICU S51

Figure S4 External calibration and recording-policy sensitivity S52

Figure S5 Locked external calibration and risk distribution S53

Supplementary methods and interpretation boundaries

Historical methods: 10 October 2026; completion analyses added 11 October 2026. This document describes the implemented exploratory analysis, including inherited restrictions. It does not present a prospective protocol.

S1. Source cohort rules

MIMIC-IV

The source base retained the first ICU stay per subject, ordered by ICU entry time and stay identifier. Age was calculated from anchor age plus ICU-entry year minus anchor year; the older cohort required age ≥65. This preserves the database's deidentification of the oldest ages and is not an exact recovered age for every patient.

The local author-code eligibility reconstruction searched antibiotic prescriptions and microbiology records within 24 hours before and after ICU admission using subject-level joins. Antibiotics used the inherited author-name list, MAIN or ADDITIVE prescription types, and route/form exclusions. Culture time used chart time, chart date, storage time or storage date in that fallback order. Eligibility required both antibiotic and culture evidence, no antibiotic or culture evidence before ICU entry within the search window, and evidence after entry. Records precisely at the entry time did not alone satisfy the after-entry condition. These rules differ from an unrestricted clinical diagnosis of sepsis and from other operational suspected-infection algorithms.

The service rule selected the last recorded service between ICU entry and ICU exit plus one day. The excluded service codes were CSURG, VSURG, TSURG, DENT, ENT, EYE, NB, NBB, NMED, NSURG, OBS, ORTHO, OMED, PSURG, PSYCH and TRAUM. Missing service was not itself an exclusion. The preliminary source construction required ICU stay ≥6 hours; the present landmark subsequently required residence beyond 24 hours. Because the service window extends beyond the landmark, this is retrospective cohort selection.

Further joins required staged local physiology, a local first-day eligibility SOFA ≥2, subtype/output records inherited from the parent pipeline and outcome records. The present SOFA analysis did not use the predicted RI/RW label as a predictor, matching stratum or treatment modifier. Recorded steroid orders starting before −10 hours were excluded according to the available extraction. Hospital discharge time and nonmissing discharge destination were also required in MIMIC-IV. Thus, future outcome-data completeness contributed to eligibility, and the manuscript does not claim a fully real-time inception cohort.

All three remaining steroid categories were retained: no relevant order during −10 to +24 hours, an order with extracted hydrocortisone-equivalent daily dose below 160 mg, or one meeting the threshold. The dose category came from summed record-derived daily-equivalent values within the window and does not establish actual dose administration. Restoring the subthreshold group added 77 landmark stays; other inherited restrictions remained.

eICU

The source base used unitvisitnumber=1, adult recorded age, and the available unit encounter identifiers. The recorded age category > 89 was represented as 90. This first-unit-visit definition is not identical to MIMIC-IV first lifetime subject stay and does not ensure only one encounter per person across all admissions.

The local eICU infection rule required both an antibiotic prescription and culture evidence within 0–1,440 minutes. Cultures could be identified in microbiology or culture-related treatment records. Prior infection was operationalized as both earlier antibiotic and culture evidence; this differs from the MIMIC-IV author-code exclusion of either earlier form of evidence in its window. CSICU and CTICU unit types were excluded. The pipeline also required staged first-day physiology, local eligibility SOFA ≥2, assignment/subtype/outcome joins,

absence of extracted pre-window steroid orders and nonmissing hospital discharge time. Missing discharge location was mapped to “Other”, rather than excluded as in MIMIC-IV.

Landmark and selection interpretation

The MIMIC-IV and eICU source-selected cohorts contained 4,069 and 1,647 stays. Landmark eligibility required recorded ICU exit >24 hours and no recorded death ≤24 hours. Mutually exclusive exclusions yielded 3,689 and 1,395 stays (Table S2). These counts start after inherited source selection; they are not full-database screening denominators. A later completion audit reconstructed the full inherited screening flow (Table S15). This remains different from a prospectively fixed common eligibility rule.

S2. Local score construction

The following describes the implemented local six-hour algorithm. It must not be labelled an unmodified standard daily SOFA or SOFA-2. Each organ was scored 0–4, and totals were summed. Criteria were evaluated from highest to lowest; when no condition was activated, zero points were assigned.

Component	Implemented positive-point conditions
Respiratory	PaO_2/FiO_2 <100 with recorded invasive ventilation: 4; <200 with ventilation: 3; <300: 2; <400: 1
Coagulation	Platelets <20: 4; <50: 3; <100: 2; <150: 1; platelets in 10^9/L
Liver	Bilirubin ≥12: 4; ≥6: 3; ≥2: 2; ≥1.2: 1; bilirubin in mg/dL
CNS	GCS <6: 4; ≤9: 3; ≤12: 2; ≤14: 1
Renal	Creatinine ≥5 mg/dL or positive six-hour urine ≤50 mL: 4; creatinine ≥3.5 or positive urine ≤125 mL: 3; creatinine ≥2: 2; creatinine ≥1.2: 1
Cardiovascular	Highest satisfied condition in the explicit rule below

FiO_2 values above 1 were divided by 100. The PaO_2/FiO_2 denominator was retained only in the range 0.2–1.0. Window-level extrema can combine measurements taken at different moments; this does not ensure a simultaneously measured gas/oxygen pair. For MIMIC-IV GCS, the three component minima were summed only if all three were observed in the window before carrying GCS forward. The eICU score used total GCS. Missing invasive-ventilation flags were treated as zero in the baseline reconstruction.

Cardiovascular scoring used normalized dopamine, dobutamine, epinephrine, norepinephrine, phenylephrine and vasopressin rates. Four points were assigned for dopamine >15, epinephrine >0.1 or norepinephrine >0.1, or the combination of a higher vasopressin/phenylephrine condition with a catecholamine-three condition. The former was vasopressin >0.4 or phenylephrine ≥200; the latter was dopamine 5–15, or positive epinephrine/norepinephrine ≤0.1. Three points were assigned for dopamine ≥5, any positive epinephrine, norepinephrine or vasopressin, or phenylephrine ≥200. Two points were assigned for positive dopamine, dobutamine or phenylephrine. One point was assigned for mean arterial pressure <70 mmHg. Catecholamine rates used μg/kg/min, phenylephrine μg/min and vasopressin units/min after the source-specific normalization. These vasopressin/phenylephrine rules are adaptations, not canonical SOFA criteria; their source-specific mappings remain a transport limitation.

The algorithm carried PaO_2, FiO_2, platelets, bilirubin, creatinine, mean arterial pressure, GCS and six-hour urine values forward within each stay, using only earlier windows. It did not fill from later windows. In particular, carrying urine forward is an inherited implementation choice and does not reconstruct a newly observed interval urine volume. Zero-valued raw urine entries were removed before historical aggregation; the score additionally used positive-urine oliguria predicates. This two-stage handling was explicitly revisited in the completion sensitivities. Missing values that did not meet any score condition contributed zero points. Missingness was therefore not ignorable, and none of these choices establishes normal physiology.

S3. Time policies and remaining uncertainty

The raw-record reconstruction reproduced all 14,756 MIMIC-IV and 5,580 eICU windows before modifying time policies. MIMIC-IV restricted laboratory, nursing and urine storage timestamps to ICU hour 24. Original support intervals were retained. The eICU policies were nested: half-open point-observation windows; additional infusion-entry-window alignment; additional entry/revision offset restrictions.

The eICU lab policy used labresultrevisedoffset, whereas observation was located using labresultoffset. Nursing, respiratory and urine entries used their corresponding entry-offset columns. Infusion entries used infusionoffset, documented as entry time. Periodic/aperiodic vital-sign records did not provide an independent entry timestamp. Laboratory revision can occur after an earlier result was known; no inference of absent clinical knowledge was made from the revision gate alone. Official documentation: lab, infusionDrug, respiratoryCharting, intakeOutput.

A support sensitivity in MIMIC-IV triangulated late-stored ventilation procedures with contemporaneous endotracheal-tube and ventilator-mode records. Of 1,349 attained-window ventilation intervals, 210 procedures were stored by 24 hours. Of the remaining 1,139, 999 had that contemporaneous documentation. This corroboration is incomplete and is not a gold-standard ventilator-state reconstruction. The support sensitivity produced 121 strict matched pairs and a 365-day risk difference of −12.4 percentage points (95% CI −23.1 to −1.7), with care-adjusted OR 0.98 (0.60–1.62) in 601 strict candidates. The 116-pair analysis remains the physiological-storage-policy result; the two samples should not be mixed.

S4. Statistical implementation details

For the broad attained-score analysis, current integers 2–13 entered as indicator variables, with net-stable change as the reference category. Age entered as a centered continuous term per 10 years, sex as a binary recorded-sex indicator, and the three steroid-record categories as indicators. Restricted-policy model size was 3,344; the broad categorical 365-day model was separate from the 600-person strict-path analysis. The strict model additionally adjusted for first-window direct-input count. Neither the exact pairing nor the ordinary regression included a validated frailty score or a complete pre-ICU disease history.

Matching used a linear assignment within exact strata of attained score, recorded sex, steroid category and initial direct-input count. Eligible age gaps were ≤5 years. The cost was age gap plus twice the attained-window input-count gap, with a deterministic tie-breaking term. Dummy assignments permitted unmatched stays. Pair-bootstrap intervals resampled 5,000 sets of pairs with seed 20261007. Exact discordant-pair tests used the binomial distribution. The day-28 analysis retained only pairs with both members still alive and therefore changed the target population.

The prediction models used scikit-learn standardization followed by L2-regularized logistic regression with C=1. Standardization was fitted separately within each training fold and then within the complete MIMIC-IV training cohort for external evaluation. Five-fold stratification and paired hospital-bootstrap sampling used seed 20261009. All 1,000 hospital resamples had both outcome classes. The same draws were used across model comparisons. The bootstrap estimates evaluation-sample uncertainty conditional on fitted development models.

Calibration intercept and slope in the source table came from jointly regressing outcome on an intercept and the logit of predicted risk. That intercept is not a separately estimated calibration-in-the-large intercept with slope fixed to one. No model was recalibrated using these estimates.

S5. Legacy analyses retained as context

Earlier MIMIC-IV analyses used a 3,612-person two-category steroid subset, original event-time scores and a 122-pair strict comparison. Those figures are not the current restricted-policy primary descriptive tables. The later full-risk-set strict comparison included 600 candidates and 116 pairs. All main-text descriptions use the latter where applicable.

An earlier MIMIC-III CareVue transport sensitivity used 1,151 possible eligible landmark stays and incompletely aligned infection, treatment-time, GCS and support definitions. For attained scores 2–13 and age/sex adjustment, improvement versus net stability had a 365-day OR of 0.830 (0.605–1.139). Its 28-day OR was 0.680 (0.464–0.998), changing to 0.708 (0.477–1.051) when requiring at least five direct initial inputs. Those findings did not confirm the long-term association and do not provide a same-definition validation of the present restricted-policy analysis. They are disclosed as exploratory history rather than pooled with eICU short-term evaluation.

The corticosteroid target-trial results and heart-failure fluid analyses concern separate questions and estimands. No corticosteroid mortality hazard ratio or fluid-dose threshold is used to establish the present manuscript's mechanism or primary conclusion.

S6 Version and source mapping

The main figures retain the accepted flow design, recorded-mortality Figure 3 and individual component-profile Figure 4. All four use the archived availability-gated adapted score, explicitly identified in their legends. Tables S1–S14 retain the historical audited values; new completion tables begin at S15. Source Data contains full-precision aggregate outputs. The accepted figures were not silently regenerated under a different score policy.

S7 Completion analysis of score construction

Before refitting outcome models on 11 October 2026, a completion plan fixed five scoring policies: the archived availability-gated reconstruction; no carrying forward of interval urine volume; inclusion of recorded zero-valued urine entries with no urine carrying; coherent GCS totals; and the combined urine/GCS policy. The reference policy was retained to connect the accepted figures with the earlier results. No policy was selected on its mortality association. These are recording sensitivities for an adapted six-hour score, not competing validated implementations of standard daily SOFA.

Urine extraction was checked before aggregation. Positive and zero-valued observed entries were distinguished from absent records and negative corrections. The zero-inclusive policy considered nonnegative entries recorded in each six-hour interval and available by hour 24, with no carry of that interval total to a later interval. A charted zero is not proof of complete six-hour anuria; completeness of fluid recording and interval versus cumulative labels remain relevant. Missing urine was not replaced with zero. Creatinine criteria and the other inherited components were held constant.

In MIMIC-IV, coherent GCS required a complete valid eye, verbal and motor triplet at the same observation timestamp. The worst such total within each window was used; forward-only carrying of a previously observed coherent total was allowed. Values stored after hour 24 were excluded. Historical scoring minimized the three components separately within each window. eICU recorded total GCS was retained. RASS values, airway-related GCS text, selected sedative intervals and ventilation records described the measurement context; no sedative record was interpreted as a measured sedation depth or actual drug effect.

All policies used their own recomputed first and attained scores and path classes. Exact candidate eligibility, matching and base/care-adjusted models were then repeated without outcome-based policy changes. A coverage-selected combined-policy sensitivity required at least five direct inputs and direct GCS in both endpoint windows. This changes the selected population and cannot isolate measurement error by itself. Full four-window direct, carried and missing input counts accompany the analysis.

S8 Clinical provenance and follow-up audit

Full-source screening was reconstructed from raw patient, admission, ICU, infection and inherited extraction records (Table S15). MIMIC-IV contains 94,458 ICU stays; the first older stay and inherited infection, service, score, prior-steroid and discharge-data gates lead to 4,069 source-selected stays and 3,689 at the landmark. The analogous eICU flow begins with 200,859 unit stays and ends at 1,395 stays from 1,334 people. These are source-specific retrospective rules and are not clinically interchangeable.

An entry-service sensitivity was evaluated separately. The MIMIC-IV inherited last-service rule searches through ICU exit plus one day and allows missing values. Applying the recorded entry service changes selection substantially. Among 2,409 candidate older stays remaining alive and in ICU at hour 24 under that entry-service route, 439 have no inherited staged eligibility score. They cannot be classified as SOFA below two. The 1,819 scored candidates meeting the score threshold are a lower-bound selected set, with 1,766 overlapping the archived risk set. eICU has 1,419 analogous score-eligible candidates, including the 1,395 archived stays. The limited outcome sensitivity is restricted to the overlapping current risk set; it does not recover all patients excluded by the previous pipeline or establish an independent common-entry validation cohort.

Pre-index MIMIC-IV disease history was identified only in completed prior admissions, with conservative ICD mappings and an explicit observation-coverage indicator. Same-admission discharge diagnoses were counted separately and not treated as baseline knowledge. eICU qualifying past-history entries required both observation and entry by hour 24. These sources measure documented history, not absence of disease when a code is not found. Frailty and baseline functional status were not validated in this extraction. Early care-limitation records were available for 1,313/1,395 eICU stays; 364 had DNR or no-CPR wording, and 170 had both full-therapy and limiting wording during the day. These are time-varying documentation, not a fixed baseline DNR variable suitable for automatic adjustment.

Index hospital and ICU timestamps were checked against raw admissions and ICU tables using the exact hospital and stay identifiers. All 3,689 matches agreed with the staged records. Hospital discharge preceded ICU exit in 608 records (median difference 1.92 h; 29 differences exceeded 24 h; none exceeded 48 h). Recorded death followed hospital discharge in 32 records, all within 24 h and all with a hospital-death flag. These are cross-system timestamp inconsistencies, not a discovered subject-join error. Taking the earlier of ICU and hospital departure as an administrative sensitivity leaves 3,682 landmark stays, with 132 classified ICU deaths and four uncertain exits by hour 72. The original definition remains identified separately.

MIMIC-IV version 3.1 truncates death dates at one year after the last discharge. A local last-admission discharge proxy places ICU day 365 within that nominal horizon for the complete main risk set, but is not a linkage-completeness flag and does not include every possible ED discharge. Date-only death intervals produce 1,079–1,081 deaths by ICU day 365, around the historical midpoint count of 1,080. The horizon 365 days after the 24-hour landmark is one day later and has six nominally uncovered records under the local proxy. The primary origin was not changed. In eICU, ICU death and hospital status are different endpoints; alive ICU departure was not assigned confirmed all-cause survival. A source-status crosswalk and all timestamp checks are provided in Source Data.

S9 Precision and independence in external prediction

All archived external linear-model predictions reproduced to maximum absolute difference below 1e-12. For the completion analysis, 1,000 hospital-cluster percentile resamples (seed 20261011) estimated absolute AUC, average precision, Brier score, CITL and joint intercept/slope intervals. The same draws were used for model differences. All 76 hospitals were represented in the source sample. These intervals condition on fitted MIMIC-IV models and omit development training uncertainty. Fixed out-of-fold internal intervals likewise do not refit training data or represent complete model-development uncertainty.

eICU contains 57 people with multiple stays and two people represented at more than one hospital. A deterministic outcome-blind sensitivity kept minimum unit-stay identifier per person, leaving 1,334 people and

59 events; identifier order was not assumed chronological. The all-stay hospital bootstrap does not fully represent the known cross-hospital dependence of the two people. The sensitivity changes the target from stays to representative persons and is labelled accordingly.

CITL fits one intercept with predicted logit fixed as an offset; joint calibration fits a free intercept and slope. They are separate diagnostics. No diagnostic coefficient was used to recalibrate test predictions. All point estimates were checked against an independent GLM calculation. Calibration slopes for subgroups with fewer than 20 events or nonevents were omitted; this rule is numerical, not a guarantee of stability above the threshold or a fairness test.

One functional-form comparison used a restricted cubic spline for attained SOFA with fixed knots 2, 6, 10 and 14 and three basis terms with linear tails. Age and recorded sex remained linear, and the second variant added the first score linearly. Standardization was fitted within the same training folds, followed by L2 logistic regression with C=1. There was no test-set hyperparameter search. Changing basis dimensionality changes effective shrinkage even at the same C. Full/fold parameters, formulas and preprocessing values are retained locally for authorized reproduction. This comparison was prespecified for completion after previous evaluation of the external cohort; it is not fresh validation.

The first-score Brier difference has the same point estimate in the historical and completion analyses (−0.00047723). The historical 1,000-resample seed 20261009 interval was −0.00108173 to −0.00002385, whereas seed 20261011 gave −0.00108608 to +0.00002054. This shift across zero is Monte Carlo sensitivity near the interval boundary, not a change of cohort or fitted model. The manuscript uses the completion interval and describes the gain as small and uncertain. Historical intervals remain labelled in Table S7 for traceability; no favorable seed was selected.

S10 Reporting scope

The prediction comparisons are an auxiliary evaluation within a retrospective measurement study. TRIPOD+AI is used to map reporting elements for this component; it does not certify clinical readiness. No prespecified treatment threshold or intervention utility supports DCA in this study. No SHAP analysis is used to infer causal effects. INSPIRE is a perioperative database with age and numerical-value transformations that require a separate feasibility and harmonization question; adding it solely as another database was not part of this completion analysis. Model discrimination, nominal associations and database count do not establish treatment benefit, fairness, or publication tier.

All scoring variants fixed in the completion plan are reported. Holm correction, added during revision rather than prespecified, is confined to the 23 revised Table 1 tests and three Table 2 omnibus tests as separate families; the archived between-database Table S29 retains its own original 10-test family. Other P values and all confidence intervals remain nominal. Raw records remain under source access controls. Author, ethics, funding, competing-interest and contribution statements require factual author input.

S11 Completion results for scoring and selection

The archived total scores, six component scores, direct-input flags and strict matching were numerically reproduced before changing policy. No-urine-carry, zero-inclusive/no-carry, coherent-GCS and combined policies produced 117, 118, 117 and 119 pairs, respectively, compared with 116 archived pairs. Their recorded 365-day death counts were 27/44, 27/45, 24/44 and 26/44. Care-adjusted ORs were 0.907 (0.552–1.493), 0.951 (0.580–1.558), 0.839 (0.508–1.386) and 0.925 (0.564–1.516). These are separate rematched samples; differences do not quantify a causal measurement effect.

The combined coverage-selected subgroup contained 155 strict candidates and 24 pairs (365-day deaths 5 versus 9). Removing the structurally absent subthreshold-steroid nuisance column restored full rank without outcome-based variable selection. Its base OR was 0.445 (0.211–0.941) and care-adjusted OR 0.619 (0.266–1.438). The fixed-pair risk difference was −16.7 percentage points (−41.7 to 8.3). The nominal

association in the base model of this small selected subgroup does not supersede the care sensitivity or constitute confirmation of an independent effect.

MIMIC-IV archived net-zero stays had 15/248 versus 6/245 ICU deaths during hours 24–72 for opposing versus unchanged components; recorded day-28 deaths were 60/248 versus 56/245 and day-365 deaths 86/248 versus 100/245. Day-365 base OR was 0.509 (0.329–0.786); care OR was 0.676 (0.423–1.080), nominal P=0.102. The short-term multivariable models failed the prespecified support rule and remained descriptive. The six eICU ICU deaths among net-zero stays likewise supported descriptive reporting only. All policies and estimability decisions remain in Source Data.

MIMIC-IV same-time complete GCS was available in 3,335 first and 3,116 attained windows; independent-minimum GCS was directly available in 3,336 and 3,117. Their direct totals differed in 59 and 19 windows. No conflicting duplicate component values at the same timestamp were found. Chosen triplets contained airway-marked text in 1,990 first and 1,139 attained windows. Nearby RASS ≤−3 occurred in 1,175 first and 374 attained windows. RASS could be recorded after the selected GCS within 60 minutes and before hour 24; it was not used as a pre-GCS prediction input.

The inherited minute-boundary convention differs from exact elapsed-second binning for 162 urine and 232 GCS/RASS records near boundaries. It was retained across policies, while raw event and hour-24 storage gates used exact timestamps. No negative urine entries occurred in these selected caches. Zero-only observed windows numbered 102 in MIMIC-IV and 68 in eICU. eICU positive entries reached 21,090 mL; source labels did not certify interval completeness or rule out cumulative/duplicate recording. The zero-inclusive analysis therefore remains a recording sensitivity rather than adjudicated renal physiology. Free-text custom source labels were excluded from the aggregate manuscript export.

The entry-service-retained subset within the archived MIMIC-IV cohort contained 1,766 stays and 322 strict candidates. Using the same base model, recorded 365-day OR was 1.053 (0.594–1.866), nominal P=0.860; adding care yielded 1.466 (0.767–2.804), P=0.247. Rematching produced 42 pairs: recorded day-28 deaths 9 versus 13, risk difference −9.5 points (−26.2 to 7.1); day-365 deaths 16 versus 18, difference −4.8 points (−26.2 to 16.7). Exact McNemar P values were 0.4240 and 0.8318. This extension was triggered during source-audit review and is exploratory. The wide intervals, changes in case mix and inherited conditioning prohibit either confirmation of benefit or a claim of no association. It does not substitute for a newly extracted full common-entry cohort.

S12. Exact pairwise score comparisons for Figure 2a

Supplementary Table S28. Nominal pairwise comparisons of local six-hour SOFA score distributions by time window in the MIMIC-IV hour-24 landmark cohort.

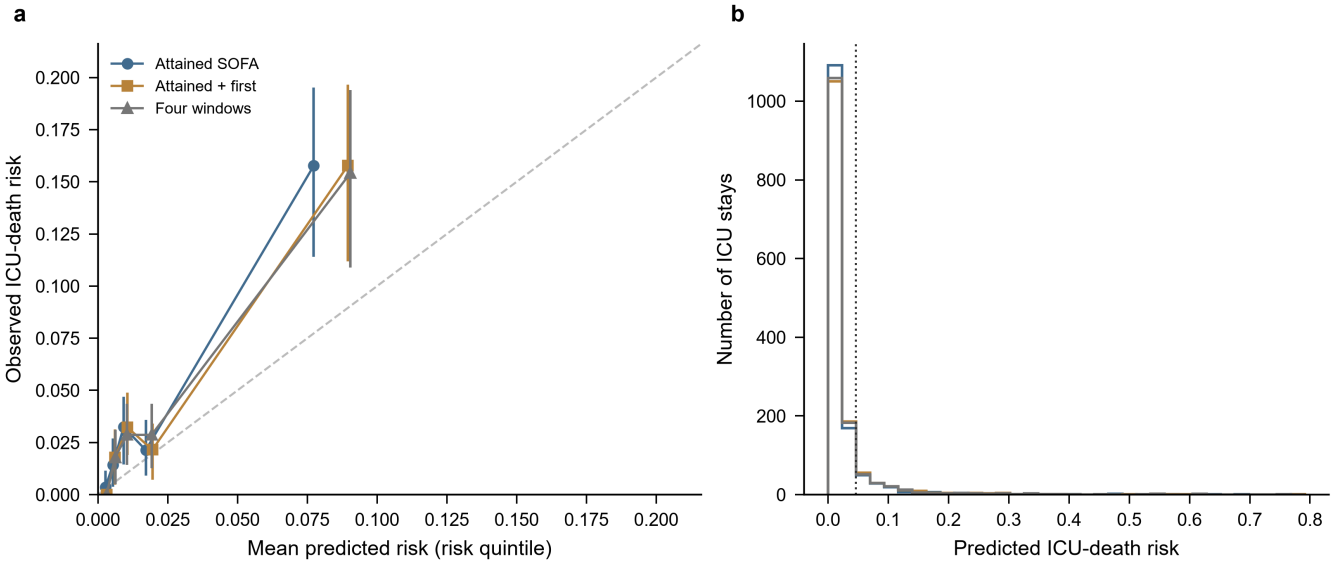
Time window	Decrease vs net stable P	Decrease vs increase P	Net stable vs increase P
0–6 h	5.41e-131	4.09e-178	4.55e-26
6–12 h	2.61e-09	3.30e-06	4.90e-14
12–18 h	0.0270	1.92e-42	1.23e-22
18–24 h	2.15e-10	1.66e-102	3.22e-42

The decrease, net-stable and increase groups comprise 1,600, 1,282 and 807 stays, respectively. Each P value is from a two-sided asymptotic Mann-Whitney U test with correction for tied ranks and continuity. The four time windows were tested separately; repeated windows from the same stay were not treated as independent observations within a pooled test. P values were not adjusted for covariates or multiple comparisons. These are conditional descriptive comparisons: the first and last scores also define the change groups. Full-precision test statistics and P values are in Figure_2_trajectory_pairwise_tests.csv; medians and interquartile ranges are in Figure_2_trajectory_quantiles.csv.

Additional figure legend

Figure S5 Locked external calibration and risk distribution

a, Mean predicted and observed risks in five groups formed by fixed predicted-risk quantiles. Hospital-bootstrap 95% intervals describe observed risk; lines guide the eye and are not fitted recalibration curves. **b,** Predicted-risk distributions for the three linear models, with overall observed event frequency marked at 4.59%. The sample is 1,395 eICU stays from 76 hospitals, with 64 ICU deaths during hours 24–72. The ideal calibration diagonal and risk distribution show that useful ranking can coexist with underestimation of average risk. No hypothesis test is displayed.



S13 Revised descriptive and balance tables

The descriptive-table revision retained the source cohort, outcome labels, fitted mortality models and matching. Figures S1 and S2 were subsequently redrawn using the same data. Main Table 1 now compares the three MIMIC-IV score-change groups. It includes 23 omnibus tests, with Holm adjustment within this family. All categorical tests met the expected-count criterion for Pearson chi-square; the specified sparse-table fallback was not needed. Table 2 retains its three-test Holm family. The archived between-database comparison remains Table S29 and retains its original 10-test adjustment. Main-table hypothesis counts must not be combined across versions.

History entries count positive diagnosis codes from hospital admissions discharged before the index admission; they are documentation frequencies rather than adjudicated disease prevalence. The source contains 1,033 stays with observed prior-admission diagnosis records. Percentages among all stays and among those with this coverage are supplied separately in source_data/revision_20261012/prior_history_denominators.csv. Zero-valued documentation flags indicate that no qualifying positive code was found; they do not prove absence of disease. No frailty score was constructed from these limited records.

Table S30 supplies all six organ-component summaries by database, time window and score-change group as a machine-readable table. Table S31 supplies all pre-match and post-match covariate summaries and standardized differences. Its standardization denominator is fixed using the full 600-candidate sample, with pooled within-group variance and the corresponding Bernoulli formula for binary indicators. The selected post-match rows below illustrate the remaining care imbalance. A difference of zero in exact-matching variables reflects the design; it does not imply exchangeability.

Characteristic	Decrease 3	Net zero	SMD
Age, years	76.0 [72.0, 82.0]	76.0 [71.0, 83.0]	-0.068
Male sex	65 (56.0)	65 (56.0)	0.000
Attained total SOFA	5.0 [4.0, 7.0]	5.0 [4.0, 7.0]	0.000
Sedative record at 0–6 h	90 (77.6)	48 (41.4)	0.848
Invasive ventilation at 0–6 h	90 (77.6)	48 (41.4)	0.841
Direct inputs in first window, /8	7.0 [6.0, 7.0]	7.0 [6.0, 7.0]	0.000

No new mortality regression or external prediction model was fitted for these tables. The revised narrative retains both the observed mortality differences and the attenuation with early-care adjustment.

S14 Rule crosswalk and criterion-validation status

Table S32 records the distinction between local-score reproducibility and clinical validity. The current analysis reproduces the archived algorithm and compares alternative recording policies. It has not established an independently adjudicated clinical reference standard.

Time windows. Four nonoverlapping six-hour windows; first versus attained score. Not a standard 24-hour daily SOFA; endpoint difference is determined by first score at fixed attained score.

Respiratory. Minimum PaO2 divided by available FiO2; invasive-ventilation flag activates scores 3–4. PaO2 and FiO2 may not be simultaneous; retrospective support intervals remain.

Cardiovascular. Local dopamine boundary and vasopressin/phenylephrine rules are retained in score_sofa(). Does not establish equivalence to original SOFA vasopressor criteria; dose units and interval duration require reference review.

Renal. Creatinine plus six-hour urine cutoffs of 50 and 125 mL; primary policy excludes zero entries and carries prior positive totals. Alternative policies retain zero and remove carry; recording completeness and cumulative entries remain unresolved.

Neurological. MIMIC-IV uses independently minimized GCS components; eICU uses recorded total. Same-time GCS is a sensitivity in MIMIC-IV. Sedation and airway-related verbal entries are not independently adjudicated. Same-time agreement alone is not clinical validation.

Liver and coagulation. Bilirubin maximum and platelet minimum within each local window, with the documented threshold code. Units, timing and carried values need to be distinguished from direct observations.

Missingness and entry. Forward carry; missing values that trigger no condition receive zero; point-entry gates use source-specific fields. Zero score is not proof of normal physiology; storage/revision time is not verified first bedside availability.

A separate outcome-blinded pilot review packet has been prepared locally. No clinician reference scores have yet been entered and no agreement coefficient or accuracy claim can be reported. The packet and a proposed independent-cohort protocol are research preparation documents, not completed validation results.

Table S29 Archived between-database characteristics

Characteristic	MIMIC-IV (n=3,689)	eICU (n=1,395)	P	Holm-adjusted P
Age, years	75.0 [70.0, 82.0]	77.0 [70.0, 84.0]	8.54×10^{-5}	2.56×10^{-4}
Male sex, n (%)	2,196 (59.5)	759 (54.4)	9.61×10^{-4}	9.61×10^{-4}
First-window score, 0–6 h	8.0 [4.0, 10.0]	3.0 [1.0, 6.0]	6.97×10^{-139}	6.28×10^{-138}
Attained score, 18–24 h	5.0 [3.0, 8.0]	4.0 [3.0, 7.0]	1.06×10^{-23}	5.32×10^{-23}
Direct inputs in first window, /8	7.0 [6.0, 7.0]	4.0 [3.0, 6.0]	1.21×10^{-251}	1.21×10^{-250}
Direct inputs in attained window, /8	4.0 [3.0, 5.0]	3.0 [2.0, 4.0]	1.74×10^{-119}	1.39×10^{-118}
Direct GCS in first window, n (%)	3,336 (90.4)	1,073 (76.9)	8.63×10^{-37}	5.18×10^{-36}
Direct GCS in attained window, n (%)	3,117 (84.5)	799 (57.3)	3.69×10^{-94}	2.58×10^{-93}
Threshold steroid category, n (%)	332 (9.0)	172 (12.3)	3.92×10^{-4}	7.85×10^{-4}
Subthreshold steroid category, n (%)	77 (2.1)	109 (7.8)	2.90×10^{-22}	1.16×10^{-21}

Values are median [IQR] or n (%). P is the unadjusted between-database P value. Holm-adjusted P controls multiplicity across the 10 comparisons in this table as one family. This adjustment was added during the exploratory revision, not prespecified. Age, scores and input counts use two-sided asymptotic Mann–Whitney U tests with tie and continuity corrections; binary rows use Pearson chi-square tests without Yates correction. Each steroid-category row compares that category against all other categories. No covariate or hospital/person-cluster adjustment is applied; tests treat stays as independent and describe these archived samples. Source eligibility and measurement differences limit generalization. Direct observations precede carrying. Timestamp fields are database-specific proxies, not verified bedside availability. Steroid categories describe orders, not administered dose.

Supplementary manuscript tables

Fourteen supplementary tables report matching, recording-rule audits, regression adjustment and prediction evaluation. Tables S13 and S14 were moved from the main text. Numerical results, confidence intervals and nominal P values are unchanged.

Table S1 MIMIC-IV characteristics by restricted-policy SOFA change

Characteristic	Decrease ≥2 n = 1,600	Net stable n = 1,282	Increase ≥2 n = 807
Age, years	74.0 [69.0, 80.0]	77.0 [71.0, 83.0]	76.0 [70.0, 83.0]
Male sex, n (%)	988 (61.8)	729 (56.9)	479 (59.4)
First-window score, 0–6 h	9.0 [8.0, 11.0]	5.0 [3.0, 9.0]	3.0 [1.0, 7.0]
Attained score, 18–24 h	4.0 [3.0, 6.0]	5.0 [3.0, 9.0]	8.0 [5.0, 11.0]
Direct inputs in first window, /8	7.0 [7.0, 7.0]	7.0 [5.0, 7.0]	6.0 [4.0, 7.0]
Direct inputs in attained window, /8	3.0 [3.0, 5.0]	4.0 [3.0, 5.0]	5.0 [3.0, 6.0]
Direct GCS in first window, n (%)	1,549 (96.8)	1,180 (92.0)	607 (75.2)
Direct GCS in attained window, n (%)	1,359 (84.9)	1,075 (83.9)	683 (84.6)
Threshold steroid category, n (%)	84 (5.2)	152 (11.9)	96 (11.9)
Subthreshold steroid category, n (%)	16 (1.0)	37 (2.9)	24 (3.0)

Values are median [IQR] or n (%). Direct inputs are observed values before carrying forward (maximum eight); available timestamp fields are database-specific proxies and do not establish exact bedside availability. GCS, Glasgow Coma Scale. Net stability denotes total change from −1 to +1.

Steroid categories use the inherited order extraction for −10 to +24 hours: no relevant order, subthreshold (<160 mg/day hydrocortisone equivalent), or threshold (≥160 mg/day). These are record-derived categories and do not establish administered doses.

Scores are locally adapted six-hour SOFA, not an unmodified standard daily SOFA or SOFA-2. ΔSOFA is the 18–24-hour score minus the 0–6-hour score. Physiological storage records are restricted to hour 24 in MIMIC-IV; original support intervals are retained.

Table S2 Flow from source selection to the 24-hour landmark

Flow step	MIMIC-IV	eICU
Source-selected stays	4,069	1,647
Excluded: ICU exit by hour 24	365	252
Excluded: death by hour 24 with later ICU exit	15	0
Landmark risk set	3,689	1,395
Recorded ICU deaths during hours 24–72	134	64

Counts refer to source-selected stays, not all database admissions. Exclusion rows are mutually exclusive. The risk set includes patients aged ≥65 years alive and in ICU at hour 24; all three steroid-record categories are retained. Source selection includes retrospective restrictions; the two database eligibility algorithms are not fully harmonized.

Table S3 Broad path associations conditional on attained SOFA

Recorded outcome	Contrast	n / events	OR (95% CI)	Nominal P
365-day death	Decrease ≥2 vs net stable	3,344 / 912	0.43 (0.35, 0.53)	4.37×10^{-16}
365-day death	Increase ≥2 vs net stable	3,344 / 912	0.83 (0.67, 1.04)	0.1008
ICU death 24–72 h	Decrease ≥2 vs net stable	3,344 / 72	0.49 (0.24, 0.99)	0.0460
ICU death 24–72 h	Increase ≥2 vs net stable	3,344 / 72	1.33 (0.77, 2.29)	0.3035

Restricted physiological-storage policy, all three steroid categories. Models condition on exact attained SOFA integers 2–13 and adjust for age, recorded sex and steroid category. Net stable denotes ΔSOFA from −1 to +1. The 3,344-person supported-score population differs from the full landmark risk set and the strict-path samples. Intervals are HC3 95% Wald CIs; P values are nominal.

Table S4 Strict-path eligibility and matched comparisons across policies

A Eligibility and pair construction

Measure	Original	Physiology gate	Support sensitivity
Full landmark risk set	3,689	3,689	3,689
Strict candidates	598	600	601
Candidates with ΔSOFA = −3	233	235	240
Candidates with ΔSOFA = 0	365	365	361
Subthreshold steroid candidates	13	12	12
Minimum count in attained-score by path cells	15	14	12
Matched pairs	122	116	121
New subthreshold steroid pairs	0	0	0
Exact pair overlap with the 116-pair reference	63	116	78
Pairs meeting the direct-GCS sensitivity restriction	41	36	44

Strict paths are ΔSOFA = −3 versus 0 at attained scores 2–9. Matching is exact on attained score, recorded sex, steroid category and first-window input count, with age gap ≤5 years and a cost for age and attained-window input-count gaps. All stays meet the 24-hour landmark.

Original: original chart-time score policy. Physiology gate: storage timestamps restricted by hour 24 for physiological inputs, retaining original support intervals. Support sensitivity: the physiology gate plus support information restricted or corroborated by contemporaneous documentation. These are sensitivity policies, not bedside gold standards.

The pair-overlap reference is the current 116-pair physiological-storage sample. Count labels retain the inherited audit definitions; the direct-GCS restriction is not a guarantee of unsedated assessment.

Table S4 continued Matched outcomes and early care records

B Matched comparisons

Measure	Original	Physiology gate	Support sensitivity
Matched pairs	122	116	121
365-day deaths, ΔSOFA = −3	25	26	29
365-day deaths, ΔSOFA = 0	45	45	44
Risk difference, percentage points (95% CI)	−16.4 (−27.0, −5.7)	−16.4 (−27.6, −5.2)	−12.4 (−23.1, −1.7)
Discordant-pair OR (95% CI)	0.41 (0.20, 0.79)	0.44 (0.22, 0.83)	0.50 (0.25, 0.96)
Nominal exact paired P	0.0055	0.0094	0.0357
Early sedative record, ΔSOFA = −3	95	90	95
Early sedative record, ΔSOFA = 0	46	48	54
Early invasive ventilation, ΔSOFA = −3	97	90	92
Early invasive ventilation, ΔSOFA = 0	47	48	55

Risk difference is the ΔSOFA = −3 group minus the ΔSOFA = 0 group. Risk-difference CIs use 5,000 pair-bootstrap resamples. Paired ORs compare discordant outcomes; nominal P values use the exact binomial discordant-pair test. Early care records refer to the first six-hour window. None of these comparisons establishes a causal benefit of score reduction.

Original: original chart-time score policy. Physiology gate: storage timestamps restricted by hour 24 for physiological inputs, retaining original support intervals. Support sensitivity: the physiology gate plus support information restricted or corroborated by contemporaneous documentation. These are sensitivity policies, not bedside gold standards.

Table S5 Matched mortality at 28 90 and 365 days

A Original

Day	Pairs	Deaths Δ = −3	Deaths Δ = 0	RD pp	Discordant −3 / 0	Nominal P
28	122	15	30	−12.3	11 / 26	0.0201
90	122	19	39	−16.4	10 / 30	0.0022
365	122	25	45	−16.4	14 / 34	0.0055

B Physiology gate

Day	Pairs	Deaths Δ = −3	Deaths Δ = 0	RD pp	Discordant −3 / 0	Nominal P
28	116	16	33	−14.7	11 / 28	0.0095
90	116	22	41	−16.4	13 / 32	0.0066
365	116	26	45	−16.4	15 / 34	0.0094

C Support sensitivity

Day	Pairs	Deaths Δ = −3	Deaths Δ = 0	RD pp	Discordant −3 / 0	Nominal P
28	121	15	26	−9.1	9 / 20	0.0614
90	121	23	37	−11.6	12 / 26	0.0336
365	121	29	44	−12.4	15 / 30	0.0357

Δ denotes total ΔSOFA. RD, risk difference in percentage points (Δ = −3 minus Δ = 0). Discordant counts show pairs with death only in the indicated member. P values use the exact binomial discordant-pair test. The same matched sample within each policy is followed cumulatively from ICU admission; all participants survived to the 24-hour landmark.

Original: original chart-time score policy. Physiology gate: storage timestamps restricted by hour 24 for physiological inputs, retaining original support intervals. Support sensitivity: the physiology gate plus support information restricted or corroborated by contemporaneous documentation. These are sensitivity policies, not bedside gold standards.

Table S6 Additional deaths after the day 28 landmark

Policy	Surviving pairs	New deaths Δ = −3	New deaths Δ = 0	RD pp	Nominal P
Original	81	7	11	−4.9	0.4807
Physiology gate	72	6	10	−5.6	0.4545
Support sensitivity	86	10	16	−7.0	0.2863

Only pairs whose two members were both alive at day 28 are retained. New recorded deaths occur after day 28 and through day 365. This conditional analysis changes the target population. Δ denotes total ΔSOFA; RD is the difference in percentage points (Δ = −3 minus Δ = 0). P values use the exact binomial discordant-pair test.

Original: original chart-time score policy. Physiology gate: storage timestamps restricted by hour 24 for physiological inputs, retaining original support intervals. Support sensitivity: the physiology gate plus support information restricted or corroborated by contemporaneous documentation. These are sensitivity policies, not bedside gold standards.

Table S7 External prediction differences with hospital bootstrap intervals

A Original

Model comparison	ΔAUC (95% CI)	ΔBrier × 1,000 (95% CI)
First added vs attained	−0.0072 (−0.0187, 0.0017)	−0.496 (−1.106, −0.030)
All prior added vs first added	−0.0034 (−0.0120, 0.0074)	0.173 (−0.066, 0.446)
All prior added vs attained	−0.0106 (−0.0231, 0.0032)	−0.323 (−1.014, 0.173)

B Point windows aligned

Model comparison	ΔAUC (95% CI)	ΔBrier × 1,000 (95% CI)
First added vs attained	−0.0071 (−0.0186, 0.0018)	−0.525 (−1.142, −0.048)
All prior added vs first added	−0.0036 (−0.0121, 0.0072)	0.188 (−0.046, 0.460)
All prior added vs attained	−0.0107 (−0.0234, 0.0032)	−0.338 (−1.025, 0.169)

C Available timestamp gates

Model comparison	ΔAUC (95% CI)	ΔBrier × 1,000 (95% CI)
First added vs attained	−0.0062 (−0.0179, 0.0032)	−0.477 (−1.082, −0.024)
All prior added vs first added	−0.0025 (−0.0103, 0.0072)	0.189 (−0.030, 0.454)
All prior added vs attained	−0.0087 (−0.0214, 0.0049)	−0.288 (−0.941, 0.181)

Every comparison uses 76 eICU hospital clusters and 1,000 valid paired bootstrap resamples, with shared hospital draws across models. Differences are new minus comparator. Higher ΔAUC and lower ΔBrier favour the new model. The CIs are percentile intervals conditional on fitted MIMIC-IV models. All model comparisons and both performance measures are retained.

Outcome is recorded ICU death during hours 24–72. Every model includes age and recorded sex. Attained SOFA uses the 18–24-hour score; + first-window SOFA adds 0–6 hours; + all prior SOFA windows adds 0–6, 6–12 and 12–18 hours. MIMIC-IV uses five-fold out-of-fold predictions (3,689 stays, 134 events); eICU uses locked MIMIC-IV models (1,395 stays, 64 events, 76 hospitals), without refitting or recalibration.

Table S8 All prediction metrics by recording policy

A Original

Evaluation	Model	AUC	Average precision	Brier
MIMIC-IV	Attained SOFA	0.8416	0.3305	0.02877
MIMIC-IV	+ First-window SOFA	0.8487	0.3492	0.02850
MIMIC-IV	+ All prior SOFA windows	0.8487	0.3612	0.02832
eICU	Attained SOFA	0.8216	0.2926	0.03936
eICU	+ First-window SOFA	0.8144	0.2919	0.03886
eICU	+ All prior SOFA windows	0.8109	0.2820	0.03903

Risks and calibration

Evaluation	Model	Predicted / observed (%)	Calibration intercept	Calibration slope
MIMIC-IV	Attained SOFA	3.62 / 3.63	−0.0038	0.9972
MIMIC-IV	+ First-window SOFA	3.63 / 3.63	0.0029	1.0002
MIMIC-IV	+ All prior SOFA windows	3.63 / 3.63	−0.0060	0.9968
eICU	Attained SOFA	2.26 / 4.59	0.8419	1.0016
eICU	+ First-window SOFA	2.61 / 4.59	0.5537	0.9598
eICU	+ All prior SOFA windows	2.63 / 4.59	0.4709	0.9364

Outcome is recorded ICU death during hours 24–72. Every model includes age and recorded sex. Attained SOFA uses the 18–24-hour score; + first-window SOFA adds 0–6 hours; + all prior SOFA windows adds 0–6, 6–12 and 12–18 hours. MIMIC-IV uses five-fold out-of-fold predictions (3,689 stays, 134 events); eICU uses locked MIMIC-IV models (1,395 stays, 64 events, 76 hospitals), without refitting or recalibration.

Calibration intercept and slope were estimated jointly by regressing outcome on an intercept and logit predicted risk; the intercept is not calibration-in-the-large with slope fixed to one. These diagnostics were not used to recalibrate the evaluated predictions. Recording restrictions approximate clinical availability; see Supplementary Methods S3.

Table S8 continued All prediction metrics by recording policy

B Point windows aligned

Evaluation	Model	AUC	Average precision	Brier
MIMIC-IV	Attained SOFA	0.8416	0.3305	0.02877
MIMIC-IV	+ First-window SOFA	0.8487	0.3492	0.02850
MIMIC-IV	+ All prior SOFA windows	0.8487	0.3612	0.02832
eICU	Attained SOFA	0.8215	0.2923	0.03936
eICU	+ First-window SOFA	0.8144	0.2921	0.03883
eICU	+ All prior SOFA windows	0.8108	0.2820	0.03902

Risks and calibration

Evaluation	Model	Predicted / observed (%)	Calibration intercept	Calibration slope
MIMIC-IV	Attained SOFA	3.62 / 3.63	−0.0038	0.9972
MIMIC-IV	+ First-window SOFA	3.63 / 3.63	0.0029	1.0002
MIMIC-IV	+ All prior SOFA windows	3.63 / 3.63	−0.0060	0.9968
eICU	Attained SOFA	2.27 / 4.59	0.8376	1.0006
eICU	+ First-window SOFA	2.62 / 4.59	0.5499	0.9596
eICU	+ All prior SOFA windows	2.64 / 4.59	0.4652	0.9354

Outcome is recorded ICU death during hours 24–72. Every model includes age and recorded sex. Attained SOFA uses the 18–24-hour score; + first-window SOFA adds 0–6 hours; + all prior SOFA windows adds 0–6, 6–12 and 12–18 hours. MIMIC-IV uses five-fold out-of-fold predictions (3,689 stays, 134 events); eICU uses locked MIMIC-IV models (1,395 stays, 64 events, 76 hospitals), without refitting or recalibration.

Calibration intercept and slope were estimated jointly by regressing outcome on an intercept and logit predicted risk; the intercept is not calibration-in-the-large with slope fixed to one. These diagnostics were not used to recalibrate the evaluated predictions. Recording restrictions approximate clinical availability; see Supplementary Methods S3.

Table S8 continued All prediction metrics by recording policy

C Available timestamp gates

Evaluation	Model	AUC	Average precision	Brier
MIMIC-IV	Attained SOFA	0.8393	0.3138	0.02910
MIMIC-IV	+ First-window SOFA	0.8457	0.3327	0.02880
MIMIC-IV	+ All prior SOFA windows	0.8455	0.3437	0.02861
eICU	Attained SOFA	0.8254	0.2922	0.03947
eICU	+ First-window SOFA	0.8192	0.2882	0.03900
eICU	+ All prior SOFA windows	0.8167	0.2822	0.03919

Risks and calibration

Evaluation	Model	Predicted / observed (%)	Calibration intercept	Calibration slope
MIMIC-IV	Attained SOFA	3.62 / 3.63	−0.0020	0.9980
MIMIC-IV	+ First-window SOFA	3.62 / 3.63	0.0049	1.0005
MIMIC-IV	+ All prior SOFA windows	3.63 / 3.63	−0.0046	0.9972
eICU	Attained SOFA	2.24 / 4.59	0.8933	1.0130
eICU	+ First-window SOFA	2.58 / 4.59	0.5928	0.9676
eICU	+ All prior SOFA windows	2.59 / 4.59	0.5147	0.9434

Outcome is recorded ICU death during hours 24–72. Every model includes age and recorded sex. Attained SOFA uses the 18–24-hour score; + first-window SOFA adds 0–6 hours; + all prior SOFA windows adds 0–6, 6–12 and 12–18 hours. MIMIC-IV uses five-fold out-of-fold predictions (3,689 stays, 134 events); eICU uses locked MIMIC-IV models (1,395 stays, 64 events, 76 hospitals), without refitting or recalibration.

Calibration intercept and slope were estimated jointly by regressing outcome on an intercept and logit predicted risk; the intercept is not calibration-in-the-large with slope fixed to one. These diagnostics were not used to recalibrate the evaluated predictions. Recording restrictions approximate clinical availability; see Supplementary Methods S3.

Table S9 eICU score and component sensitivity across timestamp policies

Measure	Original	Half-open windows	Entry windows	24 h gates	Gates vs entry windows
Landmark stays	1,395	1,395	1,395	1,395	1,395
First-window SOFA changed	0	7	18	24	6
Attained SOFA changed	0	3	3	66	63
ΔSOFA changed	0	10	21	82	62
Broad path class changed	0	4	7	37	30
Total ΔSOFA = 0	319	317	317	323	323
Opposing component changes	109	107	107	107	107
Original opposing cases retained	109	106	106	103	104
Original opposing cases lost	0	3	3	6	3
New opposing cases	0	1	1	4	3

All entries are counts. Policy order follows the source: legacy; observation_halfopen; point_entry_window; entry_revision_gate24; entry_gate_vs_point_window. The final column is a comparison of the gated policy with the entry-window policy, rather than a fifth independent cohort. Other change counts use the original policy as comparator. Opposing changes denote at least one component increasing and another decreasing among zero-total-change stays.

Half-open policies align point-observation boundaries; entry windows additionally align infusion entry records; 24-hour gates additionally restrict available entry/revision timestamps. Laboratory revision after hour 24 does not prove the earlier result was unknown to clinicians.

Table S10 eICU input record timestamps

Source	Input	Rows	Stays	Late or missing rows	Stays with any late or missing row	Missing entry rows
Lab	Creatinine	2,368	1,349	41	39	0
Lab	FiO_2	1,261	580	46	35	0
Lab	PaO_2	1,587	714	8	8	0
Lab	Platelets	1,907	1,328	20	20	0
Lab	Total bilirubin	966	798	14	14	0
Nursing	Arterial-line MAP	295	9	0	0	0
Nursing	GCS eyes	5,938	1,253	113	94	0
Nursing	GCS total	6,430	1,276	118	100	0
Nursing	GCS motor	5,934	1,253	113	94	0
Nursing	GCS verbal	5,856	1,237	113	94	0
Nursing	MAP	90	3	0	0	0
Resp.	Respiratory entries	7,535	876	47	29	0
Urine	Urine	8,517	1,066	469	365	0

Counts describe the extracted input records and source-specific entry/revision fields. Rows are not independent patients; a stay can contribute to multiple inputs. Late means after the 24-hour landmark under the implemented field-specific gate. Missing entry offsets are counted separately. GCS, Glasgow Coma Scale; MAP, mean arterial pressure; FiO_2, inspired oxygen fraction; PaO_2, arterial oxygen partial pressure. These audit fields do not provide a uniform bedside availability time across record types.

Table S11 MIMIC-IV score and component sensitivity to storage restrictions

Measure	All steroid categories	Original two steroid categories
Landmark stays	3,689	3,612
First-window total changed	1	1
Attained total changed	358	348
ΔSOFA changed	357	347
Zero total change before restriction	488	476
Zero total change after restriction	493	481
Opposing changes before restriction	238	234
Opposing changes after restriction	248	243
Original opposing cases retained	210	206
Original opposing cases lost	28	28
New opposing cases	38	37

All entries are counts. The all-category cohort includes the subthreshold steroid-record group; the original two-category cohort omits it. The restriction uses physiological storage timestamps by hour 24 and retains original support intervals. Opposing changes require a zero total change with at least one component increasing and another decreasing. These are scoring audit counts, not adjusted outcome estimates.

Table S12 Recording and component source values

Measure	MIMIC-IV	eICU
Full landmark denominator	3,689	1,395
Attained SOFA changed	358	66
Broad path class changed	114	37
Restricted-policy zero total change	493	323
Opposing changes within zero-total-change stays	248	107

Figure S1c divides changed-score and changed-class counts by the full landmark denominator. Figure 4a divides opposing-change counts by the restricted-policy zero-total-change denominator (493 and 323). Thus, the percentages in the two panels use different populations. Counts are descriptive; no hypothesis tests or confidence intervals are estimated from this table.

Table S13 Strict-path associations before and after early care adjustment

Recorded outcome	Adjustment	n / events	OR (95% CI)	Nominal P
28-day death	Base	600 / 102	0.42 (0.24, 0.74)	0.0028
28-day death	Base + early care	600 / 102	0.66 (0.36, 1.23)	0.1924
365-day death	Base	600 / 181	0.49 (0.32, 0.76)	0.0015
365-day death	Base + early care	600 / 181	0.86 (0.52, 1.42)	0.5641

Comparison: total SOFA decrease of exactly 3 points versus net change of 0 in all 600 strict candidates. This model population differs from the 116 matched pairs shown in Figure 3b.

Base adjustment includes attained SOFA integer fixed effects (2–9), age, recorded sex, three steroid-record categories and first-window direct-input count. Early care adds first-window selected sedative and invasive ventilation records. ORs are prognostic associations; intervals are HC3 95% Wald CIs. P values are nominal and unadjusted for multiplicity.

Scores are locally adapted six-hour SOFA, not an unmodified standard daily SOFA or SOFA-2. ΔSOFA is the 18–24-hour score minus the 0–6-hour score. Physiological storage records are restricted to hour 24 in MIMIC-IV; original support intervals are retained.

Table S14 Prediction performance after available timestamp restrictions

A MIMIC-IV internal evaluation

Model	AUC	Average precision	Brier score	Predicted / observed risk (%)
Attained SOFA	0.8393	0.3138	0.02910	3.62 / 3.63
+ First-window SOFA	0.8457	0.3327	0.02880	3.62 / 3.63
+ All prior SOFA windows	0.8455	0.3437	0.02861	3.63 / 3.63

B eICU external evaluation

Model	AUC	Average precision	Brier score	Predicted / observed risk (%)
Attained SOFA	0.8254	0.2922	0.03947	2.24 / 4.59
+ First-window SOFA	0.8192	0.2882	0.03900	2.58 / 4.59
+ All prior SOFA windows	0.8167	0.2822	0.03919	2.59 / 4.59

Outcome is recorded ICU death during hours 24–72. Every model includes age and recorded sex. Attained SOFA uses the 18–24-hour score; + first-window SOFA adds 0–6 hours; + all prior SOFA windows adds 0–6, 6–12 and 12–18 hours. MIMIC-IV uses five-fold out-of-fold predictions (3,689 stays, 134 events); eICU uses locked MIMIC-IV models (1,395 stays, 64 events, 76 hospitals), without refitting or recalibration.

AUC, area under the receiver operating characteristic curve. Lower Brier scores indicate lower mean squared probability error. Mean predicted and observed risks are percentages. Calibration coefficients and other recording policies appear in Table S8; paired hospital-bootstrap differences appear in Table S7. Clinical availability is only approximated by the available timestamps.

Supplementary tables S15–S28

Table S15 Full-source screening under inherited definitions

Database	Selection step	Stays	People
MIMIC-IV	All ICU stays	94,458	65,366
MIMIC-IV	First ICU stay and age ≥65	35,506	35,506
MIMIC-IV	Antibiotic and culture within ±24 h	16,805	16,805
MIMIC-IV	No pre-ICU evidence and evidence after entry	5,438	5,438
MIMIC-IV	Eligible last service through ICU exit +1 day	4,561	4,561
MIMIC-IV	ICU stay at least 6 h	4,535	4,535
MIMIC-IV	Local first-day eligibility SOFA ≥2	4,218	4,218
MIMIC-IV	No extracted steroid order before −10 h	4,078	4,078
MIMIC-IV	Required hospital discharge fields recorded	4,069	4,069
MIMIC-IV	Alive and in ICU at hour 24	3,689	3,689
eICU	All ICU stays	200,859	139,367
eICU	First unit visit per encounter and age ≥65	80,641	68,798
eICU	Antibiotic and culture during hours 0–24	1,839	1,757
eICU	No combined pre-ICU antibiotic/culture evidence	1,811	1,731
eICU	Entry unit not CSICU or CTICU	1,796	1,716
eICU	Local first-day eligibility SOFA ≥2	1,676	1,600
eICU	No extracted steroid order before −10 h	1,647	1,574
eICU	Hospital discharge time recorded	1,647	1,574
eICU	Alive and in ICU at hour 24	1,395	1,334

Steps are sequential within each database after first-elderly-stay selection. MIMIC-IV service and discharge gates use future information; eICU first unit per encounter does not imply one record per person. This reconstructs inherited selection, not a common prospective sepsis inception cohort.

Table S16 Documented pre-index clinical history

Characteristic	MIMIC-IV	eICU
Prior-history coding coverage	1,033/3,689 (28.0%)	1,382/1,395 (99.1%)
Heart failure	322/3,689 (8.7%)	346/1,395 (24.8%)
Chronic kidney disease	279/3,689 (7.6%)	Not comparable
Chronic lung disease	223/3,689 (6.0%)	406/1,395 (29.1%)
Cancer	191/3,689 (5.2%)	345/1,395 (24.7%)
Cerebrovascular disease	190/3,689 (5.2%)	231/1,395 (16.6%)
Dementia	35/3,689 (0.9%)	184/1,395 (13.2%)
Diabetes	385/3,689 (10.4%)	540/1,395 (38.7%)

Denominators are all 3,689 and 1,395 landmark stays. MIMIC-IV history is identified in prior completed hospital admissions; eICU history uses qualifying past-history entries by hour 24. No evidence of a condition does not establish its absence. Positive documentation and missing interfaces cannot be harmonized as true disease prevalence. Same-admission discharge diagnoses are exported separately and are not baseline history.

Table S17 Mortality horizons and date-only uncertainty

Origin / horizon	Nominal coverage	Historical deaths	Certain–possible deaths	Ambiguous date
ICU admission / 28 d	3689/3689	645	644–645	1
ICU admission / 90 d	3689/3689	820	820–821	1
ICU admission / 365 d	3689/3689	1080	1079–1081	2
24h landmark / 28 d	3689/3689	646	645–647	2
24h landmark / 90 d	3689/3689	822	821–822	1
24h landmark / 365 d	3683/3689	1081	1081–1081	0

MIMIC-IV only. Historical date-only deaths use the established midday convention. Nominal coverage uses local last-admission discharge as a proxy for the documented one-year horizon, and does not verify individual death linkage. The study endpoint is ICU day 365, not 365 days after the hour-24 landmark.

Table S18 Score and category sensitivity

Database / policy	n	Changed first / attained	Changed class	Net zero / opposing
MIMIC-IV / Archived gated score	3689	0 / 0	0	493 / 248
MIMIC-IV / No urine carrying	3689	0 / 55	20	492 / 245
MIMIC-IV / Zero-inclusive urine; no carry	3689	25 / 49	28	490 / 246
MIMIC-IV / Coherent GCS	3689	33 / 10	14	494 / 246
MIMIC-IV / Combined	3689	58 / 59	42	491 / 244
eICU / Archived gated score	1395	0 / 0	0	323 / 107
eICU / No urine carrying	1395	0 / 36	17	323 / 105
eICU / Zero-inclusive urine; no carry	1395	18 / 40	32	321 / 104
eICU / Coherent GCS	1395	0 / 0	0	323 / 107
eICU / Combined	1395	18 / 40	32	321 / 104

Changes are relative to the archived gated adapted six-hour score within the same risk set. Class means decreased ≥2, net stable −1 to +1, or increased ≥2. A change of class does not establish which score is clinically correct. eICU uses direct total GCS in both the historical and coherent policies, so these coincide by design.

Table S19 Rematched 365-day recorded mortality

Policy	Candidates / pairs	Deaths −3 / 0	Risk difference pp (95% CI)	Nominal paired P
Archived gated score	600 / 116	26 / 45	-16.4 (-27.6 to -5.2)	0.0094
No urine carrying	604 / 117	27 / 44	-14.5 (-25.6 to -3.4)	0.0186
Zero-inclusive urine; no carry	603 / 118	27 / 45	-15.3 (-27.1 to -3.4)	0.0153
Coherent GCS	604 / 117	24 / 44	-17.1 (-28.2 to -6.0)	0.0066
Combined	607 / 119	26 / 44	-15.1 (-26.1 to -3.4)	0.0153
Combined; coverage selected	155 / 24	5 / 9	-16.7 (-41.7 to 8.3)	0.3438

Exact attained SOFA 2–9 and ΔSOFA −3 vs 0. Each policy rematches without replacement using the same algorithm. Intervals resample fixed pairs 5,000 times, seed 20261011, omitting match-selection uncertainty. P is the two-sided exact McNemar test. Coverage-selected combined policy requires direct GCS and ≥5 direct inputs in both endpoint windows; this is a different selected population.

Table S20 Conditional associations across score policies

Policy	n	Base OR (95% CI)	Care OR (95% CI)	Care nominal P
Archived gated score	600	0.492 (0.317–0.762)	0.863 (0.523–1.424)	0.5641
No urine carrying	604	0.509 (0.330–0.784)	0.907 (0.552–1.493)	0.7021
Zero-inclusive urine; no carry	603	0.531 (0.347–0.814)	0.951 (0.580–1.558)	0.8413
Coherent GCS	604	0.467 (0.301–0.726)	0.839 (0.508–1.386)	0.4927
Combined	607	0.507 (0.330–0.777)	0.925 (0.564–1.516)	0.7560
Combined; coverage selected	155	0.445 (0.211–0.941)	0.619 (0.266–1.438)	0.2643

HC3 logistic intervals; outcome is recorded death by ICU day 365. Base: attained integer-score fixed effects, age, sex, steroid categories and first-window input count. Care adds first-window recorded sedation and ventilation. Structural zero/redundant nuisance columns are removed only to preserve an identifiable design; omitted columns and ranks are exported. This is not causal adjustment or mediation analysis.

Table S21 Outcomes within the archived net-zero groups

Database	Endpoint	Opposing component changes	All components unchanged
MIMIC-IV	ICU hours 24–72	15/248 (6.0%)	6/245 (2.4%)
MIMIC-IV	Recorded day 28	60/248 (24.2%)	56/245 (22.9%)
MIMIC-IV	Recorded day 365	86/248 (34.7%)	100/245 (40.8%)
eICU	ICU hours 24–72	4/107 (3.7%)	2/216 (0.9%)

This table uses the archived gated policy (MIMIC-IV net-zero n=493; eICU n=323). Opposing changes require ≥1 decreasing and ≥1 increasing subscore. MIMIC-IV 365-day base OR 0.509 (0.329–0.786), care OR 0.676 (0.423–1.080), nominal P=0.102. Short-term events are too sparse for the fixed multivariable support rule; eICU results remain descriptive. All policies and model support decisions are in Source Data.

Table S22a Absolute external prediction performance

Locked model	AUC (95% CI)	Average precision (95% CI)	Brier (95% CI)
Attained	0.825 (0.781 to 0.866)	0.292 (0.214 to 0.396)	0.03947 (0.02976 to 0.04789)
Attained + first	0.819 (0.770 to 0.859)	0.288 (0.209 to 0.402)	0.03900 (0.02910 to 0.04767)
All four windows	0.817 (0.771 to 0.855)	0.282 (0.200 to 0.397)	0.03919 (0.02926 to 0.04782)
Spline attained	0.807 (0.762 to 0.849)	0.288 (0.209 to 0.396)	0.03978 (0.02994 to 0.04842)
Spline + first	0.791 (0.734 to 0.841)	0.283 (0.199 to 0.400)	0.03936 (0.02923 to 0.04809)

1,395 eICU stays; 64 ICU deaths during hours 24–72; 76 hospitals. Intervals use 1,000 hospital-cluster percentile resamples, seed 20261011. Predictions are fixed from MIMIC-IV without eICU training or recalibration. Intervals omit model-development uncertainty. Spline variants are exploratory sensitivities on an already examined external cohort.

Table S22b External calibration and patient sensitivity

Evaluation / model	n / events	Mean predicted risk	CITL (95% CI)	Joint slope (95% CI)
All stays / Attained	1395 / 64	2.24%	0.850 (0.599 to 1.034)	1.013 (0.891 to 1.203)
All stays / Attained + first	1395 / 64	2.58%	0.695 (0.426 to 0.879)	0.968 (0.844 to 1.160)
All stays / All four windows	1395 / 64	2.59%	0.693 (0.424 to 0.882)	0.943 (0.824 to 1.119)
One per person / Attained	1334 / 59	2.25%	0.800 (0.552 to 0.990)	1.001 (0.870 to 1.201)
One per person / Attained + first	1334 / 59	2.60%	0.644 (0.378 to 0.841)	0.965 (0.829 to 1.171)
One per person / All four windows	1334 / 59	2.61%	0.640 (0.368 to 0.837)	0.941 (0.810 to 1.136)

Observed risks: all stays 64/1,395 (4.59%); one per person 59/1,334 (4.42%). CITL fixes predicted logit as an offset and estimates one intercept (ideal 0); the slope is jointly fitted with a separate intercept (ideal 1). The patient sensitivity uses minimum unit-stay identifier, not a verified chronological first admission. All absolute values and joint intercepts are in Source Data.

Table S23 Paired external model differences

New model vs reference	AUC difference (95% CI)	Brier difference (95% CI)
Attained + first vs Attained	-0.0062 (-0.0191 to 0.0033)	-0.00048 (-0.00109 to 0.00002)
All four windows vs Attained + first	-0.0025 (-0.0097 to 0.0068)	0.00019 (-0.00005 to 0.00046)
All four windows vs Attained	-0.0087 (-0.0221 to 0.0051)	-0.00029 (-0.00098 to 0.00020)
Spline + first vs Spline attained	-0.0159 (-0.0402 to 0.0020)	-0.00042 (-0.00099 to 0.00002)
Spline attained vs Attained	-0.0183 (-0.0326 to -0.0022)	0.00031 (-0.00004 to 0.00060)
Spline + first vs Attained + first	-0.0279 (-0.0458 to -0.0081)	0.00037 (-0.00000 to 0.00067)

The same hospital draws are paired across predictions. Negative AUC differences indicate lower discrimination; negative Brier differences indicate lower probability error. The old 1,000-resample seed 20261009 gave a just-below-zero first-score Brier interval; seed 20261011 crosses zero with the same point estimate. This boundary is Monte Carlo sensitive and does not justify selecting a favorable seed or claiming robust incremental benefit.

Table S24 Demographic coverage and attained-model precision

Recorded subgroup	n / events	AUC (95% CI)	CITL (95% CI)	Joint slope (95% CI)
age / 65-74	571 / 23	0.852 (0.801 to 0.919)	0.887 (0.374 to 1.245)	0.961 (0.801 to 1.258)
age / 75-84	504 / 24	0.762 (0.609 to 0.870)	0.824 (0.452 to 1.157)	0.915 (0.478 to 1.249)
age / 85+	320 / 17	0.884 (0.808 to 0.946)	0.837 (0.098 to 1.297)	Omitted (<20 events)
recorded sex / Female	636 / 32	0.819 (0.761 to 0.876)	0.730 (0.419 to 1.050)	0.964 (0.801 to 1.258)
recorded sex / Male	759 / 32	0.831 (0.755 to 0.882)	0.975 (0.399 to 1.284)	1.109 (0.840 to 1.373)

Age categories use recorded numeric age, including eICU >89 coded as 90; 115 stays have that source category. Source sex codes are not inferred gender identity. Subgroups below 20 events omit joint slope fitting. Overlapping intervals do not establish fairness, equivalence or absence of bias. All models, internal subgroup metrics and finite-resample counts are supplied in Source Data.

Table S25 Direct carried and missing inputs by window

Database / input	0–6 h	6–12 h	12–18 h	18–24 h
MIMIC-IV / PaO_2	2878 / 0 / 811	2287 / 756 / 646	1586 / 1526 / 577	1192 / 1956 / 541
MIMIC-IV / FiO_2	2643 / 0 / 1046	2614 / 274 / 801	2130 / 814 / 745	1717 / 1265 / 707
MIMIC-IV / Platelets	3283 / 0 / 406	1724 / 1819 / 146	2149 / 1486 / 54	850 / 2819 / 20
MIMIC-IV / Bilirubin	1206 / 0 / 2483	564 / 842 / 2283	360 / 1145 / 2184	193 / 1344 / 2152
MIMIC-IV / Creatinine	3205 / 0 / 484	1695 / 1882 / 112	2255 / 1401 / 33	913 / 2763 / 13
MIMIC-IV / MAP	3523 / 0 / 166	3614 / 45 / 30	3598 / 66 / 25	3585 / 85 / 19
MIMIC-IV / GCS	3336 / 0 / 353	3443 / 215 / 31	3336 / 344 / 9	3117 / 567 / 5
MIMIC-IV / 6-hour urine	3291 / 0 / 398	3495 / 71 / 123	3466 / 135 / 88	3455 / 160 / 74
eICU / PaO_2	531 / 0 / 864	276 / 345 / 774	244 / 444 / 707	156 / 558 / 681
eICU / FiO_2	812 / 0 / 583	580 / 290 / 525	580 / 343 / 472	534 / 413 / 448
eICU / Platelets	626 / 0 / 769	536 / 445 / 414	411 / 810 / 174	235 / 1085 / 75
eICU / Bilirubin	369 / 0 / 1026	262 / 318 / 815	204 / 528 / 663	90 / 698 / 607
eICU / Creatinine	733 / 0 / 662	613 / 450 / 332	494 / 760 / 141	318 / 1018 / 59
eICU / MAP	1387 / 0 / 8	1391 / 1 / 3	1392 / 2 / 1	1385 / 9 / 1
eICU / GCS	1073 / 0 / 322	824 / 382 / 189	814 / 431 / 150	799 / 462 / 134
eICU / 6-hour urine	637 / 0 / 758	675 / 125 / 595	625 / 216 / 554	592 / 265 / 538

Each cell gives direct / carried / missing counts, summing to 3,689 in MIMIC-IV and 1,395 in eICU. Archived gated policy; direct means observed in that window under the relevant database-field gate. All policies are exported. Carrying physiological values does not verify persistent physiology; carrying interval urine was removed in designated sensitivities. Missing measurements are not normal observations.

Table S26 Recorded zeros and coherent GCS

Window	Zero-only urine M4 / eICU	GCS direct old / coherent	Different direct GCS totals	Near GCS RASS ≤−3
0–6 h	28 / 22	3336 / 3335	59	1175
6–12 h	19 / 18	3443 / 3440	62	740
12–18 h	31 / 14	3336 / 3330	36	565
18–24 h	24 / 14	3117 / 3116	19	374

GCS and RASS columns concern MIMIC-IV only; GCS denominator is 3,689 windows at each time. Different totals compare independent component minima with a complete same-timestamp triplet when both exist. Nearest RASS may follow the GCS observation within 60 min and by hour 24; these are contextual records, not real-time predictors. A recorded zero-only urine window does not certify complete six-hour anuria.

Table S27 Entry-service selection within the archived risk set

Analysis	Sample	Estimate (95% CI)	Nominal P
28 d / Base	322 candidates	OR 0.881 (0.443–1.752)	0.7184
28 d / Base + care	322 candidates	OR 1.122 (0.541–2.326)	0.7576
365 d / Base	322 candidates	OR 1.053 (0.594–1.866)	0.8599
365 d / Base + care	322 candidates	OR 1.466 (0.767–2.804)	0.2472
28 d / Matched	42 pairs; deaths 9 vs 13	RD -9.5 pp (-26.2 to 7.1)	0.4240
365 d / Matched	42 pairs; deaths 16 vs 18	RD -4.8 pp (-26.2 to 16.7)	0.8318

The entry-service subset contains 1,766 of the archived 3,689 MIMIC-IV stays and 322 strict candidates (81 with Δ−3 and 241 with Δ0). This conditional subset does not recover older patients excluded by inherited selection. Models and matching use the same score and adjustment rules. Fixed-pair bootstrap: 5,000 draws, seed 20261011; no selection/matching uncertainty. Wide intervals and changed selection do not establish absence of a trajectory association or a causal service effect.

Table S28 Exact pairwise SOFA score comparisons for Figure 2a

Time window	Decrease vs net stable P	Decrease vs increase P	Net stable vs increase P
0–6 h	5.41e-131	4.09e-178	4.55e-26
6–12 h	2.61e-09	3.30e-06	4.90e-14
12–18 h	0.0270	1.92e-42	1.23e-22
18–24 h	2.15e-10	1.66e-102	3.22e-42

Decrease ≥2 points, n=1,600; net stable, n=1,282; increase ≥2 points, n=807. Each comparison uses the two-sided asymptotic Mann–Whitney U test with tie and continuity corrections. Windows are tested separately, not pooled as independent observations. Values are nominal and not adjusted for multiplicity or covariates. The first and final scores also define the groups, so these conditional comparisons do not validate independent trajectory classes. Values below 0.001 use scientific notation with three significant digits; full precision and test statistics are in Source Data.

Supplementary tables S29 to S32

Table S29. Landmark cohorts and measurement coverage

Characteristic	MIMIC-IV (n=3689)	eICU (n=1395)	P	Holm-adjusted P
Age, years	75.0 [70.0, 82.0]	77.0 [70.0, 84.0]	8.54×10^{-5}	2.56×10^{-4}
Male sex, n (%)	2196 (59.5)	759 (54.4)	9.61×10^{-4}	9.61×10^{-4}
First-window score, 0–6 h	8.0 [4.0, 10.0]	3.0 [1.0, 6.0]	6.97×10^{-139}	6.28×10^{-138}
Attained score, 18–24 h	5.0 [3.0, 8.0]	4.0 [3.0, 7.0]	1.06×10^{-23}	5.32×10^{-23}
Direct inputs in first window, /8	7.0 [6.0, 7.0]	4.0 [3.0, 6.0]	1.21×10^{-251}	1.21×10^{-250}
Direct inputs in attained window, /8	4.0 [3.0, 5.0]	3.0 [2.0, 4.0]	1.74×10^{-119}	1.39×10^{-118}
Direct GCS in first window, n (%)	3336 (90.4)	1073 (76.9)	8.63×10^{-37}	5.18×10^{-36}
Direct GCS in attained window, n (%)	3117 (84.5)	799 (57.3)	3.69×10^{-94}	2.58×10^{-93}
Threshold steroid category, n (%)	332 (9.0)	172 (12.3)	3.92×10^{-4}	7.85×10^{-4}
Subthreshold steroid category, n (%)	77 (2.1)	109 (7.8)	2.90×10^{-22}	1.16×10^{-21}

Values are median [IQR] or n (%). P is the unadjusted between-database P value. Holm-adjusted P controls multiplicity across the 10 comparisons in this table as one family. This adjustment was added during the exploratory revision, not prespecified. Age, scores and input counts use two-sided asymptotic Mann–Whitney U tests with tie and continuity corrections; binary rows use Pearson chi-square tests without Yates correction. Each steroid-category row compares that category against all other categories. No covariate or hospital/person-cluster adjustment is applied; tests treat stays as independent and describe these archived samples. Source eligibility and measurement differences limit generalization. Direct observations precede carrying. Timestamp fields are database-specific proxies, not verified bedside availability. Steroid categories describe orders, not administered dose.

Table S30 Organ component scores by change group in MIMIC-IV

Window / component	Decrease	Net stable	Increase
0-6 h / Respiratory	3.0 [2.0, 3.0]	0.0 [0.0, 2.0]	0.0 [0.0, 2.0]
0-6 h / Coagulation	1.0 [0.0, 1.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
0-6 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
0-6 h / Cardiovascular	2.0 [1.0, 2.0]	1.0 [1.0, 2.0]	1.0 [0.0, 1.0]
0-6 h / CNS	4.0 [3.0, 4.0]	1.0 [0.0, 3.0]	0.0 [0.0, 2.0]
0-6 h / Renal	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
6-12 h / Respiratory	2.0 [1.0, 3.0]	1.0 [0.0, 2.0]	2.0 [0.0, 3.0]
6-12 h / Coagulation	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
6-12 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
6-12 h / Cardiovascular	2.0 [1.0, 2.0]	1.0 [1.0, 2.0]	1.0 [1.0, 3.0]
6-12 h / CNS	2.0 [1.0, 3.0]	2.0 [0.0, 3.0]	3.0 [0.5, 4.0]
6-12 h / Renal	0.0 [0.0, 1.0]	0.0 [0.0, 2.0]	0.0 [0.0, 2.0]
12-18 h / Respiratory	2.0 [1.0, 2.0]	1.0 [0.0, 2.0]	2.0 [0.0, 3.0]
12-18 h / Coagulation	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
12-18 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
12-18 h / Cardiovascular	1.0 [1.0, 2.0]	1.0 [1.0, 2.0]	1.0 [1.0, 3.0]
12-18 h / CNS	1.0 [0.0, 2.0]	1.0 [0.0, 3.0]	2.0 [0.0, 3.0]
12-18 h / Renal	0.0 [0.0, 1.0]	1.0 [0.0, 2.0]	1.0 [0.0, 3.0]
18-24 h / Respiratory	2.0 [1.0, 2.0]	1.0 [0.0, 2.0]	2.0 [0.0, 3.0]
18-24 h / Coagulation	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
18-24 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
18-24 h / Cardiovascular	1.0 [1.0, 2.0]	1.0 [1.0, 2.0]	1.0 [1.0, 3.0]
18-24 h / CNS	0.0 [0.0, 2.0]	1.0 [0.0, 3.0]	2.0 [1.0, 3.0]
18-24 h / Renal	0.0 [0.0, 1.0]	1.0 [0.0, 2.0]	1.0 [0.0, 3.0]

Values are median [interquartile range]. Groups are defined by attained minus first-window total: decrease ≤−2, net stable −1 to +1, and increase ≥2. Component scores use the local adapted six-hour algorithm. No hypothesis tests are assigned to these descriptive component summaries.

Table S30 Organ component scores by change group in eICU

Window / component	Decrease	Net stable	Increase
0-6 h / Respiratory	1.0 [0.0, 3.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
0-6 h / Coagulation	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
0-6 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
0-6 h / Cardiovascular	2.0 [1.0, 4.0]	1.0 [0.8, 1.0]	1.0 [0.0, 1.0]
0-6 h / CNS	2.0 [0.0, 3.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
0-6 h / Renal	1.0 [0.0, 3.0]	0.0 [0.0, 2.0]	0.0 [0.0, 0.0]
6-12 h / Respiratory	0.0 [0.0, 2.0]	0.0 [0.0, 1.0]	0.0 [0.0, 2.0]
6-12 h / Coagulation	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 1.0]
6-12 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
6-12 h / Cardiovascular	1.0 [1.0, 3.0]	1.0 [0.0, 1.0]	1.0 [0.0, 1.0]
6-12 h / CNS	1.0 [0.0, 2.0]	0.0 [0.0, 1.0]	0.0 [0.0, 2.0]
6-12 h / Renal	1.0 [0.0, 2.0]	0.5 [0.0, 2.0]	1.0 [0.0, 2.0]
12-18 h / Respiratory	0.0 [0.0, 2.0]	0.0 [0.0, 1.0]	0.0 [0.0, 3.0]
12-18 h / Coagulation	0.0 [0.0, 1.0]	0.0 [0.0, 0.0]	0.0 [0.0, 1.0]
12-18 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
12-18 h / Cardiovascular	1.0 [0.0, 3.0]	1.0 [0.0, 1.0]	1.0 [0.0, 2.2]
12-18 h / CNS	1.0 [0.0, 2.0]	0.0 [0.0, 1.0]	1.0 [0.0, 2.0]
12-18 h / Renal	1.0 [0.0, 2.0]	1.0 [0.0, 2.0]	1.0 [0.0, 3.0]
18-24 h / Respiratory	0.0 [0.0, 2.0]	0.0 [0.0, 1.0]	1.0 [0.0, 3.0]
18-24 h / Coagulation	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]	0.0 [0.0, 1.0]
18-24 h / Liver	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
18-24 h / Cardiovascular	1.0 [0.0, 1.0]	1.0 [0.0, 1.0]	1.0 [1.0, 3.0]
18-24 h / CNS	1.0 [0.0, 2.0]	0.0 [0.0, 1.0]	1.0 [0.0, 2.0]
18-24 h / Renal	1.0 [0.0, 1.0]	1.0 [0.0, 2.0]	2.0 [1.0, 3.0]

Values are median [interquartile range]. Groups are defined by attained minus first-window total: decrease ≤−2, net stable −1 to +1, and increase ≥2. Component scores use the local adapted six-hour algorithm. No hypothesis tests are assigned to these descriptive component summaries.

Table S31 Matching balance before matching

Variable	Decrease (n=235)	Net zero (n=365)	SMD	Absolute SMD
Age, years	75.0 [69.0, 81.0]	78.0 [71.0, 85.0]	-0.385	0.385
Male sex	141 (60.0)	208 (57.0)	0.061	0.061
First-window total SOFA	8.0 [7.0, 10.0]	4.0 [3.0, 7.0]	1.687	1.687
Attained total SOFA	5.0 [4.0, 7.0]	4.0 [3.0, 7.0]	0.281	0.281
First-window CNS subscore	4.0 [3.0, 4.0]	1.0 [0.0, 2.0]	1.469	1.469
Attained CNS subscore	1.0 [0.0, 2.0]	1.0 [0.0, 2.0]	0.001	0.001
First-window non-CNS sum	5.0 [4.0, 7.0]	3.0 [2.0, 5.0]	0.809	0.809
Attained non-CNS sum	4.0 [3.0, 6.0]	4.0 [2.0, 5.0]	0.333	0.333
Sedative record at 0–6 h	193 (82.1)	117 (32.1)	1.173	1.173
Invasive ventilation at 0–6 h	190 (80.9)	115 (31.5)	1.146	1.146
Direct inputs in first window, /8	7.0 [7.0, 7.0]	6.0 [5.0, 7.0]	0.683	0.683
Direct inputs in attained window, /8	4.0 [3.0, 5.0]	4.0 [3.0, 5.0]	0.141	0.141
Direct GCS in first window	222 (94.5)	326 (89.3)	0.190	0.190
Direct GCS in attained window	194 (82.6)	297 (81.4)	0.031	0.031
No relevant order	223 (94.9)	317 (86.8)	0.282	0.282
Threshold category	10 (4.3)	38 (10.4)	-0.238	0.238
Below threshold	2 (0.9)	10 (2.7)	-0.143	0.143
Prior-admission diagnosis coverage	63 (26.8)	116 (31.8)	-0.109	0.109
Documented heart failure	25 (10.6)	41 (11.2)	-0.019	0.019
Documented chronic kidney disease	16 (6.8)	32 (8.8)	-0.073	0.073
Documented chronic lung disease	14 (6.0)	19 (5.2)	0.033	0.033
Documented cancer	10 (4.3)	26 (7.1)	-0.124	0.124
Documented cerebrovascular disease	11 (4.7)	21 (5.8)	-0.048	0.048
Documented dementia	2 (0.9)	6 (1.6)	-0.071	0.071
Documented diabetes	20 (8.5)	42 (11.5)	-0.100	0.100

SMD denotes the signed standardized mean difference, using the pooled pre-match standard deviation for both comparisons. Binary variables use the corresponding pooled Bernoulli standard deviation. The stable comparator has exactly zero total score change. Values and standardizers are retained at full precision in the source CSV.

Table S31 Matching balance after matching

Variable	Decrease (n=116)	Net zero (n=116)	SMD	Absolute SMD
Age, years	76.0 [72.0, 82.0]	76.0 [71.0, 83.0]	-0.068	0.068
Male sex	65 (56.0)	65 (56.0)	0.000	0.000
First-window total SOFA	8.0 [7.0, 10.0]	5.0 [4.0, 7.0]	1.406	1.406
Attained total SOFA	5.0 [4.0, 7.0]	5.0 [4.0, 7.0]	0.000	0.000
First-window CNS subscore	4.0 [2.8, 4.0]	2.0 [0.0, 3.0]	0.878	0.878
Attained CNS subscore	1.0 [0.0, 2.0]	1.0 [0.0, 3.0]	-0.322	0.322
First-window non-CNS sum	5.0 [4.0, 7.0]	3.0 [2.0, 4.0]	0.912	0.912
Attained non-CNS sum	4.0 [3.0, 6.0]	4.0 [3.0, 5.0]	0.221	0.221
Sedative record at 0–6 h	90 (77.6)	48 (41.4)	0.848	0.848
Invasive ventilation at 0–6 h	90 (77.6)	48 (41.4)	0.841	0.841
Direct inputs in first window, /8	7.0 [6.0, 7.0]	7.0 [6.0, 7.0]	0.000	0.000
Direct inputs in attained window, /8	4.0 [3.0, 5.0]	4.0 [3.0, 5.0]	-0.112	0.112
Direct GCS in first window	105 (90.5)	107 (92.2)	-0.063	0.063
Direct GCS in attained window	96 (82.8)	96 (82.8)	0.000	0.000
No relevant order	115 (99.1)	115 (99.1)	0.000	0.000
Threshold category	1 (0.9)	1 (0.9)	0.000	0.000
Below threshold	0 (0.0)	0 (0.0)	0.000	0.000
Prior-admission diagnosis coverage	35 (30.2)	36 (31.0)	-0.019	0.019
Documented heart failure	16 (13.8)	9 (7.8)	0.193	0.193
Documented chronic kidney disease	9 (7.8)	8 (6.9)	0.032	0.032
Documented chronic lung disease	9 (7.8)	5 (4.3)	0.150	0.150
Documented cancer	5 (4.3)	8 (6.9)	-0.112	0.112
Documented cerebrovascular disease	5 (4.3)	9 (7.8)	-0.155	0.155
Documented dementia	1 (0.9)	2 (1.7)	-0.078	0.078
Documented diabetes	9 (7.8)	9 (7.8)	0.000	0.000

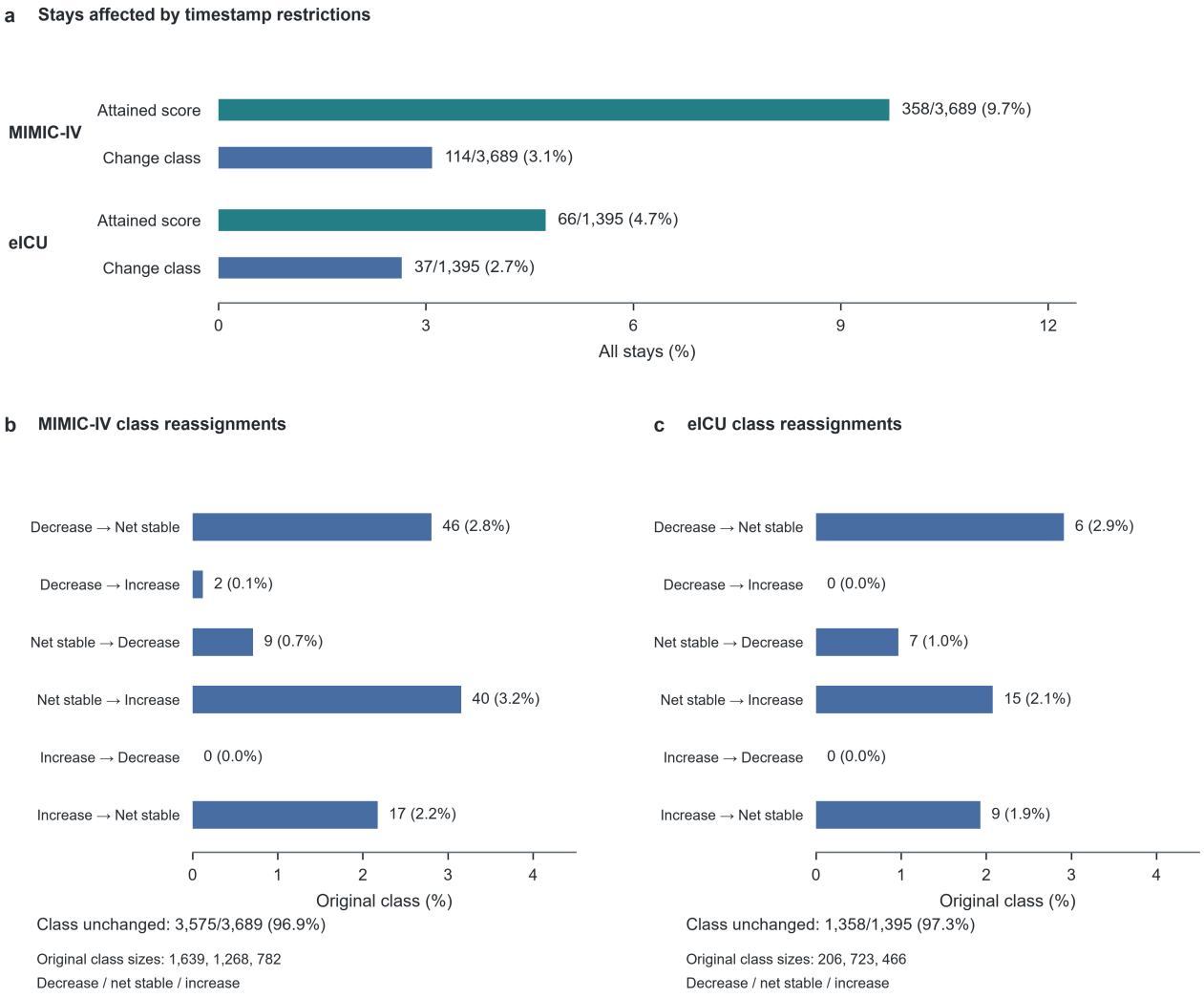
SMD denotes the signed standardized mean difference, using the pooled pre-match standard deviation for both comparisons. Binary variables use the corresponding pooled Bernoulli standard deviation. The stable comparator has exactly zero total score change. Values and standardizers are retained at full precision in the source CSV.

Table S32 Local scoring rules and interpretation

Domain	Local rule	Interpretation and remaining measurement issue
Time windows	Four nonoverlapping six-hour windows; first versus attained score.	Not a standard 24-hour daily SOFA; endpoint difference is determined by first score at fixed attained score.
Respiratory	Minimum PaO2 divided by available FiO2; invasive-ventilation flag activates scores 3–4.	PaO2 and FiO2 may not be simultaneous; retrospective support intervals remain.
Cardiovascular	Local dopamine boundary and vasopressin/phenylephrine rules are retained in score_sofa().	Does not establish equivalence to original SOFA vasopressor criteria; dose units and interval duration require reference review.
Renal	Creatinine plus six-hour urine cutoffs of 50 and 125 mL; primary policy excludes zero entries and carries prior positive totals.	Alternative policies retain zero and remove carry; recording completeness and cumulative entries remain unresolved.
Neurological	MIMIC-IV uses independently minimized GCS components; eICU uses recorded total. Same-time GCS is a sensitivity in MIMIC-IV.	Sedation and airway-related verbal entries are not independently adjudicated. Same-time agreement alone is not clinical validation.
Liver and coagulation	Bilirubin maximum and platelet minimum within each local window, with the documented threshold code.	Units, timing and carried values need to be distinguished from direct observations.
Missingness and entry	Forward carry; missing values that trigger no condition receive zero; point-entry gates use source-specific fields.	Zero score is not proof of normal physiology; storage/revision time is not verified first bedside availability.

The crosswalk describes the actual adapted scoring algorithm. It does not assert agreement with independently adjudicated clinical SOFA scores.

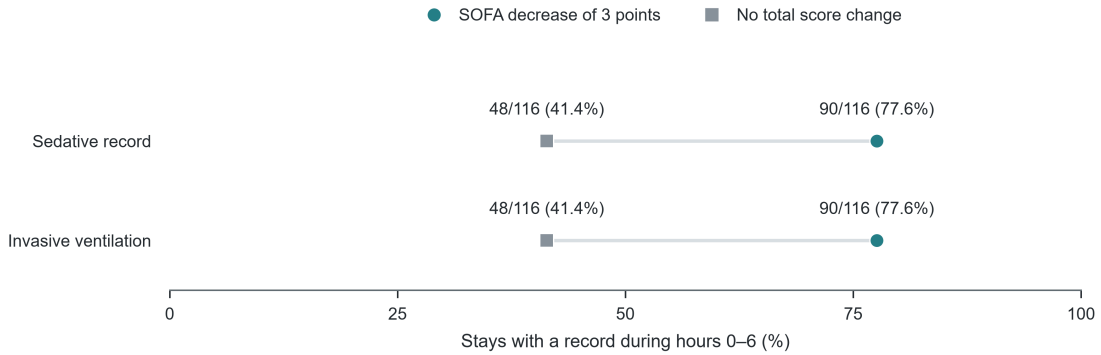
Figure S1 | Score changes and class reassignments under timestamp restrictions



a, Proportions of all stays whose attained SOFA score or broad score-change class changed between the original and restricted recording policies. Counts and cohort denominators are printed beside the bars. **b,c,** Every off-diagonal transition for MIMIC-IV and eICU, respectively. Arrows run from the original to the restricted class. Percentages use the original class as the denominator, as listed below each panel; these differ from the full-cohort denominator in panel a. Unchanged-class totals are also reported. Decrease denotes ΔSOFA≤−2, net stable −1 to +1 and increase ≥2. All bars start at zero, and panels b and c share the same scale. MIMIC-IV restricts physiological storage timestamps while retaining retrospective support intervals; eICU cumulatively aligns point windows and entry/revision timestamps. The displays are descriptive; no hypothesis tests or confidence intervals are shown. Sources: aggregate score-policy counts and complete classification cross-tabulations.

Figure S2 | Early care records and adjusted mortality associations

a **Early care records in 116 matched pairs**

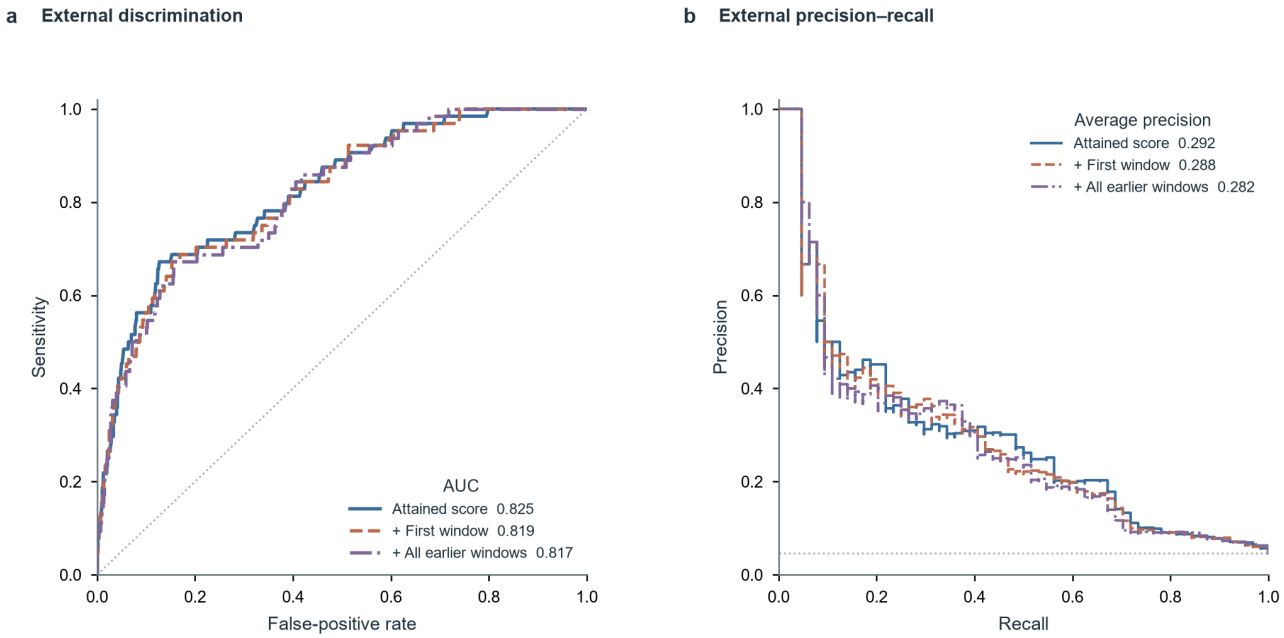


b **Mortality associations in all 600 strict-comparison candidates**

Recorded mortality	Base adjustment		Additional early-care adjustment	
	OR (95% CI)	P	OR (95% CI)	P
Day 28 (102 events)	0.42 (0.24–0.74)	0.0028	0.66 (0.36–1.23)	0.1924
Day 365 (181 events)	0.49 (0.32–0.76)	0.0015	0.86 (0.52–1.42)	0.5641

a, Observed proportions with selected sedative or invasive ventilation records during ICU hours 0–6 in 116 matched pairs. Circles indicate an exact SOFA decrease of 3 points and squares no total change. Connecting lines join the two group summaries and do not represent individual pairs or confidence intervals. Counts are displayed for each group; identical aggregate counts across the two care variables do not imply identical patient-level records. No hypothesis tests are attached to these descriptive proportions. **b,** Logistic associations comparing an exact decrease of 3 points with no change among all 600 strict-comparison candidates, with 102 day-28 and 181 day-365 recorded deaths. Base adjustment includes attained-score fixed effects, age, sex and steroid-record category. The second model additionally includes first-window sedative and invasive ventilation records. Intervals use HC3 robust covariance; P values are nominal and unadjusted for multiplicity. These candidate-level models use a larger population than the matched sample in panel a. The full estimates are retained in Table S13.

Figure S3 | Discrimination of locked models in eICU

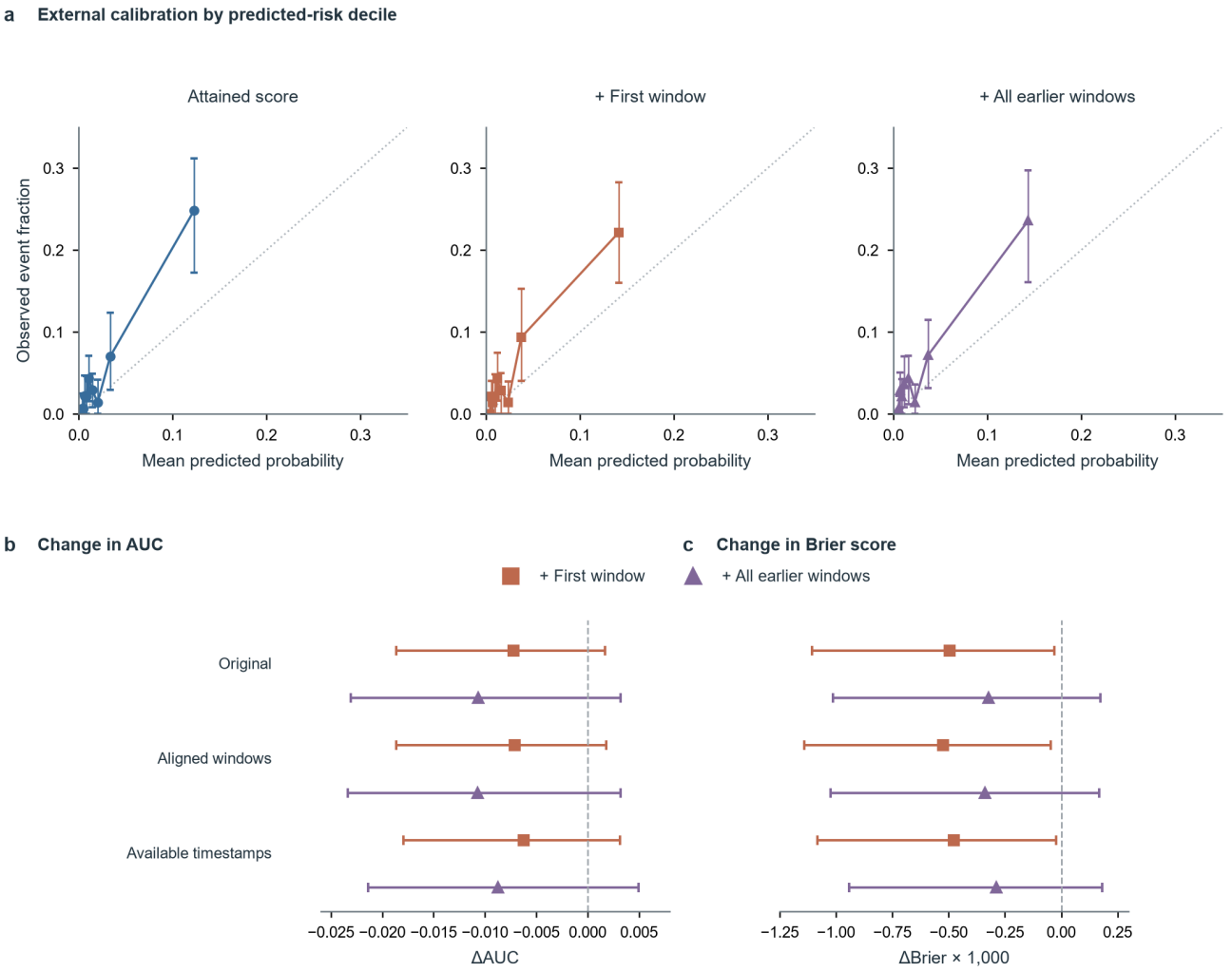


eICU: 1,395 stays, 64 ICU deaths during hours 24–72. Models locked in MIMIC-IV without external refitting.

Available timestamp restrictions. The precision–recall reference line is the observed event prevalence.

a, Receiver operating characteristic curves and AUCs. **b,** Precision–recall curves and average precision, with a horizontal reference at the observed event prevalence of 64/1,395. Average precision is computed from the threshold sequence rather than the trapezoidal area under the plotted curve. All models include age, sex and attained SOFA; expanded models add the first window or all three earlier windows. All panels use the same 1,395 stays across 76 eICU hospitals and the endpoint of recorded ICU death during hours 24–72. Models were fitted in MIMIC-IV and transferred without external fitting or recalibration. Curves are computed directly from frozen probabilities under the available-timestamp policy; no confidence bands or hypothesis tests are attached to these descriptive curves. Sources: Table S14 and the frozen prediction file; exported curve coordinates contain no stay identifiers.

Figure S4 | External calibration and recording-policy sensitivity

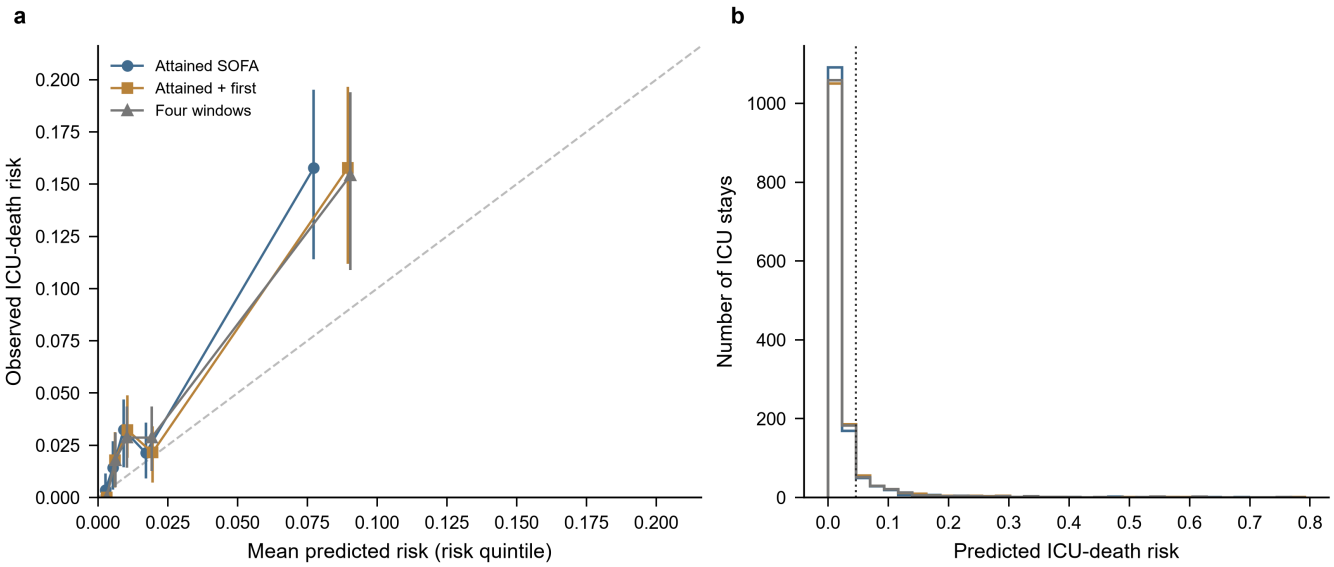


Calibration and metric differences use hospital-bootstrap 95% CIs. Comparator: attained-score model.

No external recalibration was applied. Lower Brier and higher AUC favour the expanded model.

a, Calibration for each locked model, using deciles of its predicted risk and a common axis range. Points show mean predictions and observed event fractions, joined in predicted-risk order. Each bin contains 136–143 stays because tied predictions remain together. Error bars are hospital-cluster bootstrap percentile 95% intervals from 1,000 resamples, holding the original fitted model and bin boundaries fixed (seed 20261010). The diagonal denotes perfect calibration. **b,c,** Changes in AUC and Brier score relative to the attained-score model under each recording policy, using the same 1,395 eICU stays. Squares add the first window and triangles all earlier windows. Error bars are the previously audited paired hospital-bootstrap percentile 95% intervals from 1,000 valid resamples, conditional on fitted development models. Brier differences are multiplied by 1,000. The null difference is marked by a dashed line; higher ΔAUC and lower ΔBrier favour the expanded model. All AUC and Brier results are retained, including inconclusive discrimination differences and the small first-window Brier improvement. No external recalibration was performed. Sources: Tables S7, S8 and S14 and the aggregate calibration bins.

Figure S5 Locked external calibration and risk distribution



a, Mean predicted and observed risks in five groups formed by fixed predicted-risk quantiles. Hospital-bootstrap 95% intervals describe observed risk; lines guide the eye and are not fitted recalibration curves. **b,** Predicted-risk distributions for the three linear models, with overall observed event frequency marked at 4.59%. The sample is 1,395 eICU stays from 76 hospitals, with 64 ICU deaths during hours 24–72. The ideal calibration diagonal and risk distribution show that useful ranking can coexist with underestimation of average risk. No hypothesis test is displayed.