Research Square

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

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

Lei Wang
Affiliated Hospital of Hangzhou Normal University
Yu Song
songyu@zcmu.edu.cn

The Second Affiliated Hospital of Zhejiang Chinese Medical University

Article

Keywords: sepsis, older adults, SOFA, electronic health records, measurement, prognosis, transportability

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11387486/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Patients who reach the same Sequential Organ Failure Assessment (SOFA) score after 24 hours of intensive care may have different preceding score changes. We examined the prognostic information in these changes and their relationship to organ components, recording rules and early care in older adults with sepsis.

Methods

We analysed selected cohorts of adults aged ≥ 65 years who were alive and in the ICU at 24 hours in MIMIC-IV and eICU. Locally adapted SOFA scores were reconstructed in four six-hour windows. We compared recorded mortality across score-change groups and matched patients with a three-point decrease versus no change at the same attained score. Component changes, timestamp restrictions and adjustment for early sedative and ventilation records were evaluated. Models developed in MIMIC-IV were tested in eICU for ICU death during hours 24–72.

Results

The cohorts included 3,689 MIMIC-IV and 1,395 eICU stays. In MIMIC-IV, recorded mortality by day 365 was 16.3%, 37.4% and 42.3% in the decrease, net-stable and increase groups. Among 116 matched pairs, mortality was 22.4% versus 38.8% (paired difference, − 16.4 percentage points; 95% CI, − 27.6 to − 5.2; nominal P = 0.0094). In all 600 strict-comparison candidates, the odds ratio changed from 0.49 (95% CI, 0.32–0.76) to 0.86 (0.52–1.42) after adding early care records. Opposing component changes occurred in 50.3% of MIMIC-IV and 33.1% of eICU stays with zero total change. Timestamp restrictions changed broad classifications in 3.1% and 2.7%, respectively. Adding the first score to the attained-score model yielded an external AUC difference of − 0.0062 (95% CI, − 0.0191 to 0.0033).

Conclusions

Early SOFA decreases were associated with lower recorded mortality, including among older patients reaching the same attained score. This association varied with early care records, while unchanged totals frequently concealed opposing component changes. Interpretation of first-day SOFA change should include the component scores and the care context in which they were measured.

Background

Sepsis is defined by organ dysfunction arising from a dysregulated response to infection [1]. Global estimates identify a substantial burden of sepsis-associated illness and death [2]. For older survivors,

the consequences can include sustained cognitive impairment and functional disability [3]. Analyses of two international trials also found persistent limitations in quality of life after severe sepsis [4]. Early assessment in this population therefore concerns both the severity of the acute illness and the course that follows.

The Sequential Organ Failure Assessment (SOFA) score summarizes dysfunction across six organ systems and has been evaluated prospectively in multicenter ICU populations [5]. Serial SOFA assessments have long been associated with mortality, supporting interest in changes beyond the initial score [6]. By the end of the first ICU day, a patient may reach the same total score through several different combinations of starting severity and component change. A stable total can also result from a decrease in one component offset by an increase in another.

These comparisons depend on how measurements enter the score. GCS may be recorded during sedation or airway support, and event timestamps may differ from the time a result is entered. We examined recorded mortality among older adults reaching the same 24-hour SOFA through different early score changes, and the frequency of opposing component changes when the total was unchanged. We then assessed sensitivity to recording rules, concurrent care and cohort selection. A locked external model comparison evaluated whether earlier scores added information for short-term ICU mortality beyond the attained score.

Methods

Study design and data sources

MIMIC-IV is a single-center database derived from routine electronic hospital records [7]. The eICU Collaborative Research Database contains granular critical-care records from multiple US hospitals [8]. We used local MIMIC-IV version 3.1, with hospital coverage during 2008–2022, and eICU version 2.0, with coverage during 2014–2015. MIMIC-IV provided development data and recorded mortality through ICU day 365. eICU provided a separate evaluation population for the common short-term ICU-death endpoint. Version-specific coverage, field definitions and availability constraints are documented in Supplementary Methods and the accompanying database-source notes; the journal articles describe the underlying resources.

STROBE provided the framework for reporting the observational cohort analyses [9]. RECORD informed the description of source-data selection, coding and processing in these routinely collected records [10]. A reporting map accompanies the manuscript. Supplementary methods, results, Tables S1-S32 and Figures S1-S5 are provided in Additional file 1; aggregate source data and analysis code are provided in Additional file 2.

The study population was inherited from a local corticosteroid-replication pipeline, with all three steroid-record categories restored for this separate prognostic analysis. Source selection required recorded antibiotic and culture evidence in the initial ICU day, a locally reconstructed eligibility SOFA score ≥ 2, and

the applicable infection, service, prior-steroid and outcome-data restrictions. MIMIC-IV and eICU applied different operational infection and service rules. In particular, MIMIC-IV selection used a service record drawn from a window extending through ICU exit plus one day and required discharge information. Thus, eligibility was determined retrospectively and was not fully observable at the prediction time. Full-database screening counts were subsequently reconstructed and retained alongside the inherited flow (Table S15). Supplementary Methods specifies these inherited rules; this cohort must not be described as all older patients with sepsis or as an exclusively admission-defined inception cohort.

We restricted the analysis to age ≥ 65 years, ICU exit after hour 24 and no recorded death by hour 24. MIMIC-IV used the first ICU stay per subject in its source construction; eICU used the first unit visit within its source encounter definition. We use “stays” for the analytic units and do not assume identical patient-level selection across databases. The eICU cohort contained 57 people with multiple stays, including two represented at more than one hospital. An outcome-blind sensitivity retained the minimum unit-stay identifier per person; identifier order was not assumed to be chronological. No exclusion based on survival or continued ICU residence at 72 hours was applied. The analysis included no matched clones from the earlier treatment analysis. The study used all available stays meeting these source-specific and landmark criteria; no prospective sample-size or power target determined inclusion.

Six-hour score reconstruction

Guidance on serial SOFA assessment emphasizes explicit and consistent scoring conventions [11]. We reconstructed scores for 0–6, 6–12, 12–18 and 18–24 hours after ICU admission. The first and attained scores were the 0–6-hour and 18–24-hour values. The local algorithm aggregated physiological extrema and support records, carried observed physiological inputs forward from earlier windows without backward filling, and assigned zero points when missing values did not activate a scoring condition. An eight-input count captured direct observations of PaO_2, FiO_2, platelets, bilirubin, creatinine, mean arterial pressure, GCS and urine volume within each window, before carrying values forward.

The reconstruction included adaptations that require explicit distinction from a standard daily SOFA score. Urine thresholds were 0–50 mL and > 50–125 mL over six hours, combined with creatinine criteria; zero-valued urine records were removed before aggregation, and recorded zero totals did not activate oliguria conditions in the inherited implementation. The cardiovascular algorithm incorporated local vasopressin and phenylephrine rules. MIMIC-IV GCS was constructed from the window-specific minima of three components, which need not have been observed simultaneously; eICU used recorded total GCS. The full scoring rules, units, missing-data behavior and source differences are given in Supplementary Methods. A rule-by-rule crosswalk identifies which calculations reproduce the local algorithm and which still require clinical reference assessment (Table S32). Agreement with archived scores is a reproducibility check and does not establish criterion validity. We did not implement SOFA-2, which revised component definitions and thresholds through multicohort development and validation [12].

Score change (ΔSOFA) was defined as attained minus first-window score. The broad groups were decrease (ΔSOFA ≤ − 2), net stable (− 1 to + 1) and increase (ΔSOFA ≥ 2). Historical source tables use the corresponding labels improved, stable and worsened. These are score-change categories, not adjudicated physiological states. In a separate analysis, net zero meant ΔSOFA exactly 0; opposing component changes required at least one component decrease and one component increase. Net stable and net zero therefore denote different populations.

Recording-time policies

Original scoring used event timestamps. Before applying new restrictions, we reconstructed the original scores from raw physiological records and checked agreement with stored outputs. For MIMIC-IV, the restricted policy retained physiological laboratory, nursing and urine records stored by ICU hour 24. Organ-support intervals remained retrospective. For eICU, point-observation windows were changed to left-closed, right-open intervals, infusion entries were assigned to corresponding intervals within [0,24 hours), and laboratory revision, nursing entry, respiratory entry and intake/output entry offsets were restricted to ≤ 1,440 minutes.

These fields are proxies for availability. The eICU laboratory revision offset does not identify first result availability; its infusion offset denotes entry time and does not independently establish actual administration intervals. Later-entered records can describe care that was already clinically known. Therefore, the policies are recording-rule sensitivity analyses, not a verified reconstruction of all information available at the bedside. Source-field documentation is listed in Supplementary Methods.

MIMIC-IV inherited window assignment counted calendar-minute boundaries using DATE_DIFF before six-hour binning, rather than exact elapsed seconds. All five score policies retained that convention for reproducibility; 162 urine and 232 GCS/RASS records differed from exact-second bin assignment near boundaries.

We kept the landmark risk sets fixed across scoring policies. Restricted-policy results are presented first for manuscript clarity, with original and intermediate policies reported alongside them. This presentation order was selected during the exploratory audit and should not be described as preregistered.

Outcomes

For prediction, the endpoint was recorded ICU death after hour 24 and by hour 72. A patient discharged alive from the ICU before hour 72 was classified as not experiencing ICU death during the specified interval; this does not establish all-cause survival to hour 72. We did not pool long-term mortality across databases.

For MIMIC-IV association analyses, outcomes were recorded death by ICU days 28 and 365, conditional on landmark eligibility; day 90 was an additional descriptive horizon for matched pairs. These binary

endpoints describe recorded deaths, not independently verified complete vital-status ascertainment. MIMIC-IV includes hospital and linked death information, with death dates truncated at one year after the last discharge under the version 3.1 documentation. The local last-admission discharge proxy placed ICU day 365 within that nominal horizon for all 3,689 stays; this did not verify completeness of individual death linkage. Date-only deaths were examined as intervals and could change two ICU-day-365 classifications. A horizon measured 365 days after the 24-hour landmark differs from ICU day 365 and was not substituted for it. We report the database endpoint and its coverage rather than assume that a missing death record confirms survival. Version-specific death-data documentation is listed in the database-source notes.

Completion scoring and clinical-history sensitivities

A completion plan fixed urine-carry, recorded-zero urine, coherent same-timestamp GCS and combined scoring sensitivities before their outcome refits. Direct, carried and missing inputs were counted by window. Pre-index disease documentation, full-source screening and death-date precision were audited separately. Detailed rules and all variants are reported in Supplementary Methods S7–S8. These audits cannot recover undocumented physiology or verify actual bedside knowledge.

Descriptive comparisons and multiplicity

Table 1 describes demographics, first and attained scores, concurrent care, measurement coverage and positive prior-admission history records across the three MIMIC-IV score-change groups. Continuous and ordinal variables use the Kruskal–Wallis test with tie correction; categorical variables use Pearson chi-square tests without continuity correction when all expected counts are at least five. A Fisher–Freeman–Halton Monte Carlo test with 99,999 resamples and seed 20261012 is specified for sparse tables. The three steroid-record categories contribute one omnibus test, rather than three separate binary tests. Raw P values are followed by Holm-adjusted P values for the 23 comparisons in Table 1 as one family. The three outcome comparisons in Table 2 form a separate Holm family. These adjustments were added during exploratory revision, do not adjust for covariates and do not correct the entire analysis program. The archived between-database Table 1 is retained as Table S29 with its original 10-test family. Its stay-level tests do not adjust for person or hospital clustering. Figure 2 pairwise comparisons and other analyses retain nominal P values.

Table 1
Characteristics and measurement context by MIMIC-IV score-change group

Characteristic	Decrease ≥ 2 (n = 1600)	Net stable (n = 1282)	Increase ≥ 2 (n = 807)	P	Holm-adjusted P
Age, years	74.0 [69.0, 80.0]	77.0 [71.0, 83.0]	76.0 [70.0, 83.0]	8.83e-16	1.06e-14
Male sex	988 (61.8)	729 (56.9)	479 (59.4)	0.0292	0.2631
First-window total SOFA	9.0 [8.0, 11.0]	5.0 [3.0, 9.0]	3.0 [1.0, 7.0]	1.42e-228	3.13e-227
Attained total SOFA	4.0 [3.0, 6.0]	5.0 [3.0, 9.0]	8.0 [5.0, 11.0]	1.34e-95	2.14e-94
First-window CNS subscore	4.0 [3.0, 4.0]	1.0 [0.0, 3.0]	0.0 [0.0, 2.0]	5.17e-270	1.19e-268
Attained CNS subscore	0.0 [0.0, 2.0]	1.0 [0.0, 3.0]	2.0 [1.0, 3.0]	3.37e-103	6.07e-102
First-window non-CNS sum	6.0 [5.0, 7.0]	4.0 [2.0, 6.0]	2.0 [1.0, 5.0]	1.40e-129	2.67e-128
Attained non-CNS sum	4.0 [3.0, 5.0]	4.0 [2.0, 6.0]	6.0 [4.0, 8.0]	5.41e-66	8.12e-65
Sedative record at 0–6 h	1403 (87.7)	572 (44.6)	282 (34.9)	1.86e-186	3.71e-185
Invasive ventilation at 0–6 h	1403 (87.7)	583 (45.5)	270 (33.5)	4.38e-189	9.20e-188
Direct inputs in first window, /8	7.0 [7.0, 7.0]	7.0 [5.0, 7.0]	6.0 [4.0, 7.0]	2.32e-100	3.94e-99
Direct inputs in attained window, /8	3.0 [3.0, 5.0]	4.0 [3.0, 5.0]	5.0 [3.0, 6.0]	4.40e-48	5.72e-47
Direct GCS in first window	1549 (96.8)	1180 (92.0)	607 (75.2)	8.67e-65	1.21e-63
Direct GCS in attained window	1359 (84.9)	1075 (83.9)	683 (84.6)	0.7211	1.0000

Values are median [IQR] or n (%); steroid category has one omnibus P value. Each group denominator includes all stays in that score-change group. History entries are positive records from completed prior admissions, not estimates of disease prevalence; absence of a code is not confirmed absence of disease. The available-history row makes coverage explicit. Scores, components and care are measured during the first 24 hours and are not all pretreatment covariates. The first and attained windows are 0–6 h and 18–24 h. P values describe three-group comparisons; Holm correction covers 23 tests in this table only. No covariate adjustment is applied. Continuous and ordinal rows use Kruskal–Wallis tests; categorical rows use Pearson chi-square tests when expected counts are adequate, otherwise the specified Monte Carlo exact test. Full test identities, missingness and exact values are in Source Data. The six component scores and history-specific coverage denominators appear in the Supplement. Between-database characteristics from the preceding version are retained in Table S29.

Characteristic	Decrease ≥ 2 (n = 1600)	Net stable (n = 1282)	Increase ≥ 2 (n = 807)	P	Holm-adjusted P
Steroid-record category				7.28e-14	8.01e-13
No relevant order	1500 (93.8)	1093 (85.3)	687 (85.1)	—	—
Threshold category	84 (5.2)	152 (11.9)	96 (11.9)	—	—
Below threshold	16 (1.0)	37 (2.9)	24 (3.0)	—	—
Prior-admission diagnosis coverage	417 (26.1)	383 (29.9)	233 (28.9)	0.0633	0.4043
Documented heart failure	134 (8.4)	117 (9.1)	71 (8.8)	0.7747	1.0000
Documented chronic kidney disease	103 (6.4)	110 (8.6)	66 (8.2)	0.0730	0.4043
Documented chronic lung disease	80 (5.0)	85 (6.6)	58 (7.2)	0.0578	0.4043
Documented cancer	54 (3.4)	82 (6.4)	55 (6.8)	7.96e-05	7.96e-04
Documented cerebrovascular disease	72 (4.5)	72 (5.6)	46 (5.7)	0.2932	1.0000
Documented dementia	8 (0.5)	18 (1.4)	9 (1.1)	0.0389	0.3110
Documented diabetes	157 (9.8)	144 (11.2)	84 (10.4)	0.4639	1.0000

Values are median [IQR] or n (%); steroid category has one omnibus P value. Each group denominator includes all stays in that score-change group. History entries are positive records from completed prior admissions, not estimates of disease prevalence; absence of a code is not confirmed absence of disease. The available-history row makes coverage explicit. Scores, components and care are measured during the first 24 hours and are not all pretreatment covariates. The first and attained windows are 0–6 h and 18–24 h. P values describe three-group comparisons; Holm correction covers 23 tests in this table only. No covariate adjustment is applied. Continuous and ordinal rows use Kruskal–Wallis tests; categorical rows use Pearson chi-square tests when expected counts are adequate, otherwise the specified Monte Carlo exact test. Full test identities, missingness and exact values are in Source Data. The six component scores and history-specific coverage denominators appear in the Supplement. Between-database characteristics from the preceding version are retained in Table S29.

Table 2
MIMIC-IV outcomes by restricted-policy score change

Outcome	Decrease ≥ 2 (n = 1600)	Net stable (n = 1282)	Increase ≥ 2 (n = 807)	P	Holm-adjusted P
ICU death, 24–72 h	16/1600 (1.0%)	51/1282 (4.0%)	67/807 (8.3%)	1.29×10^{-18}	1.29×10^{-18}
Recorded death by 28 d	130/1600 (8.1%)	288/1282 (22.5%)	227/807 (28.1%)	2.27×10^{-40}	4.54×10^{-40}
Recorded death by 365 d	260/1600 (16.3%)	479/1282 (37.4%)	341/807 (42.3%)	3.00×10^{-52}	8.99×10^{-52}

Decrease: ≥2 points; net stable: −1 to + 1; increase: ≥2 points. Cells show events/group denominator (%). Mortality horizons are measured from ICU admission, conditional on hour-24 eligibility. P is the unadjusted three-group Pearson chi-square P value (df = 2); all expected counts exceed five and no Yates correction is used. Holm-adjusted P controls multiplicity across the three omnibus outcome comparisons in this table as one family. This adjustment was added during the exploratory revision, not prespecified. Neither column adjusts for covariates. Pairwise P values in Fig. 2b–d remain nominal and are outside this family. These recorded-event comparisons do not establish independent trajectory effects; mortality horizons overlap.

The path groups are defined from the first and attained scores; tests of those scores and their components are conditional descriptions, not evidence that independent classes have been discovered. History rows count positive diagnoses recorded during admissions ending before the index hospital admission, using each full group as the denominator. Lack of a prior code does not establish absence of disease. Ascertainment coverage and frequencies among those with observed prior diagnoses are reported separately. The fields do not measure frailty, functional dependence or premorbid cognition. Full component summaries appear in Table S30.

For the strict comparison, standardized mean differences describe balance before and after matching (Table S31). The standardizer is the pooled pre-match standard deviation, retained for the post-match comparison; binary indicators use the analogous pooled Bernoulli variance. Balance is interpreted from magnitudes and distributions, not P values. No balance statistic converts a prognostic comparison into a causal treatment contrast.

Associations at comparable attained scores

Exploratory logistic models compared score-change categories within attained integer scores 2–13, with attained-score fixed effects and adjustment for age, sex and three steroid-record categories. The historical steroid categories were no relevant order, an order below the local threshold, and orders meeting the threshold of 160 mg/day hydrocortisone equivalent within the inherited − 10 to + 24-hour window. These were record-derived covariates and were not interpreted as actual administered treatment effects.

A stricter illustrative comparison selected attained scores 2–9 and exact changes of − 3 versus 0. This range was chosen in the earlier exploratory analysis for cell support and retained through the timing audits. Matching without replacement was exact on attained score, sex, steroid category and first-window direct-input count, with an age difference ≤ 5 years. An assignment algorithm prioritized eligible pair counts and then smaller age and current-input-count differences. A further measurement sensitivity required direct GCS in both windows and identical current-input counts.

Models in all strict candidates adjusted for the same covariates; a second model added 0–6-hour records of selected sedatives and invasive ventilation. Sedative records included lorazepam, midazolam, ketamine, propofol and dexmedetomidine. These indicators capture recorded care and are not equivalent to sedation depth. Available Richmond Agitation–Sedation Scale (RASS) entries informed a separate descriptive neurological-context audit; the scale was developed and validated to assess agitation and sedation in adult ICU patients [13]. Because care and first-window scoring occurred within the same period, coefficient changes were interpreted as conditional model sensitivity, not as mediation or proof of causal confounding. Logistic confidence intervals used HC3 robust covariance estimates.

Matched risk differences used 5,000 pair-level bootstrap samples and percentile intervals, a resampling approach whose interpretation depends on the selected sampling unit and interval method [14]. Unadjusted group proportions were displayed with Wilson score 95% intervals [15]. For Fig. 2, all three pairs of groups were compared for each recorded-death endpoint using Pearson chi-square tests of two-by-two tables (one degree of freedom, without Yates continuity correction); all expected cell counts exceeded five. SOFA score distributions were compared for each pair within each six-hour window using two-sided asymptotic Mann-Whitney U tests with tie and continuity corrections, without pooling repeated measurements. Figure 2a summarizes the 12 score-distribution comparisons, with exact nominal P values in Supplementary Table S28; Fig. 2b–d displays nine mortality comparisons. These Fig. 2 comparisons have no covariate or multiplicity adjustment; omnibus comparisons with Holm adjustment are reported separately in Table 2. These nominal tests are conditional descriptive comparisons because the same first and attained scores define group membership; they do not establish independent trajectory classes. Discordant-pair odds ratios and two-sided exact McNemar tests were secondary descriptions. The exact P values used a binomial test of the two discordant-pair counts under equal probabilities. Figure 3 annotations report four decimal places. Figure 2 mortality comparisons and Supplementary Table S28 use four decimal places at or above 0.001 and scientific notation with three significant digits below 0.001. Full-precision values are retained in the source tables. We also restricted pairs to those in which both members were alive at day 28 and counted subsequent recorded deaths through day 365. This selected survivor analysis was exploratory and potentially imprecise.

Prediction comparison and transport

Three fixed logistic models included age, sex and, sequentially, the attained score; the attained and first scores; or all four scores. Standardization was fitted within each training split. The regularization parameter was C = 1, with a maximum of 1,000 iterations. MIMIC-IV performance used five-fold stratified

out-of-fold predictions with a fixed seed. Each model was subsequently fitted on all MIMIC-IV observations and applied unchanged to eICU, following the principle that external validation evaluates predictions without refitting the model [16]. There was no external hyperparameter selection or recalibration.

Performance assessment covered discrimination, probability error and calibration rather than relying on a single metric [17]. Metrics were the area under the receiver-operating-characteristic curve (AUC), average precision, Brier score and mean predicted risk compared with observed event frequency. Average precision complemented ROC-based discrimination because precision–recall measures help characterize performance when outcome classes are imbalanced [18]. Calibration-in-the-large (CITL) was estimated with the predicted logit as a fixed offset; joint calibration intercepts and slopes were fitted separately. Ideal values are zero for CITL and joint intercept and one for slope. These diagnostic fits did not recalibrate predictions. Calibration was considered separately from discrimination because useful ranking does not ensure accurate estimated probabilities [19]. Model differences in eICU used 1,000 paired bootstrap samples of its 76 hospitals, retaining observations within sampled hospitals and comparing the same sampled observations across models. These intervals condition on the fitted MIMIC-IV models and omit development-sample uncertainty. A new completion analysis used seed 20261011 for 1,000 hospital-cluster resamples of absolute metrics and paired model differences. Demographic results described coverage and uncertainty. A fixed four-knot restricted cubic spline for attained SOFA (knots 2, 6, 10, 14), with or without linear first-window SOFA, used the same folds and penalties. These were prespecified for the completion analysis, after the validation dataset had already been examined; they were not fresh untouched validation. Except for the table-specific Holm-adjusted P columns in Tables 1, 2 and S29, P values and confidence intervals were nominal, without multiplicity adjustment. Analyses used Python with DuckDB, pandas, NumPy, statsmodels, scikit-learn and SciPy. The reporting map also addresses TRIPOD + AI items for the prediction component [20].

Computational assistance

Large-language-model assistance was used for analysis-code preparation, manuscript drafting and language editing. Numerical results were obtained by executing the documented statistical scripts. Responsibility for the analyses and manuscript rests with the authors.

Results

Cohorts and measurement coverage

The raw databases contained 94,458 MIMIC-IV ICU stays and 200,859 eICU stays. The inherited pipeline selected 4,069 and 1,647 older sepsis stays, respectively (Table S15). Of 4,069 retrospectively source-selected MIMIC-IV stays, 365 ended by hour 24 and a further 15 had recorded death by hour 24 despite a later recorded ICU exit, leaving 3,689 landmark stays. Of 1,647 eICU stays, 252 ended by hour 24, leaving

1,395 stays from 1,334 people at 76 hospitals. There were 134 and 64 subsequent ICU deaths by hour 72, respectively (Fig. 1).

Median age was 75 [interquartile range 70–82] years in MIMIC-IV and 77 [70–84] in eICU. With the restricted recording policy, median first-window scores were 8 [4–10] and 3 [1–6], whereas attained scores were 5 [3–8] and 4 [3–7]. Median first-window direct-input counts were 7/8 and 4/8; direct GCS was available in 90.4% and 76.9% of first windows and 84.5% and 57.3% of attained windows (Table S29). These measurements document source differences, without assigning them solely to disease severity or recording practices.

Table 1 describes the care and measurement context within score-change groups. First-window sedative records occurred in 87.7%, 44.6% and 34.9% of the decrease, net-stable and increase groups, respectively; invasive ventilation was recorded in 87.7%, 45.5% and 33.5%. These indicators were observed during the same period as the first score and cannot all be treated as antecedent confounders.

Documented prior-admission diagnosis history was available for 1,033/3,689 MIMIC-IV stays (28.0%). Missing prior codes were not classified as absence of comorbidity. Source-time anomalies were confirmed against raw records; an administrative exit-time sensitivity changed the short-term risk set to 3,682 stays and 132 classified ICU deaths, with four uncertain exits. These checks do not replace the primary source-specific endpoint (Tables S16–S17; Source Data).

Early SOFA paths and recorded outcomes

Under the restricted recording policy, the 3,689 MIMIC-IV stays comprised 1,600 with a decrease of at least two points, 1,282 with a net change from − 1 to + 1, and 807 with an increase of at least two points. Four-window median and interquartile summaries describe these classified samples (Fig. 2a); because the categories use their first and attained scores, the plotted separation is not independent evidence of new trajectory classes.

The decrease group had the lowest recorded mortality at each reported horizon. Deaths during ICU hours 24–72 were 16/1,600 (1.0%), 51/1,282 (4.0%) and 67/807 (8.3%) in the decrease, net-stable and increase groups, respectively. Recorded mortality was 130/1,600 (8.1%), 288/1,282 (22.5%) and 227/807 (28.1%) by day 28, and 260/1,600 (16.3%), 479/1,282 (37.4%) and 341/807 (42.3%) by day 365 (Fig. 2b–d; Table 2). The three-group comparisons remained significant after Holm correction across the three outcomes (adjusted $P = 1.29 \times 10^{-18}$, 4.54×10^{-40} and 8.99×10^{-52}, respectively). These tests describe group differences without covariate adjustment; the overlapping horizons are not independent replications.

Different paths to the same attained score

The exact − 3 versus 0 comparison included 600 candidates: 235 with a three-point decrease and 365 with no net change. Matching yielded 116 pairs with identical attained-score distributions. Recorded deaths at days 28, 90 and 365 were 16 versus 33, 22 versus 41, and 26 versus 45, respectively; the

corresponding nominal two-sided exact paired P values were 0.0095, 0.0066 and 0.0094. At day 365, recorded mortality was 22.4% versus 38.8%, a paired difference of − 16.4 percentage points (95% CI − 27.6 to − 5.2). Thus, the two selected groups had different recorded outcomes despite reaching the same attained-score distribution. Figure 3 displays cumulative recorded-event proportions with a fixed denominator, rather than censoring-validated survival estimates. Paired odds ratios and exact tests are retained in Table S5 and Supplementary Results.

Early sedation and invasive ventilation remained imbalanced: each was recorded in 90 decrease-group and 48 net-zero-group members of the 116 pairs. Post-match standardized mean differences were 0.848 and 0.841, respectively, despite exact balance in the matched score and categorical variables (Table S31). In the full 600-person candidate sample, the 365-day mortality OR was 0.49 (95% CI 0.32–0.76) before and 0.86 (0.52–1.42) after adding these concurrent care indicators. The two models were fitted in the same candidates; their difference is a conditional sensitivity result, not evidence that sedation caused the association. Among the 72 pairs in which both members were alive at day 28, subsequent deaths through day 365 were 6 versus 10 (nominal exact paired P = 0.4545). The cumulative one-year difference therefore did not establish continued separation in this selected survivor sample. Adjusted estimates, recording-policy comparisons and survivor results are presented in Supplementary Figure S2 and Tables S3–S6 and S13.

In the additional entry-service sensitivity, 1,766 of the archived 3,689 MIMIC-IV stays were retained. This conditional subset had 322 strict candidates (81 with ΔSOFA − 3 and 241 with ΔSOFA 0). The same matching algorithm produced 42 pairs, with recorded 365-day deaths of 16 versus 18 and a paired risk difference of − 4.8 percentage points (95% CI − 26.2 to 16.7). Candidate-level ORs were 1.05 (0.59–1.87) before and 1.47 (0.77–2.80) after adding early care. This analysis changed the selected population and was imprecise; it did not recover all patients excluded by the inherited pipeline. It shows that the original matched association cannot be assumed to generalize across service-selection rules (Table S27).

Opposing component changes despite zero total change

A net-zero total frequently concealed changes in opposite directions. Among stays with ΔSOFA exactly 0, 248/493 in MIMIC-IV (50.3%) and 107/323 in eICU (33.1%) had at least one component decrease and one component increase. Only the remaining 245 and 216 stays had unchanged values in all six recorded components. Figure 4 shows the signed component profiles; every column sums to zero, and one stay can contribute changes in several components. This finding describes information lost when component changes are reduced to their net total.

Under the archived policy, opposing-component and unchanged-component groups had 365-day recorded deaths of 86/248 and 100/245 in MIMIC-IV. The base OR of 0.51 (0.33–0.79) became 0.68 (0.42–1.08) after early-care adjustment. eICU had only six ICU deaths among its 323 net-zero stays. Thus the component profiles provide a descriptive measurement finding, without establishing a reproducible independent prognostic phenotype (Table S21).

Measurement sensitivity and external prediction

Recording-rule restrictions changed attained scores and category membership in both databases, and changed membership of the strict matched comparison. Complete reassignment counts and policy-specific results are reported in Supplementary Figure S1 and Tables S4 and S9–S12. These restrictions are database-field sensitivities rather than a verified reconstruction of bedside information availability.

The combined zero-inclusive, non-carried urine and coherent-GCS policy changed broad classification in 42/3,689 MIMIC-IV and 32/1,395 eICU stays. Across the four subsequent scoring variants, rematching yielded 117–119 pairs; 365-day paired differences ranged from − 17.1 to − 14.5 percentage points, with all four nominal 95% intervals below zero (Table S19). This consistency was specific to the matched comparisons: all five candidate-level policies had care-adjusted intervals spanning one (Table S20). The combined policy yielded 607 candidates and 119 pairs, with deaths of 26 versus 44. Requiring direct GCS and at least five direct inputs in both endpoint windows left 24 pairs, with deaths of 5 versus 9 and a difference interval spanning zero. Same-timestamp GCS totals differed from independently minimized component totals in 59 first and 19 attained windows; recorded zero-only urine occurred in 102 MIMIC-IV and 68 eICU windows across the day (Tables S18–S20 and S25–S26). These overlapping rematched samples are sensitivity analyses, not independent replications.

In MIMIC-IV, adding the first-window score increased out-of-fold AUC from 0.8393 to 0.8457 (difference 0.0063, 95% interval 0.0005–0.0121). The corresponding average-precision and Brier difference intervals included zero. In eICU, the locked attained-score model had AUC 0.825 (95% hospital-cluster interval 0.781–0.866), average precision 0.292 (0.214–0.396) and Brier score 0.03947 (0.02976–0.04789). Mean predicted risk was 2.24%, compared with 4.59% observed risk; CITL was 0.850 (0.599–1.034), and the joint calibration slope was 1.013 (0.891–1.203). Adding the first score changed external AUC by − 0.0062 (− 0.0191 to 0.0033) and Brier score by − 0.00048 (− 0.00109 to 0.00002). The one-record-per-person sensitivity retained 1,334 people and 59 events, with AUC 0.821 and persistent average-risk underestimation. The spline comparison did not establish an incremental discrimination gain. Tables S22–S24, Figure S5 and Source Data report absolute precision, calibration and subgroup results. The small internal AUC increment did not establish external incremental benefit or clinical utility.

Discussion

In this study of older adults with sepsis, patients with an early decrease in SOFA had lower recorded mortality than those with stable or increasing scores. The difference was also present among matched patients who reached the same score at 24 hours. Early sedative and ventilation records were more frequent in the decrease group, and including these records in the candidate-level model moved the association towards the null. A complementary finding was that unchanged total scores often combined improving and worsening subscores. These observations show why the meaning of a first-day SOFA change depends on both its component changes and the clinical setting in which the measurements were obtained.

Previous multicohort research identified four 72-hour SOFA trajectory groups with different patterns of organ dysfunction and mortality [21]. Predicted trajectories have subsequently been used to stratify corticosteroid target-trial emulations [22]. Our analysis addresses an earlier question: whether the score reached at 24 hours adequately describes the preceding course. The 116-pair comparison showed a 16.4-percentage-point difference in recorded mortality by day 365 despite identical attained-score distributions. Fixing the attained score makes the two-point change a comparison of starting scores; the four-window analysis then tests whether the intervening measurements contribute further information. The observed association supports examining the preceding course alongside the current score, while its dependence on early care helps explain which patients contribute to that association.

The neurological component offers a clinically plausible explanation for part of this pattern. The early decrease group frequently entered the ICU with high CNS subscores and contemporaneous sedative or ventilation records. GCS can change as consciousness improves, sedative exposure changes or airway management alters the recorded verbal response. A database study comparing SOFA with and without GCS found that its contribution to mortality prediction varied across settings [23]. In a separate automated-SOFA validation study, GCS during ongoing sedation was treated as not assessable and algorithmic scores were compared with independent clinician assessments [24]. In our strict candidates, the day-365 odds ratio changed from 0.49 to 0.86 after early care records were included. The expanded model estimates the score-change association conditional on these concurrent care records. Illness severity, care decisions and GCS measurement occurred within the same six-hour period; distinguishing their contributions requires more detailed temporal and clinical information. For clinical interpretation, the reason for a low GCS and its relation to sedative exposure are relevant to the meaning of a subsequent SOFA decrease.

The component analysis provides a second reason to examine more than the total. Dusaj and colleagues found different organ-component patterns among patients sharing the same SOFA score, with similar overall prognosis across the component-based groups [25]. We extended this comparison to change over time: among stays with exactly zero total change, opposing subscore changes occurred in half of the MIMIC-IV cohort and one third of the eICU cohort. A fall in one subscore can therefore conceal a rise in another. Describing these patients as net stable is arithmetically correct, but a clinical review still needs to identify which component improved and which worsened. The different proportions across databases should be interpreted in the context of their case mix and scoring inputs.

Recording practices also shaped the reconstructed course. Research using electronic health records has shown that the timing and frequency of testing can carry information about prognosis through the processes of care [26]. In our data, timestamp restrictions changed the attained score more often than they changed the broad score-change category. Most reassignments were between adjacent categories, indicating that small numerical differences can matter near classification boundaries. Event time and entry time describe different aspects of a record: a later entry may refer to information already available at the bedside. The supporting ventilation audit, for example, found contemporaneous tube and ventilator-mode documentation for many intervals whose procedure records were stored later.

Reproducible dynamic scoring therefore requires both an explicit time rule and attention to how the underlying clinical observation is recorded.

The longer follow-up places the early mortality difference in context. The matched groups had already separated by day 28, and the cumulative difference remained visible at day 365. Among the 72 pairs in which both patients survived to day 28, subsequent deaths were six versus ten. These data are compatible with a difference established mainly during the acute illness; they provide limited information about additional risk among survivors. In older patients, the interpretation of later mortality would also benefit from premorbid function, frailty and treatment preferences. Prior-admission diagnoses were documented for only 28.0% of the MIMIC-IV cohort, leaving these aspects of baseline health incompletely measured.

The prediction analysis addressed a separate, shorter-term endpoint. Adding earlier SOFA values did not improve external discrimination for ICU death during hours 24–72. The attained-score model retained an AUC of 0.825, but its mean predicted risk was 2.24% compared with an observed event proportion of 4.59%. External evaluation of an implemented sepsis alert has similarly demonstrated the need to test performance in the intended population [27]. Our findings suggest that adjustment of absolute risk deserves attention before adding more score measurements to the tested model. With 64 external events, precision remains important when interpreting differences between models, particularly their calibration [28]. A future clinical use would also need a defined decision and an evaluation of the consequences of acting on its predicted risks, as formalized by decision-curve analysis [29].

The study combines a comparison at the same attained score with organ-level decomposition and explicit audits of timestamps and early care. Keeping each database's risk set fixed across recording policies allowed us to identify changes attributable to score reconstruction within that cohort. External testing used frozen development models and hospital-level resampling, and aggregate source data make the main comparisons reproducible. These features help connect the mortality findings to the records from which the scores were constructed.

The principal limitations concern cohort selection and measurement. The cohorts originated from a corticosteroid study and retained source-specific infection, service and prior-treatment restrictions; some eligibility information was recorded after hour 24. Selection based on information arising during care can affect observed associations [30]. The estimates consequently apply to the included population, and the admission-service sensitivity produced a smaller, imprecise mortality difference. The local six-hour scores used adapted support and urine rules, forward carrying and zero assignment when missing inputs did not trigger a scoring condition. Their clinical agreement with independently assessed scores remains unmeasured. Mortality reflects recorded events, with date precision and linkage coverage limiting ascertainment. The analyses were exploratory, and the selected score ranges, exact-change contrast and repeated model comparisons increase uncertainty beyond the table-specific multiplicity corrections. The eICU evaluation concerns short-term ICU death and provides no direct validation of the one-year association.

For older adults assessed after the first ICU day, the attained SOFA score is more informative when read with the starting score, the direction of individual component changes and the accompanying neurological and care records. The observed mortality differences make this context clinically relevant. A stable total can coexist with substantial component change, and a falling total can combine physiological improvement with changes in sedation, support and documentation. Future evaluations should retain this component-level information when testing early SOFA changes in independently assembled cohorts.

Conclusions

Early SOFA decreases were associated with lower recorded mortality among selected older adults with sepsis, including patients reaching the same attained score. The association varied with early care records, and net-zero totals often concealed opposing component changes. These findings support interpreting first-day SOFA change alongside the starting score, its organ components and the clinical context of measurement.

Abbreviations

AUC
area under the receiver operating characteristic curve
CI
confidence interval
CITL
calibration-in-the-large
CNS
central nervous system
GCS
Glasgow Coma Scale
ICU
intensive care unit
MIMIC
Medical Information Mart for Intensive Care
OR
odds ratio
RASS
Richmond Agitation-Sedation Scale
RECORD
REporting of studies Conducted using Observational Routinely-collected health Data
SOFA
Sequential Organ Failure Assessment
STROBE

Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate

This study was a secondary analysis of deidentified records from MIMIC-IV and the eICU Collaborative Research Database. The Beth Israel Deaconess Medical Center Institutional Review Board approved sharing of MIMIC-IV and waived informed consent. The eICU database was deidentified in accordance with the Safe Harbor provisions of the Health Insurance Portability and Accountability Act. Yu Song accessed the databases as a credentialed PhysioNet user under the applicable data-use agreements. No separate local ethics approval was obtained for this analysis, which involved no participant contact or collection of new identifiable data.

Consent for publication

Not applicable. The manuscript and its journal additional files contain aggregate results and no individual participant details.

Competing interests

The authors declare that they have no competing interests.

Funding

Author Contribution

Lei Wang conceived the study and analyzed the data. Ying-chao Liu contributed clinical interpretation and critical revision. Yu Song supervised the work and drafted the manuscript. All authors read and approved the final manuscript and agreed to the author order.

Acknowledgement

The authors acknowledge the investigators, clinicians, data scientists and institutions involved in the development and maintenance of MIMIC-IV and the eICU Collaborative Research Database. We thank the Massachusetts Institute of Technology Laboratory for Computational Physiology, Beth Israel Deaconess Medical Center, PhysioNet and the eICU Research Institute for making these deidentified critical care datasets available to the research community.

Data Availability

MIMIC-IV version 3.1 (https://doi.org/10.13026/kpb9-mt58) and the eICU Collaborative Research Database version 2.0 (https://doi.org/10.13026/C2WM1R) are available through PhysioNet to credentialed researchers who meet the required training and data-use conditions. Aggregate source data, analysis code and reproduction instructions are included in Additional file 2. Individual-level extracts and restricted derivatives, including match identifiers, fitted model parameters and individual predictions, are excluded from the journal additional files and remain subject to the source data-use agreements.

References

1. Singer, M. et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315** (8), 801–810. 10.1001/jama.2016.0287 (2016).
2. Rudd, K. E. et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* **395** (10219), 200–211. 10.1016/S0140-6736(19)32989-7 (2020).
3. Iwashyna, T. J. et al. Long-term Cognitive Impairment and Functional Disability Among Survivors of Severe Sepsis. *JAMA* **304** (16), 1787–1794. 10.1001/jama.2010.1553 (2010).
4. Yende, S. et al. Long-Term Quality of Life Among Survivors of Severe Sepsis: Analyses of Two International Trials. *Crit. Care Med.* **44** (8), 1461–1467. 10.1097/CCM.0000000000001658 (2016).
5. Vincent, J. L. et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Crit. Care Med.* **26** (11), 1793–1800. 10.1097/00003246-199811000-00016 (1998).
6. Ferreira, F. L. et al. Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. *JAMA* **286** (14), 1754–1758. 10.1001/jama.286.14.1754 (2001).
7. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*. **10** (1), 1. 10.1038/s41597-022-01899-x (2023).
8. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data*. **5** (1), 180178. 10.1038/sdata.2018.178 (2018).
9. von Elm, E. et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLoS Med.* **4** (10), e296. 10.1371/journal.pmed.0040296 (2007).
10. Benchimol, E. I. et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med.* **12** (10), e1001885. 10.1371/journal.pmed.1001885 (2015).
11. Lambden, S. et al. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. *Crit. Care*. **23** (1), 374. 10.1186/s13054-019-2663-7 (2019).

12. Ranzani, O. T. et al. Development and Validation of the Sequential Organ Failure Assessment (SOFA)-2 Score. *JAMA* **334** (23), 2090–2103. 10.1001/jama.2025.20516 (2025).
13. Sessler, C. N. et al. The Richmond Agitation–Sedation Scale: Validity and Reliability in Adult Intensive Care Unit Patients. *Am. J. Respir. Crit Care Med.* **166** (10), 1338–1344. 10.1164/rccm.2107138 (2002).
14. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* **19** (9), 1141–1164. 10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F (2000).
15. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **22** (158), 209–212. 10.1080/01621459.1927.10502953 (1927).
16. Riley, R. D. et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* **384**, e074820. 10.1136/bmj-2023-074820 (2024).
17. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21** (1), 128–138. 10.1097/EDE.0b013e3181c30fb2 (2010).
18. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. **10** (3), e0118432. 10.1371/journal.pone.0118432 (2015).
19. Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17** (1), 230. 10.1186/s12916-019-1466-7 (2019).
20. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378. 10.1136/bmj-2023-078378 (2024).
21. Xu, Z. et al. Sepsis subphenotyping based on organ dysfunction trajectory. *Crit. Care*. **26** (1), 197. 10.1186/s13054-022-04071-4 (2022).
22. Rajendran, S. et al. Multicenter target trial emulation to evaluate corticosteroids for sepsis stratified by predicted organ dysfunction trajectory. *Nat. Commun.* **16** (1), 4450. 10.1038/s41467-025-59643-z (2025).
23. Wang, L. et al. SOFA in sepsis: with or without GCS. *Eur. J. Med. Res.* **29** (1), 296. 10.1186/s40001-024-01849-w (2024).
24. Helleberg, J. et al. Automated Calculation of Sequential Organ Failure Assessment (SOFA) Score in the Intensive Care Unit: Algorithm Development, Validation, and Association With 30-Day Mortality. *Acta Anaesthesiol. Scand.* **70** (3), e70205. 10.1111/aas.70205 (2026).
25. Dusaj, N. et al. Lost in a number: concealed heterogeneity within the sequential organ failure assessment (SOFA) score. *Crit. Care*. **28** (1), 6. 10.1186/s13054-023-04782-2 (2024).
26. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479. 10.1136/bmj.k1479 (2018).

27. Wong, A. et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern. Med.* **181** (8), 1065–1070. 10.1001/jamainternmed.2021.2626 (2021).
28. Riley, R. D. et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40** (19), 4230–4251. 10.1002/sim.9025 (2021).
29. Vickers, A. J. & Elkin, E. B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med. Decis. Making*. **26** (6), 565–574. 10.1177/0272989X06295361 (2006).
30. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A Structural Approach to Selection Bias. *Epidemiology* **15** (5), 615–625. 10.1097/01.ede.0000135174.63482.43 (2004).

Figures

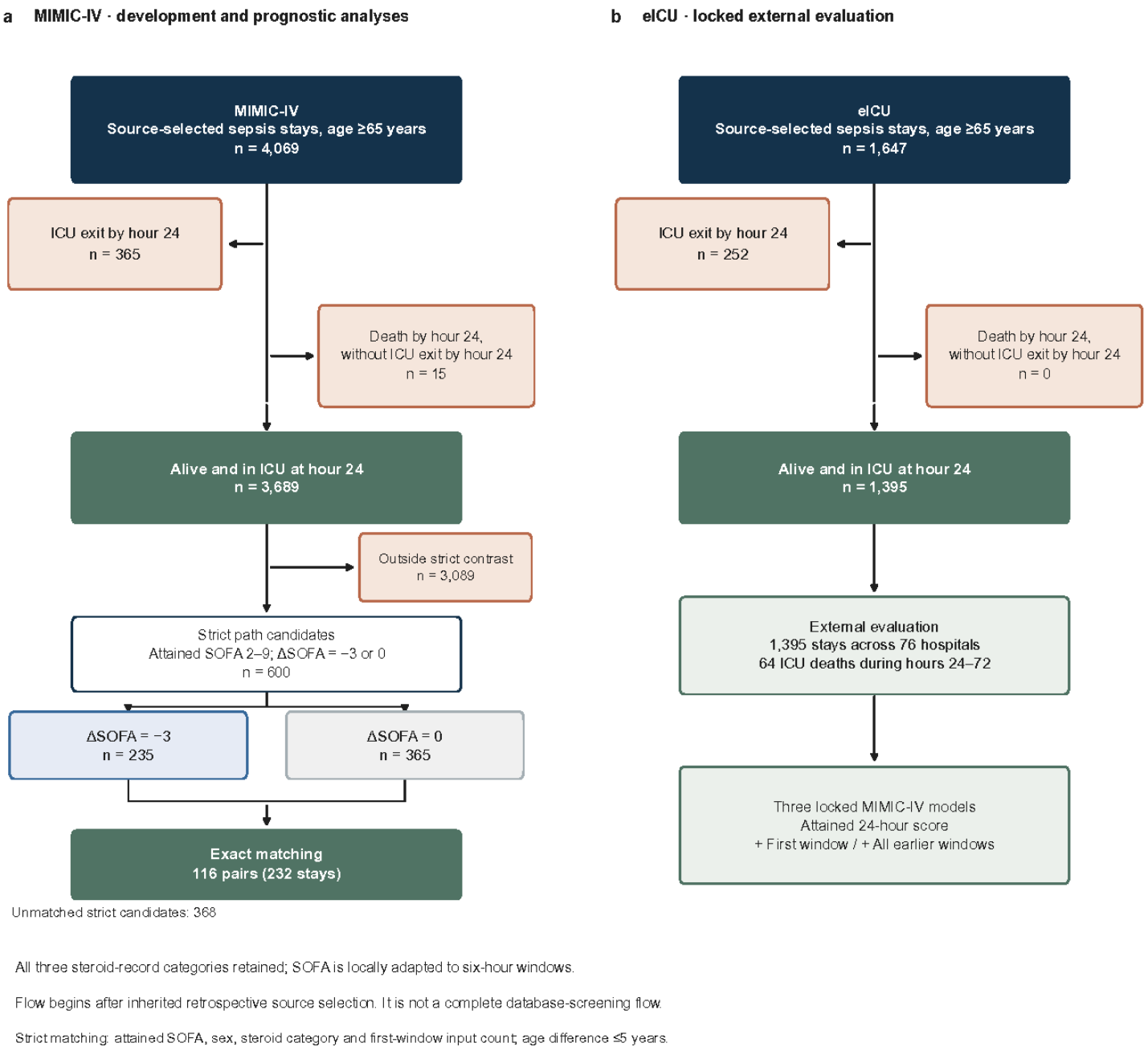


Figure 1

Source-selected cohorts and analysis sets

a, MIMIC-IV source-selected sepsis stays in adults aged ≥65 years, mutually exclusive exclusions before the 24-hour landmark, and the strict same-attained-score comparison. The 3,689-stay landmark cohort provides 600 strict candidates and 116 matched pairs. The 368 unmatched candidates remain in the 600-person regression analyses in the Supplement. Matching is exact on attained score, sex, steroid-record category and first-window direct-input count, with age difference ≤5 years. **b,** eICU landmark selection and locked external model evaluation. The flow begins after inherited retrospective source

selection, rather than at all database admissions. All three steroid-record categories are retained. Counts are descriptive. Sources: Tables S2 and S4.

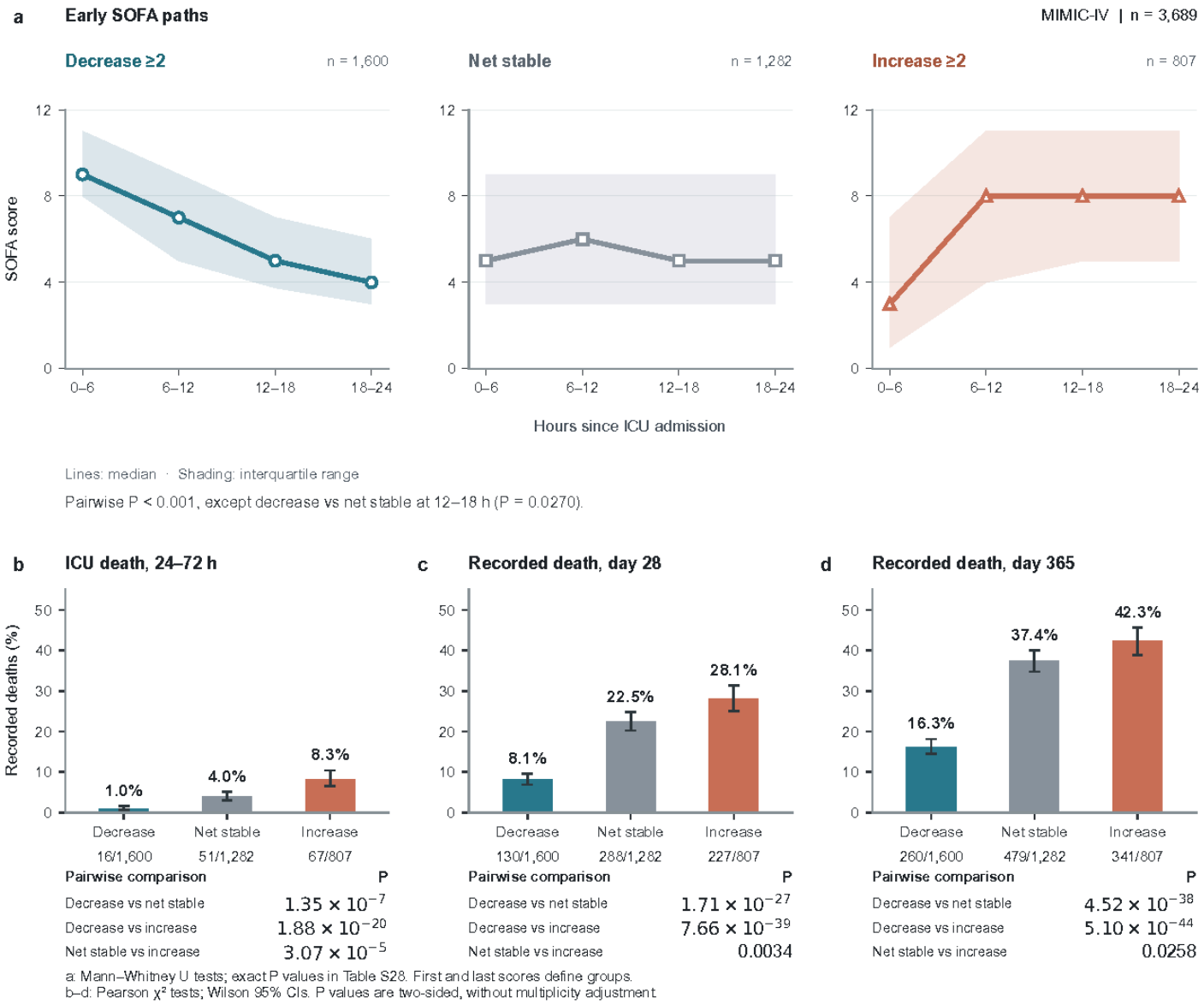
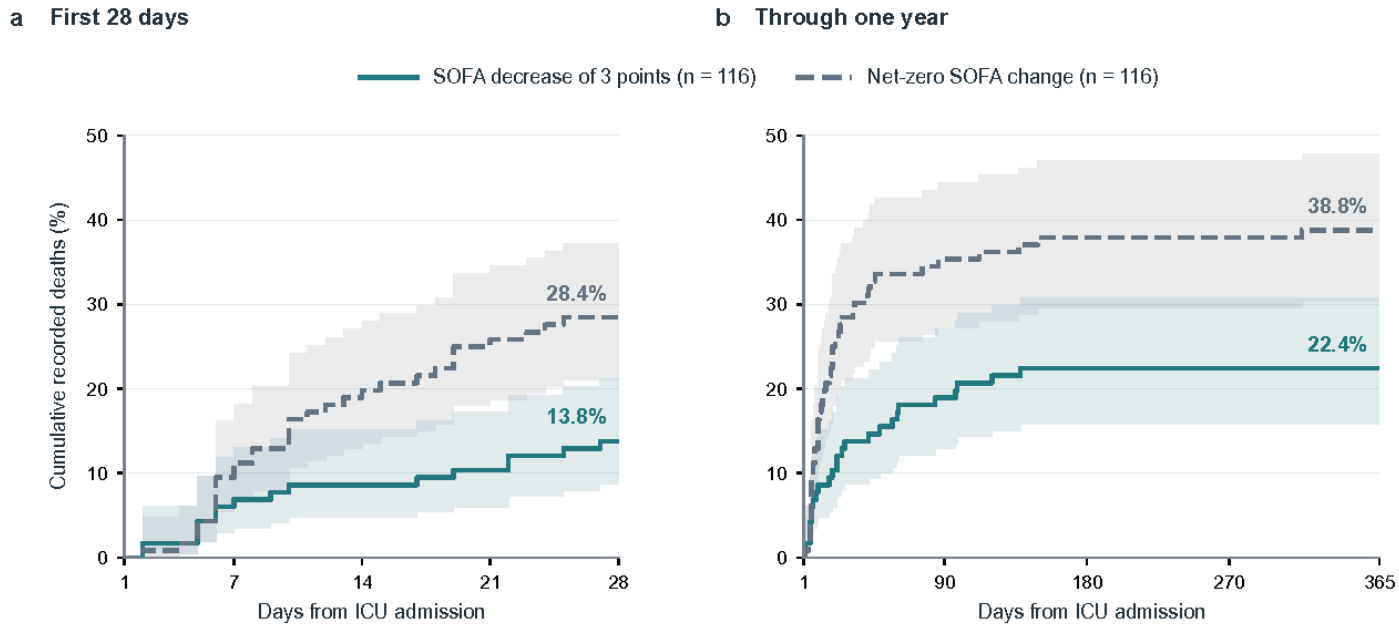


Figure 2

Early SOFA paths and subsequent recorded mortality

a, SOFA medians and interquartile ranges in four six-hour windows among 3,689 MIMIC-IV stays alive and in ICU at hour 24. Aligned panels share axes; shading represents interquartile ranges. Groups use the 18–24-hour minus 0–6-hour score: decrease ≤−2 (n=1,600), net stable −1 to +1 (n=1,282), and increase ≥2 (n=807). Within each window, two-sided asymptotic Mann–Whitney U tests compare each pair of groups, with tie and continuity corrections. Eleven of 12 comparisons have P<0.001; decrease versus net stable at 12–18 hours has P=0.0270. Exact values and group quantiles appear in Table S28 and Source Data. Each stay contributes once per window. Comparisons are descriptive because first and last scores define the groups.

b–d, Recorded ICU death during hours 24–72 and recorded death by days 28 and 365. Bars show group proportions with Wilson binomial 95% intervals; labels give percentages and events/denominators. Mortality panels share a vertical scale and zero baseline. Each panel names three pairwise comparisons, tested using two-sided Pearson chi-square tests of 2×2 tables without Yates correction. P values have no covariate or multiplicity adjustment. Overlapping mortality horizons are measured from ICU admission, conditional on hour-24 eligibility. Absence of a death record does not establish complete vital-status ascertainment. Physiological storage times are restricted to hour 24 while original support intervals are retained. The scores are locally adapted six-hour SOFA. Sources: audited scored windows, Table 2, Table S28 and accompanying aggregate data.



Recorded deaths, n/N (%)	Day 28	Day 90	Day 365
SOFA decrease of 3 points	16/116 (13.8)	22/116 (19.0)	26/116 (22.4)
Net-zero SOFA change	33/116 (28.4)	41/116 (35.3)	45/116 (38.8)
Paired P value	**0.0095**	**0.0066**	**0.0094**

Day-365 paired difference: −16.4 percentage points (95% CI −27.6 to −5.2).

Shading: pointwise 95% CI. P: exact McNemar test at the stated day; no multiplicity adjustment.

Conditional on ICU hour-24 eligibility; recorded outcomes with a fixed denominator, not verified complete follow-up.

Figure 3

Recorded mortality over time in matched patients attaining the same SOFA

a,b, Cumulative proportions with recorded death through ICU day 28 and day 365 in the same 116 MIMIC-IV matched pairs. The groups have exact SOFA changes of −3 and 0 between 0–6 h and 18–24 h and identical attained-score distributions. Each curve is computed from recorded death offsets using a fixed denominator of 116, conditional on survival and ICU stay at hour 24. Daily aggregation places an

event at the first displayed integer day on or after its recorded offset; there is no interpolation between the three reported endpoints. Shading shows pointwise Wilson 95% intervals for recorded-event proportions, not simultaneous confidence bands. At hour 24, curves and bands are anchored at zero by the survival eligibility rule. These descriptive curves are not censoring-validated Kaplan–Meier estimates, and absence of a death record does not confirm survival. The endpoint summary reports deaths and two-sided exact McNemar P values calculated from 11/28, 13/32 and 15/34 discordant pairs, without multiplicity adjustment. P values test the stated endpoints, not whole curves. The day-365 paired difference is decrease minus net-zero mortality; its previously audited percentile interval uses 5,000 fixed-pair bootstrap resamples (seed 20261007), excluding match-selection uncertainty. The panels overlap in time; the horizons are not independent replications. Detailed care-balance, adjusted-model and day-28 landmark results are provided in the supplement. The figure does not establish a causal or independent recovery effect.

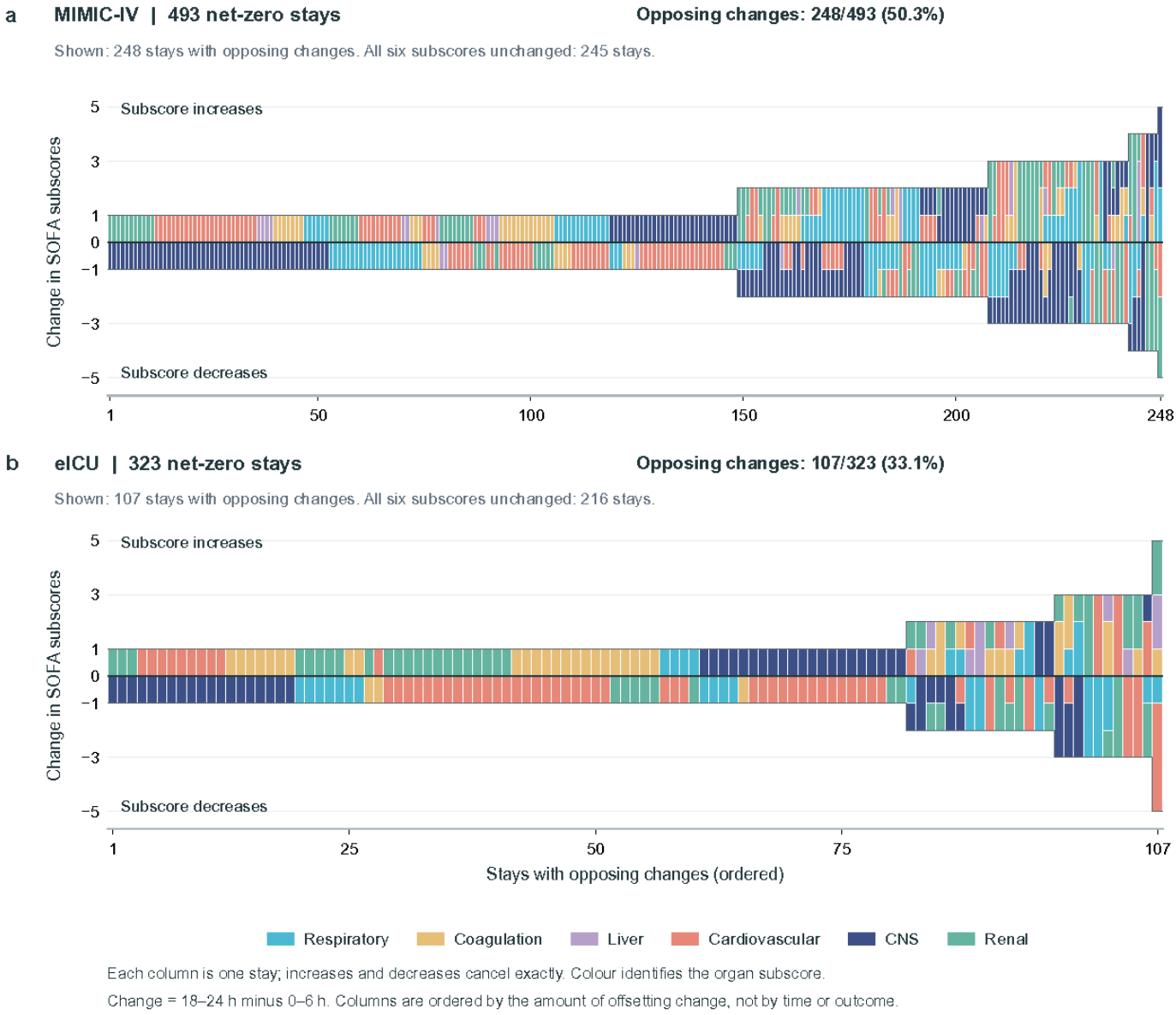


Figure 4

Opposing organ-subscore changes concealed by net-zero total SOFA

a,b, Individual profiles for the 248 MIMIC-IV and 107 eICU stays with opposing subscore changes among 493 and 323 stays, respectively, with exactly zero total SOFA change. Each column represents one stay. Coloured segments show the changes in six organ subscores between 0–6 h and 18–24 h. Positive segments indicate subscore increases; negative segments indicate subscore decreases. Their signed sum is zero in every column. The plots use identical vertical scales. Profiles are ordered by the total increasing points, then CNS change direction and the full component pattern; this ordering is descriptive, not time, an inferred cluster or a prognostic gradient. All observed opposing-change patterns are included with their exact frequencies. The 245 MIMIC-IV and 216 eICU stays with unchanged values in all six components are reported above the panels but are not drawn as zero-height columns. The headings retain the full net-zero denominators. No hypothesis tests or model estimates are displayed. MIMIC-IV uses the existing physiologic availability-at-24-h policy with retrospectively ascertained support; eICU uses its entry/revision gate. Source-specific scoring policies retain limitations and are not asserted to be fully harmonised. Subscore changes do not establish true organ recovery, deterioration or a causal mechanism. Full joint-pattern counts and display-order profiles are supplied as aggregate Source Data.

Supplementary Files

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

- Additionalfile1.pdf