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.

SOFA-2 versus SOFA for mortality risk assessment in ICU patients with acute cholangitis: a retrospective cohort study using MIMIC-IV and eICU

Yunxiao LYU
Dongyang People's Hospital
Bin Wang
Dongyang People's Hospital
Zhong Lyu
dylz3399@163.com

Dongyang People's Hospital

Research Article

Keywords: acute cholangitis, SOFA-2, SOFA, intensive care unit, mortality, comparative evaluation, MIMIC-IV

Posted Date: August 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10288690/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

Acute cholangitis (AC) can progress rapidly to organ failure and death. The Sequential Organ Failure Assessment-2 (SOFA-2) is a recently revised organ-dysfunction score, but whether it adds disease-specific value beyond SOFA in AC is unknown. We compared SOFA-2 with SOFA for mortality risk assessment in ICU patients with AC.

Methods

We conducted a retrospective cohort study using MIMIC-IV v3.1 as the primary dataset and eICU for cross-database assessment. Adults with AC meeting Tokyo Guidelines 2018 diagnostic criteria were included. Day-1 SOFA-2 and SOFA were reconstructed over the first 24 h after ICU admission. The primary outcome was ICU mortality; secondary outcomes were in-hospital and 28-day mortality.

Results

The MIMIC-IV ICU cohort included 644 patients, with 71 ICU deaths, 102 in-hospital deaths, and 133 deaths within 28 days. SOFA-2 showed good discrimination for ICU mortality (AUROC, 0.834; 95% CI, 0.777–0.883), but its performance was similar to SOFA (AUROC, 0.838; 95% CI, 0.784–0.886; difference, − 0.005; 95% CI, − 0.020 to 0.010; P = 0.514). For in-hospital mortality, AUROCs were 0.786 for SOFA-2 and 0.784 for SOFA. For 28-day mortality, AUROCs were 0.760 and 0.756, respectively. Calibration, reclassification and decision-curve analyses did not show a clinically meaningful advantage for SOFA-2. ICU mortality increased across SOFA-2 strata, from 1.9% at scores 0–2 to 27.9% at scores ≥ 9. In eICU, SOFA-2 and SOFA again showed similar discrimination (AUROC, 0.833 vs 0.835), and mortality increased from 1.2% to 22.9% across the same score strata.

Conclusions

SOFA-2 stratified mortality risk in ICU patients with AC, but it did not outperform SOFA. These findings support SOFA-2 as a feasible disease-specific research measure, but they do not support replacing SOFA with SOFA-2 for clinical risk assessment in AC without further independent validation.

Introduction

Acute cholangitis (AC) is a life-threatening biliary infection caused by ductal obstruction and bacterial contamination.[1–4] Elevated intraductal pressure drives bacteraemia, systemic inflammation and progressive organ failure. Most patients respond to antibiotics and biliary drainage, but a clinically important minority deteriorates rapidly and requires intensive care. In this group, early risk stratification informs drainage urgency and the intensity of organ support.

The Tokyo Guidelines 2018 (TG18) provide the standard framework for diagnosing and grading AC severity.[1, 2] TG18 Grade III identifies patients with overt organ dysfunction and supports urgent source control, but the grading system was not designed to generate a continuous mortality-risk estimate and does not directly compare organ-dysfunction scores.

The Sequential Organ Failure Assessment (SOFA) score quantifies dysfunction across neurological, respiratory, cardiovascular, hepatic, renal and haemostatic domains and underpins the Sepsis-3 definition.[5, 6] SOFA-2 was recently developed to update organ-dysfunction scoring with modern critical-care data.[7] Early studies suggest that SOFA-2 is feasible and prognostic, but its advantage over SOFA has not been consistent across clinical settings.[8–10]

AC provides a clinically relevant setting for this comparison. Biliary sepsis can involve shock, respiratory dysfunction, kidney injury, coagulation abnormalities and jaundice, yet outcome is also shaped by source control and disease-specific management. A revised organ-dysfunction score may therefore stratify risk well without necessarily improving on SOFA.

Large electronic health record databases allow paired patient-level comparison of the two scores. MIMIC-IV provides detailed ICU measurements for score reconstruction, whereas the eICU Collaborative Research Database (eICU) provides a multicentre cohort for exploratory cross-database assessment. Because eICU contributed to the development of SOFA-2, it was not treated as a fully independent validation dataset.

We therefore evaluated whether day-1 SOFA-2 provides prognostic information beyond SOFA in ICU patients with AC. We focused on ICU mortality as the primary outcome, examined in-hospital and 28-day mortality as secondary outcomes, assessed calibration and clinical utility, and explored whether the risk gradient was reproduced in eICU.

Methods

Data sources

We performed a retrospective cohort study using MIMIC-IV v3.1 and eICU.[11–13] MIMIC-IV served as the primary cohort; eICU provided cross-database assessment rather than external validation, because it contributed to the derivation of SOFA-2.[7] Reporting followed TRIPOD and STROBE principles where applicable.[14–16]

MIMIC-IV cohort identification

We included adult patients with AC meeting operationalised TG18 diagnostic criteria. Because symptoms and imaging findings are not consistently available as structured fields in MIMIC-IV, TG18 criteria were operationalised using diagnosis codes in conjunction with structured evidence of systemic inflammation and hepatobiliary abnormality. Systemic inflammation was defined as an abnormal white blood cell count within 24 h of hospital admission; hepatobiliary abnormality required elevated bilirubin, aminotransferase or

alkaline phosphatase within 48 h. Every qualifying ICU admission additionally carried an explicit cholangitis diagnosis code.

Inclusion required age ≥ 18 years, a qualifying AC diagnosis with structured TG18-supporting evidence, and a first qualifying hospital admission. For the primary ICU analysis, at least one ICU stay was additionally required. Patients were excluded if aged under 18 years, if the admission represented a repeated encounter from the same individual, if concurrent acute pancreatitis codes were present, if structured TG18-supporting evidence was absent, or if no ICU admission was recorded.

eICU cross-database cohort

In eICU, AC was identified from diagnosis text and ICD-9 fields; malignant or chronic biliary diagnoses were excluded where identifiable. Inclusion criteria were age ≥ 18 years, first ICU stay, and a diagnosis-supported AC phenotype with evidence compatible with TG18-defined cholangitis. The final operational TG18-definite cohort included 289 ICU stays. A diagnosis-defined broader cohort of 558 adult first ICU stays was retained only for sensitivity analysis and was not used as the primary eICU cohort.

Data harmonisation and variable extraction

To reduce between-database heterogeneity, variables were harmonised before analysis. Laboratory units were converted to common clinical units as required, including creatinine, bilirubin and lactate. All time windows were aligned to the first ICU day: scores incorporated the most abnormal value recorded within the first 24 h of ICU admission, and baseline physiological and laboratory variables were summarised from the same window when available. Variable definitions were kept consistent across databases wherever the underlying structured fields permitted direct mapping.

Baseline variables encompassed demographics, vital signs, laboratory tests, comorbidities and early organ-support markers. Vital signs included heart rate, systolic and mean arterial pressure, respiratory rate, temperature and peripheral oxygen saturation. Laboratory variables comprised white blood cell count, haemoglobin, platelet count, international normalised ratio, prothrombin time, activated partial thromboplastin time, total bilirubin, alanine aminotransferase, aspartate aminotransferase, albumin, creatinine, urea nitrogen, bicarbonate, electrolytes and lactate. Comorbidities and coded complications — including hypertension, diabetes mellitus, chronic kidney disease, chronic obstructive pulmonary disease, congestive heart failure, chronic liver disease, malignancy, acute kidney injury and sepsis — were extracted from diagnosis-code tables.

Scores and covariates

The primary exposure was the day-1 SOFA-2 score, derived according to Ranzani and colleagues.[7] SOFA was reconstructed over the same 24 h window using the Vincent and colleagues specification.[5] Both scores used the most abnormal value recorded within the window. Unobserved components were assigned

0 points, in keeping with standard practice for structured-database score reconstruction. Detailed component availability is reported in the supplement.

Missing data

Each score domain used the worst recorded value during the scoring window. SpO_2/FiO_2 or oxygen-delivery data substituted for arterial blood gas values when the latter were unavailable. Domains with neither a recorded measurement nor a usable substitute received 0 points. This approach preserves score definitions but may underestimate dysfunction when monitoring is incomplete.

For non-score baseline variables, those with more than 20% missing data were excluded from primary descriptive comparisons and sensitivity analyses; variables with 20% or less missingness were summarised using available cases. Multiple imputation was not performed, as the analysis evaluated reconstructed clinical scores and score-to-risk mappings rather than developing a new composite multivariable prediction model. Component availability, outcome-dependent missingness patterns and eICU completeness-restricted sensitivity analyses are reported in the supplement.

Outcomes and statistical analysis

The primary outcome was ICU mortality. Secondary outcomes were in-hospital mortality and 28-day all-cause mortality in the MIMIC-IV ICU cohort.

Continuous variables are summarised as medians with interquartile ranges and compared using the Wilcoxon rank-sum test. Categorical variables are presented as counts and percentages and compared using the chi-square test or Fisher exact test when cell counts were small. Baseline comparisons were descriptive.

Discrimination was quantified using AUROC with 5,000 patient-level bootstrap samples.[17, 18] The primary analysis compared SOFA-2 with SOFA as paired bedside organ-dysfunction scores. Separate univariable logistic score-to-outcome models were fitted for SOFA-2 and SOFA. We assessed calibration intercepts and slopes, Brier scores, reclassification, and decision-curve analysis (DCA) across clinically relevant threshold probabilities.[19, 20] A descriptive SOFA-2 cutoff was explored using maximally selected rank statistics, and Kaplan-Meier curves were used to display survival separation.[21, 22] No new composite prediction model was developed, because the aim was comparative score evaluation rather than model derivation.

Secondary analyses encompassed in-hospital and 28-day mortality, phenotype sensitivity, component completeness, admission context and subgroup comparisons. The eICU cohort was analysed without model refitting and treated as a cross-database assessment. These analyses were exploratory and hypothesis-generating; no correction for multiple testing was applied. All tests were two-sided, and $P < 0.05$ was considered statistically significant for the primary paired comparison. All analyses were performed in Python using DuckDB.

Results

Baseline characteristics

Cohort selection identified 1,562 first qualifying MIMIC-IV admissions with AC, of whom 644 patients (41.2%) had at least one ICU stay and formed the primary analysis cohort (Fig. 1a). ICU mortality was 11.0% (71/644), in-hospital mortality 15.8% (102/644) and 28-day mortality 20.7% (133/644). The eICU cross-database cohort comprised 289 patients from 105 hospitals, with 17 ICU deaths (Fig. 1b).

Baseline characteristics by ICU survival status are summarised in Table 1. Median age was 72 years, and 284 patients (44.1%) were female. ICU nonsurvivors were less often female than survivors (31.0% vs 45.7%) and more frequently had TG18 Grade III disease (88.7% vs 71.4%). They also required more frequent invasive ventilation or advanced respiratory support (66.2% vs 26.7%), vasopressors (67.6% vs 36.0%) and renal replacement therapy (8.5% vs 1.4%).

Table 1
Baseline characteristics of the MIMIC-IV ICU cohort by ICU survival status

Characteristic	Overall (N = 644)	ICU survivors (N = 573)	ICU deaths (N = 71)	P value
Demographics				
Age, years	72.0 (61.0–82.0)	72.0 (61.0–83.0)	72.0 (62.5–78.0)	0.367
Female sex	284 (44.1%)	262 (45.7%)	22 (31.0%)	0.018
White race	456 (70.8%)	410 (71.6%)	46 (64.8%)	0.237
Black race	38 (5.9%)	33 (5.8%)	5 (7.0%)	0.665
Asian race	34 (5.3%)	29 (5.1%)	5 (7.0%)	0.481
Other/unknown race	116 (18.0%)	101 (17.6%)	15 (21.1%)	—
Medicare insurance	456 (70.8%)	406 (70.9%)	50 (70.4%)	0.940
Medicaid insurance	69 (10.7%)	65 (11.3%)	4 (5.6%)	0.159
Clinical status				
ICU length of stay, days	1.9 (1.1–3.5)	1.9 (1.2–3.3)	2.2 (0.9–5.8)	0.906
TG18 Grade I	71 (11.0%)	67 (11.7%)	4 (5.6%)	0.159
TG18 Grade II	101 (15.7%)	97 (16.9%)	4 (5.6%)	0.014
TG18 Grade III	472 (73.3%)	409 (71.4%)	63 (88.7%)	0.002
Invasive ventilation or advanced respiratory support	200 (31.1%)	153 (26.7%)	47 (66.2%)	< 0.001
Vasopressor use	254 (39.4%)	206 (36.0%)	48 (67.6%)	< 0.001
Renal replacement therapy	14 (2.2%)	8 (1.4%)	6 (8.5%)	< 0.001
Vital signs during first ICU day				
Maximum heart rate, beats/min	105.0 (92.0-120.0)	103.0 (91.0-118.0)	117.0 (102.5-130.5)	< 0.001
Minimum systolic blood pressure, mmHg	87.0 (79.0–95.0)	87.0 (80.0–96.0)	80.0 (70.0-87.5)	< 0.001
Minimum mean arterial pressure, mmHg	56.0 (51.0–63.0)	57.0 (51.0–63.0)	53.0 (44.5–57.0)	< 0.001
Maximum respiratory rate, breaths/min	28.0 (24.0–32.0)	27.0 (24.0–31.0)	31.0 (28.0-35.5)	< 0.001

Data are median (interquartile range) or n (%). P values are descriptive and compare ICU survivors with ICU deaths. TG18 grades are operationalised from structured database fields. Other/unknown race includes patients not recorded as White, Black, or Asian. ICU, intensive care unit; INR, international normalized ratio; SOFA, Sequential Organ Failure Assessment; TG18, Tokyo Guidelines 2018.

Characteristic	Overall (N = 644)	ICU survivors (N = 573)	ICU deaths (N = 71)	P value
Minimum SpO2, %	92.0 (90.0–94.0)	92.0 (90.0–94.0)	90.0 (82.0–92.0)	< 0.001
Maximum temperature, deg C	37.2 (36.9–37.7)	37.2 (36.9–37.6)	37.2 (36.8–37.8)	0.986
Laboratory tests during first ICU day				
White blood cell count, K/uL	16.0 (11.8–22.2)	15.7 (11.7–21.8)	19.1 (13.7–27.3)	0.022
Hemoglobin, g/dL	9.9 (8.4–11.4)	10.0 (8.6–11.4)	9.1 (7.5–10.7)	0.002
Platelets, K/uL	162.0 (106.8-230.2)	162.5 (110.2-231.8)	126.5 (81.8-214.5)	0.012
INR	1.5 (1.3-2.0)	1.5 (1.3–1.9)	2.0 (1.6–3.1)	< 0.001
PT, s	16.6 (14.4–21.4)	16.3 (14.3–20.4)	21.5 (17.0-34.2)	< 0.001
PTT, s	34.2 (29.9–41.8)	33.6 (29.6–40.0)	44.3 (37.7–56.4)	< 0.001
Total bilirubin, mg/dL	4.1 (2.3–7.1)	4.0 (2.3–6.6)	7.0 (3.2–11.8)	< 0.001
ALT, IU/L	126.0 (61.5–237.0)	126.0 (65.0-231.0)	121.5 (44.5-335.2)	0.962
AST, IU/L	133.0 (75.0-277.2)	131.0 (74.5-259.5)	221.0 (94.0-888.0)	< 0.001
Albumin, g/dL	2.8 (2.4–3.2)	2.8 (2.4–3.2)	2.4 (2.0-2.7)	< 0.001
Creatinine, mg/dL	1.1 (0.8–1.8)	1.1 (0.8–1.7)	1.9 (1.3-3.0)	< 0.001
Urea nitrogen, mg/dL	24.0 (15.0–40.0)	23.0 (14.0–36.0)	42.0 (26.0–73.0)	< 0.001
Lactate, mmol/L	2.1 (1.4–3.4)	2.0 (1.4-3.0)	6.5 (3.1–10.1)	< 0.001
Sodium, mmol/L	137.0 (134.0-140.0)	137.0 (135.0-140.0)	136.0 (132.0-140.0)	0.102
Potassium, mmol/L	4.2 (3.8–4.6)	4.2 (3.8–4.5)	4.5 (4.0-5.2)	< 0.001
Bicarbonate, mmol/L	21.0 (17.0–23.0)	21.0 (18.0–23.0)	16.0 (10.0–20.0)	< 0.001

Data are median (interquartile range) or n (%). P values are descriptive and compare ICU survivors with ICU deaths. TG18 grades are operationalised from structured database fields. Other/unknown race includes patients not recorded as White, Black, or Asian. ICU, intensive care unit; INR, international normalized ratio; SOFA, Sequential Organ Failure Assessment; TG18, Tokyo Guidelines 2018.

Characteristic	Overall (N = 644)	ICU survivors (N = 573)	ICU deaths (N = 71)	P value
Comorbidities and coded complications				
Hypertension	413 (64.1%)	374 (65.3%)	39 (54.9%)	0.087
Diabetes mellitus	207 (32.1%)	182 (31.8%)	25 (35.2%)	0.557
Chronic kidney disease	133 (20.7%)	119 (20.8%)	14 (19.7%)	0.837
Chronic obstructive pulmonary disease	99 (15.4%)	85 (14.8%)	14 (19.7%)	0.282
Congestive heart failure	150 (23.3%)	133 (23.2%)	17 (23.9%)	0.890
Chronic liver disease	122 (18.9%)	96 (16.8%)	26 (36.6%)	< 0.001
Malignancy	222 (34.5%)	182 (31.8%)	40 (56.3%)	< 0.001
Acute kidney injury	302 (46.9%)	248 (43.3%)	54 (76.1%)	< 0.001
Sepsis diagnosis code	465 (72.2%)	405 (70.7%)	60 (84.5%)	0.014
Scores				
SOFA-2	7.0 (4.0–10.0)	6.0 (4.0–9.0)	13.0 (10.0–16.0)	< 0.001
SOFA	7.0 (5.0–11.0)	6.0 (4.0–9.0)	14.0 (11.0–17.0)	< 0.001

Data are median (interquartile range) or n (%). P values are descriptive and compare ICU survivors with ICU deaths. TG18 grades are operationalised from structured database fields. Other/unknown race includes patients not recorded as White, Black, or Asian. ICU, intensive care unit; INR, international normalized ratio; SOFA, Sequential Organ Failure Assessment; TG18, Tokyo Guidelines 2018.

Physiological and laboratory profiles reflected greater shock and organ dysfunction among nonsurvivors. Heart rate was higher, arterial pressures lower, respiratory rate faster and SpO_2 lower than in survivors; white blood cell count, INR, bilirubin, creatinine, urea nitrogen and lactate were all elevated among those who died, while albumin and platelet counts were reduced. Chronic liver disease, malignancy, acute kidney injury and coded sepsis were more prevalent in ICU deaths. Median SOFA-2 was 7.0 overall — 6.0 among survivors and 13.0 among those who died; median SOFA followed the same pattern (7.0, 6.0 and 14.0, respectively).

SOFA-2 performance across mortality outcomes

Median day-1 SOFA-2 was 7.0 (IQR, 4.0–10.0). ROC curves for ICU, in-hospital and 28-day mortality are shown in Fig. 2a–c. For ICU mortality, discrimination was good (AUROC, 0.834; 95% CI, 0.777–0.883; Table 2; Fig. 2a). For in-hospital mortality, AUROC was 0.786 (95% CI, 0.734–0.833; Fig. 2b). For 28-day mortality, AUROC was 0.760 (95% CI, 0.714–0.804; Fig. 2c). These results indicate consistent risk stratification, with lower discrimination for longer-horizon outcomes.

Table 2
Primary score-performance comparison for SOFA-2 and SOFA

Outcome	Score	Events/N	Event rate, %	AUROC (95% CI)	Calibration intercept	Calibration slope	Brier score	Validation
ICU mortality	SOFA-2	71/644	11.0	0.834 (0.777–0.883)	-0.02	0.989	0.081	10-fold cross-validated risk predictions
ICU mortality	SOFA	71/644	11.0	0.838 (0.784–0.886)	-0.019	0.989	0.079	10-fold cross-validated risk predictions
In-hospital mortality	SOFA-2	102/644	15.8	0.786 (0.734–0.833)	-0.045	0.973	0.114	10-fold cross-validated risk predictions
In-hospital mortality	SOFA	102/644	15.8	0.784 (0.731–0.831)	-0.047	0.972	0.114	10-fold cross-validated risk predictions
28-day mortality	SOFA-2	133/644	20.7	0.760 (0.714–0.804)	-0.061	0.953	0.144	10-fold cross-validated risk predictions
28-day mortality	SOFA	133/644	20.7	0.756 (0.711–0.801)	-0.06	0.953	0.143	10-fold cross-validated risk predictions

AUROC, area under the receiver operating characteristic curve. Calibration and Brier statistics are based on tenfold cross-validated risk predictions in MIMIC-IV.

Comparison with SOFA and incremental value

SOFA-2 and SOFA were strongly correlated (Spearman ρ = 0.947). ICU mortality discrimination was virtually identical: AUROC 0.834 for SOFA-2 versus 0.838 for SOFA (difference, − 0.005; 95% CI, − 0.020 to 0.010; P = 0.514; Table 2). Differences were also negligible for in-hospital mortality (0.786 vs 0.784) and 28-day mortality (0.760 vs 0.756). Calibration curves and Brier scores were similar for both scores (Fig. 3a,c,e). DCA showed overlapping net benefit across threshold probabilities (Fig. 3b,d,f). Reclassification analyses did not identify incremental value for SOFA-2 over SOFA.

Risk stratification by SOFA-2

ICU mortality rose with SOFA-2 score: 1.9% at 0–2, 2.1% at 3–5, 3.2% at 6–8 and 27.9% at ≥ 9. The maximally selected cutoff scan identified SOFA-2 ≥ 10 as the optimal day-1 separation threshold (Fig. 4); Kaplan-Meier curves confirmed markedly lower ICU survival above this value in MIMIC-IV (Fig. 5a). The same threshold also separated ICU survival in eICU (Fig. 5b).

Sensitivity and subgroup analyses

Sensitivity analyses consistently supported the primary comparison. Discrimination estimates remained stable after requiring non-bilirubin hepatobiliary evidence, excluding patients with recorded chronic liver disease, removing those requiring first-day invasive ventilation or extracorporeal membrane oxygenation, and restricting to emergency or urgent admissions without a surgical-service flag. No subgroup in which SOFA-2 clearly outperformed SOFA was identified (Fig. 6; Supplementary Tables S5, S17 and S22).

eICU cross-database assessment

In eICU, SOFA-2 and SOFA had virtually identical discrimination (AUROC, 0.833 vs 0.835; 95% CI, 0.709–0.933 vs 0.712–0.936; Fig. 7a). Mortality rose from 1.2% in the 0–2 stratum to 22.9% at ≥ 9, and the ≥ 10 survival separation was reproduced in the cross-database cohort (Fig. 7b). Transported calibration is shown in Fig. 7c. With only 17 deaths and incomplete respiratory data, these results provide cross-database corroboration rather than independent validation.

Discussion

In this cohort of ICU patients with AC, SOFA-2 stratified mortality risk but did not improve on SOFA. Discrimination was good for ICU mortality and moderate for in-hospital and 28-day mortality. The same pattern was seen in eICU: patients with higher SOFA-2 scores had higher mortality, but SOFA-2 and SOFA had nearly identical performance. The main message is therefore comparative and deliberately restrained: SOFA-2 is usable in AC, but current evidence does not show that it should replace SOFA.

This interpretation is consistent with early SOFA-2 literature. Wei and colleagues reported overlapping discrimination curves for SOFA-2 and SOFA in pneumonia-associated sepsis.[8] Subsequent multicentre analyses also suggested that SOFA-2 may update score construction without producing a clear universal gain in prognostic accuracy.[9, 10] Our disease-specific analysis extends this pattern to AC, a condition in which biliary obstruction, infection and organ dysfunction interact closely.

The similarity between the two scores is clinically plausible. SOFA-2 updates mainly the respiratory and cardiovascular components, whereas AC prognosis is strongly influenced by cholestasis, obstructive jaundice, renal injury and timely biliary drainage. In this setting, the revised components may improve score modernisation without changing overall risk ranking enough to affect AUROC, calibration or DCA.

A new composite prediction model was intentionally not developed. Such a model could probably improve discrimination by adding age, comorbidity, laboratory variables and treatment factors, but that would

answer a different question. The clinically relevant question here was whether the updated organ-dysfunction score performs differently from SOFA as a simple bedside measure. It did not.

SOFA-2 ≥ 10 provided the clearest survival separation in MIMIC-IV, and a similar pattern appeared in eICU. This cutoff should be viewed as a descriptive risk marker, not a treatment trigger. Prior work in severe AC has also evaluated ICU scores for mortality prediction, but risk scores should complement, not replace, source-control judgment.[23] In AC, escalation decisions depend on source control, haemodynamic response, infection severity and procedural feasibility, not on a score threshold alone.

Limitations

This study has important constraints. MIMIC-IV is a single tertiary centre with an intensive monitoring environment that may not reflect broader ICU practice. TG18 phenotyping relied on structured database fields; imaging findings were not reviewed, and some misclassification is likely. The eICU cohort — 17 ICU deaths across 105 hospitals — is too small to function as a validation dataset. Source-control timing, drainage method and antibiotic choice were not consistently captured, and the interval between biliary decompression and ICU admission almost certainly influenced both scores and outcomes. Respiratory measurements were frequently absent, particularly in eICU; assigning zero to unobserved domains may have compressed score variance and narrowed the observable difference between the two scores. Prospective studies should measure both scores serially around the time of confirmed biliary drainage.

Conclusion

SOFA-2 provided clinically interpretable mortality risk stratification in ICU patients with acute cholangitis, but it did not improve discrimination, calibration or decision-curve performance compared with SOFA. Current data support SOFA-2 as a feasible research-compatible organ-dysfunction score in AC, but not as a replacement for SOFA in routine clinical risk assessment. Further independent, disease-specific validation is needed before clinical adoption.

Abbreviations

AC
acute cholangitis
AUROC
area under the receiver operating characteristic curve
DCA
decision-curve analysis
eICU
eICU Collaborative Research Database
ICU
intensive care unit
MIMIC-IV

Medical Information Mart for Intensive Care IV
SOFA
Sequential Organ Failure Assessment
SpO_2
peripheral oxygen saturation
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology
TG18
Tokyo Guidelines 2018.

Declarations

Ethics approval and consent to participate. This study used deidentified, publicly available data from the MIMIC-IV and eICU databases hosted on PhysioNet. The original data collection was approved by the relevant institutional review boards, and all authors completed required human-subjects training and data-use agreements. Because the present study used only deidentified public data, additional institutional review board approval and informed consent were not required.

Consent for publication. Not applicable.

Availability of data and materials. MIMIC-IV v3.1 is available from PhysioNet at https://physionet.org/content/mimiciv/3.1/. The eICU Collaborative Research Database is available from PhysioNet at https://physionet.org/content/eicu-crd/. Access requires completion of the required training and data-use agreements.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by the Medical and Health Science and Technology Plan of Zhejiang Province [Grant Number 2025HY1368]. The funder had no role in study design, data analysis, interpretation of results, or manuscript preparation.

Authors' contributions. Yunxiao Lyu contributed to conceptualization, methodology, analysis planning and drafting of the manuscript. Bin Wang contributed to data curation, statistical analysis and verification of results. Zhong Lyu supervised the study and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements. None.

References

1. Kiriyama S, Kozaka K, Takada T, et al. Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholangitis. J Hepatobiliary Pancreat Sci. 2018;25(1):17–30. https://doi.org/10.1002/jhbp.512.

2. Miura F, Okamoto K, Takada T, et al. Tokyo Guidelines 2018: initial management of acute biliary infection and flowchart for acute cholangitis. J Hepatobiliary Pancreat Sci. 2018;25(1):31–40. https://doi.org/10.1002/jhbp.509.
3. Kimura Y, Takada T, Kawarada Y, et al. Definitions, pathophysiology, and epidemiology of acute cholangitis and cholecystitis: Tokyo Guidelines. J Hepatobiliary Pancreat Surg. 2007;14(1):15–26. https://doi.org/10.1007/s00534-006-1152-y.
4. Gomi H, Solomkin JS, Schlossberg D, et al. Tokyo Guidelines 2018: antimicrobial therapy for acute cholangitis and cholecystitis. J Hepatobiliary Pancreat Sci. 2018;25(1):3–16. https://doi.org/10.1002/jhbp.518.
5. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med. 1996;22(7):707–10. https://doi.org/10.1007/BF01709751.
6. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801–10. https://doi.org/10.1001/jama.2016.0287.
7. Ranzani OT, Singer M, Salluh JIF, et al. Development and validation of the Sequential Organ Failure Assessment (SOFA)-2 score. JAMA. 2025;334(23):2090–103. https://doi.org/10.1001/jama.2025.20516.
8. Wei S, Wang T, Li C, et al. SOFA-2 score predicts mortality in pneumonia-associated sepsis: a retrospective cohort study. Crit Care. 2026;30:302. https://doi.org/10.1186/s13054-026-06027-4.
9. Zhu H, Li P, Wang B, et al. Impact of the updated SOFA-2 score on sepsis diagnosis and prognosis: a retrospective multicenter cohort study. Crit Care Published online May. 2026;6. https://doi.org/10.1186/s13054-026-06070-1.
10. Ellen J, Hao S, Gao CA, et al. Evaluation of SOFA-2 score performance across demographic subgroups: an external validation study using MIMIC-IV. medRxiv. 2026. https://doi.org/10.64898/2026.03.10.26348061. Preprint; not peer reviewed.
11. Johnson A, Bulgarelli L, Shen L et al. MIMIC-IV (version 3.1). PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58
12. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x.
13. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. https://doi.org/10.1038/sdata.2018.178.
14. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63. https://doi.org/10.7326/M14-0697.
15. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73. https://doi.org/10.7326/M14-0698.

16. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344–9. https://doi.org/10.1016/j.jclinepi.2007.11.008.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747.
18. Efron B, Hastie T. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press; 2016. https://doi.org/10.1017/CBO9781316576533.
19. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.
20. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74. https://doi.org/10.1177/0272989X06295361.
21. Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics. 1992;48(1):73–85. https://doi.org/10.2307/2532740.
22. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81. https://doi.org/10.1080/01621459.1958.10501452.
23. Elshimi E, Fouad TR, Abdel-Razek WM, et al. Risk stratification and validation of ICU scores in prediction of post-ERCP mortality for critically ill patients with severe acute cholangitis based on Tokyo Guidelines 2018. J Clin Gastroenterol Published online Febr. 2026;9. https://doi.org/10.1097/MCG.0000000000002344.

Figures

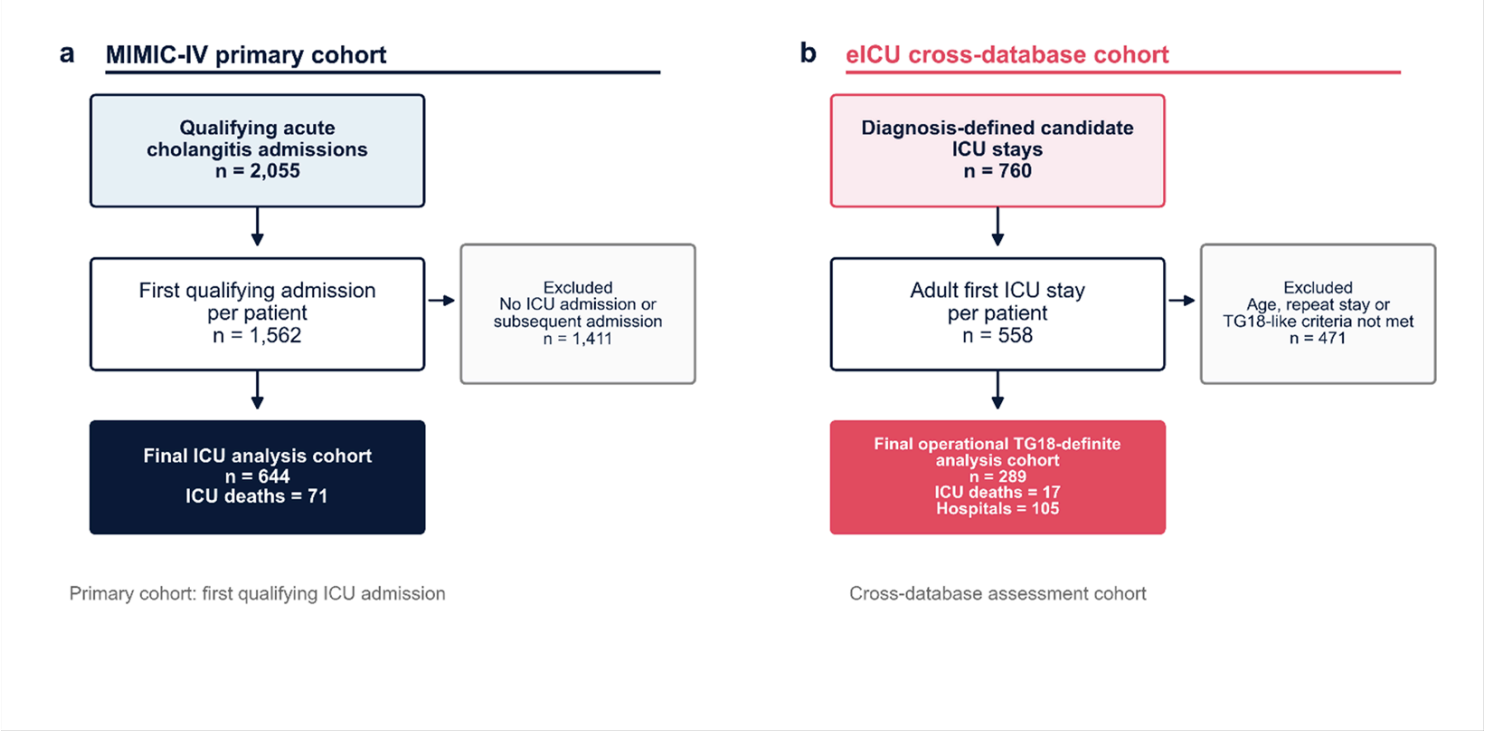


Figure 1

Cohort flow diagram. a, MIMIC-IV primary ICU cohort. b, eICU cross-database cohort.

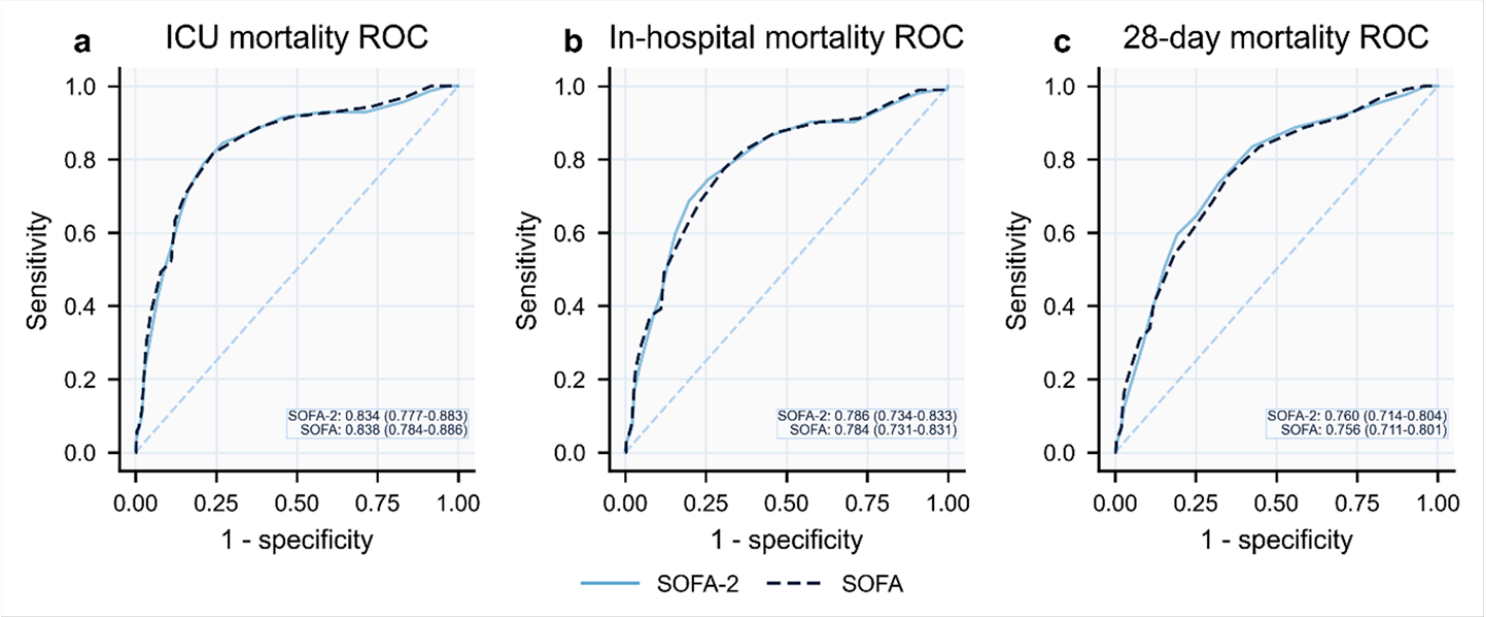


Figure 2

ROC analysis of SOFA-2 and SOFA in MIMIC-IV. a, ROC curves for ICU mortality. b, ROC curves for in-hospital mortality. c, ROC curves for 28-day mortality. AUROC values with 95% confidence intervals are shown within each panel.

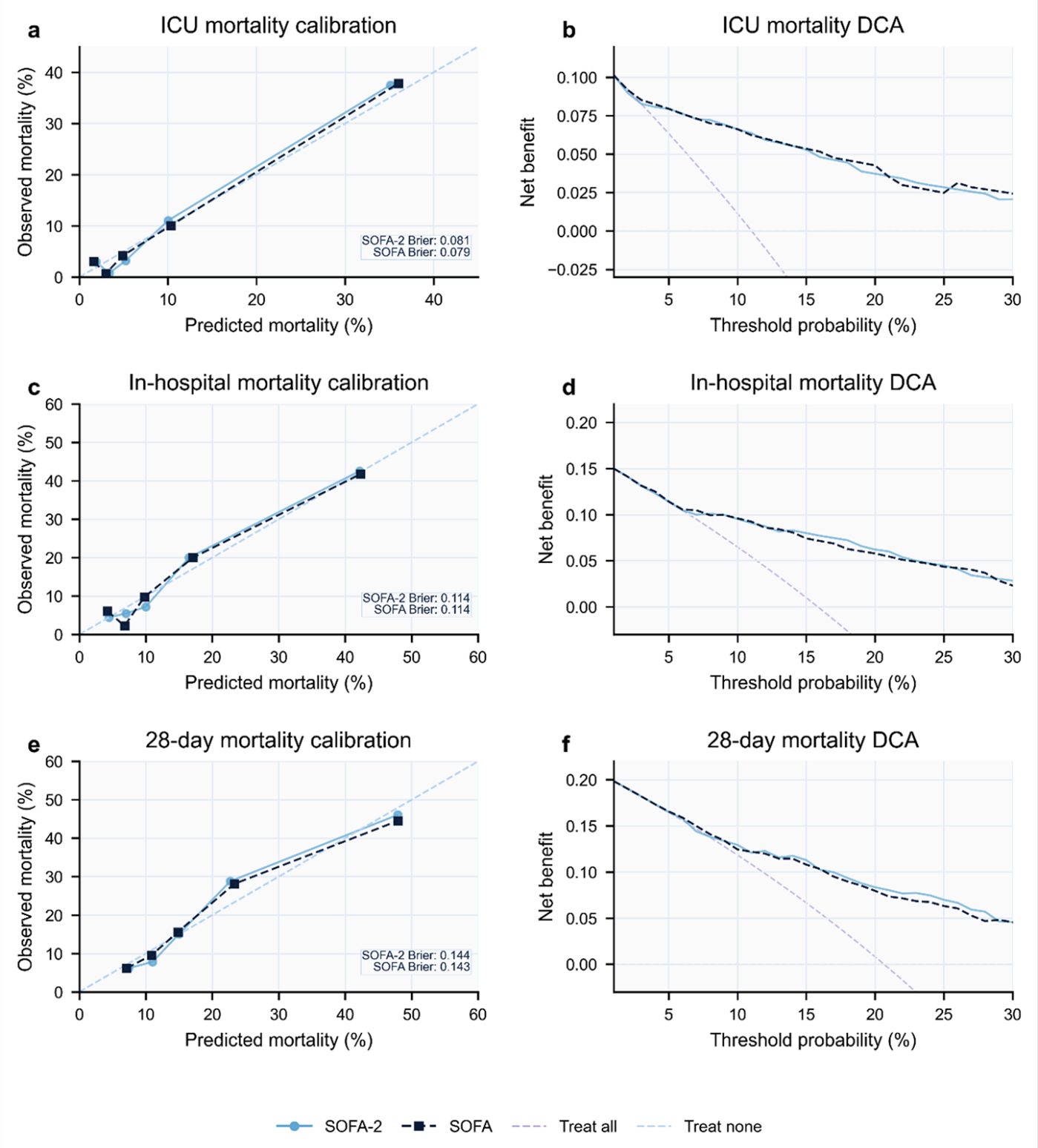


Figure 3

Calibration and decision-curve analysis for SOFA-2 and SOFA in MIMIC-IV. a, ICU mortality calibration. b, ICU mortality decision-curve analysis. c, in-hospital mortality calibration. d, in-hospital mortality decision-curve analysis. e, 28-day mortality calibration. f, 28-day mortality decision-curve analysis. Brier scores are shown within the calibration panels.

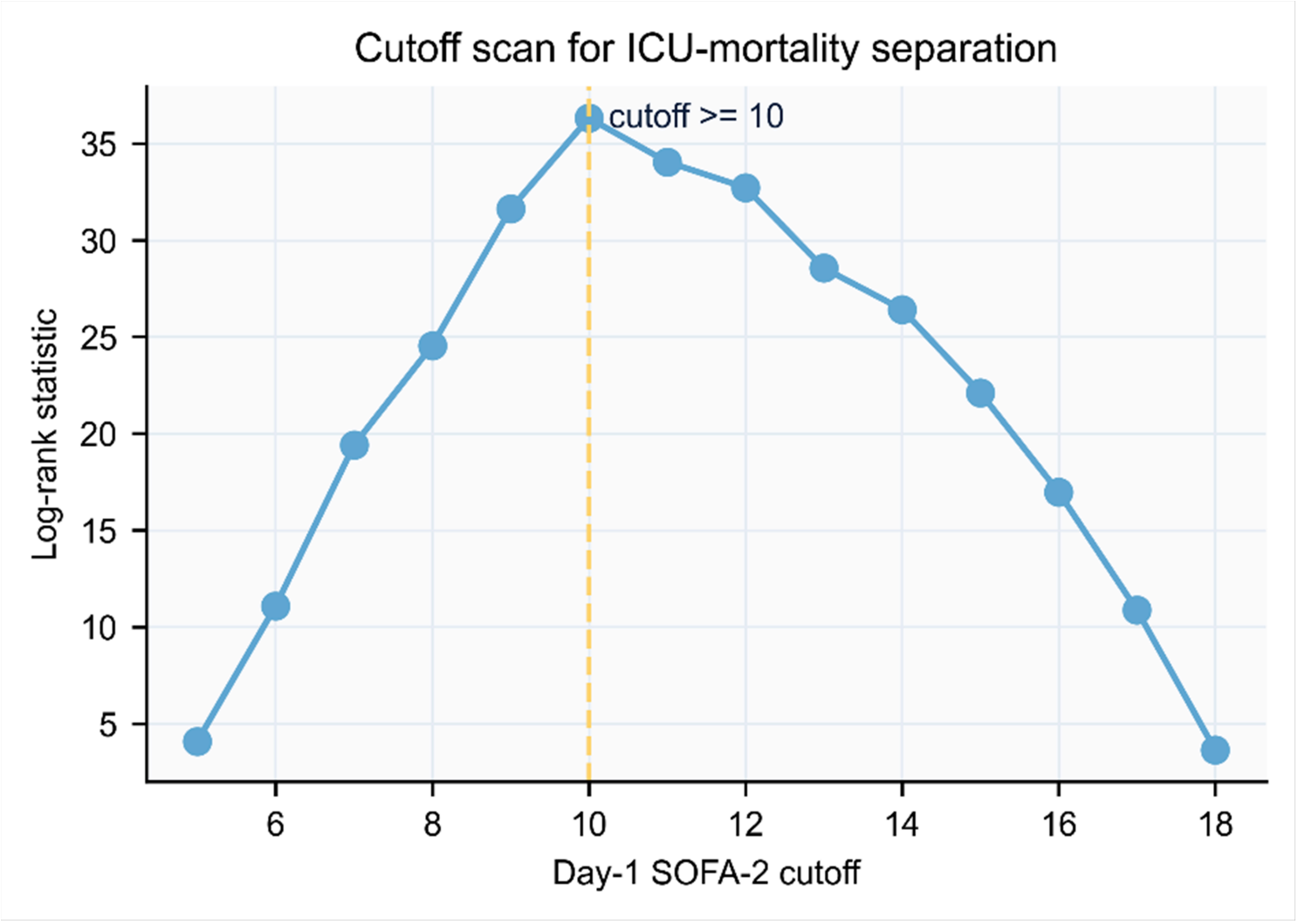


Figure 4

Cutoff-based risk separation for day-1 SOFA-2 in MIMIC-IV.

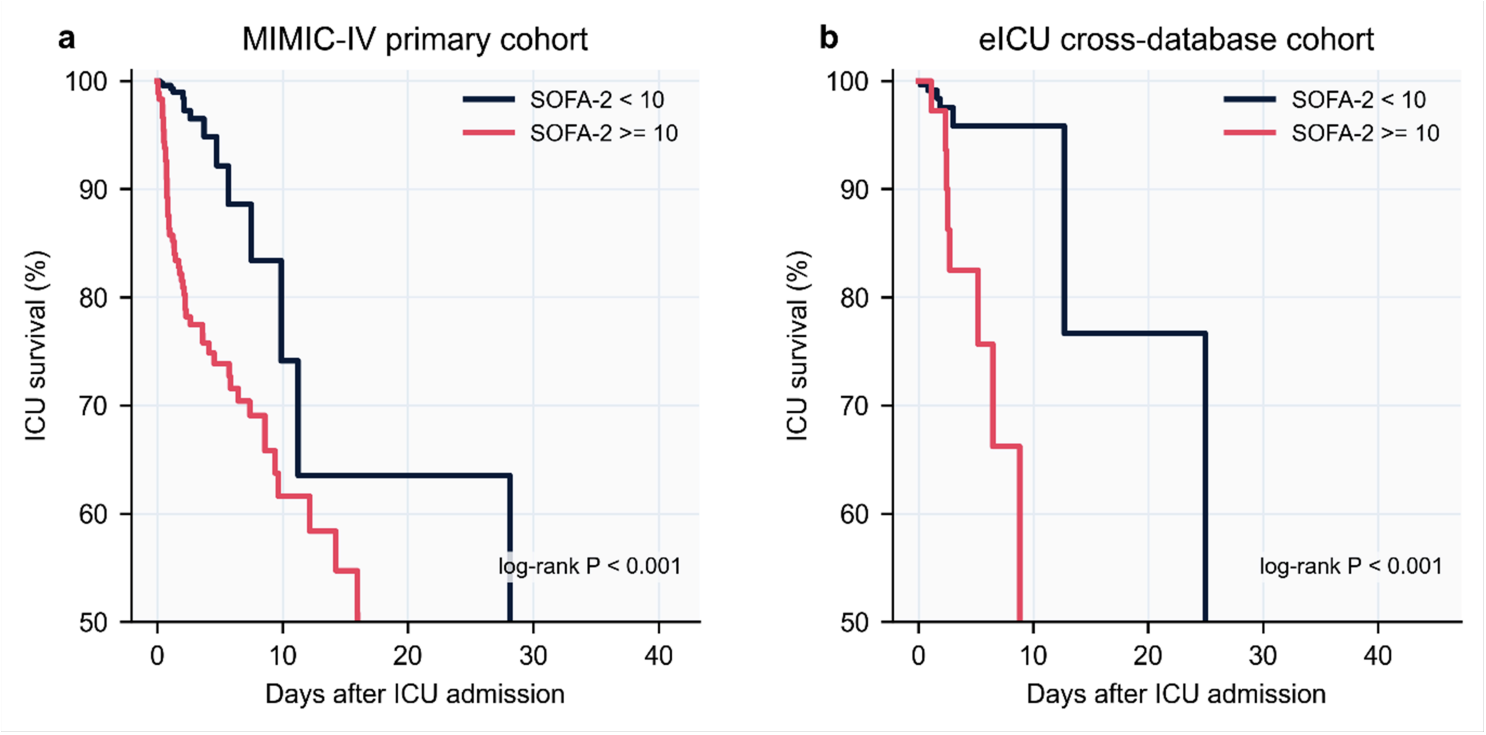


Figure 5

Kaplan-Meier ICU survival curves according to the SOFA-2 threshold. a, MIMIC-IV. b, eICU.

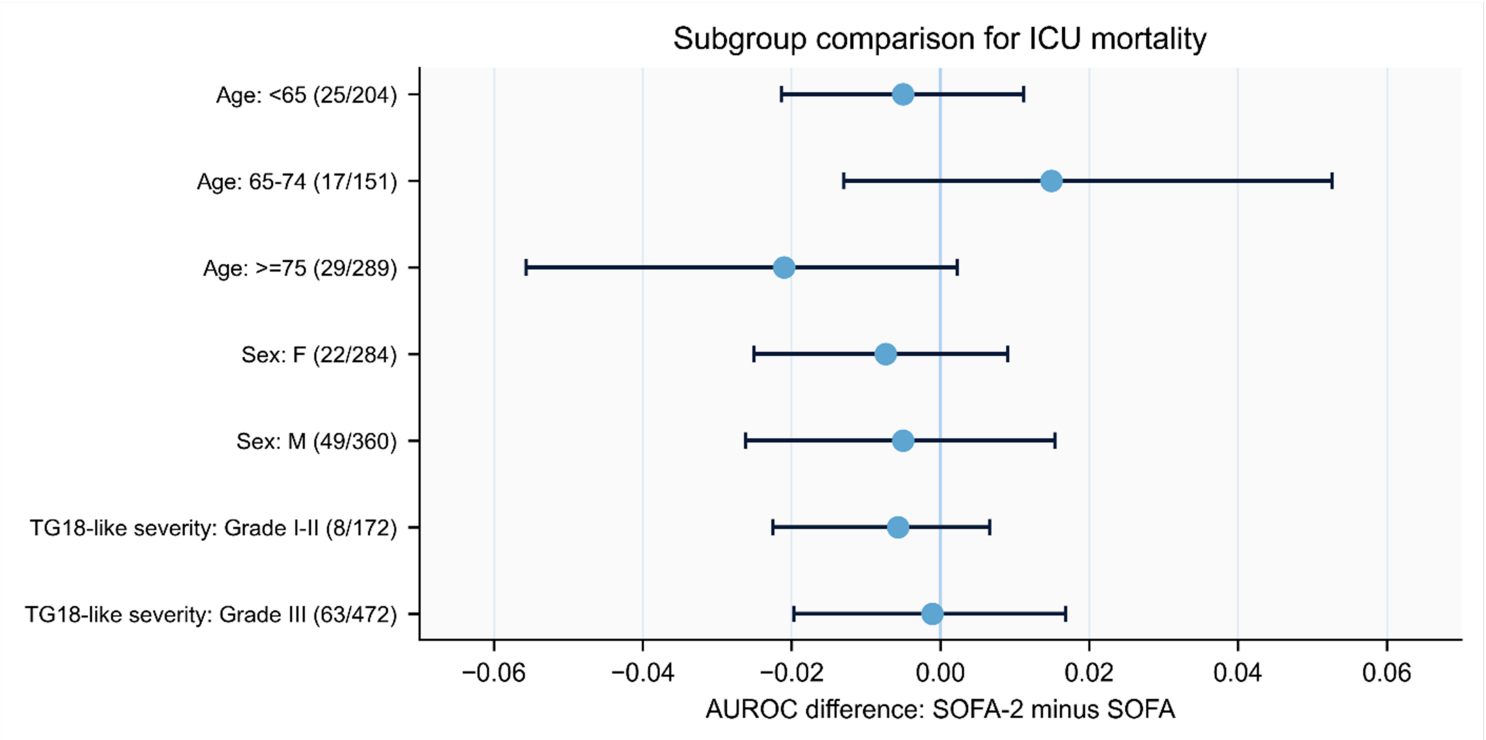


Figure 6

Subgroup analysis comparing SOFA-2 with SOFA.

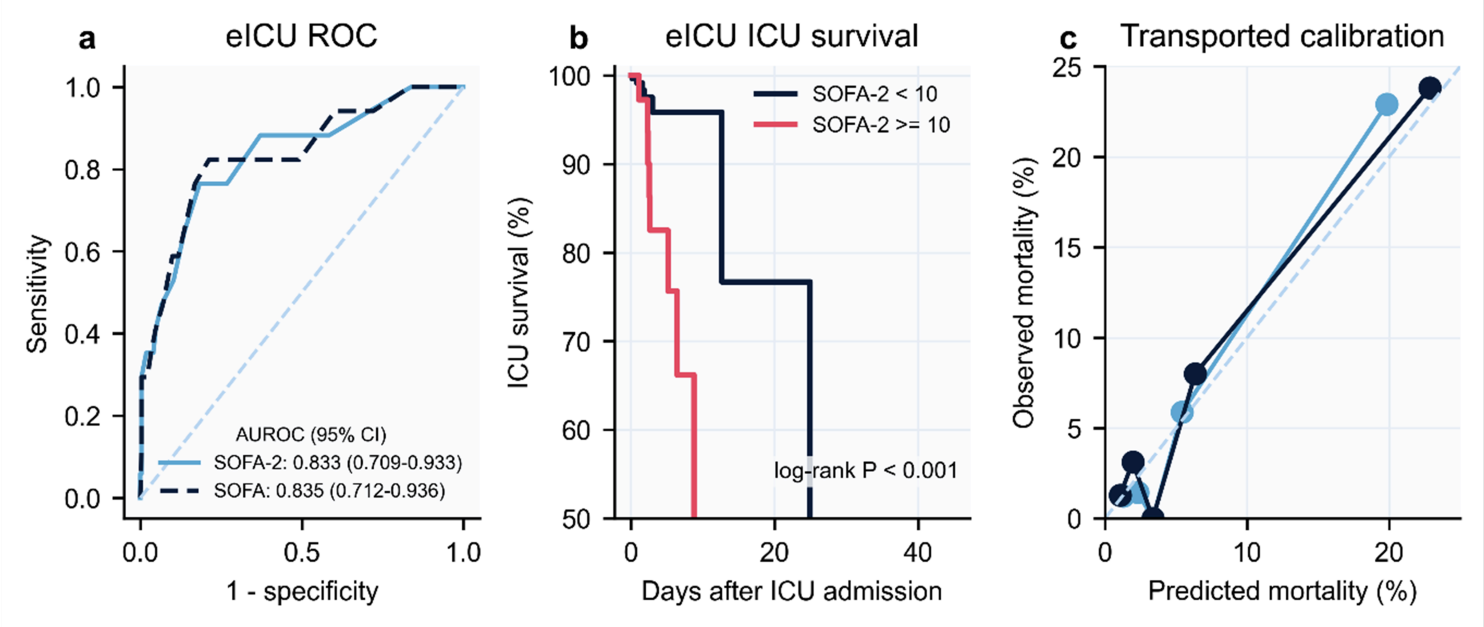


Figure 7

Cross-database assessment of SOFA-2 and SOFA in eICU. a, ROC curves with AUROC values and 95% confidence intervals. b, ICU survival by SOFA-2 <10 versus ≥10. c, Transported calibration.

Supplementary Files

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

- figures3timedependentauc.tif
- figures2extendedanalyses.tif
- figures1dynamicsofa2.tif
- BMCGsupplementarymaterial.docx