Clinical Medicine 26 (2026) 100601

Contents lists available at ScienceDirect

Clinical Medicine

journal homepage: www.sciencedirect.com/journal/clinical-medicine

Original Research

Comparative prognostic performance of SOFA-1 and SOFA-2 in patients with sepsis at ICU admission

Lifang Lin[a,1], Zhifei Fu[b,1], Hongxia Sun[c], Yongye Wang[d], Sheng Zhang[e,*], Songjie Bai[f,*], Xuehuan Wen[g,*]

[a] Department of Surgery, Cangnan Hospital of Traditional Chinese Medicine, Wenzhou, Zhejiang, China
[b] Department of Emergency, Zhejiang Medical & Health Group Hangzhou Hospital of Hangzhou Medical College, Hangzhou, Zhejiang, China
[c] Wenzhou Medical University, Wenzhou, Zhejiang, China
[d] Department of Traditional Chinese Medicine, The People's Hospital of Cangnan Zhejiang, Wenzhou Medical University, Wenzhou, Zhejiang, China
[e] Department of Critical Care Medicine, Taizhou Hospital of Zhejiang Province, affiliated to Wenzhou Medical University, Taizhou, China
[f] Department of Cardiovascular Surgery ICU, The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, Jiangxi, China
[g] Department of Oncology, The People's Hospital of Cangnan Zhejiang, Wenzhou Medical University, Wenzhou, Zhejiang, China

ARTICLE INFO

Keywords:
Sepsis
SOFA-1
SOFA-2
Mortality prediction
Risk stratification

ABSTRACT

Study objective: SOFA-1 underpins Sepsis-3 definitions but may no longer optimally reflect contemporary ICU practice; SOFA-2 was recently developed from large multinational datasets, yet its prognostic performance in septic populations remains unknown. We aimed to compare the prognostic performance of SOFA-1 and SOFA-2 in ICU patients with sepsis.

Design: Retrospective cohort study.

Setting and participants: This study was conducted using the MIMIC-IV database. A total of 30,667 adult patients (≥18 years) with first ICU admissions meeting Sepsis-3 criteria were included.

Main outcome measures: ICU mortality and in-hospital mortality. We compared discrimination (AUC), calibration, net reclassification improvement (NRI), integrated discrimination improvement (IDI), decision curve analysis and time-dependent survival.

Results: Among 30,667 patients, 87.0% met both SOFA-1 and SOFA-2 criteria, 12.0% met only SOFA-2 criteria, and 1.0% met only SOFA-1 criteria. SOFA-2 total scores were systematically higher than SOFA-1, driven primarily by updated renal, cardiovascular and brain scoring criteria. SOFA-2 demonstrated modestly but consistently better discrimination for ICU mortality (AUC 0.779 vs. 0.755; $P < 0.001$) and hospital mortality (AUC 0.745 vs. 0.725; $P < 0.001$), with significant reclassification improvement for ICU mortality (NRI 0.143; IDI 0.020). In multivariable Cox models, SOFA-2 achieved higher AUCs for 7-day (0.790 vs. 0.771), 28-day (0.709 vs. 0.697) and 90-day mortality (0.682 vs. 0.671; all $P < 0.001$), with consistent advantages across all prespecified subgroups.

Conclusions: SOFA-2 provides modestly but consistently better prognostic performance than SOFA-1 in septic ICU patients. Whether this incremental gain justifies the added scoring complexity warrants prospective evaluation.

Introduction

Sepsis remains a leading cause of morbidity and mortality in critically ill patients. A recent global analysis estimated 166 million incident cases and 21.4 million sepsis-related deaths in 2021, accounting for approximately 31.5% of all global deaths.[1] In-hospital mortality among ICU patients with infection reaches 30%.[2] Accurate early assessment of disease severity is essential for guiding clinical decision-making, allocating resources and informing prognosis.[3]

The Sequential Organ Failure Assessment (SOFA) score underpins the Sepsis-3 consensus definitions[4] and is firmly established as the reference framework for organ dysfunction assessment in sepsis.[5]

* Corresponding authors.
E-mail addresses: zhangs@enzemed.com (S. Zhang), songjie.bai@ncu.edu.cn (S. Bai), wxh2024@zuaa.zju.edu.cn (X. Wen).
[1] Lifang Lin and Zhifei Fu contributed equally to this work.

https://doi.org/10.1016/j.clinme.2026.100601
Received 11 February 2026; Received in revised form 12 May 2026; Accepted 21 May 2026
1470-2118/© 2026 The Author(s). Published by Elsevier Ltd on behalf of Royal College of Physicians. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

However, the original SOFA score (hereafter SOFA-1) was developed more than two decades ago,[6] and several of its components may no longer optimally reflect contemporary ICU practice. Advances in vasopressor therapy, respiratory support and overall sepsis management raise concerns about its calibration and discriminatory performance in modern populations.[7,8] These limitations have prompted efforts to refine prognostic assessment, including modifications of SOFA components and machine learning-based prediction models,[9,10] though many remain constrained by limited interpretability or inconsistent validation.[11]

In response, a revised scoring framework, SOFA-2, has recently been developed using large, multinational ICU datasets and data-driven methodology.[12] SOFA-2 incorporates updated organ dysfunction variables and revised thresholds that better reflect modern organ support practices. However, SOFA-2 was derived and validated in unselected ICU cohorts rather than specifically in patients meeting Sepsis-3 criteria, for which diagnosis and severity stratification remain anchored to SOFA-1.[4] It remains unknown whether SOFA-2 meaningfully alters sepsis identification, risk stratification or mortality prediction in septic ICU populations.

We therefore hypothesised that SOFA-2 would demonstrate superior prognostic performance compared with SOFA-1 in septic ICU patients. To test this, we conducted a large-scale retrospective cohort study using the MIMIC-IV database, comparing the discrimination, calibration and clinical utility of both scoring systems for ICU and hospital mortality prediction, and further evaluating time-dependent prognostic performance at 7, 28 and 90 days. We quantified reclassification improvement using NRI and IDI, examined clinical outcomes among patients classified discordantly by the two scoring systems, and explored the prognostic contributions of individual organ-system components.

Methods

Study design and data source

This retrospective cohort study used data from MIMIC-IV (version 3.1),[13] a publicly available de-identified critical care database maintained by the MIT Laboratory for Computational Physiology and hosted on PhysioNet. MIMIC-IV contains data from patients admitted to Beth Israel Deaconess Medical Center. We obtained access after completing the CITI programme and executing a data use agreement with PhysioNet (credentialled access). As the database contains only de-identified data, the study was exempt from institutional review board approval and individual patient consent was not required.

Study population and sepsis definition

Adult patients (age ≥18 years) at first ICU admission were eligible. All patients were admitted to units tracked within the MIMIC-IV ICU module, which encompasses both intensive care units (ICUs; 95.9%) and step-down units (SDUs; 4.1%). All units provided continuous physiological monitoring, and SOFA-1/SOFA-2 scores were calculated identically regardless of unit designation.

We identified sepsis according to Sepsis-3 criteria,[4] defined as suspected infection accompanied by an acute increase ≥2 points from baseline in SOFA-1 or SOFA-2. We identified suspected infection using the official MIMIC-IV-derived suspicion_of_infection table,[14] requiring co-occurrence of antibiotic administration and microbiological culture within a defined time window (details in Supplementary Methods). SOFA-1 and SOFA-2 scores were calculated within the first 24 h of ICU admission based on published definitions,[6,12] using a 48-h lookback window for laboratory values to accommodate infrequent bilirubin measurements (36.6% of intervals > 24 h; see Supplementary Methods). A concise summary of key operationalisation assumptions, full SOFA-1 vs SOFA-2 specification differences, and a sensitivity analysis restricted to documented vasopressor exposure are reported in Supplementary Methods and Supplementary Tables S1 and S4. The complete SQL implementation is available at https://github.com/wzwxh2023/SaAki_Sofa_benchmark.

We applied a union-based cohort definition: patients meeting the organ dysfunction threshold by either scoring system were classified as septic, then categorised into three mutually exclusive groups (Both-positive, SOFA-1-only, SOFA-2-only) based on diagnostic concordance (Fig. 1). The ≥2-point cutoff was applied to SOFA-2 because it was developed as a direct iterative replacement for SOFA-1 within the same 0–24 scoring framework, and no alternative threshold has been recommended by the SOFA-2 authors.

Outcomes and covariates

The primary outcomes were ICU mortality and in-hospital mortality. Secondary outcomes included mortality at 7, 28 and 90 days, and resource utilisation (mechanical ventilation, RRT, vasopressors, length of stay). Covariates for adjustment included age, sex, Charlson Comorbidity Index (CCI) and extremes of vital signs within the first 24 h.

Statistical analysis

Continuous variables were summarised as medians with interquartile ranges; categorical variables as counts with percentages. Group differences were quantified using standardised mean differences (SMD), with |SMD| > 0.1 indicating clinically meaningful imbalances.[15] Missing score components were scored as zero per SOFA-2 Footnote b[12]; patients with missing covariates were excluded from multivariable modelling (N = 29,246).

Discriminative performance was evaluated using receiver operating characteristic (ROC) curve analysis. Area under the curve (AUC) values were calculated for binary mortality outcomes, and we performed paired comparisons using the DeLong test. Model calibration was assessed using the Brier score and calibration curves plotting observed versus predicted mortality probabilities across deciles of predicted risk. Clinical utility was evaluated using decision curve analysis (DCA), with net benefit calculated across threshold probabilities from 0.01 to 0.30. Incremental prognostic value was quantified using NRI and IDI with 95% confidence intervals estimated by bootstrap resampling (1,000 iterations).

Survival probabilities were estimated using the Kaplan–Meier method and compared across risk strata using the log-rank test. We defined risk groups using optimal cutpoints derived from maximally selected rank statistics. Multivariable Cox proportional hazards models were fitted, adjusted for age, sex, CCI and vital signs. We compared concordance statistics between Cox models using the method of Kang *et al.*[16] The cohort was chronologically divided using admission_year from icu_intime into a training set (earliest 70%; n = 20,597) and a temporal validation set (most recent 30%; n = 8,649).[17] Prespecified subgroup analyses were stratified by age (≤65 vs. > 65 years, based on the WHO definition of older persons), sex, and CCI (≤3 vs. > 3). Sensitivity analyses used alternative cohort definitions (intersection, SOFA-1-only, SOFA-2-only). All analyses were performed using R (version 4.4.3).

Results

Baseline characteristics and patient groups

Among 30,667 patients meeting Sepsis-3 criteria, 26,684 (87.0%) met both SOFA-1 and SOFA-2 criteria (Both-positive), 3,677 (12.0%) met only SOFA-2 criteria (SOFA-2-only) and 306 (1.0%) met only SOFA-1 criteria (SOFA-1-only). Extreme group validation comparing sepsis patients with non-sepsis ICU patients (n = 34,699) confirmed clinical validity, with the sepsis group exhibiting 2.8-fold higher ICU mortality (11.1% vs. 3.9%), particularly in lower SOFA-1 score strata

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

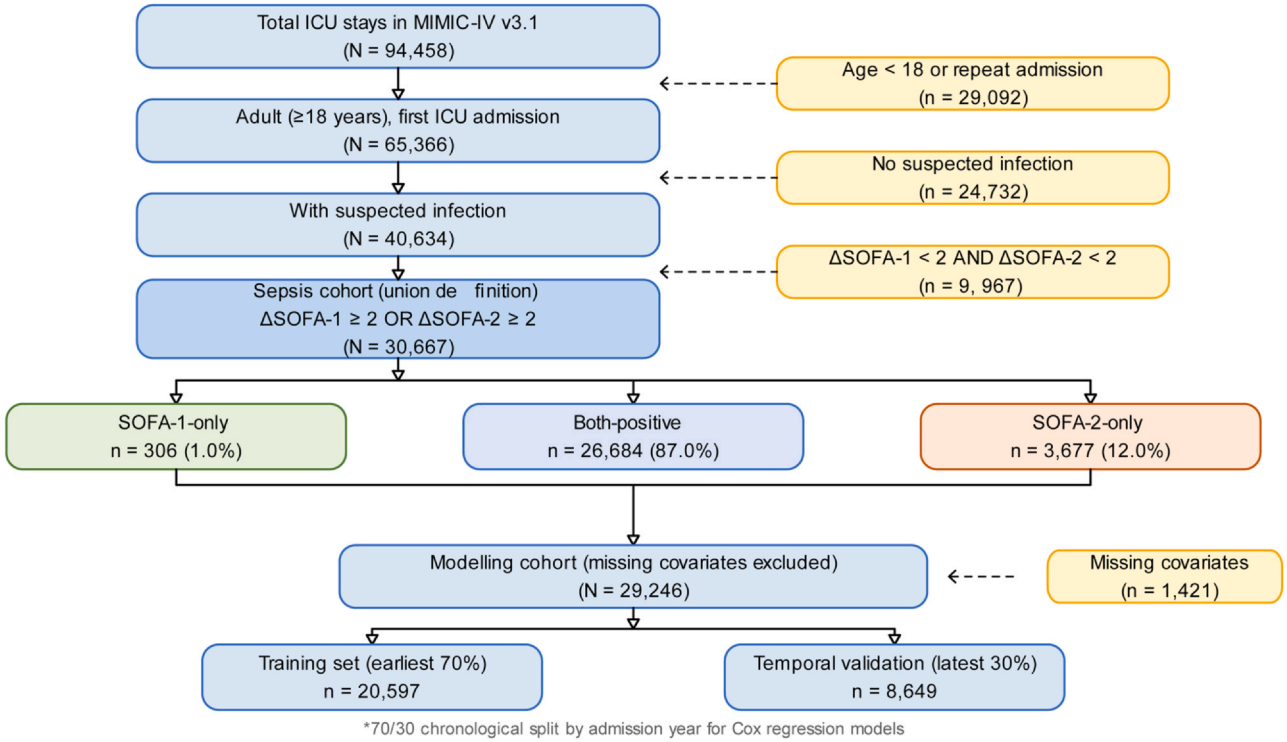

Fig. 1. Study flow diagram and analytical framework. From 94,458 ICU stays in MIMIC-IV v3.1, 65,366 adult patients with first ICU admissions were identified. Of these, 40,634 had suspected infection, and 30,667 met sepsis criteria (suspected infection accompanied by an acute increase ≥2 points in SOFA-1 or SOFA-2). Patients were classified into three groups: Both-positive (n = 26,684, 87.0%), SOFA-1-only (n = 306, 1.0%), and SOFA-2-only (n = 3,677, 12.0%). After excluding patients with missing covariates (n = 1,421), the modelling cohort (N = 29,246) was chronologically divided into a training set (earliest 70%; n = 20,597) and a temporal validation set (most recent 30%; n = 8,649).

(Supplementary Table S2). Baseline characteristics are presented in Table 1. The Both-positive group was older (median 68 years [IQR 56–79]; SMD 0.31), had greater comorbidity burden (CCI 5 [3–7]; SMD 0.27), and higher organ dysfunction scores than single-criterion groups. The largest inter-group differences were in minimum mean blood pressure (SMD 0.73), total SOFA-1 (SMD 1.31), and total SOFA-2 (SMD 1.03) scores. Men predominated in the SOFA-1-only (60%) and Both-positive (58%) groups, whereas women were more prevalent in the SOFA-2-only group (52%; SMD 0.24). Emergency admissions were most frequent in the SOFA-2-only group (60%) compared with the Both-positive group (55%; SMD 0.18).

Comparison of SOFA-1 and SOFA-2 scores

In the overall screened ICU population (N = 65,366), the two scoring systems showed 93.9% diagnostic agreement (Fig. 2A). Despite this concordance, SOFA-2 total scores were systematically higher than SOFA-1 scores ($P < 2 \times 10^{-16}$), with SOFA-2 yielding fewer patients clustering at low values (0–1) and a higher proportion at the upper range of the scale (Fig. 2B). Distribution analysis further illustrated this rightward shift in SOFA-2 values (Fig. 2C). Bland–Altman analysis confirmed that most score differences fell within the 95% limits of agreement, with a positive mean difference indicating systematically higher SOFA-2 values (Fig. 2D). The two scoring systems correlated strongly ($r = 0.78$; $P < 0.001$; Fig. 2E). Organ-level subscore comparisons revealed the largest divergences in the renal (+0.58), cardiovascular (+0.39) and brain (+0.27) domains, driven by updated thresholds and the incorporation of RRT and mechanical circulatory support in SOFA-2; in contrast, respiratory (+0.02) and liver (−0.07) scores were essentially unchanged (Fig. 2F; Supplementary Table S3 and Supplementary Fig. S1). Among patients scoring 4 in the renal domain, RRT accounted for 55.3%; ECMO contributed to only 3.0% of respiratory score 4 assignments.

All subsequent analyses are restricted to the sepsis cohort (N = 30,667 for descriptive analyses; N = 29,246 for prediction modelling).

Prognostic outcomes among sepsis groups

Clinical outcomes differed markedly across the three groups (Table 2). The Both-positive group had the highest ICU mortality (12.2%), hospital mortality (17.2%) and 28-day ICU mortality (20.3%), whereas the SOFA-2-only group had the lowest rates (3.3%, 6.3% and 9.7%, respectively); the SOFA-1-only group (n = 14 ICU deaths, 4.6%; n = 23 hospital deaths, 7.5%) occupied an intermediate position (all $P < 0.001$), though the small sample size (n = 306, 1.0%) warrants cautious interpretation of this subgroup's estimates. Resource utilisation followed concordant patterns (Table 2; Supplementary Fig. S2). ICU and hospital length of stay were longest in the Both-positive group (median 3.0 and 8.7 days), followed by the SOFA-1-only (2.2 and 6.5 days) and SOFA-2-only (1.9 and 6.8 days) groups ($P < 0.001$). The Both-positive group also had the highest rates of RRT use (8.8%), advanced respiratory support (62.4%) and vasopressor use (48.9%), compared with substantially lower rates in the SOFA-1-only (5.2%, 49.7% and 19.6%) and SOFA-2-only (1.5%, 32.8% and 13.4%) groups.

Predictive performance and reclassification

SOFA-2 achieved higher discrimination for ICU mortality (AUC 0.779 [95% CI: 0.771–0.787] vs. 0.755 [0.746–0.764]) and hospital mortality (AUC 0.745 [0.737–0.753] vs. 0.725 [0.716–0.733]; both $P < 0.001$ by DeLong test) (Fig. 3A, D). Calibration curves for both scores aligned closely with the ideal diagonal, indicating adequate agreement between predicted and observed probabilities (Fig. 3B, E). Decision curve analysis revealed a consistently higher net benefit for SOFA-2 across the range of clinically relevant threshold probabilities

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

Table 1
Baseline characteristics of the sepsis cohort by SOFA-1/SOFA-2 concordance group (N = 30,667).

Variables	Overall n = 30,667	SOFA-1-only n = 306	SOFA-2-only n = 3,677	Both-positive n = 26,684	SMD
Age, years	67.00 (56.00, 78.00)	62.00 (50.00, 74.00)	65.00 (52.00, 76.00)	68.00 (56.00, 79.00)	0.31
Gender					0.24
F	13,214 (43.1)	123 (40.2)	1,914 (52.1)	11,177 (41.9)	
M	17,453 (56.9)	183 (59.8)	1,763 (47.9)	15,507 (58.1)	
Race					0.12
Asia	886 (2.9)	10 (3.3)	101 (2.7)	775 (2.9)	
Black	2,618 (8.5)	29 (9.5)	337 (9.2)	2,252 (8.4)	
Others	7,115 (23.2)	80 (26.1)	787 (21.4)	6,248 (23.4)	
White	20,048 (65.4)	187 (61.1)	2,452 (66.7)	17,409 (65.2)	
Admission Type					0.18
Elective	4,102 (13.4)	44 (14.4)	349 (9.5)	3,709 (13.9)	
Emergency	16,914 (55.2)	156 (51.0)	2,206 (60.0)	14,552 (54.5)	
Observation	3,434 (11.2)	39 (12.7)	440 (12.0)	2,955 (11.1)	
Urgent	6,217 (20.3)	67 (21.9)	682 (18.5)	5,468 (20.5)	
Unit					0.22
ICU	29,400 (95.9)	278 (90.8)	3427 (93.2)	25,695 (96.3)	
SDU	1,267 (4.1)	28 (9.2)	250 (6.8)	989 (3.7)	
Vital Signs					
MBP, minimum (mmHg)	58.00 (52.00, 65.00)	66.00 (60.00, 74.38)	61.50 (54.00, 71.00)	58.00 (51.00, 64.00)	0.73
MBP, maximum (mmHg)	101.00 (90.00, 114.00)	105.00 (95.00, 117.75)	102.00 (92.00, 114.00)	100.00 (90.00, 114.00)	0.20
Respiratory Rate, minimum	12.00 (10.00, 15.00)	12.00 (10.00, 14.00)	13.00 (11.00, 15.00)	12.00 (10.00, 15.00)	0.24
Respiratory Rate, maximum	27.00 (24.00, 32.00)	27.00 (23.00, 31.00)	27.00 (24.00, 31.00)	28.00 (24.00, 32.00)	0.09
Temperature, minimum (°C)	36.44 (36.06, 36.67)	36.44 (36.17, 36.67)	36.50 (36.22, 36.72)	36.44 (36.00, 36.67)	0.24
Temperature, maximum (°C)	37.33 (37.00, 37.89)	37.33 (37.00, 37.86)	37.22 (36.94, 37.67)	37.33 (37.00, 37.90)	0.22
Heart Rate, minimum (bpm)	70.00 (61.00, 81.00)	69.00 (60.00, 81.00)	70.00 (61.00, 81.00)	70.00 (61.00, 81.00)	0.04
Heart Rate, maximum (bpm)	104.00 (91.00, 119.00)	104.00 (91.00, 118.00)	102.00 (90.00, 116.00)	104.00 (91.00, 119.00)	0.09
Glucose, minimum (mg/dL)	104.00 (87.00, 126.00)	101.00 (87.00, 121.25)	108.00 (93.00, 130.00)	103.00 (87.00, 126.00)	0.17
Glucose, maximum (mg/dL)	164.00 (130.00, 212.00)	157.00 (122.00, 196.25)	146.00 (118.00, 189.00)	167.00 (133.00, 215.00)	0.04
Comorbidities					
CCI score	5.00 (3.00, 7.00)	4.00 (2.00, 7.00)	4.00 (2.00, 6.00)	5.00 (3.00, 7.00)	0.27
Congestive Heart Failure	8,465 (27.6)	58 (19.0)	769 (20.9)	7,638 (28.6)	0.23
Chronic Pulmonary Disease	7,793 (25.4)	61 (19.9)	1,016 (27.6)	6,716 (25.2)	0.18
Renal Disease	6,263 (20.4)	51 (16.7)	364 (9.9)	5,848 (21.9)	0.33
Diabetes Mellitus	7,205 (23.5)	66 (21.6)	699 (19.0)	6,440 (24.1)	0.12
Diabetes with Complications	3,022 (9.9)	22 (7.2)	251 (6.8)	2,749 (10.3)	0.12
Malignancy	4,245 (13.8)	37 (12.1)	529 (14.4)	3,679 (13.8)	0.07
SOFA-1 Total	5.00 (3.00, 8.00)	3.00 (2.00, 5.00)	1.00 (1.00, 3.00)	5.00 (3.00, 8.00)	1.31
SOFA-1 Respiration	0.00 (0.00, 3.00)	0.00 (0.00, 2.00)	0.00 (0.00, 0.00)	0.00 (0.00, 3.00)	0.69
SOFA-1 Coagulation	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	0.72
SOFA-1 Liver	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.48
SOFA-1 Cardiovascular	1.00 (1.00, 3.00)	1.00 (0.00, 1.00)	1.00 (0.00, 1.00)	1.00 (1.00, 3.00)	0.79
SOFA-1 Brain	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	0.59
SOFA-1 Renal	0.00 (0.00, 2.00)	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	1.00 (0.00, 2.00)	0.54
SOFA-2 Total	6.00 (4.00, 9.00)	3.00 (1.00, 5.00)	4.00 (3.00, 5.00)	7.00 (5.00, 9.00)	1.03
SOFA-2 Respiration	0.00 (0.00, 2.00)	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	1.00 (0.00, 2.00)	0.62
SOFA-2 Coagulation	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	0.69
SOFA-2 Liver	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.44
SOFA-2 Cardiovascular	2.00 (1.00, 3.00)	1.00 (0.00, 2.00)	1.00 (0.00, 2.00)	2.00 (1.00, 3.00)	0.79
SOFA-2 Brain	1.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	1.00 (0.00, 2.00)	0.46
SOFA-2 Renal	2.00 (1.00, 2.00)	0.00 (0.00, 2.00)	2.00 (1.00, 2.00)	2.00 (1.00, 2.00)	0.59

Note: Values are presented as median (IQR) for continuous variables and n (%) for categorical variables. Group differences are quantified using the maximum pairwise standardised mean difference (SMD) across the three concordance groups; |SMD| > 0.1 indicates clinically meaningful imbalances. Missing data proportions are reported in Supplementary Table S1. 'Elective' refers to scheduled admissions, predominantly planned surgical procedures. Analysis population: sepsis cohort (N = 30,667).

Abbreviations: CCI, Charlson Comorbidity Index; ICU, intensive care unit; MBP, mean blood pressure; RRT, renal replacement therapy; SDU, step-down unit; SMD, standardised mean difference; SOFA-1, Sequential Organ Failure Assessment (original); SOFA-2, Sequential Organ Failure Assessment 2.0.

(Fig. 3C, F). Sensitivity analyses using alternative cohort definitions confirmed these findings, with SOFA-2 showing higher discrimination, adequate calibration, and higher net benefit across the SOFA-1-only (ICU AUC 0.931 vs. 0.839), SOFA-2-only (0.757 vs. 0.673), and Both-positive (0.768 vs. 0.741) subgroups (Supplementary Figs. S3–S5). A sensitivity analysis restricted to documented vasopressor exposure (excluding the MAP-based fallback pathway) confirmed that, in the sepsis cohort, mean total SOFA-2 remained higher than SOFA-1 and the discrimination advantage of SOFA-2 for both ICU and hospital mortality was preserved (Supplementary Methods §6, Supplementary Table S4).

In the full sepsis cohort (N = 30,667), SOFA-2 yielded significant reclassification improvement for ICU mortality (NRI 0.143 [95% CI: 0.107–0.177]; IDI 0.020, P < 0.001) and hospital mortality (NRI 0.133 [0.102–0.164]; IDI 0.017, P < 0.001) (Supplementary Fig. S6). Reclassification improvement persisted across all three concordance cohorts (Supplementary Figs. S7–S9).

Time-dependent survival analysis and risk stratification

In multivariable Cox regression models adjusted for age, sex, comorbidities and vital signs, the SOFA-2-based model achieved higher AUCs for 7-day (0.790 vs. 0.771), 28-day (0.709 vs. 0.697) and 90-day mortality (0.682 vs. 0.671; all P < 0.001) (Fig. 4A–D). The C-index for the SOFA-2 model was uniformly higher across all time points: 7-day (0.799 [95% CI: 0.791–0.808] vs. 0.781 [0.772–0.789]), 28-day (0.751 [0.742–0.761] vs. 0.736 [0.726–0.745]) and 90-day (0.738

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

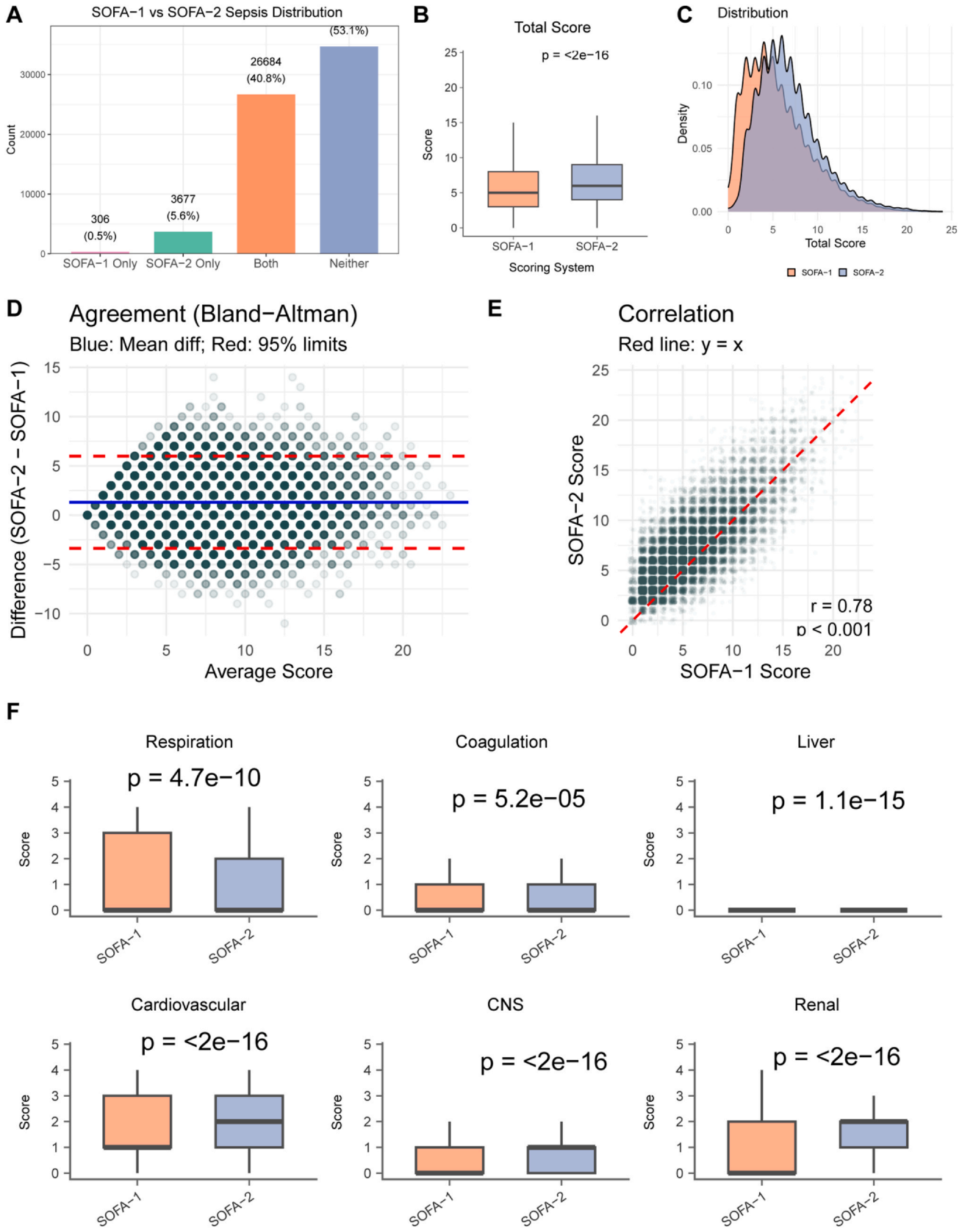

Fig. 2. Comparison of SOFA-1 and SOFA-2 scoring systems. (A) Distribution of sepsis classification by SOFA-1 and SOFA-2 criteria in the full ICU cohort (N = 65,366). (B–F) Analyses restricted to the sepsis cohort (N = 30,667): (B) Box plots comparing total scores between SOFA-1 and SOFA-2. (C) Density distribution of total scores, demonstrating a rightward shift in SOFA-2 values. (D) Bland–Altman plot showing agreement between scoring systems; blue line indicates mean difference, red dashed lines indicate 95% limits of agreement. (E) Correlation between SOFA-1 and SOFA-2 total scores; red dashed line represents identity (y = x). (F) Organ-specific subscore comparisons for all six components.

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

Table 2
Clinical outcomes by SOFA-1/SOFA-2 concordance group (sepsis cohort, N = 30,667).

Variables	Overall (n = 30,667)	SOFA-1-only (n = 306)	SOFA-2-only (n = 3,677)	Both-positive (n = 26,684)	P-value
Primary outcomes					
ICU mortality, n (%)	3,403 (11.1)	14 (4.6)	123 (3.3)	3,266 (12.2)	< 0.001
Hospital mortality, n (%)	4,843 (15.8)	23 (7.5)	232 (6.3)	4,588 (17.2)	< 0.001
28-day ICU mortality, n (%)	5,795 (18.9)	32 (10.5)	356 (9.7)	5,407 (20.3)	< 0.001
Resource utilisation					
RRT required, n (%)	2,429 (7.9)	16 (5.2)	54 (1.5)	2,359 (8.8)	< 0.001
RRT duration, hours[a]	57.9 (5.0, 167.2)	35.3 (0.8, 97.2)	24.0 (1.6, 53.9)	59.6 (5.1, 169.3)	< 0.001
ICU LOS, days	2.8 (1.4, 5.9)	2.2 (1.2, 4.0)	1.9 (1.1, 3.5)	3.0 (1.5, 6.4)	< 0.001
Hospital LOS, days	8.4 (5.0, 15.1)	6.5 (4.6, 11.8)	6.8 (4.0, 11.1)	8.7 (5.2, 15.9)	< 0.001
Advanced life support					
Advanced respiratory support, n (%)	17,997 (58.7)	152 (49.7)	1,205 (32.8)	16,640 (62.4)	< 0.001
Advanced respiratory support, hours[a]	56.0 (20.0, 186.0)	21.8 (12.1, 63.5)	34.0 (14.2, 76.0)	59.9 (21.3, 198.6)	< 0.001
Vasopressor use, n (%)	13,592 (44.3)	60 (19.6)	492 (13.4)	13,040 (48.9)	< 0.001
Vasopressor duration, hours[a]	21.1 (6.3, 57.4)	7.9 (2.2, 35.0)	12.0 (3.6, 29.0)	21.7 (6.5, 58.7)	< 0.001

Note: Continuous variables are presented as median (interquartile range). Categorical variables are presented as n (%). P-values were calculated using Kruskal–Wallis test for continuous variables and chi-square test for categorical variables.
Abbreviations: ICU, intensive care unit; LOS, length of stay; RRT, renal replacement therapy; SOFA-1, Sequential Organ Failure Assessment (original); SOFA-2, Sequential Organ Failure Assessment 2.0.
[a] Reported among patients who received the corresponding treatment. Analysis population: sepsis cohort (N = 30,667).

[0.729–0.748] vs. 0.723 [0.714–0.732]) (Fig. 4E). The kidney, brain and cardiovascular components of SOFA-2 were independently associated with 28-day mortality, with the brain component exhibiting the largest effect size (Fig. 4F). SOFA-2 achieved greater mortality separation between high-risk and low-risk groups at all time horizons (7-day: 14.5% vs. 12.9%; 28-day: 26.6% vs. 25.0%; 90-day: 31.6% vs. 29.2%) (Fig. 4G). Kaplan–Meier analysis using optimal cutoffs (score ≥10) showed that SOFA-2 identified a larger high-risk cohort than SOFA-1 (n = 5,922 vs. n = 4,351), with both scoring systems significantly separating risk groups (log-rank P < 0.0001) (Supplementary Fig. S10A–B). Time-dependent AUC curves confirmed that SOFA-2 maintained higher discrimination than SOFA-1 throughout the observation period (Supplementary Fig. S10C–D). The prognostic advantage remained robust across all prespecified subgroups, including stratification by age, sex and comorbidity burden (Supplementary Fig. S10E). Both scoring systems achieved clear risk stratification across low-, intermediate- and high-risk groups, with SOFA-2 showing wider separation of survival trajectories (Supplementary Fig. S10F).

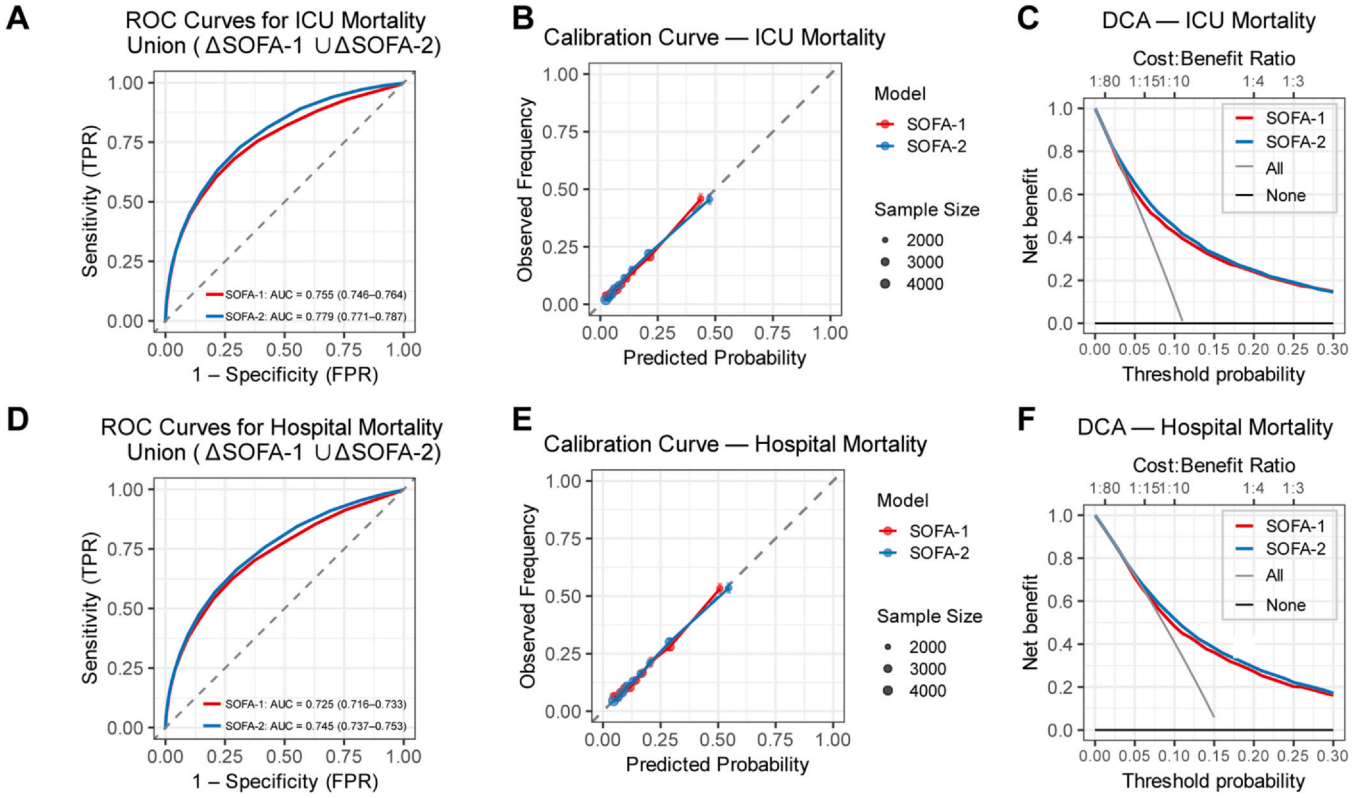

Fig. 3. Prognostic performance of SOFA-1 and SOFA-2 scores for mortality prediction in the sepsis cohort (N = 30,667). (A–C) ICU mortality prediction: receiver operating characteristic (ROC) curves with area under the curve (AUC) and 95% confidence intervals (A), calibration curves comparing predicted probabilities versus observed frequencies (B), and decision curve analysis showing net benefit across threshold probabilities (C). (D–F) Hospital mortality prediction: ROC curves (D), calibration curves (E), and decision curve analysis (F).

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

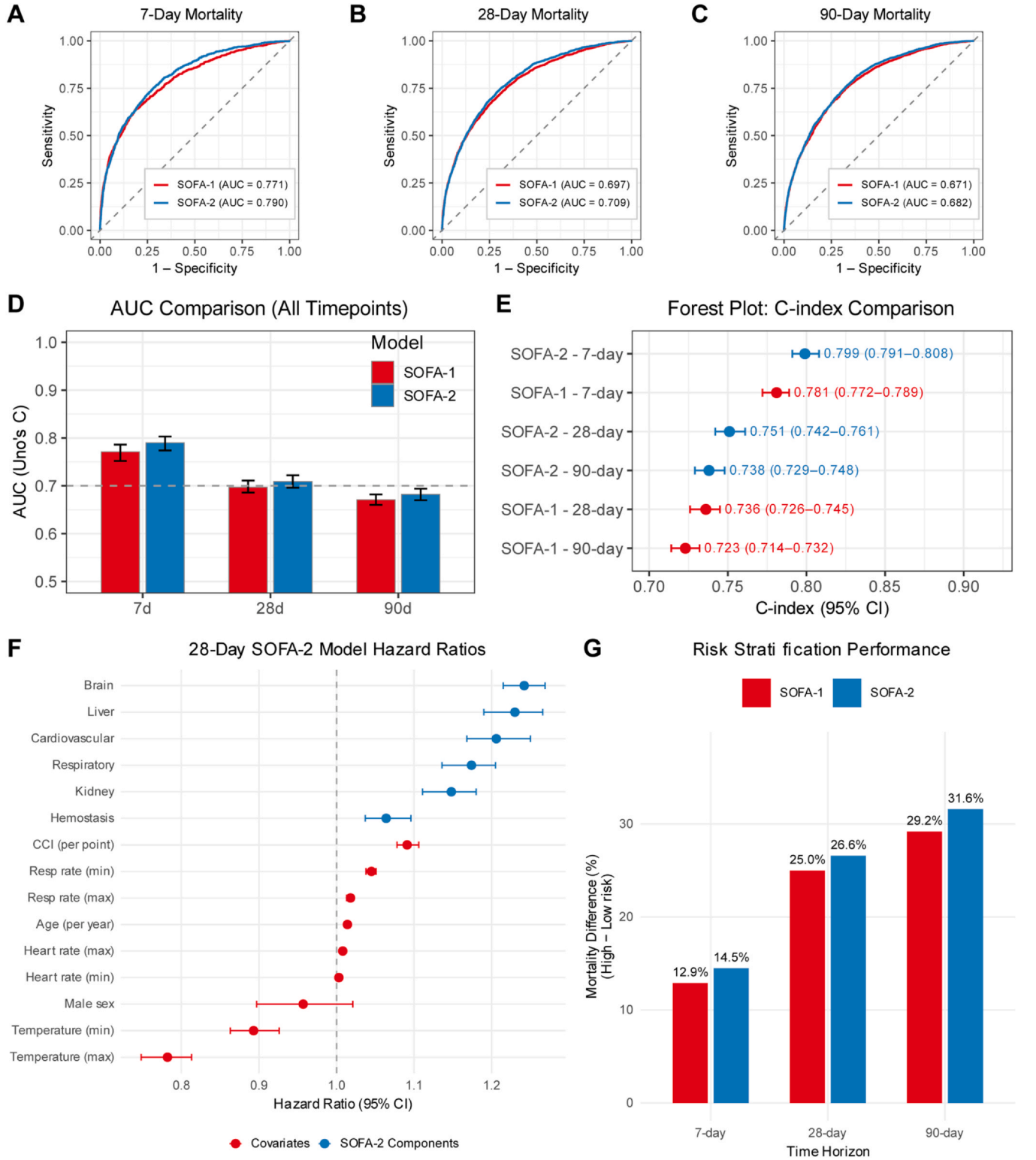

Fig. 4. Time-dependent prognostic performance of SOFA-1 and SOFA-2 scores for mortality prediction in the sepsis cohort (N = 29,246 for modelling). (A–C) ROC curves for 7-day (A), 28-day (B) and 90-day (C) mortality prediction with corresponding AUC values. (D) Bar chart comparing AUC values across time horizons. (E) Forest plot displaying concordance index (C-index) with 95% confidence intervals for survival models at each time point. (F) Hazard ratios with 95% confidence intervals for individual SOFA-2 components and adjustment covariates in the 28-day Cox regression model. (G) Risk stratification performance showing mortality difference between high-risk and low-risk groups. ***P < 0.001.

Discussion

Our analysis of 30,667 ICU patients with sepsis demonstrates that SOFA-2 provides consistently, though modestly, better prognostic discrimination, calibration and risk stratification compared with SOFA-1, with these advantages persisting across time horizons and pre-specified subgroups. Three lines of evidence support this conclusion: higher AUC values and C-indices across all mortality endpoints;

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

significant reclassification improvement (NRI 0.143 for ICU mortality); and greater separation between high-risk and low-risk groups at 7, 28 and 90 days.

These findings align with the rationale for SOFA-2. SOFA-1 cannot distinguish organ dysfunction from the effects of therapeutic interventions.[18] Ranzani *et al*[12] recalibrated the score using over 3 million encounters,[19] adding RRT, mechanical circulatory support, expanded vasoactive definitions and a MAP-based fallback. Our organ-level findings map onto these features: the renal divergence reflects RRT scoring, the brain divergence the delirium-treatment criterion, and the cardiovascular divergence a composite of recalibrated vasopressor cut-offs, the score-2 'any other vasopressor or inotrope' pathway, mechanical-support scoring, and the MAP-based pathway. Restricting to documented vasopressor exposure attenuates the cardiovascular divergence while preserving the overall discrimination advantage.

Our finding that mean SOFA-2 exceeds mean SOFA-1 in MIMIC-IV differs in direction from the pooled estimate of Ranzani *et al.*[12] This reflects the substantial inter-cohort heterogeneity documented in the derivation paper – across its 10 contributing cohorts, elective admission proportions ranged from 3% to 56.3% and ICU mortality from 4.0% to 20.5%. MIMIC-IV sits at the low-elective end (3.7%) and its high-granularity EHR fully captures the SOFA-2-specific upgrades (delirium treatment, mechanical support beyond ECMO, advanced ventilatory modes, SpO_2/FiO_2 fallback, RRT), producing a positive ΔSOFA. This is the expected behaviour of a single cohort whose case-mix and data granularity differ from the source distribution and does not contradict the pooled multi-cohort mean.

The SOFA-2 ≥ 2 cutoff for Sepsis-3 merits discussion. As SOFA-2 was developed as a direct iterative replacement for SOFA-1 within the same 0–24 scoring framework, applying the established Sepsis-3 ≥ 2-point threshold represents the most conservative approach pending formal recalibration. The SOFA-2-only group (12.0%), despite having lower mortality than the Both-positive group, had hospital mortality of 6.3% and 28-day mortality of 9.7% – clinically meaningful risk that supports their inclusion.

Notably, the discriminative improvement (ΔAUC 0.024 for ICU mortality) is modest in absolute terms, but clinically relevant given the large cohort and the inherent difficulty of outperforming established scores. Conventional AUC may underestimate clinical value, as it does not capture whether individual patients are appropriately reclassified into higher or lower risk categories.[20] We therefore employed the NRI and IDI, which directly quantify the proportion of patients correctly reclassified.[21] The NRI of 0.143 indicates a net 14.3% improvement in reclassification across event and non-event groups, with a net improvement in discrimination slope. Both scores showed adequate calibration, a prerequisite for clinical utility.[22] Decision curve analysis further revealed a higher net benefit for SOFA-2 across clinically relevant threshold probabilities, supporting its incremental value for risk-based decision-making. This modest improvement must, however, be weighed against the substantially greater scoring complexity of SOFA-2, which requires additional organ support documentation and treatment-related data elements that may not be routinely captured in all clinical settings. In institutions with mature electronic health record systems, automated calculation can mitigate this burden; in resource-limited environments, the added data requirements may offset the prognostic gain.

The discordant groups provide additional insights. Patients identified only by SOFA-2 had the lowest mortality and shortest organ support durations, suggesting that SOFA-2 expands sepsis identification at the lower-risk end of the spectrum without diluting prognostic discrimination. Rather than simply inflating severity scores, SOFA-2 appears to reshape the distribution of mortality risk by more accurately capturing the gradient of organ dysfunction. This interpretation accords with evidence that associations between SOFA components and mortality are not uniform, implying that identical total scores may represent different underlying risk profiles.[23]

Analysis of individual score components provided further biological insights. In multivariable Cox models, the renal, brain and cardiovascular components of SOFA-2 exhibited the strongest prognostic associations, with the brain component showing the largest effect size. This aligns with growing evidence establishing sepsis-associated encephalopathy as a major determinant of adverse outcomes.[24] However, the contribution of GCS within composite organ dysfunction scores may vary across settings because of sedation and intubation practices.[25] Of note, we operationalised the SOFA-2 delirium treatment criterion using prescription records; indication-level data are unavailable in MIMIC-IV. However, the SOFA-2 definition explicitly assigns one point for patients receiving 'short- or long-term' drug treatment for delirium, indicating that ongoing antipsychotic use – even without acute delirium – reflects an altered neurological state. In contrast, the coagulation component showed a relatively weaker association with mortality. Prior MIMIC-IV analyses of SOFA component contributions found that early risk models preferentially retained renal, neurological and cardiovascular components, whereas coagulation contributed less[26] – in line with our findings. The persistence of SOFA-2's modest prognostic advantage across 7-day, 28-day and 90-day horizons suggests a stable alignment between updated organ dysfunction criteria and subsequent mortality, rather than a time-point-specific improvement.[27,28] Whether SOFA-2 also provides superior dynamic prognostication through longitudinal trajectory monitoring warrants further investigation.

Our findings are broadly consistent with two concurrent studies. Pei *et al*[29] reported comparable discrimination between SOFA-2 and SOFA-1 (AUC 0.646 vs. 0.641) in 1,089 sepsis patients from the TESTS trial, with only 2.2% meeting SOFA-1 but not SOFA-2 criteria – consistent with our SOFA-1-only group of 1.0%. Our study extends their work with a 28-fold larger cohort and comprehensive modelling. The TURK-SOFA2 study (NCT07300306), a prospective multicentre validation planned in Turkey, underscores the recognised need for external validation across diverse populations.

Several limitations warrant consideration. As a retrospective single-centre study, our findings are subject to information bias inherent to secondary data analysis. MIMIC-IV reflects a single US academic centre with a predominantly White population, limiting generalisability. Scoring missing organ components as zero per SOFA-2 convention may underestimate organ dysfunction but biases conservatively towards under-identification. Prospective multicentre validation is needed to determine whether the modest prognostic improvement justifies the added scoring complexity.

In conclusion, SOFA-2 provided consistently, though modestly, better prognostic discrimination and risk stratification than SOFA-1 in septic ICU patients in this single-centre cohort, reflecting its better alignment with contemporary organ-support practice. Whether this incremental benefit justifies the added scoring complexity in routine clinical use warrants evaluation through prospective multicentre studies.

CRediT authorship contribution statement

Yongye Wang: Visualization. **Hongxia Sun:** Visualization. **Zhifei Fu:** Methodology, Formal analysis. **Lifang Lin:** Methodology. **Songjie Bai:** Supervision, Conceptualization. **Sheng Zhang:** Methodology, Conceptualization. **Xuehuan Wen:** Writing – review & editing, Writing – original draft, Conceptualization.

Ethics approval and consent to participate

The MIMIC-IV database is a publicly available de-identified critical care database hosted on PhysioNet. Access was obtained after completion of the CITI programme and execution of a data use agreement with PhysioNet (credentialled access). As the database contains only de-identified data, the study was exempt from institutional review board approval and individual patient consent was not required.

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.

Declaration of generative AI and AI-assisted technologies in the writing process

Anthropic Claude assisted with limited code drafting/debugging, and OpenAI Codex CLI was used for an independent audit of the SOFA-2 SQL implementation before resubmission. Anthropic Claude assisted with language editing. All scientific decisions, methodological choices, interpretations, and final wording were made and verified by the authors.

Funding

This work was supported by the Zhejiang Provincial Medical and Health Science and Technology Program (Grant No. 2025KY323); the Science and Technology Project of Cangnan County (Grant No. 2025S07); the 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang Province (Grant No. 2023C03083); the Zhejiang Engineering Research Center for Intelligent Medical Imaging, Sensing and Non-invasive Rapid Testing; the National Natural Science Foundation of China (No. 82502624); the Natural Science Foundation of Jiangxi Province (No. 20252BAC200111); and the Taizhou Science and Technology Program (Grant No. 25ywb06).

Data availability

The data that support the findings of this study are available from the MIMIC-IV database at https://physionet.org/content/mimiciv/.

Declaration of competing interest

All authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.clinme.2026.100601.

References

1. Gray AP, Chung E, Hsu RL, et al. Global, regional, and national sepsis incidence and mortality, 1990–2021: a systematic analysis. *Lancet Glob Health.* 2025;13(12):e2013–e2026. https://doi.org/10.1016/s2214-109x(25)00356-0
2. Vincent J-L, Sakr Y, Singer M, et al. Prevalence and outcomes of infection among patients in intensive care units in 2017. *JAMA.* 2020;323(15):1478. https://doi.org/10.1001/jama.2020.2717
3. Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Med.* 2021;47(11):1181–1247. https://doi.org/10.1007/s00134-021-06506-y
4. 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. https://doi.org/10.1001/jama.2016.0287
5. Lan L, Zhou M, Chen X, et al. Prognostic accuracy of SOFA, MEWS, and SIRS criteria in predicting the mortality rate of patients with sepsis: a meta-analysis. *Nurs Crit Care.* 2024;29(6):1623–1635. https://doi.org/10.1111/nicc.13016
6. 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):707–710. https://doi.org/10.1007/bf01709751
7. Moreno R, Rhodes A, Piquilloud L, et al. The Sequential Organ Failure Assessment (SOFA) score: has the time come for an update? *Crit Care.* 2023;27(1):15. https://doi.org/10.1186/s13054-022-04290-9
8. Lee HJ, Ko BS, Ryoo SM, et al. Modified cardiovascular SOFA score in sepsis: development and internal and external validation. *BMC Med.* 2022;20(1):263. https://doi.org/10.1186/s12916-022-02461-7
9. Moor M, Rieck B, Horn M, et al. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Front Med.* 2021;8:607952. https://doi.org/10.3389/fmed.2021.607952
10. Komorowski M, Green A, Tatham KC, et al. Sepsis biomarkers and diagnostic tools with a focus on machine learning. *eBioMedicine.* 2022;86:104394. https://doi.org/10.1016/j.ebiom.2022.104394
11. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.* 2020;46(3):383–400. https://doi.org/10.1007/s00134-019-05872-y
12. 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. https://doi.org/10.1001/jama.2025.20516
13. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10(1):1. https://doi.org/10.1038/s41597-022-01899-x
14. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis. *JAMA.* 2016;315(8):762. https://doi.org/10.1001/jama.2016.0288
15. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput.* 2009;38(6):1228–1234. https://doi.org/10.1080/03610910902859574
16. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med.* 2015;34(4):685–703. https://doi.org/10.1002/sim.6370
17. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol.* 2016;69:245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005
18. Lambden S, Laterre PF, Levy MM, Francois B. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. *Crit Care.* 2019;23(1):374. https://doi.org/10.1186/s13054-019-2663-7
19. Moreno R, Rhodes A, Ranzani O, et al. Rationale and methodological approach underlying the development of the Sequential Organ Failure Assessment (SOFA)–2 score. *JAMA Netw Open.* 2025;8(10):e2545040. https://doi.org/10.1001/jamanetworkopen.2025.45040
20. Pencina MJ, D' Agostino RB, D' Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27(2):157–172. https://doi.org/10.1002/sim.2929
21. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* 2011;30(1):11–21. https://doi.org/10.1002/sim.4085
22. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7
23. Pölkki A, Pekkarinen PT, Takala J, et al. Association of Sequential Organ Failure Assessment (SOFA) components with mortality. *Acta Anaesthesiol Scand.* 2022;66(6):731–741. https://doi.org/10.1111/aas.14067
24. Sonneville R, Benghanem S, Jeantin L, et al. The spectrum of sepsis-associated encephalopathy: a clinical perspective. *Crit Care.* 2023;27(1):386. https://doi.org/10.1186/s13054-023-04655-8
25. Wang L, Ma X, Zhou G, et al. SOFA in sepsis: with or without GCS. *Eur J Med Res.* 2024;29(1):296. https://doi.org/10.1186/s40001-024-01849-w
26. Pan X, Xie J, Zhang L, et al. Evaluate prognostic accuracy of SOFA component score for mortality among adults with sepsis by machine learning method. *BMC Infect Dis.* 2023;23(1):76. https://doi.org/10.1186/s12879-023-08045-x
27. Yang R, Han D, Zhang L, et al. Analysis of the correlation between the longitudinal trajectory of SOFA scores and prognosis in patients with sepsis at 72 h after admission based on group trajectory modeling. *J Intensive Med.* 2022;2(1):39–49. https://doi.org/10.1016/j.jointm.2021.11.001
28. Xu Z, Mao C, Su C, et al. Sepsis subphenotyping based on organ dysfunction trajectory. *Crit Care.* 2022;26(1):197. https://doi.org/10.1186/s13054-022-04071-4
29. Pei F, Gu B, Liu Z, et al. Validation of SOFA-2 score in sepsis and exploration of its extension with additional immune markers. *J Intensive Med.* 2026;6(3):294–301. https://doi.org/10.1016/j.jointm.2025.12.003

Descargado para Anonymous User (n/a) en Andrés Bello University de ClinicalKey.es por Elsevier en julio 25, 2026. Para uso personal exclusivamente. No se permiten otros usos sin autorización. Copyright ©2026. Elsevier Inc. Todos los derechos reservados.