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.

# Predicting Sepsis Mortality: An Elastic-Net Regression Model versus the Sequential Organ Failure Assessment Score

**Reza Raeesi Estabragh**
Kerman University of Medical Sciences
**Neda Hamzehnejadi**
Kerman University of Medical Sciences
**Ramita Shahabifar**
ra.shahabifar@yahoo.com

Kerman University of Medical Sciences
**Morteza Hashemian**
Kerman University of Medical Sciences
**Mahdieh Khazaneha**
Kerman University of Medical Sciences
**Mahdieh Azizian**
Kerman University of Medical Sciences
**Mohammad Rahmanian**
Shiraz University of Medical Sciences

**Research Article**

**Keywords:** Sepsis, mortality, Sequential Organ Failure Assessment, clinical prediction model, elastic net, calibration, decision curve analysis

**Posted Date:** November 12th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11141215/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

Introduction: The Sequential Organ Failure Assessment version 1 (SOFA-1) score defines organ dysfunction in Sepsis-3. However, this score describes illness severity rather than predicts mortality. This study evaluated whether routine continuous admission variables improve mortality prediction.

Methods: This retrospective cohort included 465 adults with sepsis admitted to intensive care units between March 2022 and March 2024. Overall, 114 patients died in hospital. Candidate predictor sets included six continuous SOFA-1 proxies and 14 additional routine admission variables. Models were evaluated using 20 repeats of stratified 10-fold cross-validation. The primary model was an elastic-net penalised logistic regression. Discrimination was compared using the paired DeLong test.

Results: On held-out folds, the elastic-net model achieved an area under the curve (AUC) of 0.846. In comparison, SOFA-1 achieved an AUC of 0.711 (difference 0.134, 95% confidence interval (CI) 0.074 to 0.195; $p < 0.001$). The six-proxy model improved AUC by 0.089 (95% CI 0.033 to 0.144). The Glasgow Coma Scale alone matched SOFA-1 discrimination (difference in AUC 0.035, 95% CI − 0.025 to 0.096; $p = 0.240$). Adding age and sex to SOFA-1 yielded no significant change (difference in AUC 0.000, 95% CI − 0.017 to 0.018).

Conclusions: Routinely recorded continuous admission variables predicted in-hospital mortality better than SOFA-1 in this cohort. Prospective external validation is required before clinical use.


# Introduction

Sepsis is life-threatening organ dysfunction caused by a dysregulated host response to infection [1]. Sepsis-3 defines dysfunction as an acute increase of two or more points in the Sequential Organ Failure Assessment version 1 (SOFA-1) score [2]. The 2025 international consensus update revised component variables and thresholds [3]. The updated framework retained six organ systems scored from 0 to 4. Because the Sequential Organ Failure Assessment version 2 consensus retains ordinal banding, preserving continuous prognostic information remains a practical question.

The original SOFA-1 score was developed to describe organ dysfunction over time rather than to predict mortality [4–6]. It groups continuous measurements from six systems into ordinal bands. The Glasgow Coma Scale represents the neurological component. A systematic review reported area under the curve values between 0.61 and 0.88 for Sequential Organ Failure Assessment models [7]. A large cohort study reported discrimination of approximately 0.75 [8]. Grouping continuous predictors discards information, reduces statistical power, and obscures non-linear relationships [9, 10]. Data-driven cut points introduce further bias.

Machine learning models for sepsis outcomes consistently report high discrimination [11–14]. However, studies rarely assess calibration [15], and most published models carry high risk of bias [16]. Furthermore, machine learning shows no systematic performance advantage over logistic regression

across clinical prediction tasks [17]. In addition, sepsis early-warning systems have not demonstrated reduced mortality in trials [18].

This study developed a prognostic model for admission-time risk stratification in sepsis. The primary objective was to determine whether continuous admission variables discriminate mortality better than SOFA-1. Secondary objectives evaluated calibration, decision-curve net benefit, and performance attribution. Reporting follows the Transparent Reporting of Individual Prognosis or Diagnosis guidelines with artificial intelligence extensions [19] (Additional file 1).

# Methods and Materials

Participants

This single-centre retrospective cohort included 465 consecutive adults aged ≥ 18 years who met Sepsis-3 criteria. Patients were admitted to intensive care units (ICU) at Afzalipour Hospital between March 2022 and March 2024. Afzalipour Hospital is a tertiary referral centre affiliated with Kerman University of Medical Sciences in Kerman, Iran. All eligible records during the accrual window were included. This yielded 114 events across 20 candidate variables. This effective events-per-variable ratio of 5.700 required regularized regression to mitigate overfitting [20]. The dataset included demographics, vital signs, laboratory results, and comorbidity categories. It also included the Sequential Organ Failure Assessment version 1 (SOFA-1) score. Specific ICU therapeutic interventions and vasopressor doses were not recorded in the archival database.

The data custodian attested the following collection protocol:

- Vital signs and Glasgow Coma Scale: First recorded value after ICU admission.
- Laboratory panel: Initial admission panel retrieved from the hospital information system.
- The SOFA-1 score: Worst values within the first 24 h after ICU admission.
- Outcome: Sepsis-related in-hospital mortality.

Data preparation

Across the data, 35 missing cells (0.2%) were present in laboratory measurements. Missing continuous variables were imputed using in-fold medians. Missing categorical variables were imputed using in-fold modes. Imputation was nested strictly within training folds during cross-validation to prevent information leakage.

Outcome and Predictors

The primary outcome was in-hospital mortality recorded as a binary indicator. The follow-up time horizon ended at hospital discharge. Two nested predictor sets were evaluated. Set A comprised six continuously measured variables representing SOFA-1 components. Set B comprised Set A plus 14

routine admission variables. These included demographics, vital signs, extended laboratory markers, and comorbidity categories. Right-skewed laboratory variables entered linear models as log(1 + x).

### Models

Nine models were evaluated across the cohort. Four benchmark models included the uncalibrated SOFA-1 score, logistic SOFA-1, Glasgow Coma Scale alone, and SOFA-1 with demographics. An elastic net on Set A served as the attribution model. The primary model was an elastic net on Set B [21, 22]. Three comparator models included spline-ridge regression, random forest [23], and gradient-boosted trees [24].

### Validation procedure

All models were evaluated through a unified cross-validation pipeline. The evaluation used 20 repeats of stratified 10-fold cross-validation on shared folds. Continuous predictors were standardized within each training fold. Derived scaling parameters were applied to test folds. Hyperparameters were tuned using an inner 5-fold cross-validation loop within each training fold. Class rebalancing was excluded from primary models and evaluated solely in sensitivity analysis [25].

### Performance measures

Discrimination was assessed using the area under the curve (AUC) and precision–recall curves. Paired curves were compared using the DeLong test [26]. Calibration was assessed by intercept, slope, integrated calibration index, and E90 [15, 27]. Overall accuracy was measured by Brier score, scaled Brier score, and logarithmic loss. Differences in AUC are reported as medians across 20 repeats with paired DeLong 95% confidence intervals.

Clinical utility was evaluated through sensitivity, specificity, predictive values, and decision-curve net benefit [28–30]. Net reclassification improvement and integrated discrimination improvement were calculated relative to SOFA-1 [27]. Model optimism was estimated using 300 bootstrap resamples [31]. Information criteria were computed only when model likelihood was defined. Four pre-specified checks included permuted-outcome tests, fold-disjointness verification, an AUC upper ceiling, and bootstrap agreement. Analyses were conducted in R version 4.3.3.

## Results

### Cohort

Among 465 adults with sepsis, 114 (24.5%) died in hospital and 351 (75.5%) survived to discharge. Figure 1 illustrates participant flow.

Figure 1. Flow of participants through the study. All 465 adults meeting Sepsis-3 criteria were included in model development and validation. Overall, 114 (24.5%) died in hospital.

Table 1 summarizes baseline characteristics by outcome. Non-survivors presented with higher Sequential Organ Failure Assessment version 1 (SOFA-1) scores (median 7.000 versus 5.000). Non-survivors also had lower Glasgow Coma Scale (GCS) scores (median 7.000 versus 13.000). They showed higher respiratory rate, higher mean platelet volume, and higher creatinine. The red cell distribution width coefficient of variation did not differ between groups ($p = 0.710$). Haemoglobin, blood sugar, and sex also showed no significant differences.

Table 1
Baseline characteristics by outcome.

| Variable | Overall (n = 465) | Survived (n = 351) | Died (n = 114) | p |
|---|---|---|---|---|
| Age, years | 61.000 (45.000–73.000) | 60.000 (44.000–72.000) | 65.000 (49.200–77.800) | 0.027 |
| Sequential Organ Failure Assessment score | 5.000 (3.000–7.000) | 5.000 (3.000–7.000) | 7.000 (6.000–10.000) | < 0.001 |
| Glasgow Coma Scale | 11.000 (8.000–15.000) | 13.000 (9.000–15.000) | 7.000 (4.200–11.000) | < 0.001 |
| Respiratory rate, /min | 18.000 (16.000–20.000) | 18.000 (16.000–20.000) | 20.000 (16.000–22.800) | < 0.001 |
| $SpO_2$, % | 92.0 (85.0–96.0) | 93.0 (86.0–97.0) | 90.0 (81.0–96.0) | 0.024 |
| Platelets, $\times 10^9$/L | 215.000 (146.000–305.000) | 220.000 (153.000–309.000) | 192.000 (115.000–288.000) | 0.017 |
| White cell count, $\times 10^9$/L | 12.000 (8.300–19.700) | 11.900 (8.300–18.600) | 13.400 (8.500–22.000) | 0.140 |
| Haemoglobin, g/dL | 12.300 (10.200–14.700) | 12.400 (10.100–14.700) | 12.100 (10.200–14.800) | 0.788 |
| Mean platelet volume, fL | 10.500 (9.500–11.600) | 10.300 (9.400–11.200) | 11.000 (9.600–12.400) | < 0.001 |
| Red cell distribution width (coefficient of variation), % | 14.4 (13.0–16.7) | 14.3 (13.0–16.6) | 14.8 (13.0–16.8) | 0.710 |
| C-reactive protein, mg/L | 82.000 (32.000–149.000) | 78.000 (26.000–136.000) | 86.000 (46.000–171.000) | 0.022 |
| Erythrocyte sedimentation rate, mm/h | 15.000 (9.000–35.000) | 14.000 (8.000–33.000) | 19.500 (12.000–40.000) | < 0.001 |
| Creatinine, mg/dL | 1.200 (0.900–2.100) | 1.100 (0.900–1.800) | 1.700 (1.100–3.100) | < 0.001 |
| Blood urea nitrogen, mg/dL | 52.000 (33.000–92.000) | 46.000 (31.000–84.000) | 70.000 (42.000–123.000) | < 0.001 |
| Bilirubin, mg/dL | 3.300 (2.300–5.700) | 3.300 (2.300–5.500) | 3.400 (2.500–6.000) | 0.169 |

| Variable | Overall (n = 465) | Survived (n = 351) | Died (n = 114) | p |
|---|---|---|---|---|
| Blood sugar, mg/dL | 128.000 (95.000–191.000) | 128.000 (94.000–192.000) | 128.000 (98.000–179.000) | 0.864 |
| Blood culture positive | 61 (13.1%) | 45 (12.8%) | 16 (14.0%) | 0.862 |

Continuous variables are median (interquartile range); categorical are n (%). The p-values are from Mann–Whitney U or chi-squared/Fisher tests and are descriptive. $SpO_2$, peripheral oxygen saturation.

Discrimination

Table 2 and Fig. 2 present out-of-fold performance. The elastic-net model achieved an area under the curve (AUC) of 0.846 compared to 0.711 for the SOFA-1 model.

Table 2
Out-of-fold model performance across cross-validation.

| Model | *k* | Area under curve | Brier | Calibration slope | Calibration index | Akaike criterion | Bayesian criterion |
|---|---|---|---|---|---|---|---|
| Gradient-boosted ensemble | — | 0.847 | 0.116 | 1.102 | 0.023 | — | — |
| Elastic net, Set B (primary) | 14 | 0.846 | 0.122 | 1.132 | 0.019 | 367.200 | 425.200 |
| Ridge + natural splines | 15 | 0.845 | 0.123 | 1.029 | 0.011 | 365.900 | 428.000 |
| Random forest | — | 0.839 | 0.122 | 1.209 | 0.049 | — | — |
| Elastic net, Set A (six proxies) | 7 | 0.800 | 0.144 | 1.041 | 0.019 | 415.700 | 444.700 |
| Glasgow Coma Scale alone | 2 | 0.747 | 0.155 | 0.977 | 0.025 | 448.000 | 456.300 |
| Sequential Organ Failure Assessment, unfitted score | — | 0.718 | 0.166 | 0.125 | 0.064 | — | — |
| Sequential Organ Failure Assessment with age and sex | 4 | 0.712 | 0.162 | 0.928 | 0.018 | 466.400 | 483.000 |
| Sequential Organ Failure Assessment (logistic) | 2 | 0.711 | 0.163 | 0.972 | 0.033 | 466.300 | 474.600 |

Data reflect 20 repeats of stratified 10-fold cross-validation with identical folds. The symbol *k* denotes effective parameters. Information criteria are undefined for tree ensembles.

Figure 2. Receiver operating characteristic curves, out-of-fold across 20 repeats of stratified 10-fold cross-validation on shared folds. The elastic-net model reaches an AUC of 0.846 against 0.711 for the SOFA-1 model.

Table 3 details paired comparisons against SOFA-1. The elastic-net model discriminated better than SOFA-1, with difference in AUC of 0.134 (95% confidence interval (CI) 0.074 to 0.195; p < 0.001). The gradient-boosted ensemble and spline model performed similarly to the primary model. Two comparisons were null. The GCS-only model reached an AUC of 0.747 against SOFA-1's 0.711 (difference in AUC 0.035, 95% CI − 0.025 to 0.096; p = 0.240). Adding age and sex to SOFA-1 did not improve discrimination (difference in AUC 0.000, 95% CI − 0.017 to 0.018; p = 0.029).

Table 3
Difference in AUC against the SOFA-1 model.

| Model | Difference in area under curve | 95% confidence interval | p |
|---|---|---|---|
| Elastic net, Set B (primary) | 0.134 | 0.074 to 0.195 | < 0.001 |
| Gradient-boosted ensemble | 0.135 | 0.075 to 0.195 | < 0.001 |
| Ridge + natural splines | 0.133 | 0.070 to 0.197 | < 0.001 |
| Random forest | 0.126 | 0.064 to 0.187 | < 0.001 |
| Elastic net, Set A (six proxies) | 0.089 | 0.033 to 0.144 | < 0.001 |
| Glasgow Coma Scale alone | 0.035 | −0.025 to 0.096 | 0.240 |
| Sequential Organ Failure Assessment with age and sex | 0.000 | −0.017 to 0.018 | 0.029 |

Values are median across 20 repeats with paired DeLong 95% confidence intervals. The p-values are from the paired DeLong test on pooled predictions.

Attribution of performance gain

Figure 3 shows the attribution analysis. Replacing SOFA-1 with six continuous admission proxies improved AUC by 0.089 (95% CI 0.033 to 0.144; p < 0.001). Adding 14 routine markers contributed an additional 0.045 (95% CI 0.015 to 0.075; p = 0.003). Compared with GCS alone, the full model improved AUC by 0.098 (95% CI 0.051 to 0.144).

Figure 3. Attribution of the gain over SOFA-1. Bars are paired DeLong 95% confidence intervals; the inner ticks show the Monte Carlo spread across cross-validation repeats.

Calibration and clinical utility

Figure 4 illustrates model calibration. The elastic-net model showed a calibration intercept of 0.009 and a slope of 1.132. Its integrated calibration index was 0.019 compared to 0.033 for the SOFA-1 model. The spline model achieved the closest calibration (slope 1.029, integrated calibration index 0.011). In contrast, the random forest showed the poorest calibration among fitted models (slope 1.209, integrated calibration index 0.049).

Figure 4. Calibration curves with the integrated calibration index for each model. The elastic-net model slope of 1.132 indicates mild under-fitting of the extremes.

At the Youden-optimal threshold, the elastic-net model achieved accuracy 0.805, sensitivity 0.676, and specificity 0.846. Positive predictive value was 0.589, and negative predictive value was 0.889. The SOFA-1 model reached accuracy 0.634, sensitivity 0.763, specificity 0.593, positive predictive value 0.378, and negative predictive value 0.885. The SOFA-1 model achieved higher sensitivity. However, its lower specificity misclassified 40.7% of survivors as high risk.

At a rule-out threshold (sensitivity ≥ 0.900), the elastic-net model maintained specificity of 0.597. In comparison, SOFA-1 specificity dropped to 0.168. A fixed threshold of 0.500 missed over half of all deaths (sensitivity 0.488 for elastic net, 0.192 for SOFA-1). However, it exhibited a misleadingly high overall accuracy of 0.837.

Decision curves (Fig. 5) showed higher net benefit for the elastic-net model across threshold probabilities from 0.020 to 0.600. Relative to SOFA-1, categorical net reclassification improvement was 0.470 and integrated discrimination improvement was 0.189.

Figure 5. Decision curves, net benefit against threshold probability from 0.010 to 0.600, for the elastic-net model, the SOFA-1 model, treat-all and treat-none. The elastic-net model has higher net benefit across the range from 0.020 to 0.600.

Variable importance and sensitivity analyses

Permutation importance ranked GCS highest (AUC reduction 0.140, 95% CI 0.115 to 0.156). Respiratory rate followed (0.050), then mean platelet volume (0.032), and creatinine (0.016). The contribution of GCS was approximately 2.8 times that of respiratory rate. Stability selection retained ten variables across 100.0% of folds.

Bootstrap optimism correction yielded a corrected AUC of 0.846 for the elastic-net model. Apparent optimism scaled with model complexity, ranging from 0.001 for SOFA-1 to 0.031 for the elastic net.

Excluding red cell distribution width changed AUC by 0.000. In-fold winsorizing changed AUC by 0.000. Complete-case analysis (n = 430) yielded an AUC of 0.833. Oversampling the minority class to 50.0% altered discrimination minimally (AUC loss 0.006). However, it shifted the calibration intercept from 0.008 to − 1.042 and worsened the integrated calibration index to 0.144.

In the permuted-outcome test, the primary pipeline yielded an out-of-fold AUC of 0.422. Across 36 permutations, the maximum observed AUC was 0.572. All models passed fold-disjointness checks. The highest observed out-of-fold AUC was 0.847, below the pre-specified ceiling of 0.900.

## Discussion

In sepsis, continuous admission variables predicted in-hospital mortality more accurately than Sequential Organ Failure Assessment version 1 scores. This advantage came from preserving continuous measurements of organ function rather than coarse integer categories. The Glasgow Coma Scale (GCS) alone matched the predictive discrimination of the composite score. In contrast, demographic variables added no prognostic value beyond acute physiological derangements.

The Sequential Organ Failure Assessment score was designed in 1996 to track organ dysfunction in trials [4, 5]. Dynamic trajectories capture prognosis [6]. However, Sepsis-3 adopted acute score increases as diagnostic criteria [1, 2, 32]. This entrenched the banded score as a mortality benchmark [8]. Subsequent machine learning studies claimed high discrimination [11–14, 33]. Yet systematic audits identified a high risk of bias [16, 34, 35] and poor calibration [15]. Machine learning showed no systematic benefit over regularized regression [17, 36]. Furthermore, early-warning systems have failed to lower trial mortality [18]. Recent consensus validations [3, 37–40] confirm that integer binning still bounds discrimination.

Grouping continuous predictors into integer intervals discards non-linear risk gradients and reduces statistical power [9, 10]. In the present study, modeling continuous core organ variables raised out-of-fold discrimination from 0.711 to 0.800. This single step recovered 65.8% of the total performance gap. Furthermore, summed integer scores assume equal weight across organ systems despite varying component contributions [41]. Regularized regression estimates optimal multi-system weights directly from continuous log-odds slopes [21, 42].

Admission GCS was the strongest individual predictor. It achieved an area under the curve of 0.747 alone. During acute sepsis, encephalopathy develops rapidly through microvascular thrombosis and impaired autoregulation [43]. Because unbuffered brain tissue depends on constant perfusion, the nervous system acts as an immediate sensor of failure. In contrast, renal and hepatic markers require 12 to 48 h of clearance latency to accumulate [33]. Bedside neurological assessment and compensatory tachypnoea [44] detect acute collapse earlier than laboratory panels.

Mean platelet volume was retained across all stability folds and independently predicted mortality. Severe sepsis triggers consumptive coagulopathy and microvascular sequestration. This process

stimulates the release of hyper-reactive reticulated platelets that accelerate immunothrombosis [45, 46]. In contrast, red cell distribution width was uninformative, diverging from retrospective claims [47, 48]. Because erythrocytes have a 120-day lifespan, anisocytosis reflects chronic baseline health rather than acute shock.

Complex tree ensembles showed no performance advantage over regularized linear models [17, 36]. Tree ensembles build high-variance partitions across correlated variables [22–24], whereas regularized regression fits smooth boundaries [21]. Artificial class rebalancing via oversampling altered discrimination minimally [25]. However, it caused severe calibration collapse [15, 49]. Decision-curve analysis confirmed positive net benefit for the elastic-net model across threshold probabilities from 0.020 to 0.600 [28–30, 50]. Applying prevalence-calibrated thresholds remains essential for clinical deployment [27, 30].

Several limitations warrant consideration. With 114 deaths across 20 candidate variables, this single-centre study requires external validation [20, 31, 51–53]. Predictor timing was attested rather than machine-timestamped [54, 55]. Candidate GCS was recorded upon admission. In contrast, total Sequential Organ Failure Assessment version 1 scores incorporated worst 24 h values. In addition, GCS is subject to inter-observer variability and unrecorded sedation [55]. Finally, the static admission design precluded evaluating dynamic trajectories [6, 42], penalized degrees of freedom remain approximate [56], and subgroup audits were constrained [19].

# Conclusions

In sepsis, continuous admission variables demonstrated superior discrimination and calibration compared with categorical Sequential Organ Failure Assessment scores. This approach also improved decision-curve net benefit across clinical risk thresholds. Preserving continuous organ dysfunction parameters recovered most of this performance gain. This finding illustrates the statistical penalty of integer banding in risk stratification. Prospective external validation and randomized trials remain necessary before clinical implementation.

# Declarations

# Abbreviations

Standard abbreviations are defined upon first use in the main text and table footnotes.

**Ethics approval and consent to participate.** The Ethics Committee of Kerman University of Medical Sciences approved the study (ethics registry number: 1401-221). The need for informed consent was waived because this retrospective analysis used de-identified medical records. The study was conducted in accordance with the Declaration of Helsinki and Iranian national regulations.

**Consent for publication.** Not applicable.

**Availability of data and materials.** The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

**Competing interests.** The authors declare no conflicts of interest.

**Funding.** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Authors' contributions.** R.R.E. contributed to study conception, clinical interpretation, and critical revision. R.S. supervised radiological and clinical interpretation of the data, coordinated the multidisciplinary team, and critically revised the manuscript. N.H. collected and preprocessed the data, performed statistical analyses, and drafted the manuscript. M.H. curated data and contributed to clinical interpretation. M.K. contributed to methodology, formal analysis, and manuscript revision. M.A. collected data and recruited patients. M.R. curated data and administered the project. All authors read and approved the final manuscript.

**Use of generative artificial intelligence** During preparation of this manuscript, the authors used generative artificial intelligence only to refine grammar, wording, and readability; they did not use it to generate or analyse data, design the study, or interpret the findings, and they reviewed and approved the final text and take full responsibility for its content.

**Acknowledgements.** None.

**Authors' information.** Author affiliations are provided on the title page.

**Study protocol and registration.** No formal protocol was registered. The analysis plan was documented in writing before model fitting.

**Patient and public involvement.** Patients and the public were not involved in the design, conduct, reporting, or dissemination of this study.

# References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). J Am Med Association (JAMA). 2015;315:801. https://doi.org/10.1001/jama.2016.0287.
2. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA J Am Med Assoc. 2016;315:762–74. https://doi.org/10.1001/jama.2016.0288.
3. Ranzani O, Singer M, Salluh J, Shankar-Hari M, Pilcher D, Berger-Estilita J, et al. Development and validation of the sequential organ failure assessment (SOFA)-2 score. J Am Med Association (JAMA). 2025. https://doi.org/10.1001/jama.2025.20516. 334:2090.

4. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European society of intensive care medicine. Intensive Care Med. 1996;22:707–10. https://doi.org/10.1007/BF01709751.
5. Vincent JL, de Mendonça A, Cantraine F, Moreno R, Takala J, Suter PM, 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. 1998;26:1793–800. https://doi.org/10.1097/00003246-199811000-00016.
6. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent J. Serial evaluation of the SOFA score to predict outcome in critically ill patients. J Am Med Association (JAMA). 2001;286:1754–8. https://doi.org/10.1001/jama.286.14.1754.
7. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Crit Care. 2008;12:R161. https://doi.org/10.1186/cc7160.
8. Raith E, Udy A, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. J Am Med Association (JAMA). 2017;317:290. https://doi.org/10.1001/jama.2016.20328.
9. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. Stat Med. 2006;25:127–41. https://doi.org/10.1002/sim.2331.
10. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using optimal cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829–35. https://doi.org/10.1093/jnci/86.11.829.
11. Fleuren L, Klausch T, Zwager CL, Schoonmade L, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46:383–400. https://doi.org/10.1007/s00134-019-05872-y.
12. Islam M, Poly TN, Walther B, Wu C-C, Yang H-C, Li Y. Prediction of sepsis patients using machine learning approach: A meta-analysis. Comput Methods Programs Biomed. 2019;170:1–9. https://doi.org/10.1016/j.cmpb.2018.12.027.
13. Zhang Y, Xu W, Yang P, Zhang A. Machine learning for the prediction of sepsis-related death: A systematic review and meta-analysis. BMC Med Inf Decis Mak. 2023;23:283. https://doi.org/10.1186/s12911-023-02383-1.
14. AbuHaweeleh MN, Chowdhury AT, Newaz M, Saha P, Islam KR, Kumar J, et al. Sepsis mortality prediction using machine learning and deep learning – a systematic review. BMC Med Inf Decis Mak. 2025;26:16. https://doi.org/10.1186/s12911-025-03286-z.
15. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: The achilles heel of predictive analytics. BMC Med. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.

16. Wynants L, van Calster B, Bonten M, Collins G, Debray T, de Vos M, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369:m1328. https://doi.org/10.1136/bmj.m1328.
17. Christodoulou E, Ma J, Collins G, Steyerberg E, Verbakel J, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
18. Alshaeba SE, Marhoffer EA, Holleck JL, Theisen-Toupal J, Grimshaw AA, Gunderson CG. The effect of early warning systems for sepsis on mortality: A systematic review and meta-analysis. J Gen Intern Med. 2025;40:3463–8. https://doi.org/10.1007/s11606-025-09569-5.
19. Collins GS, Moons K, Dhiman P, Riley R, Beam A, van Calster B, et al. TRIPOD + AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
20. Riley R, Snell KIE, Ensor J, Burke D, Harrell F, Moons K, et al. Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. Stat Med. 2018;38:1276–96. https://doi.org/10.1002/sim.7992.
21. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B (Statistical Methodology). 2005;67:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
22. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22. https://doi.org/10.18637/jss.v033.i01.
23. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
24. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
26. DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics J Int Biometric Soc. 1988;44:837. https://doi.org/10.2307/2531595.
27. Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiol (Cambridge Mass). 2010. https://doi.org/10.1097/EDE.0b013e3181c30fb2.
28. Vickers A, Elkin E. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.
29. Vickers A, van Calster B, Steyerberg E. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. https://doi.org/10.1136/bmj.i6.
30. van Calster B, Wynants L, Verbeek J, Verbakel J, Christodoulou E, Vickers A, et al. Reporting and interpreting decision curve analysis: A guide for investigators. Eur Urol. 2018;74:796–804. https://doi.org/10.1016/j.eururo.2018.08.038.

31. Steyerberg E, Harrell F, Borsboom G, Eijkemans M, Vergouwe Y, Habbema J. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81. https://doi.org/10.1016/s0895-4356(01)00341-9.
32. Gerlach H. Developing a new definition and assessing new clinical criteria for septic shock: For the third international consensus definitions for sepsis and septic shock (sepsis-3). Fac Opinions – Post-Publication Peer Rev Biomedical Literature. 2018;315:775–87. https://doi.org/10.3410/F.726165854.793550943.
33. Lv X, Liu D, Chen X, Chen L, Wang X, Xu X, et al. Machine learning for the prediction of mortality in patients with sepsis-associated acute kidney injury: A systematic review and meta-analysis. BMC Infect Dis. 2024;24:1454. https://doi.org/10.1186/s12879-024-10380-6.
34. Moons K, Wolff R, Riley R, Whiting P, Westwood M, Collins G, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med. 2019;170:W1–33. https://doi.org/10.7326/M18-1377.
35. Moons K, Damen J, Kaul T, Hooft L, Navarro CAA, Dhiman P, et al. PROBAST + AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505.
36. van Calster B, Verbakel J, Christodoulou E, Steyerberg E, Collins G. Statistics versus machine learning: Definitions are interesting (but understanding, methodology, and reporting are more important). J Clin Epidemiol. 2019;116:137–8. https://doi.org/10.1016/j.jclinepi.2019.08.002.
37. Helleberg J, Sundelin A, Soltani N, Rooyackers O, Mårtensson J. Predictive validity of daily sequential organ failure assessment (SOFA)-2 score for 30-day mortality. Crit Care. 2026;30:280. https://doi.org/10.1186/s13054-026-06093-8.
38. Deveci K, Boran M, Bahar KE, Sarıdağ F, Eryılmaz Eren E. Does SOFA-2 improve mortality prediction in sepsis? A retrospective single-center observational cohort study comparing SOFA-1 and SOFA-2. BMC Infect Dis. 2026. https://doi.org/10.1186/s12879-026-13829-y.
39. Bai S, Huang M, Chen H, Yang X, Chen L, Liu F, et al. SOFA-2 versus SOFA-1 for mortality prediction in critically ill patients. J Intensive Care. 2026;14:32. https://doi.org/10.1186/s40560-026-00875-6.
40. Türkkan HM, Özgür U, Dilken O, Bayrak A, Derin O. SOFA-2 versus SOFA-1 for mortality prediction in infection-triggered ICU patients. J Intensive Care Med. 2026;8850666261435547. https://doi.org/10.1177/08850666261435547.
41. Pölkki A, Pekkarinen PT, Takala J, Selander T, Reinikainen M. Association of sequential organ failure assessment (SOFA) components with mortality. Acta Anaesthesiol Scand. 2022;66:731–41. https://doi.org/10.1111/aas.14067.
42. Houthooft R, Ruyssinck J, van der Herten J, Stijven S, Couckuyt I, Gadeyne B, et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artif Intell Med. 2015;63:191–207. https://doi.org/10.1016/j.artmed.2014.12.009.
43. Shen D, Zhou S, Wu H, Zhang W, Li Z, Sun J. Associations between glasgow coma scale trajectories and 28-day survival rate in patients with sepsis-associated encephalopathy: Insights from

longitudinal group trajectory modeling. Front Neurol. 2025;16:1607946. https://doi.org/10.3389/fneur.2025.1607946.

44. Daorattanachai K, Maithong S, Phungoen P, Weschawalit S, Srivilaithon W. Respiratory national early warning score for 28-day mortality prediction in suspected sepsis patients in the emergency department. BMC Emerg Med. 2025;26:17. https://doi.org/10.1186/s12873-025-01443-1.
45. Mangalesh S, Dudani S, Malik A. Platelet indices and their kinetics predict mortality in patients of sepsis. Indian J Hematol Blood Transfus. 2021;37:600–8. https://doi.org/10.1007/s12288-021-01411-2.
46. Toro-Huamanchumo C, Cabanillas-Ramirez C, Quispe-Vicuña C, Caballero-Alvarado J, León-Figueroa DA, Cruces-Tirado N, et al. Mean platelet volume in neonatal sepsis: Meta-analysis of observational studies. Child (Washington DC). 2022;9:1821. https://doi.org/10.3390/children9121821.
47. Kim CH, Park J, Kim EJ, Han J, Han JS, Choi J, et al. An increase in red blood cell distribution width from baseline predicts mortality in patients with severe sepsis or septic shock. Crit Care. 2013;17:R282. https://doi.org/10.1186/cc13145.
48. Dankl D, Rezar R, Mamandipoor B, Zhou Z, Wernly S, Wernly B, et al. Red cell distribution width is independently associated with mortality in sepsis. Med Principles Pract. 2022;31:187–94. https://doi.org/10.1159/000522261.
49. van Calster B, Collins G, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. Lancet Digit HEALTH. 2025;7:100916. https://doi.org/10.1016/j.landig.2025.100916.
50. Vickers A, van Calster B, Steyerberg E. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Prognostic Res. 2019;3:18. https://doi.org/10.1186/s41512-019-0064-7.
51. Riley R, Ensor J, Snell K, Archer L, Whittle R, Dhiman P, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. Lancet Digit HEALTH. 2025;7:100857. https://doi.org/10.1016/j.landig.2025.01.013.
52. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA. 2002;99:6562–6. https://doi.org/10.1073/pnas.102102699.
53. van Calster B, Steyerberg E, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023;21:70. https://doi.org/10.1186/s12916-023-02779-w.
54. Whittle R, Peat G, Belcher J, Collins GS, Riley RD. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. J Clin Epidemiol. 2018;102:38–49. https://doi.org/10.1016/j.jclinepi.2018.05.008.
55. Pajouheshnia R, Smeden M, Peelen L, Groenwold R. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. J Clin Epidemiol. 2019;105:136–41. https://doi.org/10.1016/j.jclinepi.2018.09.001.

56. Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. Ann Stat. 2007;35:2173–92. https://doi.org/10.1214/009053607000000127.

# Figures

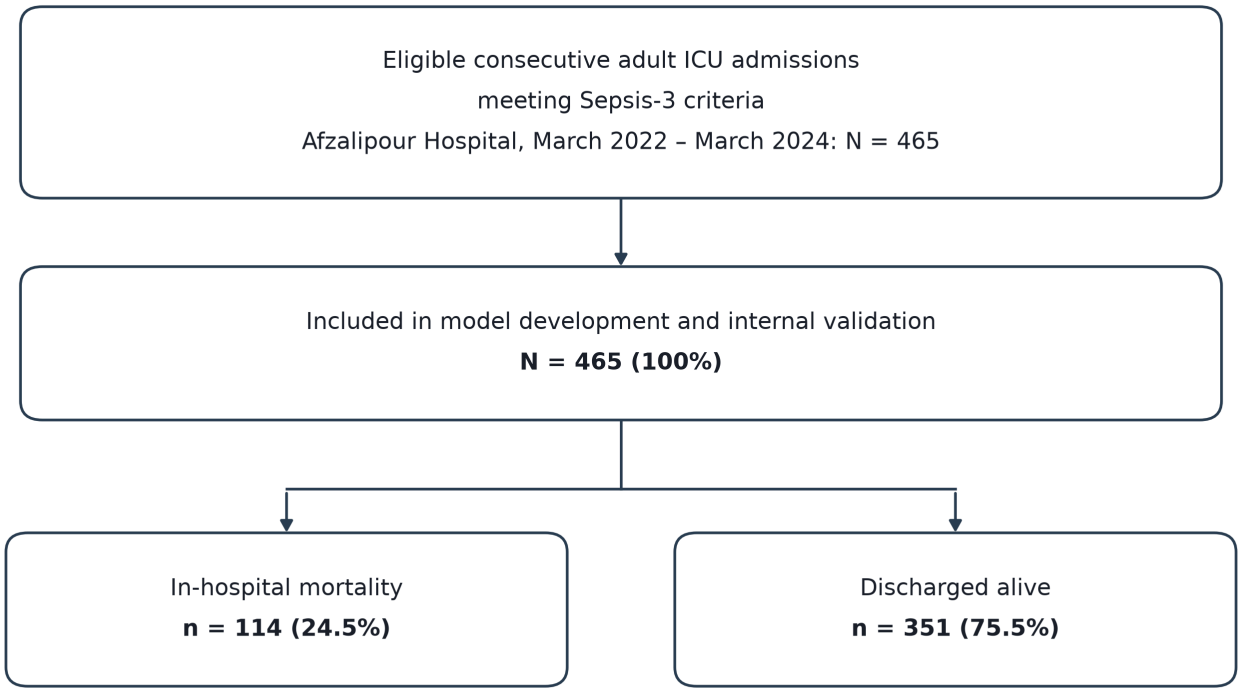



**Figure 1**

Flow of participants through the study.

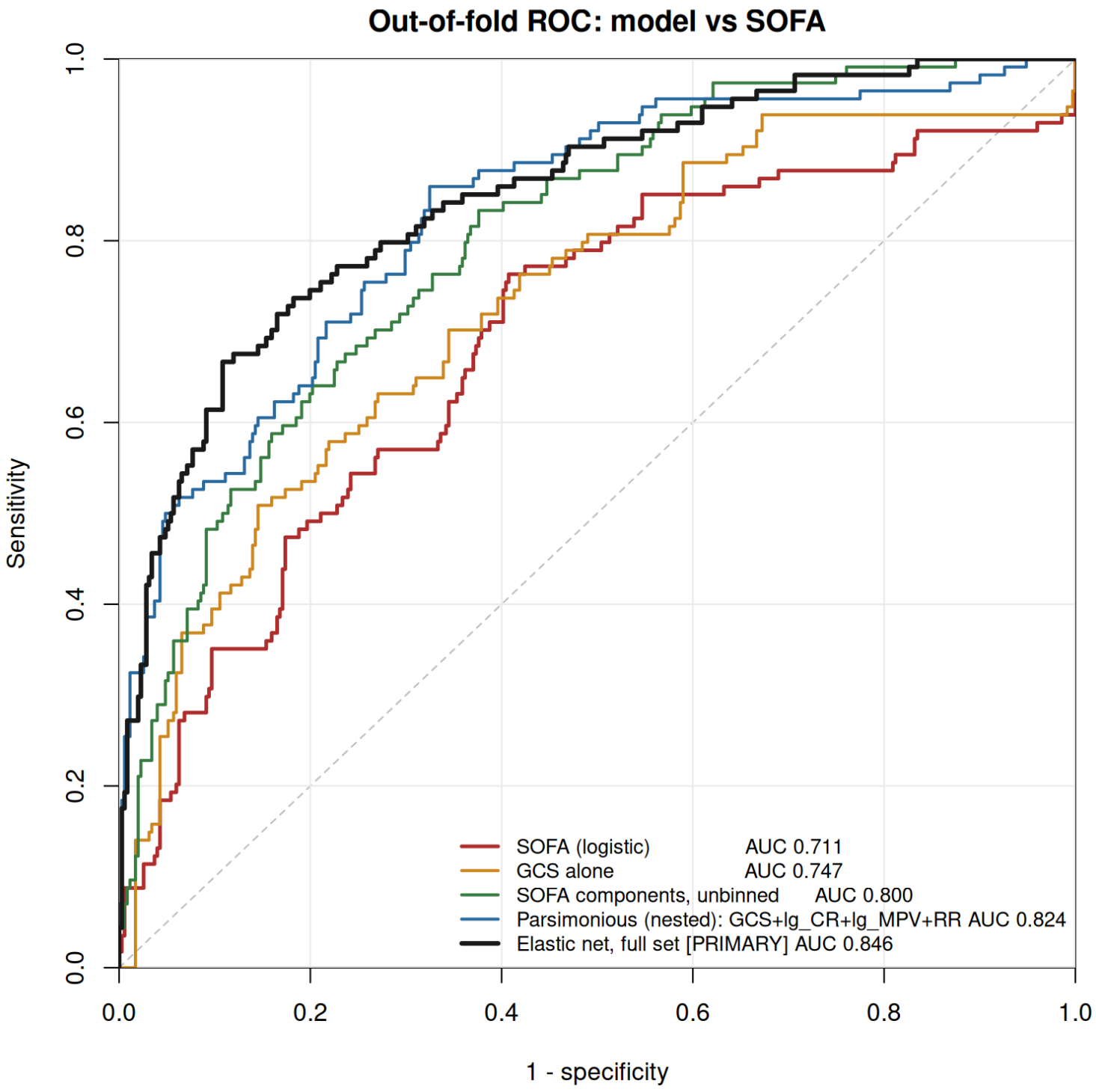



**Figure 2**

Receiver operating characteristic curves out-of-fold.

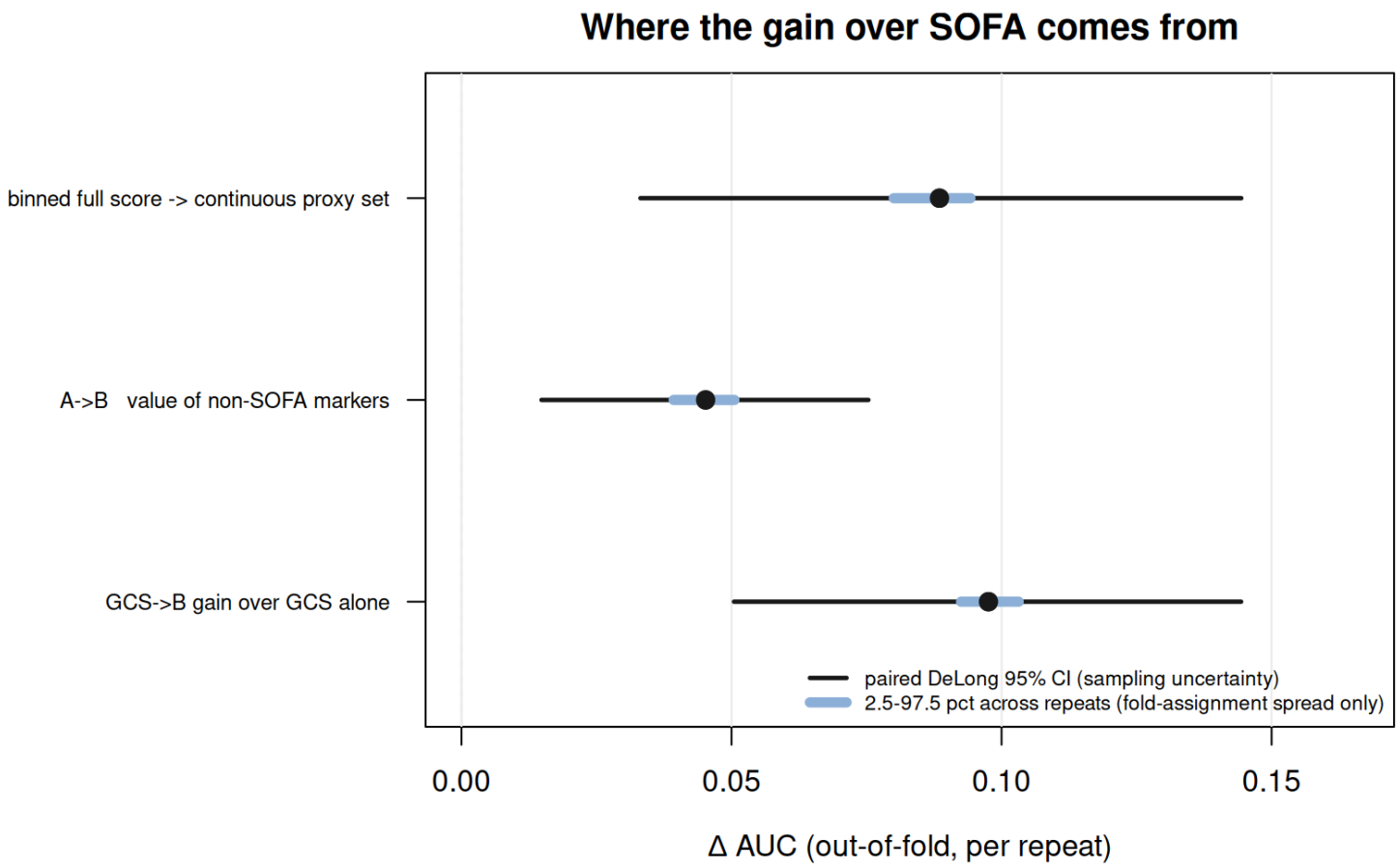



**Figure 3**

Attribution of the performance gain over SOFA-1.

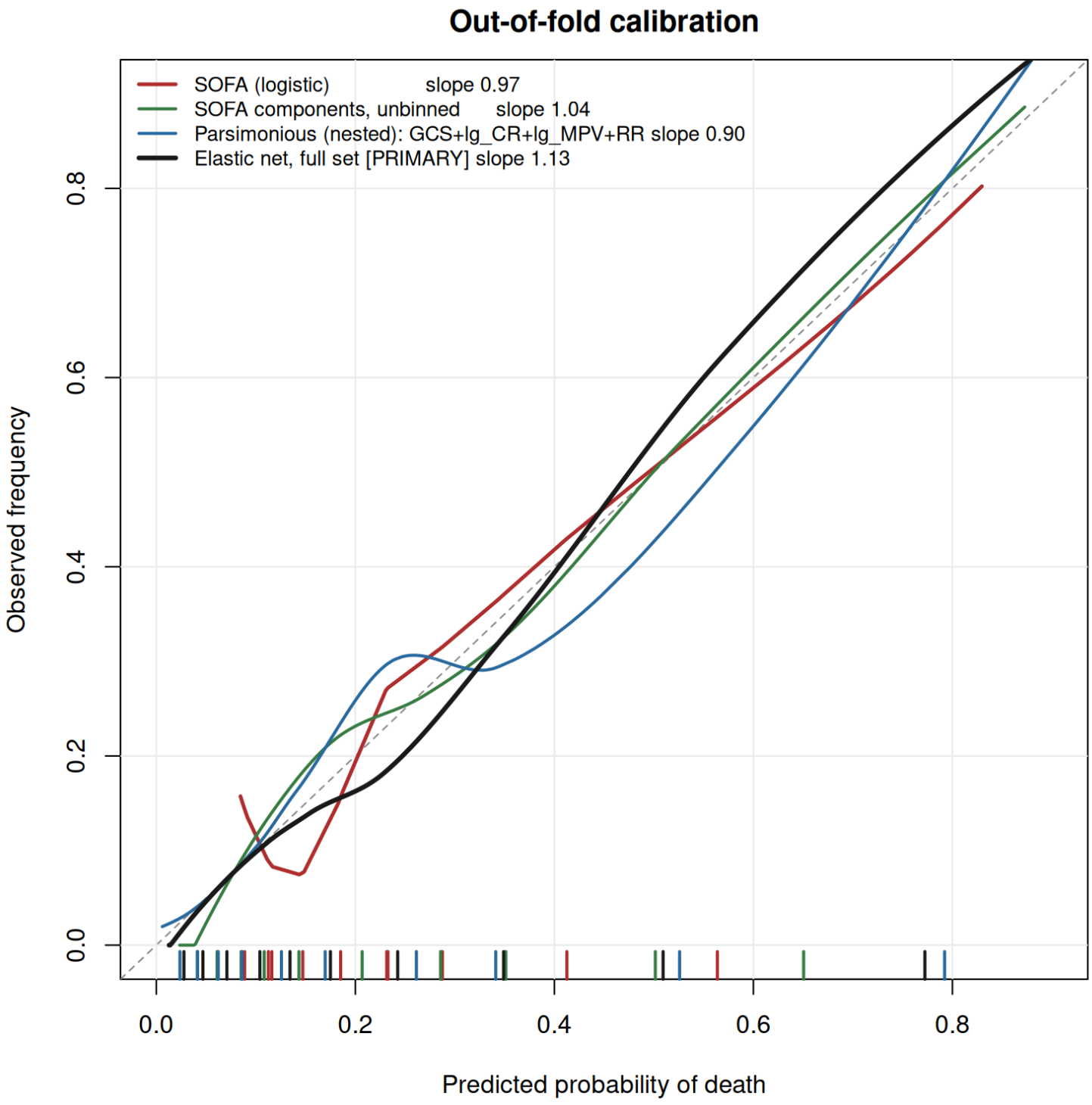



**Figure 4**

Calibration curves with the integrated calibration index for each model.

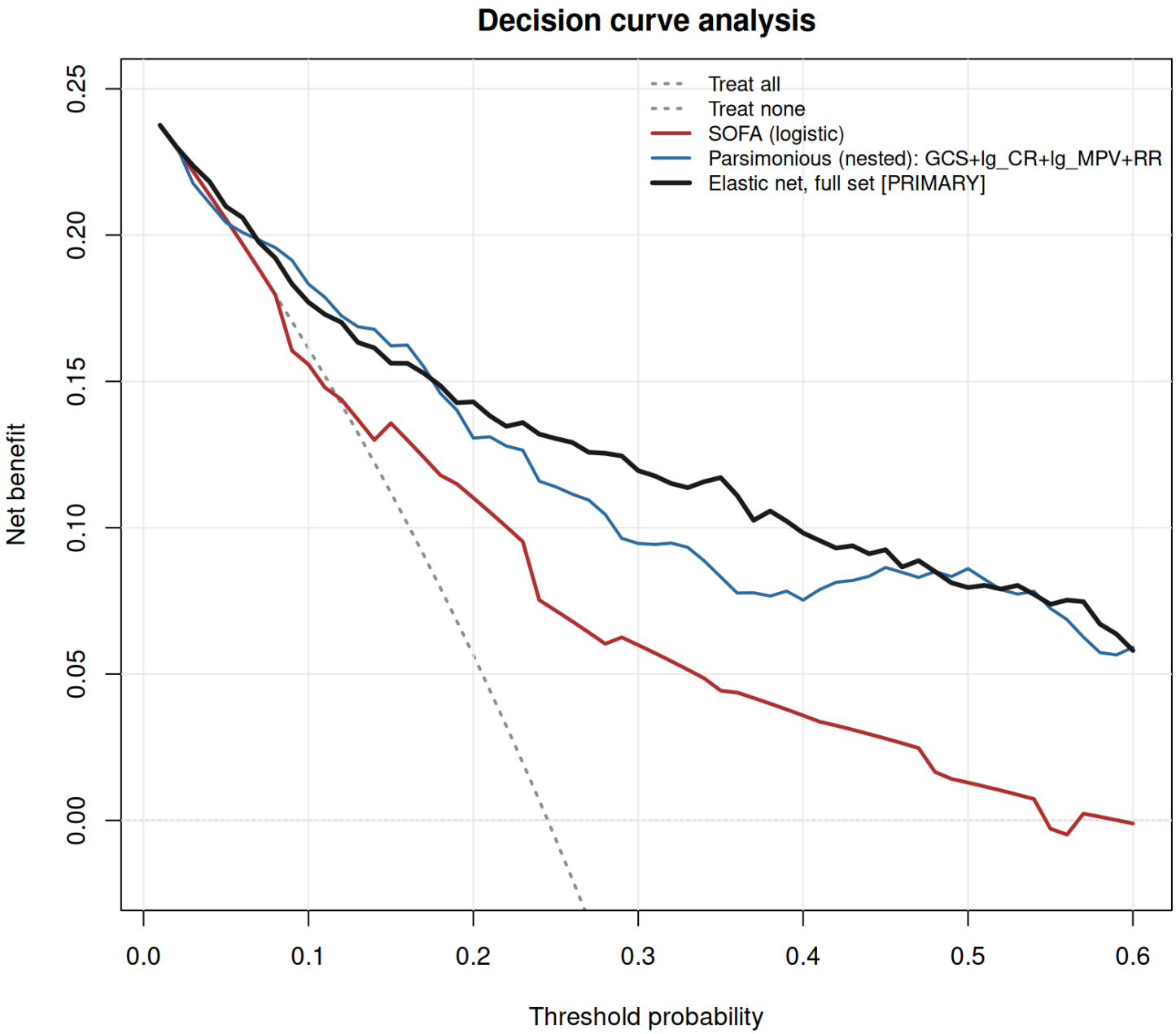



**Figure 5**

Decision curves of net benefit across threshold probabilities.

## Supplementary Files

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

- TableS1.csv
- TableS2.csv
- TRIPODAIchecklist.docx