International Journal of Clinical and Biomedical Research
Vol. 11, No. 2 (2026)
pISSN: 2521-0394 · eISSN: 2395-0471

ORIGINAL RESEARCH

APACHE II, APACHE III AND SOFA SCORES AS PREDICTORS OF MORTALITY IN ACUTE RESPIRATORY DISTRESS SYNDROME: A COMPARATIVE PROSPECTIVE STUDY WITH ROC ANALYSIS

Hakim Baset[1], **Harika K**[2*], **Mithun N. Kagalkar**[3]

[1] *Department of General Medicine, Grant Government Medical College and Sir J. J. Group of Hospitals, Maharashtra, India*

[2] *Department of General Medicine, Shri Atal Bihari Vajpayee Medical College and Research Institute, Karnataka, India*

[3] *Consultant Interventional Cardiologist, Metro United Healthcare, Shivamogga, Maharashtra, India*

*Correspondence: Harika K <harikak97@gmail.com>

ABSTRACT

Background: APACHE II, APACHE III and SOFA are widely used intensive care severity scores, but their comparative discrimination for mortality in acute respiratory distress syndrome (ARDS) is not well established in Indian cohorts.

Methods: In this prospective observational study in the medical intensive care unit of a tertiary care hospital, 100 patients fulfilling the Berlin definition of ARDS were enrolled. All three scores were calculated within 24 hours of admission and related to Berlin severity and outcome by logistic regression and receiver operating characteristic analysis.

Results: All three scores rose across severity strata and were higher in the 27 non-survivors than in the 73 survivors ($p<0.001$ throughout): APACHE II 35.8 versus 21.9, APACHE III 87.85 versus 53.14 and SOFA 9.0 versus 6.6. Each discriminated mortality better than chance — APACHE II area under the curve 0.895 (95% confidence interval 0.812–0.978), APACHE III 0.797 (0.689–0.905), SOFA 0.765 (0.651–0.879) — although the intervals overlap and no paired comparison of the curves was performed. At its optimum cut-off of ≥27, APACHE II gave 85.2% sensitivity and 78.1% specificity, with negative predictive value 93.4% but a positive predictive value of only 59.0%. On logistic regression, APACHE II (Wald 12.418), APACHE III (Wald 7.854) and mode of ventilatory support (Wald 10.29) were retained and age, sex, primary diagnosis and SOFA were not; with 27 events and 14 model degrees of freedom this model is over-parameterised.

Conclusion: APACHE II had the highest point estimate of discrimination, but overlapping confidence intervals do not establish superiority over APACHE III or SOFA.

Keywords: *Acute respiratory distress syndrome; APACHE II; APACHE III; SOFA; ROC curve; mortality prediction; severity scoring*

INTRODUCTION

Acute respiratory distress syndrome (ARDS) is defined by the acute onset of bilateral radiographic opacities and hypoxaemia not fully explained by cardiac failure or fluid overload, and is graded as mild, moderate or severe by the ratio of arterial oxygen tension to fractional inspired oxygen at a positive end-expiratory pressure of at least 5 cmH_2O.[1] It accounts for approximately one in ten intensive care unit (ICU) admissions worldwide, with hospital mortality rising across the Berlin strata and exceeding 40% in severe disease.[2] Because most deaths follow multi-organ dysfunction rather than refractory hypoxaemia alone,[3] the Berlin classification which grades oxygenation only - may not by itself capture the full prognostic picture, and general severity scores that incorporate extrapulmonary physiology remain widely used alongside it.

The Acute Physiology and Chronic Health Evaluation (APACHE) II score, derived from 5,815 ICU admissions across 13 hospitals, combines twelve physiological measurements with age and chronic health status and remains among the most widely used general ICU severity scores internationally.[4] Its successor, APACHE III, expanded the physiological variable

set and incorporated disease-category reference data to refine individual mortality-risk estimation.[5] The Sequential Organ Failure Assessment (SOFA) score, developed by consensus under the European Society of Intensive Care Medicine, takes a different approach, grading dysfunction in six organ systems on a simpler five-point scale each.[6] SOFA was designed principally to describe and track organ dysfunction over time rather than to predict mortality from a single measurement, and its prognostic value is greatest when scores are evaluated serially.[7]

Comparative evaluations in general ICU populations have generally favoured the admission-based APACHE scores for discrimination. In a systematic review of eighteen studies of SOFA-based models, admission APACHE II or III models discriminated slightly better than admission SOFA models in five of the six studies making that comparison, with areas under the curve for SOFA-based models ranging from 0.61 to 0.88.[8] Head-to-head comparisons of scoring systems in mixed critical illness have likewise found modest and often non-significant differences between them.[9] Whether these findings extend to ARDS specifically a syndrome that already carries its own oxygenation-based severity classification is less well established, and Indian data comparing all three scores head-to-head in this population are limited.

This study therefore compares APACHE II, APACHE III and SOFA as markers of Berlin severity and as predictors of in-hospital mortality among patients with ARDS admitted to a medical ICU, using logistic regression and ROC analysis to characterise their discrimination and identify practical bedside cut-offs.

MATERIALS AND METHODS

Study design and setting

This was a prospective, single-centre observational cohort study conducted in the Medical Intensive Care Unit of a tertiary care teaching hospital from 1 January 2019 to 30 April 2020. The study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[10] The demographic and aetiological profile of this cohort, its admission clinical and laboratory findings, its comorbidity burden and its interventions and ICU course are reported in companion articles.

Participants

Consecutive patients aged more than 14 years admitted to the medical ICU who fulfilled the Berlin definition of ARDS[1] were enrolled after written informed consent, to a total of 100. Patients were excluded if they or their surrogates declined participation; if ARDS was a sequel of active pulmonary tuberculosis; if there was pre-existing ischaemic heart disease; if death occurred within 24 hours of admission; or if they were surgical admissions. The exclusion of deaths within 24 hours is material to this analysis, since all three scores are computed from the worst values in the first 24 hours and the excluded patients would have been among the most severely deranged.

Severity scoring

APACHE II, APACHE III and SOFA scores were calculated using standard published methodology[4,5,6] from physiological, laboratory and clinical variables recorded within the first 24 hours of ICU admission. ARDS severity was classified independently by the Berlin oxygenation criteria as mild ($200 < PaO_2/FiO_2 \leq 300$ mmHg), moderate ($100 < PaO_2/FiO_2 \leq 200$ mmHg) or severe ($PaO_2/FiO_2 \leq 100$ mmHg). Patients were followed to ICU discharge or death, with outcome dichotomised as survivor or non-survivor.

Statistical analysis

Continuous score values are expressed as means and were compared across the three Berlin severity strata by one-way analysis of variance and between survivors and non-survivors by independent-samples testing. Binary logistic regression was performed with in-hospital mortality as the dependent variable and age, sex, primary diagnosis, mode of ventilatory support, APACHE II, APACHE III and SOFA as covariates, the Wald statistic being used to assess the significance of each predictor. Receiver operating characteristic curves were constructed for each score against mortality; the area under the

curve (AUC) is reported with a 95% confidence interval calculated by the method of Hanley and McNeil, and the optimum cut-off was identified by Youden's index, with sensitivity, specificity, predictive values and likelihood ratios derived at that threshold. Data were analysed using IBM SPSS Statistics version 21.0 (IBM Corp., Armonk, NY) with confirmatory computation in R. A two-sided p-value below 0.05 was considered statistically significant.

Two features of this analysis constrain the strength of the inferences that can be drawn and are stated here rather than left to the Discussion. First, the three scores share many of their component physiological variables APACHE III was developed within the APACHE framework and both APACHE scores incorporate the oxygenation, cardiovascular, renal, hepatic, haematological and neurological domains that constitute SOFA so entering all three in one model introduces substantial collinearity, and the non-retention of any one score cannot be read as evidence that it lacks prognostic information. Second, the model contains 14 degrees of freedom against 27 deaths, approximately 1.9 events per estimated parameter, well below the ten conventionally regarded as a minimum for stable logistic regression; coefficients from this model are correspondingly unstable and are reported as exploratory. No formal paired comparison of the three ROC curves was undertaken, so the ranking of the three areas under the curve is descriptive.

Ethical considerations

The study protocol was approved by the Institutional Ethics Committee and conducted in accordance with the Declaration of Helsinki and the Indian Council of Medical Research National Ethical Guidelines for Biomedical and Health Research Involving Human Participants. Written informed consent was obtained from all patients or their attendants.

RESULTS

Scores by severity and by outcome

One hundred patients were enrolled and all were followed to ICU discharge or death. Forty-six had mild, 31 moderate and 23 severe ARDS. Twenty-seven died in hospital, an overall mortality of 27.0%.

All three scores rose across the Berlin severity strata. Mean APACHE II rose from 18.20 in mild disease through 26.87 in moderate to 39.22 in severe disease, APACHE III from 43.39 through 63.55 to 99.35, and SOFA from 5.83 through 7.35 to 10.04 ($p<0.001$ for each; Table 1). All three were correspondingly higher among non-survivors than survivors: APACHE II 35.8 versus 21.9, APACHE III 87.85 versus 53.14 and SOFA 9.0 versus 6.6 ($p<0.001$ for each; Table 2, Figure 1). The two partitions of the cohort are mutually consistent: the severity-weighted and outcome-weighted cohort means agree to within 0.1 point for every score.

Table 1. Mean severity scores by Berlin severity stratum

Score	Mild (n=46)	Moderate (n=31)	Severe (n=23)	Whole cohort	p
APACHE II	18.20	26.87	39.22	25.72	<0.001
APACHE III	43.39	63.55	99.35	62.51	<0.001
SOFA	5.83	7.35	10.04	7.27	<0.001

Values are group means; p by one-way analysis of variance across the three strata. The whole-cohort column is the stratum-weighted mean and is given so that Tables 1 and 2 can be checked against one another. Standard deviations should be reported alongside these means at submission.

Table 2. Mean severity scores by in-hospital outcome

Score	Non-survivors (n=27)	Survivors (n=73)	Whole cohort	p
APACHE II	35.8	21.9	25.65	<0.001
APACHE III	87.85	53.14	62.51	<0.001
SOFA	9.0	6.6	7.25	<0.001

Values are group means; p by independent-samples testing. The outcome-weighted cohort means agree with the severity-weighted means in Table 1 to within 0.1 point for APACHE II and SOFA and exactly for APACHE III.

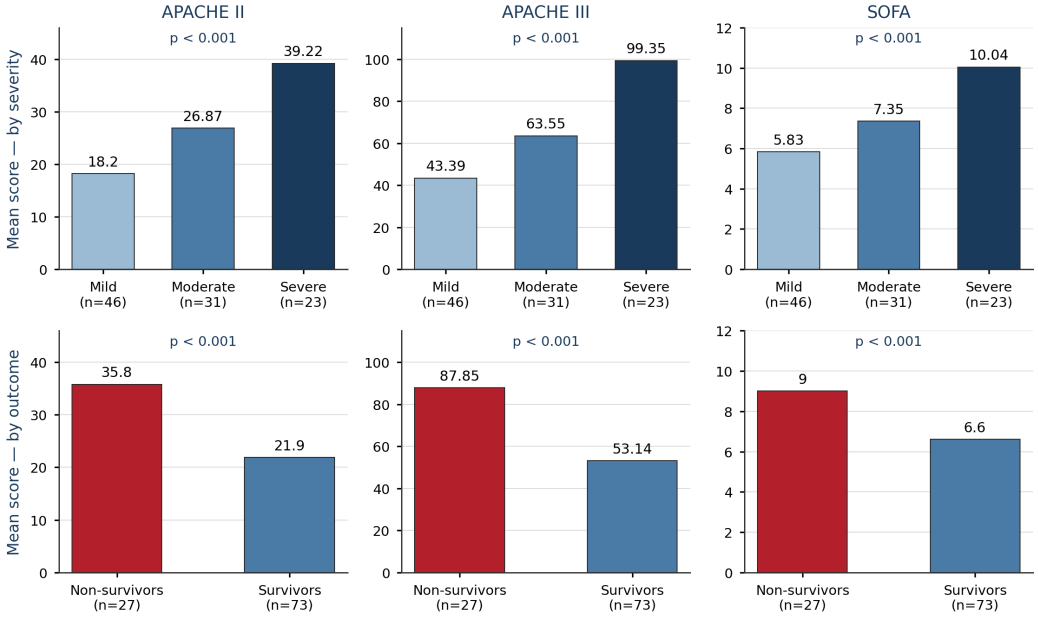

Figure 1. *Mean APACHE II, APACHE III and SOFA scores by Berlin severity stratum (upper row) and by in-hospital outcome (lower row). Each score is plotted on its own axis, since the three have different ranges.*

Multivariable analysis

On binary logistic regression, APACHE II (Wald 12.418, 1 df, p<0.001), the mode of ventilatory support (Wald 10.29, 2 df, p=0.006) and APACHE III (Wald 7.854, 1 df, p=0.005) were retained, while primary diagnosis (Wald 2.829, 7 df, p=0.900), SOFA (Wald 1.539, 1 df, p=0.215), age (Wald 0.036, p=0.849) and sex (Wald 0.013, p=0.908) were not (Table 3, Figure 3).

These results should be interpreted with the two caveats set out in the Methods. The model estimates 14 parameters from 27 deaths, and the three scores are strongly collinear; the failure of SOFA to reach significance in the presence of both APACHE scores is therefore most economically explained by shared information rather than by an absence of prognostic value, particularly since SOFA discriminated mortality well on its own (Table 4).

Table 3. Binary logistic regression: predictors of in-hospital mortality

Predictor	Wald statistic	df	p
APACHE II	12.418	1	<0.001
Mode of ventilatory support	10.29	2	0.006
APACHE III	7.854	1	0.005
Primary diagnosis	2.829	7	0.900
SOFA	1.539	1	0.215
Age	0.036	1	0.849
Sex	0.013	1	0.908

Rows are ordered by descending Wald statistic. Degrees of freedom are stated because primary diagnosis (eight categories) and mode of ventilatory support (three levels) enter the model with more than one degree of freedom; the reported p-values are only reproducible on that basis. The model estimates 14 parameters from 27 events (approximately 1.9 events per parameter) and contains three strongly collinear scores, and is therefore exploratory. Odds ratios with confidence intervals should be reported alongside the Wald statistics at submission.

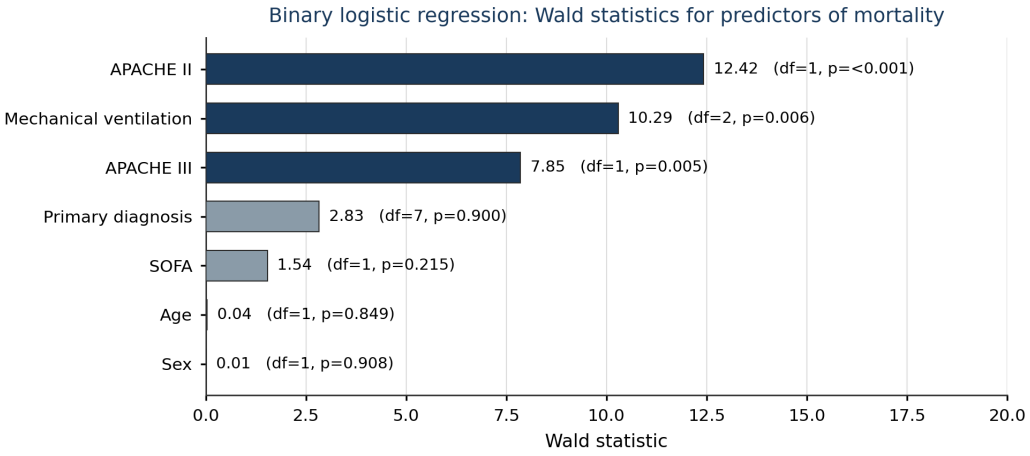

Figure 2. *Wald statistics for each covariate in the binary logistic regression model, with degrees of freedom and p-values. Predictors retained at p<0.05 are shown in dark blue.*

ROC analysis

Every score discriminated between survivors and non-survivors significantly better than chance. APACHE II had the highest area under the curve at 0.895 (95% confidence interval [CI] 0.812–0.978), followed by APACHE III at 0.797 (95% CI 0.689–0.905) and SOFA at 0.765 (95% CI 0.651–0.879); each differed from the null value of 0.5 at p<0.001. The three confidence intervals overlap substantially, however, and no paired comparison of the curves was performed, so these data do not establish that any one score discriminates better than another (Table 4, Figure 2A).

At its optimum cut-off of ≥27, APACHE II correctly identified 23 of the 27 non-survivors (sensitivity 85.2%) and 57 of the 73 survivors (specificity 78.1%), giving an overall accuracy of 80.0%, a positive likelihood ratio of 3.89 and a negative likelihood ratio of 0.19. The corresponding cut-offs for APACHE III (≥33) and SOFA (≥8) achieved lower sensitivity (77.8% and 74.1%) and specificity (68.5% and 61.6%), and lower Youden indices (0.463 and 0.357 against 0.633 for APACHE II).

Predictive values are more informative than sensitivity and specificity for bedside use, and are less favourable. Because only 27% of the cohort died, a positive APACHE II result at the ≥27 threshold carried a positive predictive value of just 59.0%: two of every five patients so classified survived. The negative predictive value was considerably better at 93.4%, so the score performs better at identifying patients unlikely to die than at identifying those who will (Table 4, Figure 2B).

Table 4. Discrimination and performance of each score for predicting in-hospital mortality

Score	AUC (95% CI)	Cut-off	Sens. (%)	Spec. (%)	PPV (%)	NPV (%)	Youden
APACHE II	0.895 (0.812–0.978)	≥27	85.2	78.1	59.0	93.4	0.633
APACHE III	0.797 (0.689–0.905)	≥33*	77.8	68.5	47.7	89.3	0.463
SOFA	0.765 (0.651–0.879)	≥8	74.1	61.6	41.7	86.5	0.357

*AUC = area under the receiver operating characteristic curve, with 95% confidence interval by the method of Hanley and McNeil; Sens. = sensitivity; Spec. = specificity; PPV and NPV = positive and negative predictive value at a mortality prevalence of 27.0%. Each AUC differs significantly from 0.5 (p<0.001); the intervals overlap and no paired comparison between curves was performed. Corresponding counts: APACHE II 23 of 27 deaths and 57 of 73 survivors correctly classified; APACHE III 21 and 50; SOFA 20 and 45. *The APACHE III cut-off of ≥33 lies well below the reported mean APACHE III of 53.14 among survivors and 62.51 in the whole cohort, and is not readily reconcilable with a specificity of 68.5%; this value requires verification against the source analysis before submission.*

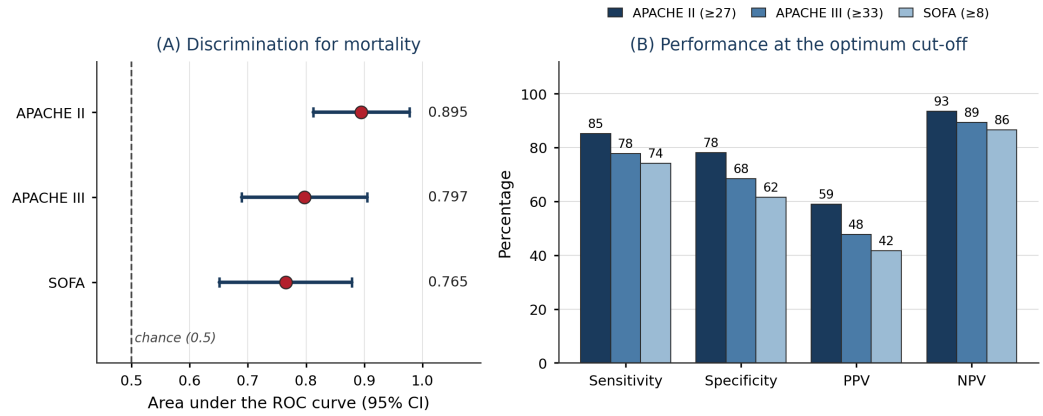

Figure 3. *(A) Area under the ROC curve for each score with 95% confidence intervals; the dashed line marks chance discrimination. The intervals overlap substantially. (B) Sensitivity, specificity and predictive values at each score's optimum cut-off, at a mortality prevalence of 27.0%.*

DISCUSSION

In this cohort of 100 patients with ARDS, all three severity scores rose with Berlin severity and were substantially higher in non-survivors, and all three discriminated mortality significantly better than chance. APACHE II achieved the highest point estimate of discrimination, with an area under the curve of 0.895 against 0.797 for APACHE III and 0.765 for SOFA.

The temptation is to conclude that APACHE II is the better score, and the submitted analysis is not sufficient to support that. The 95% confidence intervals for the three areas under the curve overlap across most of their range, and the three curves were generated from the same 100 patients and are therefore correlated, so any comparison between them requires a paired test such as that of DeLong and colleagues rather than inspection of separate intervals. What can be said is that each score discriminated well in absolute terms and that APACHE II was numerically best; what cannot be said, on these data, is that the difference is real. This distinction matters because the ordering is the paper's headline result.

The direction of the finding is nonetheless consistent with the broader literature. In a systematic review of eighteen studies, admission APACHE II or III models discriminated slightly better than admission SOFA models in five of the six studies making that comparison,[8] and head-to-head comparisons of scoring systems in mixed critical illness have generally found differences between them to be modest.[9] That APACHE III did not outperform APACHE II, despite having been developed specifically to improve upon it through an expanded variable set and disease-specific reference equations,[5] echoes the mixed record of APACHE III outside its derivation cohort, and may also reflect that APACHE III's advantage lies chiefly in calibration — the accuracy of its predicted mortality probabilities — rather than in discrimination, which is what an ROC analysis measures. We did not assess calibration, and a score can discriminate well while predicting poorly calibrated absolute risks.

SOFA showed the lowest discrimination, which is unsurprising given its design. SOFA was created to describe and track organ dysfunction over time rather than to predict death from a single measurement, and its prognostic performance is substantially better when scores are evaluated serially, the maximum or delta score carrying more information than the admission value alone.[6,7] We measured SOFA only at admission, which uses the score in the manner for which it was least designed. Its non-retention in the multivariable model should not be read as absence of prognostic value: SOFA discriminated mortality with an AUC of 0.765 on its own, and it was entered into a model alongside two scores that share most of its component variables. Collinearity, not irrelevance, is the more economical explanation.

The independent association of ventilatory support with mortality, alongside the two APACHE scores, is consistent with the companion analysis of this cohort, in which invasive ventilation carried a seventeenfold increase in the odds of death. Whether this represents genuinely complementary information or simply the same severity signal arriving by a different route cannot be settled by a model this heavily over-parameterised.

For bedside use, the predictive values matter more than sensitivity and specificity, and they temper the enthusiasm that an AUC of 0.895 invites. At the ≥27 threshold the positive predictive value of APACHE II was 59.0%, meaning that two of every five patients flagged as high risk survived to discharge. The negative predictive value of 93.4% is far more useful: a patient scoring below 27 has a high probability of surviving. Framed this way, the score is better suited to identifying patients who can safely be managed less intensively than to identifying those who will die, and it should certainly not be used to inform limitation-of-treatment decisions in an individual patient. Predictive values are also prevalence-dependent and would differ in a population with a mortality other than the 27% seen here.

Limitations

Several limitations apply. The study is single-centre with 100 patients and 27 deaths, which limits the precision of every estimate reported: the confidence interval for the APACHE II area under the curve alone spans 0.812 to 0.978. The logistic regression model estimates 14 parameters from 27 events, roughly a fifth of the events per parameter conventionally regarded as a minimum, and includes three strongly collinear scores, so its coefficients are unstable and the pattern of significance may not replicate. Odds ratios and confidence intervals for the model covariates were not available, and Wald statistics alone convey neither the direction nor the magnitude of effect. No paired test was performed between the ROC curves, so the ranking of the three scores is descriptive. Cut-offs derived by Youden's index in the same data used to fit them are optimistically biased and require external validation before use. Calibration was not assessed. Standard deviations for the score means were not available for independent verification of the ANOVA results, and the reported APACHE III cut-off of ≥33 is difficult to reconcile with the reported APACHE III means and requires checking. Finally, patients dying within 24 hours were excluded, which removes precisely the group in whom admission scores would be most extreme and is likely to have attenuated the observed discrimination.

CONCLUSION

All three severity scores rose significantly with Berlin severity and were significantly higher among non-survivors, and all three discriminated in-hospital mortality significantly better than chance in this ARDS cohort. APACHE II achieved the highest point estimate of discrimination (area under the curve 0.895) ahead of APACHE III (0.797) and SOFA (0.765), but the confidence intervals overlap and no paired comparison of the curves was performed, so superiority is suggested rather than demonstrated. At a cut-off of ≥27, APACHE II combined 85.2% sensitivity with 78.1% specificity and a negative predictive value of 93.4%, making it more useful for identifying patients at low risk of death than for identifying those who will die, since its positive predictive value at this prevalence was only 59.0%. The multivariable model, containing three collinear scores and estimating 14 parameters from 27 deaths, should be regarded as exploratory, and the non-retention of SOFA reflects shared information with the APACHE scores rather than absence of prognostic value. Confirmation requires a larger multicentre ARDS cohort with paired comparison of the ROC curves, assessment of calibration as well as discrimination, and external validation of any proposed cut-off.

DECLARATIONS

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of Interest

The authors declare that they have no competing interests. Ethics approval and consent to participate The study was approved by the Institutional Ethics Committee. Written informed consent was obtained from all patients or their attendants. Related publications This analysis forms part of a single prospective study of 100 patients with ARDS admitted to one medical ICU. The demographic and aetiological profile of the cohort, its admission clinical and laboratory findings, its comorbidity and addiction burden, and its organ-supportive interventions and ICU course are reported in companion articles. No patient data are duplicated between the analyses beyond the shared description of the cohort. Data availability The anonymised dataset supporting the conclusions of this article is available from the corresponding author on reasonable request.

Author Contributions

All named authors contributed to the conception and design of the study, participated in acquisition of data and its analysis and interpretation, took part in drafting the article or revising it critically for important intellectual content, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

References

1. ARDS Definition Task Force; Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, et al. Acute respiratory distress syndrome: the Berlin Definition. JAMA. 2012;307(23):2526–33.
2. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al.; LUNG SAFE Investigators; ESICM Trials Group. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. JAMA. 2016;315(8):788–800.
3. Ware LB, Matthay MA. The acute respiratory distress syndrome. N Engl J Med. 2000;342(18):1334–49.
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818–29.
5. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100(6):1619–36.
6. 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. Intensive Care Med. 1996;22(7):707–10.
7. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA. 2001;286(14):1754–8.
8. 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(6):R161.
9. Moseson EM, Zhuo H, Chu J, Stein JC, Matthay MA, Kangelaris KN, et al. Intensive care unit scoring systems outperform emergency department scoring systems for mortality prediction in critically ill patients: a prospective cohort study. J Intensive Care. 2014;2(1):40.
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7.