SAJEM EMERGENCY MEDICINE JOURNAL

META-ANALYSIS

The Prognostic Potential of Three Scoring Systems for Mortality Prediction in Intensive Care Unit: A Systematic review & Meta-analysis

Sanniya Khan Ghauri[1], Abdul Sattar[2], Arslaan Javaeed[3], Shafqat Husnain Khan[4], Khawaja Junaid Mustafa[5], Abdus Salam Khan[6]

Authors Affiliation:

Department of Emergency Medicine, Shifa International Hospital, Islamabad[1,2,5&6]

Department of Pathology, Poonch Medical College, Rawalakot Azad Kashmir[3-4]

Correspondence to:

Sanniya Khan Ghauri
sanniyakhan@gmail.com

ABSTRACT:

BACKGROUND

The predictive potential of scoring systems used in intensive care units (ICUs) to distinguish patients' mortality is heavily investigated but scarcely pooled.

OBJECTIVES

To statistically analyze the accuracy of three ICU generic scores, including Acute Physiology and Chronic Health Evaluation (APACHE II), Simplified Acute Physiology Score (SAPS II), and Sequential Organ Failure Assessment (SOFA), to predict mortality.

METHODS

A meta-analysis was conducted involving prospective studies published between January 2000 and February 2019 to analyze the performance of APACHE II, SAPS II, and SOFA to predict patients' mortality through pooling their discriminative indicators, such as sensitivity specificity, and the area under hierarchical summary receiver operating characteristic curve (HSROC). The inconsistency index (I2) was used to assess heterogeneity in sensitivity and specificity, while meta-regression analysis was performed to detect the potential sources of heterogeneity. Deek's funnel plots were used to assess the potential publication bias.

RESULTS

In a total of 37 studies (7612 patients, 63.58% males, 75.7% assessed in-hospital mortality), 2170 observed deaths were reported. APACHE II, SAPS II, and SOFA scoring systems showed good mortality predictive performances, where the pooled sensitivities were 0.81, 0.76, and 0.80, respectively, specificities were 0.78, 0.89, and 0.79, respectively, and pooled HSROCs were 0.87, 0.85, and 0.88, respectively. For APACHE II, SAPS II, and SOFA, there were significant in-between study heterogeneities in sensitivity (I^2=85.21%, 84.31%, and 71.67%, respectively) and specificity (I^2=93.74%, 92.65%, and 89.41%, respectively) while no publication bias was detected (P=0.689, 0.465, and 0.181, respectively). There was a significant heterogeneity among studies which investigated APACHE II with a cut-off score ≥20 (P = 0.04) and those investigating SAPS II for ≥24 months (P < .001).

CONCLUSION

Within the limitations of the study, ICU scores showed good prognostic performance. Future studies should be conducted using fixed-time endpoints of mortality, involving multiple countries and employ a combination of generic ICU scores.

KEYWORDS

APACHE II, SAPS II, SOFA, critical illness, Mortality

INTRODUCTION

It was as early as 1863 since the assessment of treatment outcomes was first addressed by Florence Nightingale.[1] Traditionally, patients' outcomes were predicted by the subjective judgement of physicians in intensive care units (ICUs). However, since critically-ill patients usually experience physiological distresses that might result in disability or mortality within days, hours, or even minutes, there was a need to predict the outcomes and to assess the effectiveness of therapies in a quantitative manner. Therefore, ICU scoring systems have been developed and widely applied. These systems rely on several clinical indicators, including tachypnea, tachycardia, hypotension, decreased urinary output, or altered consciousness.

ICU scores essentially comprise two main

parts: a severity score, which indicates the severity of the condition, and a probability model of mortality, which is calculated based on an equation to predict survival rates in the next days following hospital admission.[2] These models help in decision making via enhancing the capacity of scores to compare patients based on their treatments, triage or comparative assessments. In addition to their basic roles in clinical observation of critically-ill patients, ICU scoring systems can be used along the course of hospital admission as monitoring tools to the given treatments and for assessment of organ dysfunctions.[3]

However, the choice of a scoring system should basically depend on its ability to match the event, application, or setting. Faulty application or misuse may lead to increased costs, wasted time, unneeded extrapolations, and poor outcomes.[4] This is particularly evident for the generic scores which are extensively used in ICUs. These scores are broadly divided into scores which assess disease severity on patient's admission, such as Acute Physiology and Chronic Health Evaluation (APACHE), Mortality Probability Model (MPM), and Simplified Acute Physiology Score (SAPS), as well as those used to assess organ dysfunctions, such as Sequential Organ Failure Assessment (SOFA) and Multiple Organ Dysfunction Score (MODS).[3]

The discriminative potential of the predictive models of these scores to distinguish patients who die from those who live is an important aspect that helps in decision-making. Nevertheless, the predictive ability of the indicators of discrimination, including the sensitivity and specificity, and area under the receiver operating characteristic (ROC) curve, remains debatable. In this study, we sought to conduct a meta-analysis to investigate the accuracy of three widely used non-disease-specific ICU scores (APACHE II, SAPS II, and SOFA) to predict mortality (discriminate between survivors and non-survivors) among critically ill patients via pooling their discriminative indicators.

METHODS

This meta-analysis was conducted based on the guidelines of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)[5] and compliance to the MOOSE statement.[6] This study was registered in at the International Prospective Register of Systematic Reviews (PROSPERO) with the following registration number: CRD42019127173.

ELIGIBILITY CRITERIA

The included studies were prospective cohort studies which recruited adult patients admitted to ICU units and reported the sensitivity and specificity of either APACHE II, SAPS II, or SOFA scores to predict patients' mortality. The latter was defined as in-hospital death or mortality up to 30 days after admission. No restrictions for the cause of admission to ICUs were considered. Articles were eligible if they were published in peer-reviewed journals during the period from January 2000 until February 2019. Studies were excluded if they were written in a non-English language or employed retrospective designs, narrative reviews, systematic reviews, or randomized clinical trials.

TYPES OF OUTCOMES MEASURES

Mortality of critically-ill patients represented the main outcome measure. The discriminative ability of ICU scores under study to predict mortality was investigated based on the reported sensitivity and specificity rates. APACHE II is an indicator of disease severity which is based on values of 12 physiologic indicators. The SAPS II score measures specific variables within the first 24 hours. [7] As for the SOFA score, data on the degree of organ dysfunction is collected for six organ systems, yielding a final score ranging between 6 and 24.[8]

SEARCH STRATEGY

The whole search process was performed by two independent authors who searched the following databases for eligible articles: PubMed, Embase, Google Scholar, and Scopus. All databases were screened using specific keywords using combinations of Boolean operators, such as "AND" and "OR". An example of used search strategy in PubMed is demonstrated in Appendix 1. Furthermore, the reference lists of the screened articles were search for eligible articles to optimize the search process.

STUDY SELECTION AND DATA COLLECTION

The titles and abstracts of obtained records were screened across all databases by two independent authors. All records were then extracted and uploaded to Endnote (version X7) and all duplicates were omitted. Subsequently, the full-text versions of eligible articles were assessed for inclusion. In cases of any disagreement, the authors discussed the potential solutions until reaching a consensus. Data was extracted in a specific spreadsheet designed in Microsoft Excel. The extracted data included: 1) study data: the last name of the first author, year of publication, duration of the follow-up period, and country; 2) participants' characteristics: range of ages, gender distribution, and cause of ICU admission; 3) mortality data: number of deaths and the employed definition of mortality (in-hospital death, 28-day mortality, etc.); 4) data of ICU scores: means (standard deviations [SD]) of scores in survivors and non-survivors, sensitivity, specificity, and cut-off value.

QUALITY ASSESSMENT

The methodological quality of the included studies was assessed using the guidelines implied by the Quality

Assessment of Diagnostic Accuracy Studies (QUADAS) tool[9] and was graphically-illustrated using RevMan 5.3 software (Review Manager, the Cochrane Collaboration, Oxford, United Kingdom).

STATISTICAL ANALYSIS

The severity scores were computed as described in their original publications.[1, 7] The mean scores and SDs of scores in survivors and non-survivors were calculated from median (interquartile ranges) as previously described.[10] Mean differences (MDs) and their respective 95% CIs were used to test the statistical difference ICU scores. For all included studies, a 2x2 table was constructed containing all main variables for assessment of the discriminative ability of the three ICU scores. These variables include the numbers of true positives, false positives, true negatives, and false negatives. The extracted sensitivity and specificity rates of predictive mortality models as well as the total sample size and number of deaths were used to compute the constructed variables. All aforementioned calculations were made in RevMan 5.3 software (Review Manager, the Cochrane Collaboration, Oxford, United Kingdom).

Subsequently, the Stata statistical software (Stata Corp, College Station, TX, USA) was used to pool statistical indices to obtain summary estimates of performance statistics, including weighted estimates of sensitivity, specificity, positive likelihood ratio (PLR), and negative likelihood ratio (NLR), the area under hierarchical summary ROC curve. These estimates were computed using a bivariate random effects regression model using the *metandi* or *midas* command and the relevant statistical graphs were drawn accordingly. Additionally, the diagnostic odds ratios (DOR) with the respective 95% confidence intervals (CIs) were calculated using the *metandi* command by a random effects model. The heterogeneity by non-threshold effect was quantified by Cochrane Q test and the inconsistency index (I^2), where the heterogeneity was considered significant at $I^2 \geq 50\%$ and P value of ≤ 0.05. In the latter instance, subgroup analysis was planned and performed using meta-regression analysis, which simultaneously investigates the included covariates and their association with the diagnostic performance. The integrated covariates included year of publication (< or ≥ 2010), follow-up duration (< or ≥ 24 months), study country (Asian or non-Asian), cohort size (< or ≥ 150), and cut-off scores (< or ≥ 20 for APACHE II, < or ≥ 45 for SAPS II, and < or ≥ 8 for SOFA). Deek's funnel plot [11] was used to assess the potential publication bias and a $P < 0.05$ indicated significant publication bias.

RESULTS: Results of the search process

Figure 1 depicts the outcomes of the search process used in this study. Initially, the total number of obtained records across all databases was 1845, of which 12 duplicates were found. Additionally, we identified eight studies from the bibliographies of these records and thus a total of 1933 records were screened. Screening of titles/abstracts revealed 40 eligible articles. However, 3 articles were excluded due to lack of access to the full-article version[12, 13] and the inclusion of severe conditions (dialysis, respiratory support, etc.) with mortality as primary outcomes.[14] As such, 37 studies were ultimately included in both the qualitative and quantitative analyses.

CHARACTERISTICS OF THE INCLUDED STUDIES

As shown in Table 1, the included prospective studies were published between 2000 and 2018 with sample sizes ranging between 50 and 1670 patients and study periods of 6-72 months. The total number of patients was 7612, (4840 [63.58%] males) and they were aged 18-93 years. Five studies were conducted in European countries,[15-19] one study in Latin America,[20] while the remaining studies were based in Asian countries. Observed mortality measures included mortality up to 14 days in one study,[15] 28 days in four studies,[21-24] 30 days in four studies,[25-28] and in-hospital deaths in the remainder. A total of 2170 deaths were reported with mortality rates ranging between 8.0% - 67.5%.

QUALITY ASSESSMENT

The observed mortality was considered the reference standard in quality assessment. It was unclear whether the observed mortality was interpreted without knowledge of the results of scoring systems in all studies. Only one study [29] did not explicitly describe selection criteria of the cohort. The use of different scoring systems was merely mentioned without explaining the way by which the scoring and/or mortality prediction were executed in 11 studies,[17, 18, 21, 24, 26, 27, 30-34] accounting for 29.7% of high risk in this domain. While the causes of patient withdrawals were specifically revealed in four studies, including the development of neurological sequelae in patients with poisoning (n=5),[35] loss during follow-up (n=44),[21] patient discharge against medical advice or at the request to other facilities (n=189),[36], and patients who underwent surgeries in a study including those with spontaneous intracerebral hemorrhage (n=5),[26] the remaining studies included all patients in the analyses (Figure 2).

THRESHOLD EFFECT AND HETEROGENEITY

There was a significant heterogeneity among studies in the sensitivity of APACHE II (I^2=85.21%, p<0.01), SAPS II (I^2=84.31%, p<0.01), and SOFA (I^2=71.67%, p<0.01)

scores. Similar findings were found for pooled specificity of APACHE II (I^2=93.74%, p<0.01), SAPS II (I^2=92.65%, p<0.01), and SOFA (I^2=89.41%, p<0.01) scores (Figure

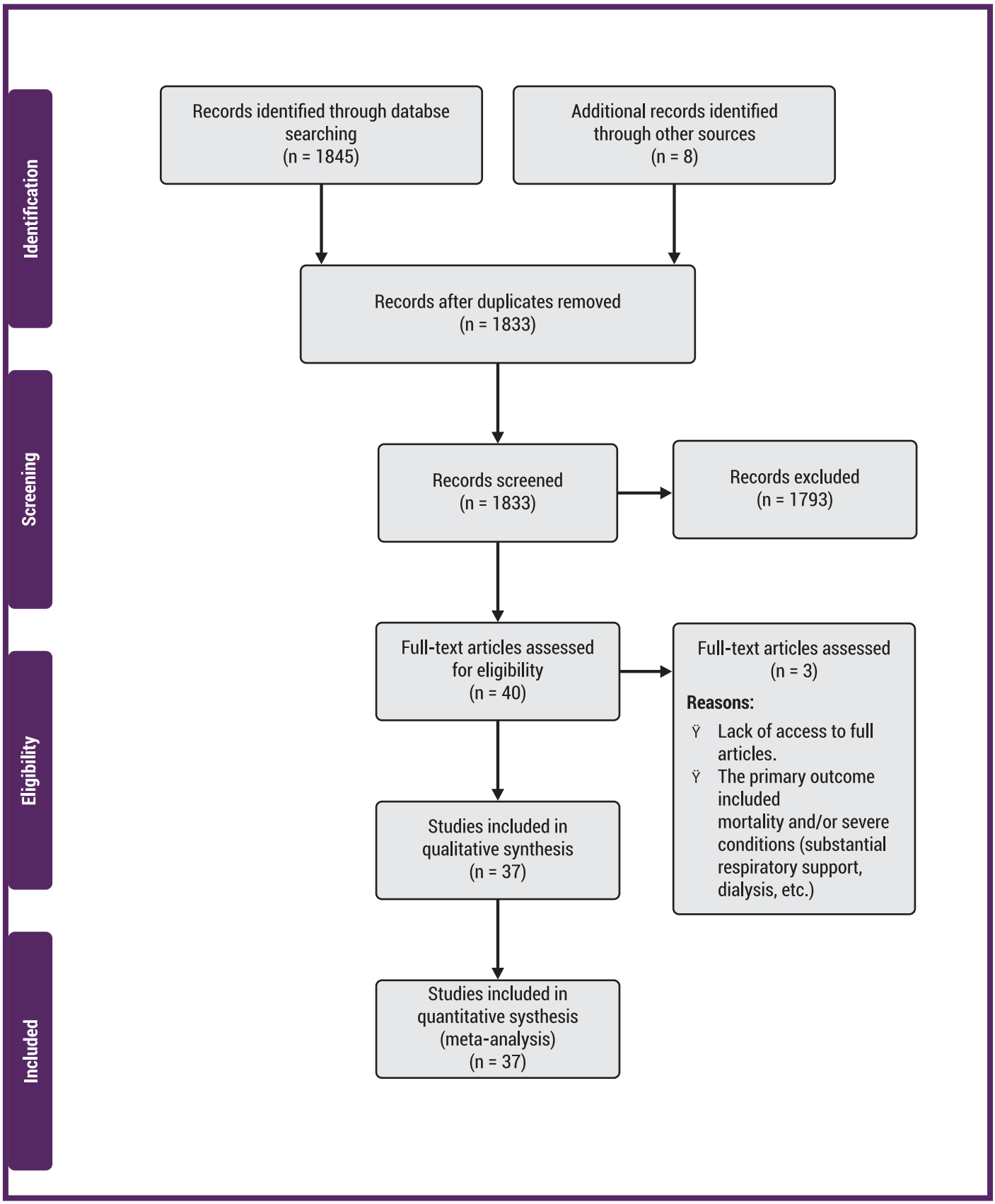

Figure 1: A flow diagram of the search process used in this study

Table 1

Author, Year	Follow-up period (months)	Country	Male/female/ total	Patients' Ages	Cause of ICU admission	Deaths	Mortality measure	Scoring system(s)
Grmec and Gašparovic 200017	24	Slovakia	142/144/286	30-78	Non-traumatic coma	80	In-hospital death	APACHE II
Chatzicostas et al. 200316	24	Greece	137/63/200	33-86	Liver cirrhosis	23	In-hospital death	APACHE II
Ho et al. 200432	13	Taiwan	103/32/135	45-71	Liver cirrhosis	90	In-hospital death	APACHE II
Wang et al. 200571	22	Taiwan	40/21/61	51-80	ARF	38	In-hospital death	APACHE II
Gursel and Demirtas 200618	26	Turkey	34/29/63	18-93	VAP	34	In-hospital death	APACHE II/SOFA
Kulkarni et al. 200766	24	India	39/11/50	31-72	Perforative peritonitis	8	In-hospital death	APACHE II
Goertz et al. 201115	10	Germany	148/120/268	39-80	Surgery	67	14d	APACHE II/SAPS II
Olmez et al. 201268	36	Turkey	130/71/201	41-61	Liver cirrhosis	84	In-hospital death	APACHE II/SOFA
Badrinath et al. 201357	NA	India	125/68/193	41-73	Sepsis	108	In-hospital death	APACHE II/SOFA
Duseja et al 201360	20	India	87/13/100	38-56	ACLF	53	In-hospital death	APACHE II/SOFA
Kim et al. 201365	48	Korea	104/27/131	42-87	Poisoning	29	In-hospital	APACHE II/SAPS II/SOFA
Oliveira et al. 201320	9	Brazil	90/62/152	40-52	Transplant (liver)	18	In-hospital death	APACHE II
Alizadeh et al. 201435	6	Iran	118/77/195	18-91	Poisoning	42	In-hospital death	APACHE II/SAPS II
Chang et al. 201431	12	Taiwan	390/153/543	61-66	AKI	116	In-hospital death	APACHE II/SOFA
Gilani et al. 201429	6	Iran	118/84/202	32-64	Surgery	81	In-hospital death	APACHE II/SAPS II
Hosseini et al. 201563	8	Iran	91/59/150	35-80	Varied	21	In-hospital death	APACHE II
Liu et al. 201524	24	China	70/67/137	59-80	Sepsis	20	28d	APACHE II/SOFA
Lopez-Delgado et al. 201519	46	Spain	38/20/58	42-60	Liver cirrhosis	12	In-hospital death	SAPS II
Zhou et al. 201527	48	China	80/55/135	50-73	VAP	39	30d	APACHE II
Baradari et al. 201658	12	Iran	189/111/300	24-64	Varied	92	In-hospital death	APACHE II/SOFA
Hosseini and Ramazani 201662	6	Iran	185/115/300	24-78	Varied	82	In-hospital death	APACHE II/SOFA

Table 1

Author, Year	Follow-up period (months)	Country	Male/female/ total	Patients' Ages	Cause of ICU admission	Deaths	Mortality measure	Scoring system(s)
Kuo et al. 201633	30	Taiwan	110/35/145	35-57	Varied	35	In-hospital death	APACHE II/SOFA
Jain et al. 201664	8	India	0/90/90	25-49	Obstetrics	30	In-hospital death	SOFA
Safari et al. 201628	12	Iran	75/65/140	18-95	Varied	72	30d	SOFA
Baradari et al. 201759	6	Iran	154/96/250	18-82	Varied	78	In-hospital death	SOFA
Feng et al. 201721	24	China	240/142/382	48-89	AECOPD	46	28d	APACHE II/SAPS II
Mohamed et al. 201767	24	India	57/23/80	50-80	Sepsis	54	In-hospital death	APACHE II/SOFA
Pan et al. 201726	17	India	69/35/104	46-74	SICH	70	30d	APACHE II
Sharma et al. 201769	36	India	65/35/100	30-71	Sepsis	34	In-hospital death	APACHE II/SAPS II/SOFA
Srinivasan et al. 201734	8	India	53/47/100	28-84	VAP	44	In-hospital death	APACHE II/SAPS II/SOFA
VijayGanapathy et al. 201770	20	India	93/85/178	37-69	Sepsis	38	In-hospital death	APACHE II
Balasubramanian 201830	12	India	28/22/50	18-70	Infection (Scrub Typhus)	4	In-hospital death	APACHE II/SOFA
Ebrahimi et al. 201861	NA	Iran	80/40/120	18-54	Poisoning	18	In-hospital death	SOFA
Goswami et al. 201825	24	India	79/21/100	22-45	CKD	39	30d	APACHE II/SAPS II/SOFA
Singh et al. 201823	20	India	72/28/100	20-70	Sepsis	63	28d	SAPS II
Venkataraman et al. 201836	21	India	1113/557/1670	31-70	Varied	374	In-hospital death	APACHE II
Yoon et al. 201822	72	Korea	94/49/143	40-72	OHCA	34	28d	APACHE II

ACLF: acute-on-chronic liver failure; AECOPD: Acute Exacerbation of Chronic Obstructive Pulmonary Disease; AKI: acute kidney injury; APACHE II: Acute Physiology and Chronic Health Evaluation; ARF: acute renal failure; CKD: chronic kidney disease; OHCA: out-of-hospital cardiac arrest survivors; RAAA: ruptured abdominal aortic aneurysm; SAPS: Simplified Acute Physiology Score; SICH: Spontaneous Intracerebral Hemorrhage; SOFA: Sequential Organ Failure Assessment; TBI: Traumatic brain injury; VAP: ventilator-associated pneumonia

Vol. 03, Issue. 01

South Asian Journal of Emergency Medicine

SAJEM EMERGENCY MEDICINE JOURNAL

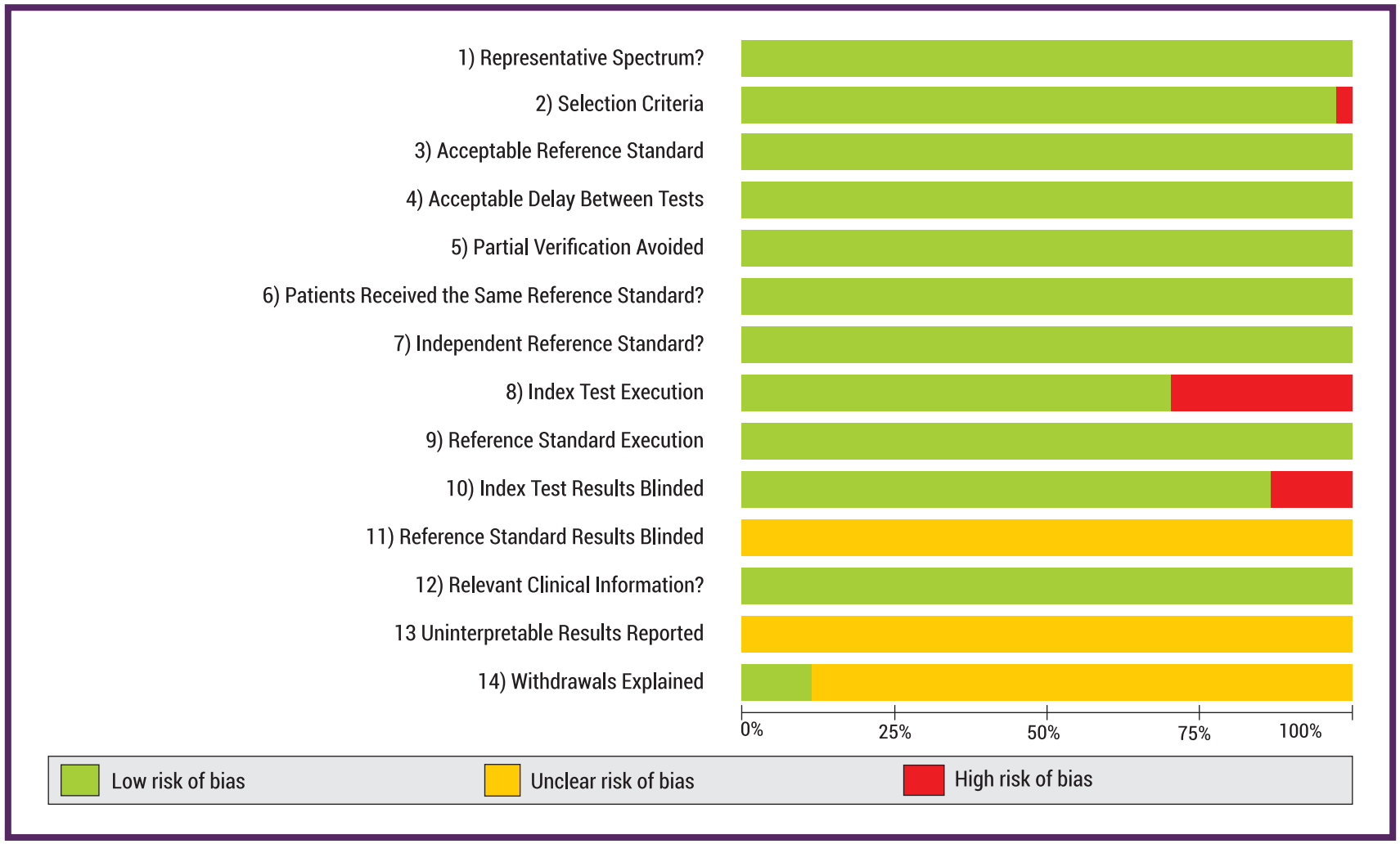

Figure 2: A summary of quality assessment of the included studies according to the QUADRAS tool. Questions answered as "yes" are labelled as "low risk of bias"

3). Threshold effect, which has been identified as an important source of heterogeneity among studies, was investigated using the *midas* command. The Spearman correlation coefficient and the P value were 0.54 and 0.29 for APACHE II, 0.56 and 0.31 for SAPS II, and 0.28 and 0.08 for SOFA, respectively. This indicates no significant threshold effect on the outcomes of all ICU scores (all P values > 0.05).

For further assessment of the sources of heterogeneity, meta-regression analysis was conducted. Results revealed that there was a significant heterogeneity among studies which investigated APACHE II with a cut-off score ≥20 (X^2 = 6.24, P = 0.04). In addition, longer study periods (≥24 months) accounted for significant heterogeneity (X^2 = 15.68, P < .001) among studies investigating SAPS II. However, no significant covariates were identified to impact heterogeneity among SOFA-investigating studies (Table 2).

Table 2

Parameter	Category	APACHE II			SAPS II			SOFA		
		No of studies	X^2	P value	No of studies	X^2	P value	No of studies	X^2	P value
Year of publication	**≥2010**	**25**	0.31	0.86	**10**	NA	NA	**18**	1.03	0.6
	<2010	**6**	--	--	**0**	--	--	**1**	--	--
Study period	**≥24 months**	**14**	3.16	0.21	**5**	15.68	<.001*	**9**	3.54	0.17
	<24 months	**17**	--	--	**5**	--	--	**10**	--	--
Country	**Asian**	**26**	2.47	0.29	**8**	0.82	0.66	**18**	1.03	0.6
	Non-Asian	**5**	--	--	**2**	--	--	**1**	--	--
Sample size	**≥150**	**14**	1.56	0.46	**4**	1.85	0.4	**6**	2.17	0.34
	<150	**17**	--	--	**6**	--	--	**13**	--	--
Cut-off value		**11 (≥20)**	6.24	0.04*	**7 (≥45)**	0.81	0.67	**9 (≥8)**	4.28	0.12
		20 (<20)	--	--	**3 (<45)**	--	--	**10 (<8)**	--	--

PUBLICATION BIAS

As illustrated using the Deek's funnel plots, there was no publication bias in all studies. This was evident for studies investigating APACHE II (P = 0.689, Figure 4A), SAPS II (P = 0.465, Figure 4B), and SOFA (P = 0.181, Figure 4C). Likewise, no obvious small-study bias was detected as shown by a P value of ≥ 0.10 for the slope coefficients, indicating no significant asymmetry between studies.[11]

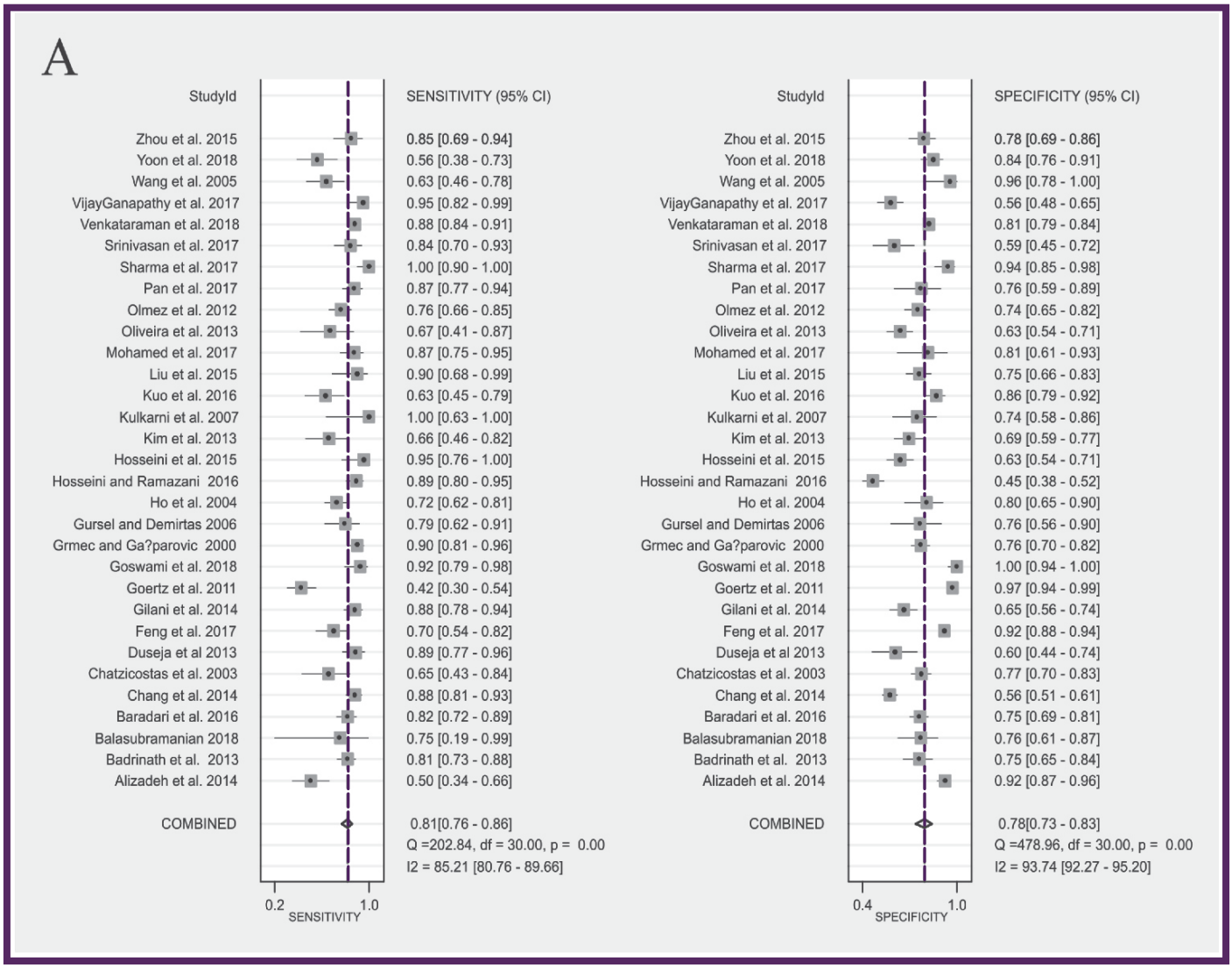

Figure 3: The pooled sensitivity and specificity of APACHE II (A)

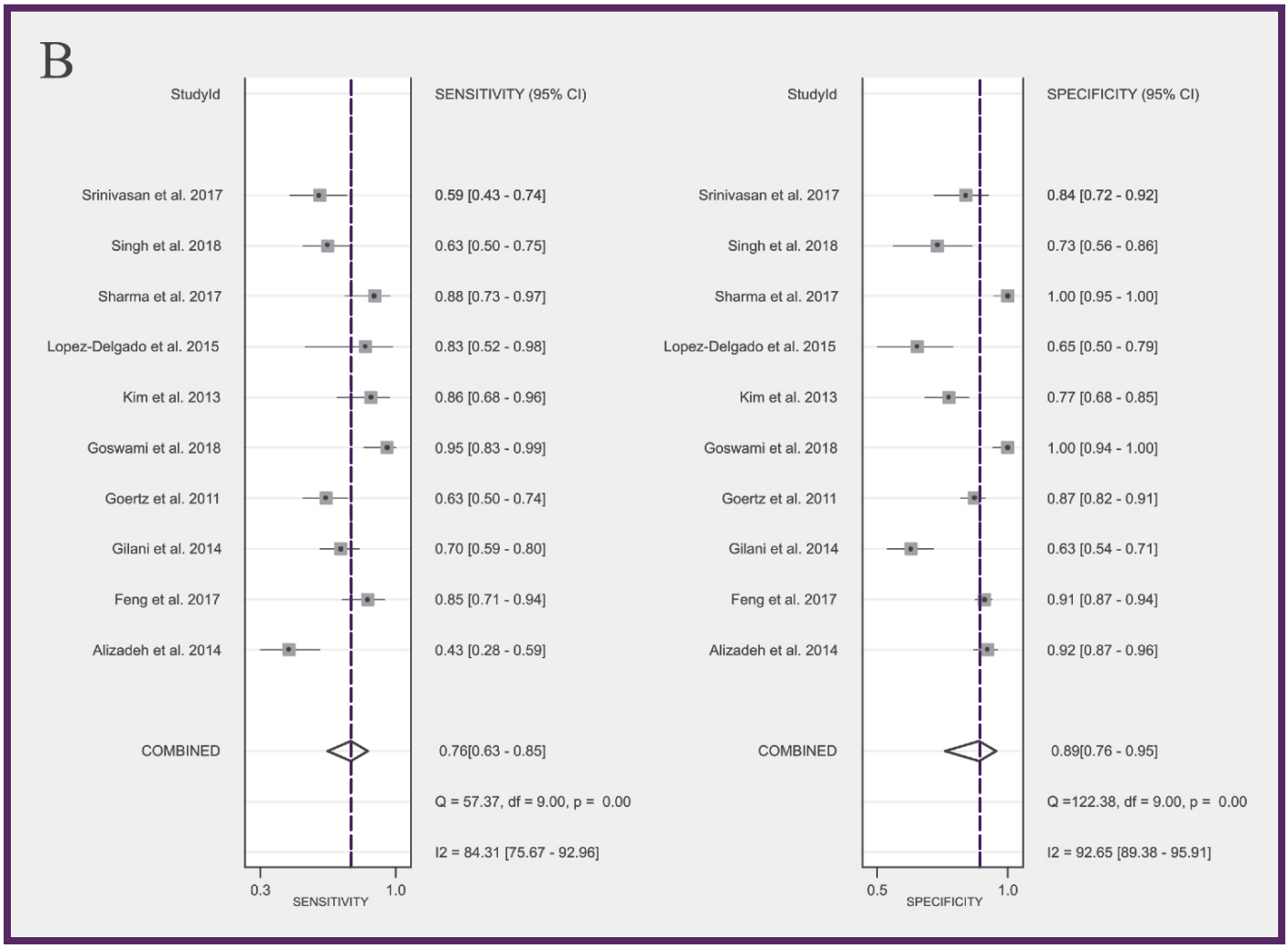

Figure 3: SAPS II (B)

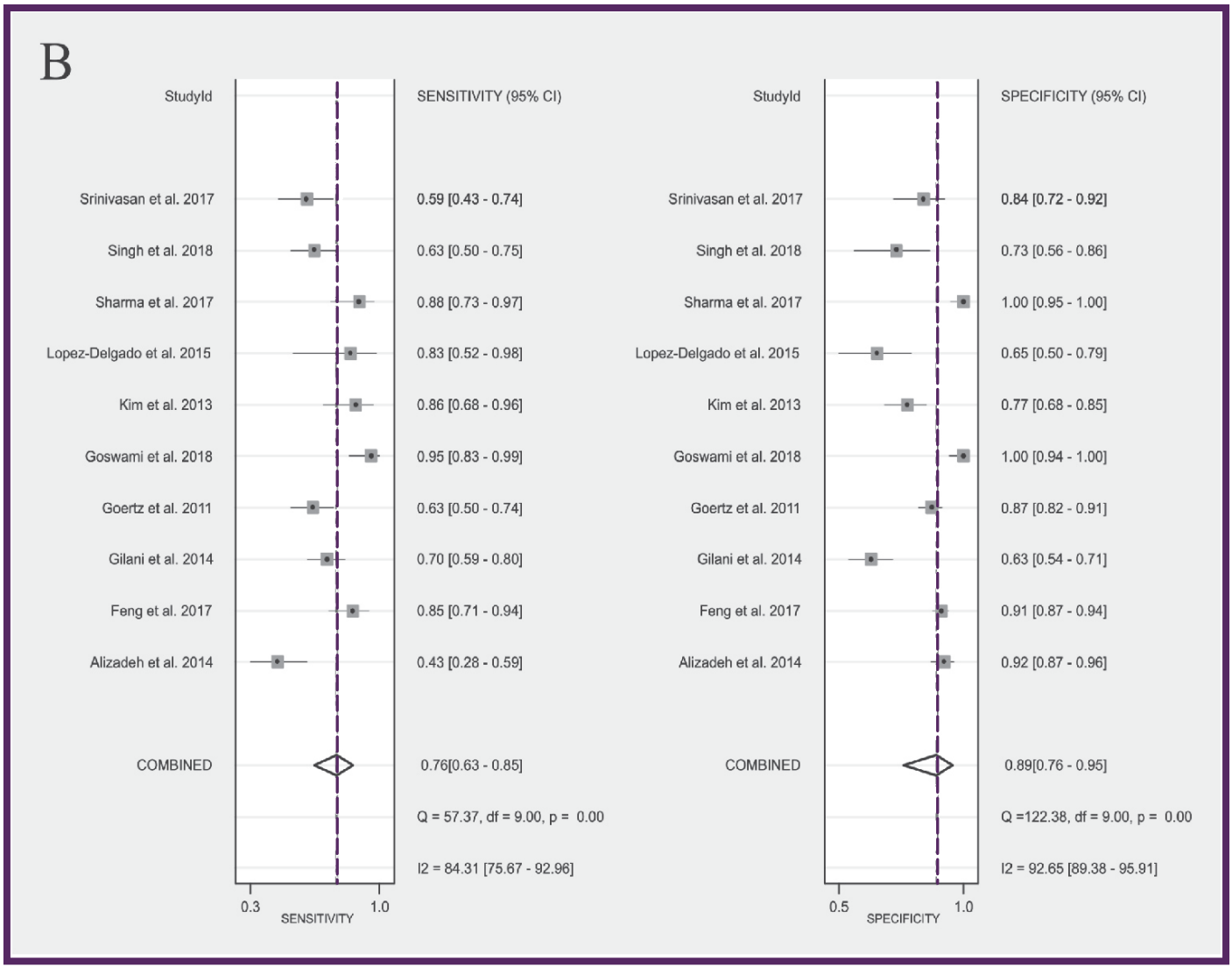

Figure 3: SOFA (C) scores as extracted from the included studies

MEAN ICU SCORES

Using random effect models, the mean scores were significantly higher among non-survivors as compared to survivors; MD of APACHE II was -8.54 (95%CI -9.62, -7.46, P < 0.001), MD of SAPS II was -20.81 (95%CI -29.26, -12.36, P < 0.001), and MD of SOFA was -3.09 (95%CI -4.03, -2.15, P < 0.001, data not shown).

DIAGNOSTIC ACCURACY MEASURES

All ICU scores showed moderate pooled sensitivity and specificity rates in differentiating mortalities. Using a bivariate binomial mixed model, the highest sensitivity rates were reported for APACHE II (0.81 [95% CI: 0.76-0.86]) and SOFA scores (0.80 [95% CI: 0.76-0.84]), while SAPS II reported the highest specificity rate (0.89 [95% CI: 0.76-0.95], Figure 3).

Other measures of accuracy included PLR, calculated as sensitivity/(1-specificity), NLR, calculated as (1-sensitivity)/specificity, and DOR (defined as PLR/NLR).[37] The pooled PLR was highest for SAPS II (6.9 [2.8-17.2]), while the best NLR was apparent for APACHE II (0.24 [0.19-0.30]). Furthermore, the best DOR for the detection of mortality was reported for SAPS II (25 [7-93], Table 3). Area under the SROC curve and Fagan's nomogram

The pooled area under the ROC curve for APACHE II was 0.87 (95% CI: 0.84-0.90), for SAPS II was 0.85 (95% CI: 0.82-0.88), and for SOFA was 0.88 (95% CI: 0.85-0.90, Figure 5). These results suggest that the diagnostic accuracy of all scores was good to discriminate patient's

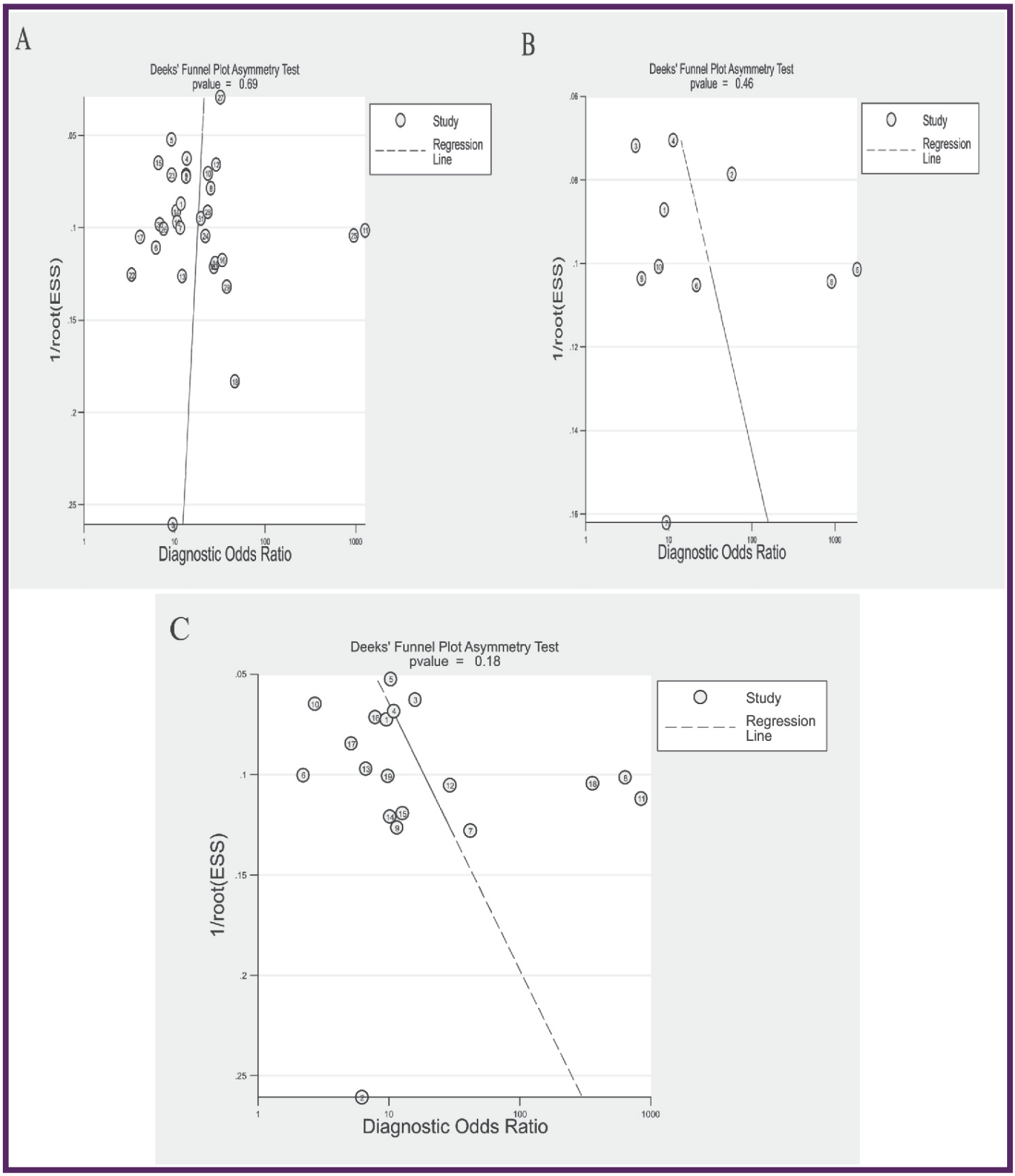

Figure 4: Deeks' funnel plots with superimposed regression lines to identify the potential publication bias in studies investigating the discriminative abilities of APACHE II (A), SAPS II (B), and SOFA (C) scores to detect patients' mortality. ESS indicates effective sample size.

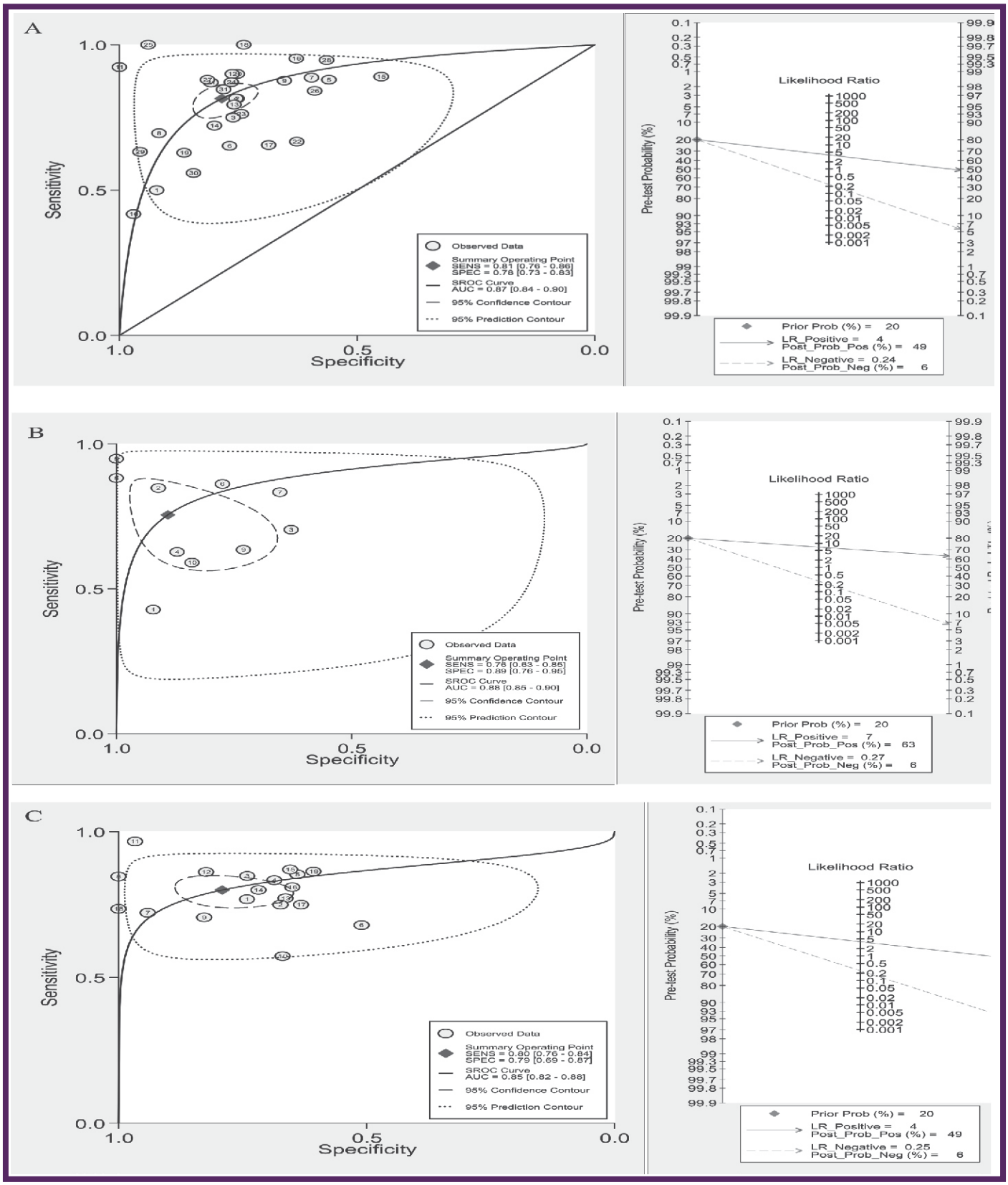

Figure 5: Hierarchical summary receiver operating characteristic curves (HSROC) and Fagan Plots of APACHE II (A), SAPS II (B), and SOFA (C).

mortality. In order to identify the diagnostic performance of each score, data was visually illustrated using Fagan plots. Results showed that the probability of death among ICU-admitted patients increased markedly to 49% when using APACHE II and SOFA and increased to 63% when using SAPS II for pretest probabilities of 20% (Figure 5).

DISCUSSION

During the past three decades, considerable efforts have been exerted in developing suitable models for predicting the risk of death among patients admitted to ICUs. Mortality prediction is an essential element to assess the severity of illness and to establish the effects of novel therapeutics and healthcare policies. SAPS II, APACHE II, and SOFA have been identified as the most commonly used scores in clinical practice. The present meta-analysis assessed the prognostic performance of these non-disease-specific scores to predict mortality in critically-ill patients. These scores showed good discriminative abilities, with a slight predominance of APACHE II in terms of sensitivity and NLR as well as SAPS II in terms of specificity, PLR, and DOR. However, there were significant levels of heterogeneity in sensitivity and specificity, which were accounted for by studies investigating APACHE II with a cut-off score of ≥20 ($P = 0.04$) and those investigating SAPS II for follow-up periods ≥24 months ($P < .001$).

Such heterogeneity might interfere with accurate interpretation of the reported outcomes although we attempted to correct such an issue by using a random-effect model. By using meta-regression analysis, we were able to reveal the sources of diversity. As with any meta-regression, the number of included studies may affect the power of regression to reveal significant effects. The total number of studies in our study was sufficient to yield reliable and trustworthy results. However, other covariates could contribute to this heterogeneity, such as patients' characteristics, methods of data collection, and specific assessment measures related to ICU scores.[38] Differences in methodological designs, settings, and subjective assessments of clinicians can all partly explain these variations.

It is important to note that the predictive models of ICU scores are not solely dependent on the "discrimination domain", but rather on the validity and calibration of these models. Ideally, a model should be well-calibrated, validated, and discriminated to predict the probability of in-hospital death. The validity of a model includes an assessment of the performance of mortality prediction via testing the used dataset that has been used for model development. On the other hand, calibration entails assessment of the degree of correspondence of the estimated mortality by the model and the actual observed mortality among the population under study. This can be performed statistically using formal goodness-of-fit tests.[39] These domains should be effectively differentiated from the discriminative ability of scoring systems to distinguish patients' mortality, which is the core principle of the "discrimination" domain.

In general, the results of the present study indicate that APACHE II, SAPS II, and SOFA had moderate discriminatory ability for mortality prediction in patients admitted to ICUs. These findings were evident in the pooled values of sensitivity and specificity (less than 0.90 for all scores). Furthermore, the correlations between sensitivity and specificity, as depicted in the hierarchical summary ROC curves (HSROC), can yield beneficial explanatory evidence of the performance of ICU scores. Ideally, a highly accurate score has an AUC close to 1, while the AUC of a poor test would be closer to 0.5. The HSROC of all scores ranged between 0.85 and 0.88, indicating good performance of these scores in differentiating patients' mortality. Indeed, the ROC curve is a better indicator of performance rather than mere values of sensitivity and specificity because ROC curves display all possible cut-off points and their results are independent on incidence estimates of mortality.[40] Based on HSROC, it can be concluded that APACHE II performed better than SAPS II when high sensitivity is required, while SAPS II performed better than other scores when high specificity is needed. Similarly, other recent meta-analyses revealed that the APACHE II score had the highest accuracy levels to predict mortality as compared to other risk prediction models in patients with ventilator-associated pneumonia[41] and acute pancreatitis.[42, 43]

In the current study, we showed that the clinical usefulness of the three scores to predict mortality was moderate. Only SAPS II reported high pooled PLR 6.9, which is higher than the acceptable threshold of clinically-relevant tests (> 5.0).[44] Actually, although the predictive ability of SAPS II showed acceptable discrimination, other studies showed poor calibration in their populations.[39, 45-47] In an early systematic review of SOFA-based models for mortality prediction, Minne et al.[48] found that the initial SOFA scores yielded good to excellent accuracy to discriminate survivors and non-survivors at the ICU. However, the diagnostic performance of SOFA was slightly worse than that of APACHE II and approximately similar to that of SAPS II. Notably, employing a combination of SOFA-based models with SAPS II and APACHE II improved the predictive performance when compared to using either model alone.[48] In our study, the discriminative ability of the initial SOFA score to predict mortality was not superior to other scores in terms of sensitivity, specificity, PLR, NLR, and DOR. This might be attributable to the relevance of SOFA score to the assessment of organ failure over time rather than specific mortality prediction.[8]

In the present study, the significant increase of pooled APACHE II scores in non-survivors as compared to survivors (-8.54 [95% CI -9.62, -7.46]) was relatively higher than that reported in a recent meta-analysis of APACHE II scores following paraquat poisoning among Chinese and Korean population (-7.29 [95% CI -8.96, -5.63]).[49] However, to the best of our knowledge, there is no evidence regarding pooled scores of SAPS II and/or SOFA to predict mortality.

STRENGTHS AND LIMITATIONS

Prospective studies were included to address the potential confounding factors that may affect the outcomes. In addition, data-gathering methods might not be homogenous in retrospective studies. Indeed, data gathering errors were common in patients with low or high APACHE II scores and this impacted the prognostic significance of their models.[50] Such erroneous reporting might have contributed to the poor calibration reported in early studies although AUC values of APACHE II were satisfactory.[7, 13, 45] In our study, the total number of studies was sufficient to provide results of a high statistical power.

However, the present study might be subject to several limitations. The outcomes should be interpreted cautiously due to the reported heterogeneity in sensitivity and specificity. In addition, the predictive ability of ICU scores might be deteriorated over time.[51, 52] For example, the performance of SAPS II was affected by national differences and case-mix when it was used in national studies.[53, 54] Nevertheless, performance has been less investigated in studies involving multiple countries. Hence, recommendations indicate that the prognostic models of these scores should be recalibrated and revalidated repeatedly and these measures should be emphasized when the models are used in a new country. Moreover, approximately three-quarters of included studies (75.7%) assessed the performance of scores to predict in-hospital mortality rather than relying on distinct time periods of patients' deaths. Actually, in-hospital mortality might be inadequate in the context of recent investigations as it might be affected by hospital discharge practices.[55, 56] As such, mortality endpoints should be based on fixed times and for long-term periods. Finally, we included studies recruiting specific patient populations, such as typhus and poisoning, which has contributed to the observed heterogeneity.

CONCLUSION

The performance of non-disease-specific ICU scores in predicting mortality of critically-ill patients was explored in this study. We showed that APACHE II, SAPS II, and initial SOFA showed good discriminative power for mortality prediction. However, the prognostic performance of APACHE II, SAPS II was slightly superior to that of initial SOFA scoring system as revealed by a trend of better pooled sensitivity, specificity, and HSROC values. The prognostic significance of ICU scores should be studied on a multinational level, including validity, calibration, and discrimination domains. In addition, the predictive performance should be assessed for long periods and using fixed-time endpoints of mortality rather than in-hospital mortality. Finally, the predictive ability of a combination of SOFA derivatives, APACHE II, and SAPS II should be heavily investigated prospectively and compared to either model alone.

KEY MESSAGES

❖ Despite the significant progress achieved in the management of critically-ill patients, there is a significant gap in the best ways to predict patients' outcomes, including disability and mortality.

❖ Scoring systems used in intensive care units can assist in prediction when used efficiently.

❖ There is a need to effectively rely on distinct scoring systems, particularly generic scores which are frequently used, to discriminate patients who die from those who live.

❖ The present study meta-analysis showed good prognostic potentials of three generic scoring systems, namely APACHE II, SAPS II, and SOFA scores.

❖ The most common limitation in meta-analyses of observational studies, heterogeneity, was also evident in the current study and we have identified some potential sources of such variation.

REFERENCES

1. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Critical Care Medicine*. 1985;13(10):818-29.

2. Bouch DC, Thompson JP. Severity scoring systems in the critically ill. *Continuing Education in Anaesthesia Critical Care & Pain*. 2008;8(5):181-5.

3. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. *Critical Care*. 2010;14(2):207.

4. Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: a compendium. *Indian Journal of Critical Care Medicine*. 2014;18(4):220-8.

5. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*. 2009;151(4):264-9.

6. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *Journal of American Medical Association*. 2000;283(15):2008-12.

7. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of American Medical Association*. 1993;270(24):2957-63.

8. Vincent JL, De Mendonca 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. *Critical Care Medicine*. 1998;26(11):1793-800.

9. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*. 2003;3(1):25.

10. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology*. 2014;14(1):135.

11. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*. 2005;58(9):882-93.

12. Khwannimit B, Bhurayanontachai R. The performance of customised APACHE II and SAPS II in predicting mortality of mixed critically ill patients in a Thai medical intensive care unit. *Anaesthesia and Intensive Care*. 2009;37(5):784-90.

13. Khwannimit B, Geater A. A comparison of APACHE II and SAPS II scoring systems in predicting hospital mortality in Thai adult intensive care units. *Journal-Medical Association of Thailand*. 2007;90(4):643-52.

14. Ahnert P, Creutz P, Horn K, Schwarzenberger F, Kiehntopf M, Hossain H, et al. Sequential organ failure assessment score is an excellent operationalization of disease severity of adult patients with hospitalized community acquired pneumonia - results from the prospective observational PROGRESS study. *Critical Care*. 2019;23(1):110.

15. Goertz O, Gharagozlou AF, Hirsch T, Homann HH, Steinau HU, Daigeler A, et al. Long-term comparison of a routine laboratory parameter-based severity score with APACHE II and SAPS II. *Journal of Trauma and Acute Care Surgery*. 2011;71(6):1835-40.

16. Chatzicostas C, Roussomoustakaki M, Notas G, Viachonikolis IG, Samonakis D, Romanos J, et al. A comparison of Child-Pugh, APACHE II and APACHE III scoring systems in predicting hospital mortality of patients with liver cirrhosis. *BMC Gastroenterology*. 2003;3(1):7.

17. Grmec S, Gasparovic V. Comparison of APACHE II, MEES and Glasgow Coma Scale in patients with nontraumatic coma for prediction of mortality. *Critical Care*. 2000;5(1):19-23.

18. Gursel G, Demirtas S. Value of APACHE II, SOFA and CPIS scores in predicting prognosis in patients with ventilator-associated pneumonia. *Respiration*. 2006;73(4):503-8.

19. Lopez-Delgado JC, Esteve F, Javierre C, Torrado H, Carrio ML, Rodriguez-Castro D, et al. Predictors of long-term mortality in patients with cirrhosis undergoing cardiac surgery. *The Journal of Cardiovascular Surgery*. 2014;56(4):647-54.

20. Oliveira VM, Brauner JS, Rodrigues Filho E, Susin RG, Draghetti V, Bolzan ST, et al. Is SAPS 3 better than APACHE II at predicting mortality in critically ill transplant patients? *Clinics*. 2013;68(2):153-8.

21. Feng Z, Wang T, Liu P, Chen S, Xiao H, Xia N, et al. Efficacy of various scoring systems for predicting the 28-day survival rate among patients with acute exacerbation of chronic obstructive pulmonary disease requiring emergency intensive care. *Canadian Respiratory Journal*. 2017;**2017**:3063510.

22. Yoon JC, Kim YJ, Lee YJ, Ryoo SM, Sohn CH, Seo DW, et al. Serial evaluation of SOFA and APACHE II scores to predict neurologic outcomes of out-of-hospital cardiac arrest survivors with targeted temperature management. *PloS one*. 2018;13(4).

23. Singh P, Pathak S, Sharma RM. A comparison of Acute Physiology and Chronic Health Evaluation III and Simplified Acute Physiology Score II in predicting sepsis outcome in intensive care unit. *Anesthesia, Essays and Researches*. 2018;12(2):592-7.

24. Liu X, Shen Y, Li Z, Fei A, Wang H, Ge Q, et al. Prognostic significance of APACHE II score and plasma suPAR in Chinese patients with sepsis: a prospective observational study. *BMC Anesthesiology*. 2015;16(1):46.

25. Goswami J, Balwani MR, Kute V, Gumber M, Patel M, Godhani U, et al. Scoring systems and outcome of chronic kidney disease patients admitted in intensive care units. *Saudi Journal of Kidney Diseases and Transplantation*. 2018;29(2):310.

26. Pan K, Panwar A, Roy U, Das BK. A comparison of the intracerebral hemorrhage score and the Acute Physiology and Chronic Health Evaluation II Score for 30-day mortality prediction in spontaneous intracerebral hemorrhage. *Journal of Stroke and Cerebrovascular Diseases*. 2017;26(11):2563-9.

27. Zhou XY, Ben SQ, Chen HL, Ni SS. A comparison of APACHE II and CPIS scores for the prediction of 30-day mortality in patients with ventilator-associated pneumonia. *International Journal of Infectious Diseases*. 2015;30:144-7.

28. Safari S, Shojaee M, Rahmati F, Bararartloo A, Hahshemi B, Forouzanfar MM, et al. Accuracy of SOFA score in prediction of 30-day outcome of critically ill patients. *Turkish Journal of Emergency Medicine*. 2016;16(4):146-50.

29. Gilani MT, Razavi M, Azad AM. A comparison of Simplified Acute Physiology Score II, Acute Physiology and Chronic Health Evaluation II and Acute Physiology and Chronic Health Evaluation III scoring system in predicting mortality and length of stay at surgical intensive care unit. *Nigerian Medical Journal*. 2014;55(2):144-7.

30. Balasubramanian P, Sharma N, Biswal M, Bhalla A, Kumar S, Kumar V, et al. Critical illness scoring systems: Sequential Organ Failure Assessment, Acute Physiology and Chronic Health Evaluation II, and Quick Sequential Organ Failure assessment to predict the clinical outcomes in scrub typhus patients with organ dysfunctions. *Indian Journal of Critical Care Medicine*. 2018;22(10):706-10.

31. Chang CH, Fan PC, Chang MY, Tian YC, Hung CC, Fang JT, et al. Acute kidney injury enhances outcome prediction ability of sequential organ failure assessment score in critically ill patients. *PloS one*. 2014;9(10):e109649-e.

32. Ho YP, Chen YC, Yang C, Lien JM, Chu YY, Fang JT, et al. Outcome prediction for critically ill cirrhotic patients: a comparison of APACHE II and Child-Pugh scoring systems. *Journal of Intensive Care Medicine*. 2004;19(2):105-10.

33. Kuo G, Yang SY, Chuang SS, Fan PC, Chang CH, Hsiao YC, et al. Using acute kidney injury severity and scoring systems to predict outcome in patients with burn injury. *Journal of Formosan Medical Association*. 2016;115(12):1046-52.

34. Srinivasan M, Shetty N, Gadekari S, Thunga G, Rao K, Kunhikatta V, et al. Comparison of the nosocomial pneumonia mortality prediction (NPMP) model with standard mortality prediction tools. *Journal of Hospital Infection*. 2017;96(3):250-5.

35. Alizadeh AM, Hassanian-Moghaddam H, Shadnia S, Zamani N, Mehrpour O. Simplified acute physiology score II/acute physiology and chronic health evaluation II and prediction of the mortality and later development of complications in poisoned patients admitted to intensive care unit. *Basic & Clinical Pharmacology & Toxicology*. 2014;115(3):297-300.

36. Venkataraman R, Gopichandran V, Ranganathan L, Rajagopal S, Abraham BK, Ramakrishnan N. Mortality prediction using Acute Physiology and Chronic Health Evaluation II and Acute Physiology and Chronic Health Evaluation IV scoring systems: is there a difference? *Indian Journal of Critical Care Medicine*. 2018;22(5):332-5.

37. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*. 2003;56(11):1129-35.

38. Harbord RM, Higgins JP. Meta-regression in stata. *The Stata Journal*. 2008;8(4):493-519.

39. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, et al. Mortality prediction using SAPS II: an update for French intensive care units. *Critical Care*. 2005;9(6):R645-52.

40. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics*. 2011;48(4):277-87.

41. Larsson J, Itenov TS, Bestle MH. Risk prediction models for mortality in patients with ventilator-associated pneumonia: a systematic review and meta-analysis. *Journal of Critical Care*. 2017;37:112-8.

42. Yang YX, Li L. Evaluating the ability of the bedside index for severity of acute pancreatitis score to predict severe acute pancreatitis: a meta-analysis. *Medical Principles and Practice*. 2016;25(2):137-42.

43. de-Madaria E, Sánchez-Payá J, Wu BU, Soler-Sala G, Lopez-Font I, Singh VK, et al. Tu1491 BISAP versus APACHE II for the prediction of mortality in acute pancreatitis: results of a cohort of patients and meta-analysis. *Gastroenterology*. 2012;142(5): S-847-S-8.

44. Lv SP, Wang Y, Huang L, Wang F, Zhou JG, Ma H. Meta-analysis of serum gastrin-releasing peptide precursor as a biomarker for diagnosis of small cell lung cancer. *Asian Pacific Journal of Cancer Prevention*. 2017;18(2):391-7.

45. Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Medicine*. 2003;29(2):249-56.

46. Desa K, Peric M, Husedzinovic I, Sustic A, Korusic A, Karadza V, et al. Prognostic performance of the Simplified Acute Physiology Score II in major Croatian hospitals: a prospective multicenter study. *Croatian Medical Journal*. 2012;53(5):442-9.

47. Haaland OA, Lindemark F, Flaatten H, Kvale R, Johansson KA. A calibration study of SAPS II with Norwegian intensive care registry data. *Acta Anaesthesiologica Scandinavica*. 2014;58(6):701-8.

48. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review. *Critical Care*. 2008;12(6):R161.

49. Huang J, Xuan D, Li X, Ma L, Zhou Y, Zou H. The value of APACHE II in predicting mortality after paraquat poisoning in Chinese and Korean population: a systematic review and meta-analysis. *Medicine*. 2017;**96**:e6838-e.

50. Holt AW, Bury LK, Bersten AD, Skowronski GA, Vedig AE. Prospective evaluation of residents and nurses as severity score data collectors. *Critical Care Medicine*. 1992;20(12):1688-91.

51. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Critical Care Medicine*. 2006;34(5):1378-88.

52. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Medicine*. 2012;38(1):40-6.

53. Strand K, Flaatten H. Severity scoring in the Sequential Organ Failure Assessment scoring systems for prognostication of outcomes among Intensive Care Unit's patients. *Saudi Journal of Anaesthesia*. 2016;10(2):168-73.

63. Hosseini M, Ramazani J. Comparison of Acute Physiology and Chronic Health Evaluation II and Glasgow Coma Score in predicting the outcomes of Post Anesthesia Care Unit's patients. *Saudi Journal of Anaesthesia*. 2015;9(2):136-41.

64. Jain S, Guleria K, Suneja A, Vaid NB, Ahuja S. Use of the Sequential Organ Failure Assessment score for evaluating outcome among obstetric patients admitted to the intensive care unit. *International Journal of Gynaecology & Obstetrics*. 2016;132(3):332-6.

65. Kim YH, Yeo JH, Kang MJ, Lee JH, Cho KW, Hong CK, et al. Performance assessment of the SOFA, APACHE II scoring system, and SAPS II in intensive care unit organophosphate poisoned patients. *Journal of Korean Medical Science*. 2013;28(12):1822-6.

66. Kulkarni SV, Naik AS, Subramanian Jr N. APACHE-II scoring system in perforative peritonitis. *The American Journal of Surgery*. 2007;194(4):549-52.

67. Mohamed AK, Mehta AA, James P. Predictors of mortality of severe sepsis among adult patients in the medical intensive care unit. *Lung India*. 2017;34(4):330-5.

68. Olmez S, Gumurdulu Y, Tas A, Karakoc E, Kara B, Kidik A, et al. Prognostic markers in cirrhotic patients requiring intensive care: a comparative prospective study. *Annals of Hepatology*. 2012;11(4):513-8.

69. Sharma S, Gupta A, Virmani SK, Lal R. Assessment and comparison of 3 mortality prediction models SAPS II, APACHE II and SOFA for prediction of mortality in patients of sepsis. *International Journal of Advances in Medicine*. 2017;4(3):623-9.

70. VijayGanapathy S, Karthikeyan VS, Sreenivas J, Mallya A, Keshavamurthy R. Validation of APACHE II scoring system at 24 hours after admission as a prognostic tool in urosepsis: A prospective observational study. *Investigative and Clinical Urology*. 2017;58(6):453-9.

71. Wang IK, Wang ST, Chang HY, Lin CL, Kuo HL, Chen TC, et al. Prognostic value of acute physiology and chronic health evaluation II and organ system failure in patients with acute renal failure requiring dialysis. *Renal Failure*. 2005;27(6):663-9.

Appendix 1: The employed search strategy in this study.

#1 “prognostic” OR “predictive” OR “survival” OR “mortality”

#2 “critical” OR “intensive”

#3 #1 AND #2

#4 “scoring system” OR “rating system” OR "APACHE" OR "SAPS" OR "SOFA" OR "GSC" OR "MPM"

#5 "acute physiology and chronic health evaluation" OR "simplified acute physiology score" OR "Sepsis-related organ failure assessment" OR "Glasgow coma scale" OR "Mortality prediction model"

#6 #4 OR #5

#7 #3 AND #6

#8 "Prospective" OR "cohort" OR "observational"

#9 Intensive Care Units OR Critical Care OR “intensive care”

#10 #8 AND #9

#11 #7 AND #10

#12 limit #11 to English language

#13 limit #12 to adults

#14 limit #13 to human

* ——————————————— *