





ORIGINAL ARTICLE

Clinical usefulness of scoring systems to predict severe acute pancreatitis: A systematic review and meta-analysis with pre and post-test probability assessment

Gabriele Capurso¹  | Ruggero Ponz de Leon Pisani¹ | Gaetano Lauri¹ |
 Livia Archibugi¹  | Peter Hegyi^{2,3,4} | Georgios I. Papachristou⁵ |
 Sanjay Pandanaboyana^{6,7} | Patrick Maisonneuve⁸ | Paolo Giorgio Arcidiacono¹  |
 Enrique de-Madaria^{9,10} 

¹Pancreato-Biliary Endoscopy and Endosonography Division, Pancreas Translational & Clinical Research Center, San Raffaele Scientific Institute IRCCS, Vita-Salute San Raffaele University, Milan, Italy

²Centre for Translational Medicine, Semmelweis University, Budapest, Hungary

³Institute of Pancreatic Diseases, Semmelweis University, Budapest, Hungary

⁴Translational Pancreatology Research Group, Interdisciplinary Centre of Excellence for Research Development and Innovation University of Szeged, Szeged, Hungary

⁵Division of Gastroenterology, Hepatology, and Nutrition, The Ohio State University, Wexner Medical Center, Columbus, Ohio, USA

⁶Department of Hepato-Pancreato-Biliary and Transplant Surgery, The Freeman Hospital, Newcastle upon Tyne, Tyne and Wear, UK

⁷Population Health Sciences Institute, Newcastle University, Newcastle, UK

⁸Division of Epidemiology and Biostatistics, IEO European Institute of Oncology, Milan, Italy

⁹Gastroenterology Department, Dr. Balmis General University Hospital, ISABIAL, Alicante, Spain

¹⁰Department of Clinical Medicine, Miguel Hernández University, Elche, Spain

Correspondence

Gabriele Capurso, Pancreato-Biliary Endoscopy and Endosonography Division, Pancreas Translational & Clinical Research

Abstract

Background: Scoring systems for severe acute pancreatitis (SAP) prediction should be used in conjunction with pre-test probability to establish post-test probability of SAP, but data of this kind are lacking.

Objective: To investigate the predictive value of commonly employed scoring systems and their usefulness in modifying the pre-test probability of SAP.

Methods: Following PRISMA statement and MOOSE checklists after PROSPERO registration, PubMed was searched from inception until September 2022. Retrospective, prospective, cross-sectional studies or clinical trials on patients with acute pancreatitis defined as Revised Atlanta Criteria, reporting rate of SAP and using at least one score among Bedside Index for Severity in Acute Pancreatitis (BISAP), *Acute Physiology and Chronic Health Examination* (APACHE)-II, RANSON, and *Systemic Inflammatory Response Syndrome* (SIRS) with their sensitivity and specificity were included. Random effects model meta-analyses were performed. Pre-test probability and likelihood ratio (LR) were combined to estimate post-test probability on Fagan nomograms. Pooled severity rate was used as pre-test probability of SAP and pooled sensitivity and specificity to calculate LR and generate post-test probability. A priori hypotheses for heterogeneity were developed and sensitivity analyses planned.

Results: 43 studies yielding 14,116 acute pancreatitis patients were included: 42 with BISAP, 30 with APACHE-II, 27 with Ranson, 8 with SIRS. Pooled pre-test probability of SAP ranged 16.6%–25.3%. The post-test probability of SAP with positive/negative score was 47%/6% for BISAP, 43%/5% for APACHE-II, 48%/5% for Ranson, 40%/12% for SIRS. In 18 studies comparing BISAP, APACHE-II, and Ranson in 6740 patients with pooled pre-test probability of SAP of 18.7%, post-test

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. United European Gastroenterology Journal published by Wiley Periodicals LLC on behalf of United European Gastroenterology.

Center, San Raffaele Scientific Institute
IRCCS, Via Olgettina 60, Milan 20132, Italy.
Email: capurso.gabriele@hsr.it

probability when scores were positive was 48% for BISAP, 46% for APACHE-II, 50% for Ranson. When scores were negative, post-test probability dropped to 7% for BISAP, 6% for Ranson, 5% for APACHE-II. Quality, design, and country of origin of the studies did not explain the observed high heterogeneity.

Conclusions: The most commonly used scoring systems to predict SAP perform poorly and do not aid in decision-making.

KEYWORDS

acute pancreatitis, APACHE-II, BISAP, meta-analysis, prediction, RANSON, Revised Atlanta Criteria, scoring system, severe, SIRS

BACKGROUND

Acute pancreatitis (AP) is a frequent condition with increasing incidence.¹ AP is a heterogeneous disease, and while most patients experience a mild course, approximately one-third have local or systemic complications that are associated with increased morbidity, and in cases of persistent (>48 h) organ failure, with high mortality risk.² Therefore, classifying AP severity is important to correctly stratify patients with extremely different disease courses. However, the most employed systems to determine AP severity, the *Revised Atlanta Classification* (RAC)³ and the *Determinant-Based Classification*⁴, take into consideration the presence of local complications and organ failure occurring at any time during the disease course, being "post-hoc" methods. While such severity classifications are useful for the final categorization of patients, they are not helpful for early management.

Predicting the severity of AP involves detecting, at an early disease stage, those patients most likely to have poor outcomes, which remains a challenge. Accurate early prediction of severity would allow the selection of patients who should be followed more closely, cared for in an intensive care unit, or transferred to tertiary centers. Severity prediction is also essential in the selection of patients to be included in trials.

Many different approaches have been developed to predict AP severity. The most employed scores include some specifically developed for AP, such as the Ranson score⁵ and the Bedside Index for Severity in Acute Pancreatitis (BISAP),⁶ and others that are not specific for AP, such as the *Acute Physiology and Chronic Health Examination* (APACHE)-II⁷ and the *Systemic Inflammatory Response Syndrome* (SIRS).⁸ Several other scoring systems and tools employing combinations of physiological, laboratory, and radiographic parameters have been developed, but they all show only moderate positive predictive values.⁹

However, as for any test, the sensitivity and specificity of these scores alone cannot be used to accurately estimate the probability of severe disease in individual patients. This depends on their combination into the likelihood ratio (LR), which should then be used in conjunction with pre-test probability to establish the post-test probability of severe AP (SAP) in a clinically meaningful manner.¹⁰ In Bayesian statistics, this concept is visually summarized by Fagan's

Key summary

Summarize the established knowledge on this subject

- Acute pancreatitis is a common, heterogeneous disease. Most patients experience a mild disease, and predicting a severe course would be of outmost clinical value.
- Many scoring systems have been used to this aim, and they all have been shown to have only moderate predictive value. However, as for any test, the sensitivity and specificity of these scores alone cannot be used to accurately estimate the probability of severe disease in individual patients.

What are the significant and/or new findings of this study?

- In this systematic review and meta-analysis, a Bayesian approach was employed for the first time to depict the combination of the likelihood ratios of scoring systems with the pre-test probability and to estimate the resulting post-test probabilities.
- We included 43 studies yielding 14,116 patients. All scoring systems had limited clinical usefulness as the actual post-test probability of severe acute pancreatitis never reached 50% when scores were predicting a severe course and ranged between 5% and 12% when they were predicting a non-severe course.
- In real-life clinical practice, the most used scoring systems to predict severe acute pancreatitis perform poorly and have the same value as tossing a coin. New approaches seem necessary.

nomogram, a graphical tool that allows the combination of the LR of a test with the pre-test probability of the outcome of interest to estimate the post-test probability.¹¹

As data on the clinical usefulness of predictive scores for SAP are sparse and heterogeneous and given the absence of previous studies of this kind, we designed a systematic review and meta-analysis to investigate the actual value of the most common predictive scoring systems in modifying the pre-test probability of developing SAP.

METHODOLOGY

The methodology of the study was developed and reviewed with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement¹² and the Meta-Analyses Of Observational Studies in Epidemiology (MOOSE)¹³ checklist. This review was registered in PROSPERO (ID CRD42022368212).

Search strategy

First, a computerized bibliographic search was performed in PubMed and the Cochrane Database of Systematic Reviews to retrieve prior systematic reviews and meta-analyses on this topic. A PubMed search was then run from inception until 10 September 2022, to identify original studies. The specific search terms are detailed in Supporting Information S1. The titles of all identified articles were screened to evaluate eligibility, and the abstracts and/or full texts of potentially relevant papers were further evaluated. We manually searched the reference lists of all the retrieved articles to identify other potentially relevant studies.

Inclusion and exclusion criteria

The selected studies had to meet these criteria: (a) be either retrospective, prospective, cross-sectional studies, or clinical trials; (b) report on AP patients defined according to RAC; (c) report SAP rate defined according to RAC or rate of organ failure, allowing SAP definition. In this view, the definition of persistent organ failure as renal, cardiovascular, or pulmonary failure lasting >48 h would be considered appropriate; (d) report on at least one of the following scores: BISAP, APACHE-II, RANSON, and SIRS; (e) report sensitivity and specificity of the score(s) and/or enough data to calculate them; (f) be in English language.

In duplicate publications, the most recent or complete were used. Two independent reviewers (Ruggero Ponz de Leon Pisani and Gaetano Lauri) completed the study identification and selection process, and disagreements were discussed with two other reviewers (Livia Archibugi and Gabriele Capurso). The excluded studies and reasons for exclusion were recorded. Case reports or series, letters, abstracts, reviews, animal, and in vitro studies were excluded, as were studies published before the RAC publication or those that did not allow calculation of severity or employed predictive scores in a non-standardized manner.

Data extraction and quality assessment

From the studies that met the eligibility criteria, the following data were extracted into a Microsoft Excel spreadsheet (Microsoft 2016, Redmond, WA, USA): (a) study—first author, publication year, setting, design, country, accrual period; (b) cases—number, sex, and age, the

rate of severity according to RAC; (c) severity score(s)—name, timing at evaluation, employed cut-off, sensitivity, specificity, +LR and –LR.

The quality of each study included in the quantitative synthesis was assessed by two independent reviewers (Ruggero Ponz de Leon Pisani and Gaetano Lauri) using a specific quality appraisal tool developed for prognostic factors.¹⁴ Disagreements were discussed with a third reviewer (Livia Archibugi).

Statistical analysis

A meta-analysis of all eligible studies was performed using the Comprehensive Meta-Analysis software package (Biostat, Englewood, N.J., USA). First, the pooled estimate of SAP was calculated to obtain the pre-test probability. Next, the pooled estimates of the sensitivity and specificity of the different scoring systems were calculated to obtain +LR and –LR. The Der Simonian-Laird method and a random-effects model were used. Random-effects models were chosen, as they consider both sampling variance within the different studies and variation in the underlying effect across studies. The assumption of variation in the underlying effect seems plausible given the different populations, designs, and etiology. Heterogeneity was assessed using the I^2 value and Cochran's Q statistics. An I^2 value $\leq 40\%$ was considered trivial heterogeneity, $I^2 > 40 < 75\%$ was considered important heterogeneity, and an $I^2 \geq 75\%$ considerable heterogeneity. Publication bias was assessed using the Begg and Mazumdar test.^{15,16} Statistical significance was set at $p < 0.05$. We developed the following a priori hypotheses that would explain heterogeneity and planned sensitivity analyses for (a) area of origin, (b) quality of the study, and (c) study design. An open-access online calculator (<http://araw.mede.uic.edu/cgi-bin/testcalc.pl>) was used to estimate the post-test probability.

RESULTS

Search results and study selection

There were no previous systematic reviews or meta-analyses in the Cochrane Database of Systematic Reviews. Out of 64 studies published in PubMed, we retrieved seven prior systematic reviews and meta-analyses on this topic. However, one only examined the performance of the Ranson score and was published before RAC,¹⁷ three only examined BISAP,^{18–20} one reviewed the performance of the Harmless Acute Pancreatitis Score (HAPS), which has the opposite aim of identifying patients who would not develop a severe disease²¹ and one focused on the role of the computed tomography index.²² Notably, the remaining study²³ differed substantially from the present one, as it aimed to investigate the net reclassification improvement using the available scores. Also, its search was terminated in mid-2016 and mortality was the main outcome.

In our search for original studies, a total of 1894 references were identified (Figure 1). After evaluation of titles, 443 records were

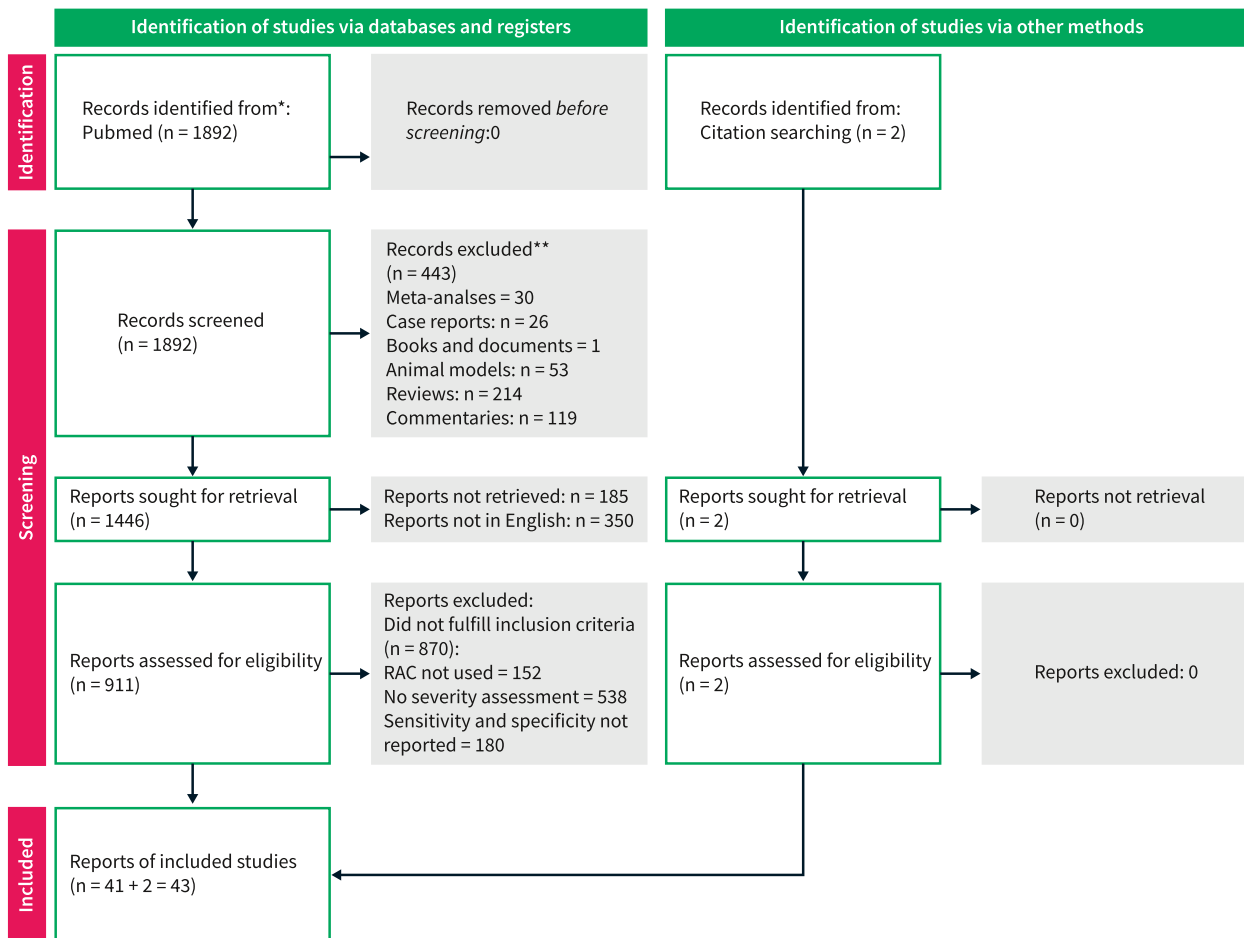


FIGURE 1 PRISMA 2020 flow diagram with included studies and reasons for exclusion.

removed as not related to the study topic. Thus, the abstracts of the remaining 1446 studies were examined and 911 checked for eligibility. Finally, 43 studies were included. There was absolute agreement among the reviewers for the assessment of eligibility and selection of studies.

Study characteristics

Table 1 presents a summary of relevant studies. The 43 studies²⁴⁻⁶⁵ included 14,116 AP patients, with mean/median age ranging widely (35–72 years) as the rate of male patients (34%–86%).

Thirty-four studies (79%) took place in Institutions in Asia,^{24-28,30-35,37,39,41-52,54-56,58-62,65} 7 in Europe,^{36,38,40,53,57,63,64} 2 in the USA.^{9,29} Twenty-one studies were prospective,^{9,25,29,32-36,38,41,44,45,47,51-54,57,59,61,65} 21 retrospective^{24,26-28,30,31,37,39,40,42,46,48-50,55,56,58,60,62-64} and 1 cross-sectional.⁴³ Three of the studies included both training and validation sets,^{9,37,39} and one two cohorts of different age groups.⁵⁰ The accrual period ranged 2005–2021.

Regarding the studies' quality, the analysis of study participation and attrition was not relevant as dealing with all patients hospitalized for the disease of interest with the availability of data on the

outcome. The QUIPS tool takes into consideration other 4 items: (a) prognostic factor measurement, (b) outcome measurement, (c) study confounding and statistical analysis, and (d) reporting. These were scored as low, moderate, or high risk of bias according to the tool. We considered 13 studies that had a low risk of bias for all items as of “high quality” and the remaining 30 as of “moderate quality” (Supplementary Table S1).

BISAP score

There were 42 studies investigating BISAP^{9,24-41,43-65} in a total of 13,944 patients. In these studies, the pooled prevalence of severity (pre-test probability) was 17.8% (14.3%–22.2%) with considerable heterogeneity ($I^2 = 96.9\%$) (Supplementary Figure S1a). There was no publication bias (Kendall's tau with continuity correction = -0.094 ; $p = 0.35$).

The pooled sensitivity and specificity of BISAP in these studies were 74.4% and 81.5%, both with considerable heterogeneity ($I^2 = 97.3$ and 95.5%) (Supplementary Figure S1b and S1c).

Figure 2a summarizes the performance of the BISAP score according to such data: with a pre-test probability of 17.8% and +LR

TABLE 1 Main characteristics of the studies eligible for the analyses.

Study	Reference	Year	Location	Country	Setting ^a	Design ^b	Subgroups	Patients	Sex (male %)	Age	Accrual period	Severe AP, N (%)	Score(s)
Mounzer R	12	2012	Pittsburgh, Boston	USA	M	P	Training	256	52%	51 (median)	2003–2010	62 (24.2)	BISAP, RANSON, APACHEII, SIRS
Mounzer R	12	2012	Pittsburgh, Boston	USA	M	P	Validation	397	49%	52 (median)	2005–2007	34 (8.6)	BISAP, RANSON, APACHEII, SIRS
Cho YS	29	2013	Uijeongbu	South Korea	U	R	None	299	69.6%	52.1 (mean)	2008–2010	22 (7.4)	BISAP, RANSON
Khanna AK	30	2013	Varanasi	India	U	P	None	72	51.4%	40.5 (mean)	2010–2012	31 (43.1)	BISAP, RANSON, APACHEII, SIRS
Park JY	31	2013	Seoul	South Korea	U	R	None	303	71.2%	52 (mean)	2007–2010	31 (10.2)	BISAP, APACHEII, RANSON
Zhang J	32	2014	Hefei	China	U	R	None	155	59%	51.8 (mean)	2010–2013	21 (13.5)	BISAP, APACHEII, RANSON
Cho JH	33	2015	Daegu	South Korea	U	R	None	161	63%	62.3 (mean)	2011–2012	21 (13)	BISAP, APACHEII, RANSON
Mok SRS	34	2015	Camden	USA	U	P	None	266	59%	48.8 (mean)	2011–2014	41 (15.4)	BISAP, APACHEII, RANSON
Qiu L	35	2015	Shanghai	China	M	R	None	129	58.9%	45.2 (mean)	2008–2014	20 (15.5)	BISAP, RANSON, SIRS
Sharma V	36	2015	Chandigarh	India	U	R	None	105	61.9%	40.6 (mean)	2013–2014	71 (67.6)	BISAP, SIRS
Yadav J	37	2016	Ranchi	India	U	P	None	119	70.6%	38.9 (mean)	2012–2014	42 (35.2)	BISAP, RANSON
Kumar AH	38	2018	Rohtak	India	U	P	None	50	34.0%	48.4 (mean)	2015–2016	14 (28)	BISAP, APACHEII, RANSON
He WH	39	2017	Nanchang	China	U	P	None	708	43.9%	51.7 (mean)	2011–2012	172 (24.3)	BISAP, APACHEII, SIRS
Shi Y	40	2017	Shenyang	China	U	P	None	56	58%	54 (median)	2015–2016	10 (13.2)	BISAP, APACHEII
Valverde-Lopez F	41	2017	Granada	Spain	U	P	None	269	49.9%	64.5 (mean)	2010–2012	17 (6.3)	BISAP, RANSON
Choi HW	42	2018	Seoul	South Korea	M	R	Training	115	69.5%	48.5 (mean)	2013–2016	17 (14.8)	BISAP, APACHEII
Choi HW	42	2018	Seoul	South Korea	M	R	Validation	77	66.2%	46.8 (mean)	2013–2016	11 (14.3)	BISAP, APACHEII
de-Madaria E	43	2018	Alicante, Barcelona	Spain	M	P	None	59	44.1%	64 (mean)	NR	13 (22)	BISAP, SIRS
Fei Y	44	2018	Nanjing	China	M	R	Training	1073	57.3%	47.3 (mean)	2013–2016	517 (48.1)	BISAP, APACHEII, RANSON
Fei Y	44	2018	Nanjing	China	M	R	Validation	326	51.5%	56.3 (mean)	2012–2016	126 (38.6)	BISAP, APACHEII, RANSON
Gravito-Soares M	45	2018	Coimbra	Portugal	U	R	None	182	54.9%	66.3 (mean)	2014–2016	91 (50)	BISAP, RANSON
Hagler S	46	2018	Assam	India	U	P	None	60	68.3%	37.1 (mean)	2015–2016	14 (23.3)	BISAP, APACHEII, RANSON
Yang WQ	47	2018	Shanghai	China	U	R	None	172	61%	48 (median)	2012–2017	11 (6.4)	APACHEII, RANSON
Arif A	48	2019	Karachi	Pakistan	U	CS	None	206	39.3%	35.2 (mean)	2015	39 (18.9)	BISAP, RANSON
Chen J	49	2019	Nanchang	China	U	P	None	113	61.1%	52.9 (mean)	2016–2018	44 (38.9)	BISAP, APACHEII
Jain D	50	2019	Rohtak	India	U	P	None	50	100%	42 (mean)	NR	10 (20)	BISAP, APACHEII, RANSON
Zhou H	51	2019	Beijing	China	U	R	None	406	59.6%	57 (mean)	2014–2017	56 (13.8)	BISAP, APACHEII, RANSON
Chatterjee R	52	2020	Mumbai	India	U	P	None	87	86.2%	37.7 (mean)	NR	20 (23)	BISAP, APACHEII
Gezer NS	53	2020	Izmir	Turkey	U	R	None	80	42.5%	55 (mean)	2015–2018	19 (23.8)	BISAP, RANSON
Li M	54	2020	Hangzhou	China	U	R	None	238	70.2%	39.7 (median)	2016–2018	60 (25.2)	BISAP, RANSON, APACHEII, SIRS
Li Y	55	2020	Nanjing	China	U	R	Elderly	368	54.6%	73.8 (mean)	2015–2018	27 (7.3)	BISAP, APACHEII, RANSON
Li Y	55	2020	Nanjing	China	U	R	Young	550	65.2%	42.1 (mean)	2015–2018	25 (4.5)	BISAP, APACHEII, RANSON
Peng R	56	2020	Panzhuhua	China	U	P	None	309	63.4%	50 (mean)	2017–2018	17 (5.5)	BISAP, APACHEII
Satis H	57	2020	Ankara	Turkey	U	P	Elderly	113	40%	73.7 (mean)	2014–2016	1 (2.5)	BISAP, APACHEII

(Continues)

TABLE 1 (Continued)

Study	Reference	Year	Location	Country	Setting ^a	Design ^b	Subgroups	Patients	Sex (male %)	Age	Accrual period	Severe AP, N (%)	Score(s)
Silva Vaz P	58	2020	Castelo Branco	Portugal	U	P	None	75	42.7%	72 (mean)	2015–2017	13 (17)	BISAP, SIRS
Venkatesh NR	59	2020	Puducherry	India	U	P	None	164	NR	45 (mean)	NR	104 (63.4)	BISAP, APACHEII, RANSON
Zhou T	60	2020	Nanchong	China	U	R	None	337	55.0%	50.4 (mean)	2016–2018	17 (5)	BISAP, APACHEII
Sun HW	61	2021	Whenzou	China	U	R	Validation	568	63.5%	50 (median)	2017–2019	162 (28.5)	BISAP, APACHEII, RANSON
Pando E	62	2021	Barcelona	Spain	U	P	None	410	51.0%	65.4 (median)	2015–2020	45 (11)	BISAP, APACHEII
Wu Q	63	2021	Nanning	China	U	R	None	1848	68.2%	48.2 (mean)	2003–2020	684 (37.0)	BISAP, RANSON
Shen D	64	2021	Changsha	China	U	P	None	143	65.7%	47 (median)	2019–2020	26 (18.1)	BISAP, APACHEII
Teng TZJ	65	2021	Singapore	Singapore	U	R	None	653	58.7%	58.7 (mean)	2009–2016	81 (12.4)	BISAP, APACHEII, RANSON
Wang Y	66	2021	Shanghai	China	U	P	None	103	58.2%	46 (median)	2019–2020	31 (30.1)	BISAP, APACHEII
Yan G	67	2021	Chongqing, Suining, Nanchong	China	M	R	None	465	54.4%	54.6 (mean)	2018–2020	27 (5.8)	BISAP, APACHEII, RANSON
Dancu GM	68	2021	Tmisoara	Romania	U	R	None	216	55.5%	56.3 (mean)	2018–2019	25 (11.5)	BISAP
Bardakcı O	69	2022	Çanakkale	Turkey	U	R	None	159	39%	68.6 (mean)	2017–2019	24 (15.1)	BISAP, APACHEII, RANSON
Wu B	70	2022	Chongqing	China	U	P	None	1046	59.4%	51.6 (mean)	2020–2021	117 (11.2)	BISAP

^aU = uncenter, M = multicenter.

^bR = retrospective, P = prospective, NR = not reported, AP = acute pancreatitis, N = number.

and –LR of 4.09 and 0.31, respectively, the post-test probability of SAP was 47% when BISAP was positive and 6% when negative.

APACHE-II score

Thirty studies^{9,25–29,33–35,37,39,41,42,44–47,49–52,54–57,59–62,64} investigated APACHE-II in 9344 patients. In these studies, the pooled prevalence of severity (pre-test probability) was 16.6% (12.6%–21.5%), with considerable heterogeneity ($I^2 = 96.8\%$) (Supplementary Figure S2a). There was no publication bias (Kendall tau with continuity correction = -0.10 ; $p = 0.37$).

The pooled sensitivity and specificity of APACHE II in these studies were 77.8% and 79.3%, both with considerable heterogeneity ($I^2 = 96.1$ and 96.3% , respectively) (Supplementary Figure S2b and S2c).

Figure 2b summarizes the performance of the APACHE II score according to such data: with a pre-test probability of 16.6% and +LR and –LR of 3.76 and 0.28, respectively, the post-test probability of SAP was 43% when APACHE was positive and 5% when negative.

Ranson score

There were 27 studies^{9,24–30,32,33,36,39–43,45,46,48–50,54,56,58,60,62,64} investigating Ranson in 10,044 patients. In these studies, the pooled prevalence of severity (pre-test probability) was 18.8% (14.3%–24.2%) with considerable heterogeneity ($I^2 = 97.4\%$) (Supplementary Figure S3a). There was no publication bias (Kendall's tau with continuity correction = -0.16 ; $p = 0.19$).

The pooled sensitivity and specificity of Ranson in these studies were 80% and 80.3%, both with considerable heterogeneity ($I^2 = 97.1$ and 96.6% , respectively) (Supplementary Figure S3b and S3c).

Figure 2c summarizes the performance of the Ranson score according to such data: with a pre-test probability of 18.8% and +LR and –LR of 4.06 and 0.25, respectively, the post-test SAP probability was 48% when Ranson was positive and 5% when negative.

SIRS score

Eight studies^{9,25,30,31,34,38,49,53} investigated SIRS in 2039 patients. The pooled prevalence of severity (pre-test probability) was 25.3% (17%–35.8%), with considerable heterogeneity ($I^2 = 94.6\%$) (Supplementary Figure S4a). There was no publication bias (Kendall's tau with continuity correction = -0.02 ; $p = 0.91$). The pooled sensitivity and specificity of SIRS in these studies were 74.4% and 62.7%, both with considerable heterogeneity ($I^2 = 97.1$ and 90.3% , respectively) (Supplementary Figure S4b,c).

Figure 2d summarizes the performance of the SIRS score according to such data, with a pre-test probability of 25.3% and +LR and –LR ratios of 1.99 and 0.41, respectively, and a post-test probability of SAP of 40% when SIRS is positive and 12% when negative.

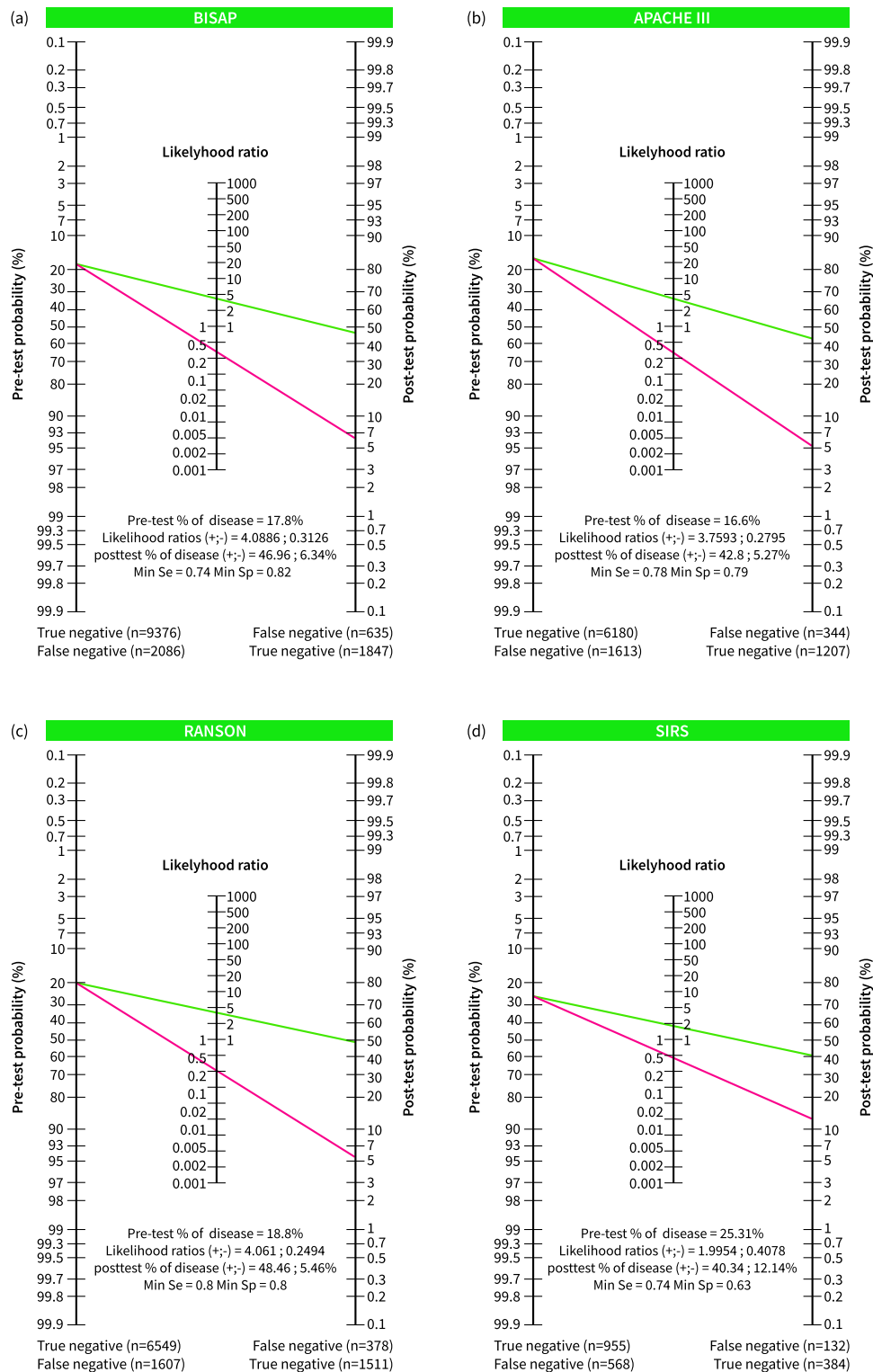


FIGURE 2 Legend on next page.

Comparison of the performance of the different scores and sensitivity analysis

Only 3 studies^{9,25,54} compared all four examined scores; four compared the BISAP, APACHE-II, and SIRS. To obtain the most comprehensive comparison of the performance of the

investigated scores, we selected 18 studies^{9,25-29,33,39,41,45,46,49,50,54,56,60,62,64} that compared the accuracy of the BISAP, APACHE-II, and Ranson scores for a total of 6740 patients. In this cohort, the pre-test SAP probability was 18.7% (13.1%–26.1%) (Supplementary Figure S5) with considerable heterogeneity ($I^2 = 97.6\%$).

Figure 3 summarizes the performances of the three scoring systems in this subgroup. Notably, with a pooled pre-test probability of 18.7%, the performance of the three Scoring Systems was very similar. The post-test probability when the scores were positive was 46% for APACHE II, 48% for BISAP, and 50% for Ranson. On the other hand, when the scores were negative, the post-test probability of a severe course was as low as 5% for APACHE-II, 6% for Ranson, and 7% for BISAP.

As for the sensitivity analyses (Supplementary Table S2), the quality, design and country of origin did not account for the observed heterogeneity. As most of the examined studies were conducted in Asia, where etiology, comorbidities and lifestyle are very different from those of Western countries (Europe and USA), we further investigated the performance of the scoring systems separately in such subgroups (Supplementary Figure S6). The performance was generally worse in studies conducted in Western Countries with post-test probabilities of a severe course when a score was positive being as low as 38% for BISAP, 19% for APACHE-II, 40% for Ranson and 27% for SIRS, compared to ,respectively, 50%, 48%, 49% and 51% in Asia.

DISCUSSION

In an individual patient, the pre-test probability of SAP is usually not higher than 20%. The purpose of the prediction scores is to generate a post-test probability of SAP that is as high as possible. Many different approaches have been developed, and many scoring systems that combine laboratory and clinical features are commonly employed for this purpose.⁶⁶

One of the limitations of scoring systems is that generalization may not be possible as they were developed and validated in certain groups of patients, but the clinicians need to make a prediction about the individual patient they are caring for. There are several ways the sensitivity and specificity can be combined into a single score. The most used method is the receiver operator characteristic curve,

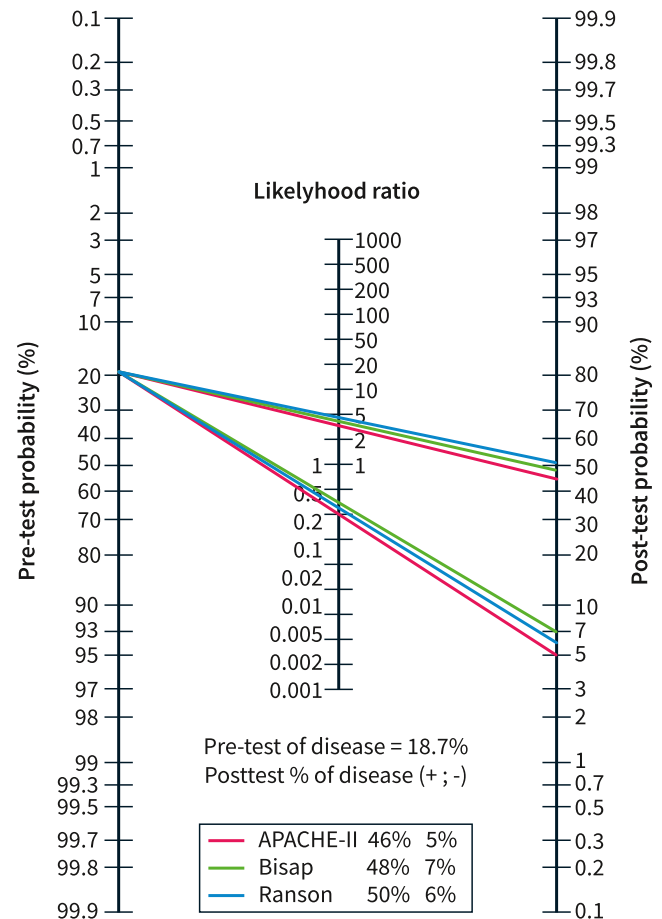


FIGURE 3 Performance of the Bedside Index for Severity in Acute Pancreatitis (BISAP), Acute Physiology and Chronic Health Examination (APACHE)-II, and Ranson scores in 18 studies with a pooled pre-test probability of 18.7% for severe acute pancreatitis. The post-test probabilities when the scores were positive were similar: 48% for BISAP, 46% for APACHE-II, and 50% for Ranson. However, when the scores were negative, the post-test probability dropped to 7% for BISAP, 6% for Ranson, and 5% for APACHE-II.

FIGURE 2 Panel (a) performance of the Bedside Index for Severity in Acute Pancreatitis (BISAP) score in 42 studies with a pre-test probability of 17.8% and positive and negative likelihood ratios of 4.09 and 0.31, respectively, and the post-test probability of severe acute pancreatitis (SAP) is 47% when BISAP is positive and 6% when it is negative. With this performance, only 1847 of the 2482 patients who eventually developed SAP would have been correctly identified as true positives, with 635 false negatives; only 9376 of the 11,462 patients experiencing non-severe AP would have been correctly classified, with 2086 having a false positive prediction of SAP. Panel (b) performance of the Acute Physiology and Chronic Health Examination (APACHE)-II score in 30 studies with a pre-test probability of 16.6% and positive and negative likelihood ratios of 3.76 and 0.28, respectively, and the post-test probability of SAP is 43% when APACHE is positive and 5% when it is negative. With this performance, only 1207 of the 1551 patients who eventually developed SAP would have been correctly identified as true positives, with 344 false negatives, and only 6180 of the 7793 patients experiencing non-severe AP would have been correctly classified, with 1613 having a false positive prediction of SAP. Panel (c) performance of the Ranson score in 27 studies with a pre-test probability of 18.8% and positive and negative likelihood ratios of 4.06 and 0.25, respectively. The post-test probability of SAP is 48% when Ranson is positive and 5% when it is negative. With this performance, only 1511 of the 1889 patients who eventually developed SAP would have been correctly identified as true positives, with 378 false negatives; only 6549 of the 8156 patients experiencing non-severe AP would have been correctly classified, with 1607 having a false positive prediction of SAP. Panel (d) performance of the *Systemic Inflammatory Response Syndrome* (SIRS) score in eight studies with a pre-test probability of 25.3% and positive and negative likelihood ratios of 1.99 and 0.41, respectively, with a post-test probability of SAP of 40% when SISR was positive and 12% when it was negative. With this performance, only 384 of the 516 patients who eventually developed SAP would have been correctly identified as true positives, with 132 false negatives, and only 955 of the 1523 patients with non-severe AP would have been correctly classified, with 568 having a false positive prediction of SAP.

which plots sensitivity and specificity presenting the performance as “area under the curve” While this may be of help in comparing the accuracy of different systems, it has limited clinical relevance for individual patients. A better approach is to derive the post-test probability by combining the expected pre-test probability for the patient population with the positive and negative LR and using a nomogram to read the post-test probability.¹¹

In the present study, we systematically retrieved literature on predictive scores of SAP defined according to RAC and calculated their performance using a Bayesian approach for the first time. We retrieved data from 43 studies conducted on >14,000 AP patients to investigate the accuracy of BISAP, APACHE-II, Ranson, and SIRS in predicting SAP. We first calculated the pre-test probability. Thereafter, the sensitivity and specificity of each score were calculated. These data were employed to generate positive and negative LR and post-test probabilities of SAP for each score. There was no publication bias.

The main result is that all scoring systems have a similar, limited, and clinical usefulness, as the post-test probability of SAP never reached 50% (Figure 2) when the score was positive and ranged between 5% and 12% when negative. Therefore, if these scoring systems are used to predict a non-severe course of AP, the HAPS score should be preferentially employed.²¹

To obtain a more reliable figure of the scoring systems' performance, we further focused on a subset of 18 studies that compared the accuracy of the BISAP, APACHE II, and Ranson scores in predicting SAP. In this cohort, the pre-test probability of SAP was 18.7%, and the performances of the three scoring systems were very similar (Figure 3), with a post-test probability $\leq 50\%$ when the scores were positive. This means that in real-life clinical practice, the use of these scores to predict SAP has the same value as tossing a coin.

The present study has strengths. This is the first systematic review with rigorous methodology to calculate the actual performance of scoring systems that have been employed for decades to predict SAP probability. However, there are limitations, mainly related to the heterogeneity of the studies. Despite pre-planned sensitivity analyses that included an evaluation of the quality, design and country of origin, no reasons for the observed high heterogeneity were found. However, there are many factors that are intrinsic to single patients, such as AP etiology,⁶⁷ age and comorbidities,⁶⁸ triglyceride and glucose levels,⁶⁹ and the setting where the patient is treated (hospital volume and resources),⁷⁰ which have an influence on AP course and may account for heterogeneity. Individual data analysis would be necessary to further investigate these aspects. Also, we separately investigated the performance of the four scoring systems in studies conducted in Asia versus Western countries, with findings of a much worse performance in the latter group. Whether this is due to the lower number of enrolled patients or to actual differences in the applicability of the systems must be established.

Our results reinforce the need for novel alternative approaches to predict an AP course. One would be to monitor the dynamic AP evolution during its course instead of focusing on a rather rare outcome, such as persistent organ failure. The *Pancreatitis Activity Scoring System* (PASS), which includes organ failure, SIRS, abdominal

pain, the need for opiates, and the ability to tolerate oral diet as variables, was developed for this aim⁷¹ and found to be able to track the clinical trajectories of an AP episode, anticipating deterioration and complications. However, its accuracy for predicting SAP at a single time point is limited.

Several novel tools are based on computed tomography (CT) imaging. The most obvious limitation of radiological approaches is that a CT scan is not required on admission in most patients.

Another novel approach uses information theory and machine learning to select the best-performing panel of circulating cytokines, which reflects the magnitude of inflammatory response. Angiopoietin-2, hepatocyte growth factor, interleukin-8, resistin, and tumor necrosis factor receptor-1 were the highest-ranking cytokines in the derivation cohort. A Random Forest classifier trained the 5-cytokine panel in the verification cohort and achieved a 10-fold cross-validated accuracy of 0.89, which significantly outperformed the prognostic accuracy of existing laboratory tests and clinical scores.⁷² As multiple factors interact in a nonlinear, complex, and unpredictable manner to determine the actual risk of developing SAP, artificial intelligence algorithms might be an appropriate tool to improve prediction ability. In a recent large cohort study, machine learning models were employed to examine simple variables such as respiratory rate, body temperature, abdominal rebound tenderness, sex, age, and glucose levels. The accuracy of the model was as high as 89% and a user-friendly web application (“EASY”) was developed for wider applications.⁷³

In conclusion, we have systematically reviewed the performance of the most commonly employed scoring systems to predict AP severity, with findings that underline their poor performance in everyday clinical practice.

It is likely that artificial intelligence will become a more common approach for rapid, early, and accurate prediction of AP severity and outcomes, which will outperform existing scoring systems.

ACKNOWLEDGEMENT

Open access funding provided by BIBLIOSAN.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest or funding to report regarding the study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Gabriele Capurso  <https://orcid.org/0000-0002-0019-8753>

Livia Archibugi  <https://orcid.org/0000-0003-3979-9553>

Paolo Giorgio Arcidiacono  <https://orcid.org/0000-0001-6692-7720>

Enrique de-Madaria  <https://orcid.org/0000-0002-2412-9541>

REFERENCES

1. Iannuzzi JP, King JA, Leong JH, Quan J, Windsor JW, Tanyingoh D, et al. Global incidence of acute pancreatitis is increasing over time: a

- systematic review and meta-analysis. *Gastroenterology*. 2022;162(1):122–34. <https://doi.org/10.1053/j.gastro.2021.09.043>
2. Sternby H, Bolado F, Canaval-Zuleta HJ, Marra-Lopez C, Hernando-Alonso A, Del-Val-Antonana A, et al. Determinants of severity in acute pancreatitis: a nation-wide multicenter prospective cohort study. *Ann Surg*. 2019;270(2):348–55. <https://doi.org/10.1097/SLA.0000000000002766>
 3. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis - 2012: revision of the Atlanta classification and definitions by international consensus. *Gut*. 2013;62(1):102–11. <https://doi.org/10.1136/gutjnl-2012-302779>
 4. Dellinger EP, Forsmark CE, Layer P, Lèvy P, Maravi-Poma E, Petrov MS, et al. Determinant-based classification of acute pancreatitis severity: an international multidisciplinary consultation. *Ann Surg*. 2012;256(6):875–80. <https://doi.org/10.1097/SLA.0b013e318256f778>
 5. Ranson JH, Rifkind KM, Roses DF, Fink SD, Eng K, Localio SA. Objective early identification of severe acute pancreatitis. *Am J Gastroenterol*. 1974;61(6):443–51.
 6. Wu BU, Johannes RS, Sun X, Tabak Y, Conwell DL, Banks PA. The early prediction of mortality in acute pancreatitis: a large population-based study. *Gut*. 2008;57(12):1698–703. <https://doi.org/10.1136/gut.2008.152702>
 7. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE-II: a severity of disease classification system. *Crit Care Med*. 1985;13(10):818–29. <https://doi.org/10.1097/00003246-198510000-00009>
 8. Mofidi R, Duff MD, Wigmore SJ, Madhavan KK, Garden OJ, Parks RW. Association between early systemic inflammatory response, severity of multiorgan dysfunction and death in acute pancreatitis. *Br J Surg*. 2006;93(6):738–44. <https://doi.org/10.1002/bjs.5290>
 9. Mounzer R, Langmead CJ, Wu BU, Evans AC, Bishehsari F, Muddana V, et al. Comparison of existing clinical scoring systems to predict persistent organ failure in patients with acute pancreatitis. *Gastroenterology*. 2012;142(7):1476–82. <https://doi.org/10.1053/j.gastro.2012.03.005>
 10. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*. 1997;15;16(9):981–91. [https://doi.org/10.1002/\(sici\)1097-0258\(19970515\)16:9<981::aid-sim510>3.0.co;2-n](https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<981::aid-sim510>3.0.co;2-n)
 11. Fagan TJ. Letter: nomogram for Bayes theorem. *N Engl J Med*. 1975;293(5):257–8. <https://doi.org/10.1056/NEJM197507312930513>
 12. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Pettycrow M, et al. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: elaboration and explanation. *BMJ*. 2015;350(jan02 1):g7647. <https://doi.org/10.1136/bmj.g7647>
 13. 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. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *JAMA*. 2000;283(15):2008–12. <https://doi.org/10.1001/jama.283.15.2008>
 14. Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Ann Intern Med*. 2006;144(6):427–37. <https://doi.org/10.1001/jama.283.15.2008>
 15. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539–58. <https://doi.org/10.1002/sim.1186>
 16. Egger M, Smith GD, Schneider M, Minder C. Papers bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629–34. <https://doi.org/10.1136/bmj.315.7109.629>
 17. De Bernardinis M, Violi V, Roncoroni L, Boselli AS, Giunta A, Peracchia A. Discriminant power and information content of Ranson's prognostic signs in acute pancreatitis: a meta-analytic study. *Crit Care Med*. 1999;27(10):2272–83. <https://doi.org/10.1097/00003246-199910000-00035>
 18. Gao W, Yang HX, Ma CE. The value of BISAP score for predicting mortality and severity in acute pancreatitis: a systematic review and meta-analysis. *PLoS One*. 2015;10(6):10. <https://doi.org/10.1371/journal.pone.0130412>
 19. 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. *Med Princ Pract*. 2016;25(2):137–42. <https://doi.org/10.1159/000441003>
 20. Chandra S, Murali A, Bansal R, Agarwal D, Holm A. The Bedside Index for Severity in Acute Pancreatitis: a systematic review of prospective studies to determine predictive performance. *J Community Hosp Intern Med Perspect*. 2017;7(4):208–13. <https://doi.org/10.1080/20009666.2017.1361292>
 21. Maisonneuve P, Lowenfels AB, Lankisch PG. The harmless acute pancreatitis score (HAPS) identifies non-severe patients: a systematic review and meta-analysis. *Pancreatology*. 2021;21(8):1419–27. <https://doi.org/10.1016/j.pan.2021.09.017>
 22. Mikó A, Vigh É, Mátrai P, Soós A, Garami A, Balaskó M, et al. Computed tomography severity index vs. Other indices in the prediction of severity and mortality in acute pancreatitis: a predictive accuracy meta-analysis. *Front Physiol*. 2019;10:1002. <https://doi.org/10.3389/fphys.2019.01002>
 23. Di MY, Liu H, Yang ZY, Bonis PAL, Tang JL, Lau J. Prediction models of mortality in acute pancreatitis in adults: a systematic review. *Ann Intern Med*. 2016;165(7):482–90. <https://doi.org/10.7326/M16-0650>
 24. Cho YS, Kim HK, Jang EC, Yeom JO, Kim SY, Yu JY, et al. Usefulness of the bedside index for severity in acute pancreatitis in the early prediction of severity and mortality in acute pancreatitis. *Pancreas*. 2013;42(3):483–7. <https://doi.org/10.1097/MPA.0b013e318267c879>
 25. Khanna AK, Meher S, Prakash S, Tiwary SK, Singh U, Srivastava A, et al. Comparison of Ranson, Glasgow, MOSS, SIRS, BISAP, APACHE-II, CTSI Scores, IL-6, CRP, and procalcitonin in predicting severity, organ failure, pancreatic necrosis, and mortality in acute pancreatitis. *HPB Surg*. 2013;2013:367581–610. <https://doi.org/10.1155/2013/367581>
 26. Park JY, Jeon TJ, Ha TH, Hwang JT, Sinn DH, Oh TH, et al. Bedside index for severity in acute pancreatitis: comparison with other scoring systems in predicting severity and organ failure. *Hepatobiliary Pancreat Dis Int*. 2013;12(6):645–50. [https://doi.org/10.1016/s1499-3872\(13\)60101-0](https://doi.org/10.1016/s1499-3872(13)60101-0)
 27. Zhang J, Shahbaz M, Fang R, Liang B, Gao C, Gao H, et al. Comparison of the BISAP scores for predicting the severity of acute pancreatitis in Chinese patients according to the latest Atlanta classification. *J Hepatobiliary Pancreat Sci*. 2014;21(9):689–94. <https://doi.org/10.1002/jhpb.118>
 28. Cho JH, Kim TN, Chung HH, Kim KH. Comparison of scoring systems in predicting the severity of acute pancreatitis. *World J Gastroenterol*. 2015;21(9):2387–94. <https://doi.org/10.3748/wjg.v21.i8.2387>
 29. Mok SRS, Mohan S, Elfant AB, Judge TA. The acute physiology and chronic health evaluation IV, a new scoring system for predicting mortality and complications of severe acute pancreatitis. *Pancreas*. 2015;44(8):1314–9. <https://doi.org/10.1097/MPA.0000000000000432>
 30. Qiu L, Sun RQ, Jia RR, Ma XY, Cheng L, Tang MC, et al. Comparison of existing clinical scoring systems in predicting severity and prognoses of hyperlipidemic acute pancreatitis in Chinese patients: a retrospective study. *Med (United States)*. 2015;94(23):e957. <https://doi.org/10.1097/MD.0000000000000957>
 31. Sharma V, Rana SS, Sharma RK, Kang M, Gupta R, Bhasin DK. A study of radiological scoring system evaluating extrapancreatic inflammation with conventional radiological and clinical scores in predicting outcomes in acute pancreatitis. *Ann Gastroenterol*. 2015;28(3):399–404.

32. Yadav J, Yadav SK, Kumar S, Baxla RG, Sinha DK, Bodra P, et al. Predicting morbidity and mortality in acute pancreatitis in an Indian population: a comparative study of the BISAP score, Ranson's score and CT severity index. *Gastroenterol Rep*. 2016;4(3):216–20. <https://doi.org/10.1093/gastro/gov009>
33. Kumar AH, Griwan MS. A comparison of APACHE-II, BISAP, Ranson's score and modified CTSI in predicting the severity of acute pancreatitis based on the 2012 revised Atlanta Classification. *Gastroenterol Rep*. 2018;6(2):127–31. <https://doi.org/10.1093/gastro/gox029>
34. He WH, Zhu Y, Zhu Y, Jin Q, Xu HR, Xion ZJ, et al. Comparison of multifactor scoring systems and single serum markers for the early prediction of the severity of acute pancreatitis. *J Gastroenterol Hepatol*. 2017;32(11):1895–901. <https://doi.org/10.1111/jgh.13803>
35. Shi Y, Liu Y, Liu YQ, Gao F, Li JH, Li QJ, et al. Early diagnosis and severity assessment of acute pancreatitis (AP) using MR elastography (MRE) with spin-echo echo-planar imaging. *J Magn Reson Imag*. 2017;46(5):1311–9. <https://doi.org/10.1002/jmri.25679>
36. Valverde-López F, Matas-Cobos AM, Alegria-Motte C, Jiménez-Rosales R, Úbeda-Muñoz M, Redondo-Cerezo E. BISAP, RANSON, lactate and others biomarkers in prediction of severe acute pancreatitis in a European cohort. *J Gastroenterol Hepatol*. 2017;32(9):1649–56. <https://doi.org/10.1111/jgh.13763>
37. Choi HW, Park HJ, Choi SY, Do JH, Yoon NY, Ko A, et al. Early prediction of the severity of acute pancreatitis using radiologic and clinical scoring systems with classification tree analysis. *Am J Roentgenol*. 2018;211(5):1035–43. <https://doi.org/10.2214/AJR.18.19545>
38. de-Madaria E, Molero X, Bonjoch L, Casas J, Cardenas-Jaen K, Montenegro A, et al. Oleic acid chlorohydrin, a new early biomarker for the prediction of acute pancreatitis severity in humans. *Ann Intensive Care*. 2018;8(1):1. <https://doi.org/10.1186/s13613-017-0346-6>
39. Fei Y, Gao K, Tu J, Wang W, Zong GQ, Li WQ. Predicting and evaluation the severity in acute pancreatitis using a new modeling built on body mass index and intra-abdominal pressure. *Am J Surg*. 2018;216(2):304–9. <https://doi.org/10.1016/j.amjsurg.2017.04.017>
40. Gravito-Soares M, Gravito-Soares E, Gomes D, Almeida N, Tomé L. Red cell distribution width and red cell distribution width to total serum calcium ratio as major predictors of severity and mortality in acute pancreatitis. *BMC Gastroenterol*. 2018;18(1):108. <https://doi.org/10.1186/s12876-018-0834-7>
41. Hagjer S, Kumar N. Evaluation of the BISAP scoring system in prognostication of acute pancreatitis – a prospective observational study. *Int J Surg*. 2018;54(Pt A):76–81. <https://doi.org/10.1016/j.ijss.2018.04.026>
42. Yang WQ, Yang Q, Chen WJ, Zhang XB, Xu QQ, Qiao Y, et al. Low FT3 is a valuable predictor of severe acute pancreatitis in the emergency department. *J Dig Dis*. 2018;19(7):431–8. <https://doi.org/10.1111/1751-2980.12609>
43. Arif A, Jaleel F, Rashid K. Accuracy of BISAP score in prediction of severe acute pancreatitis. *Pak J Med Sci*. 2019;35(4):1008–12. <https://doi.org/10.12669/pjms.35.4.1286>
44. Chen J, Wan J, Shu W, Yang X, Xia L. Association of serum levels of silent information regulator 1 with persistent organ failure in acute pancreatitis. *Dig Dis Sci*. 2019;64(11):3173–81. <https://doi.org/10.1007/s10620-019-05647-x>
45. Jain D, Bhaduri G, Jain P. Different scoring systems in acute alcoholic pancreatitis: which one to follow? an ongoing dilemma. *Arq Gastroenterol*. 2019;56(3):280–5. <https://doi.org/10.1590/S0004-2803.201900000-53>
46. Zhou H, Mei X, He X, Lan T, Guo S. Severity stratification and prognostic prediction of patients with acute pancreatitis at early phase. *Med (United States)*. 2019;98(16):e15275. <https://doi.org/10.1097/MD.00000000000015275>
47. Chatterjee R, Parab N, Sajjan B, Nagar VS. Comparison of acute physiology and chronic health evaluation ii, modified computed tomography severity index, and bedside index for severity in acute pancreatitis score in predicting the severity of acute pancreatitis. *Indian J Crit Care Med*. 2020;24(2):99–103. <https://doi.org/10.5005/jp-journals-10071-23343>
48. Gezer NS, Bengi G, Baran A, Erkmn PE, Topalak OS, Altay C, et al. Comparison of radiological scoring systems, clinical scores, neutrophil-lymphocyte ratio and serum C-reactive protein level for severity and mortality in acute pancreatitis. *Rev Assoc Med Bras*. 2020;66(6):762–70. <https://doi.org/10.1590/1806-9282.66.6.762>
49. Li M, Xing XK, Lu ZH, Guo F, Su W, Lin YJ, et al. Comparison of scoring systems in predicting severity and prognosis of hypertriglyceridemia-induced acute pancreatitis. *Dig Dis Sci*. 2020;65(4):1206–11. <https://doi.org/10.1007/s10620-019-05827-9>
50. Li Y, Zhang J, Zou J. Evaluation of four scoring systems in prognostication of acute pancreatitis for elderly patients. *BMC Gastroenterol*. 2020;20(1):165. <https://doi.org/10.1186/s12876-020-01318-8>
51. Peng R, Zhang L, Zhang ZM, Wang ZQ, Liu GY, Zhang XM. Chest computed tomography semi-quantitative pleural effusion and pulmonary consolidation are early predictors of acute pancreatitis severity. *Quant Imaging Med Surg*. 2020;10(2):451–63. <https://doi.org/10.21037/qims.2019.12.14>
52. Satiş H, Kayahan N, Sargin ZG, Karataş A, Çeliker D. Evaluation of the clinical course and prognostic indices of acute pancreatitis in elderly patients: a prospective study. *Acta Gastroenterol Bel*. 2020;83(3):413–7.
53. Silva-Vaz P, Abrantes AM, Morgado-Nunes S, Castelo-Branco M, Gouveia A, Botelho MF, et al. Evaluation of prognostic factors of severity in acute biliary pancreatitis. *Int J Mol Sci*. 2020;21(11):1–18. <https://doi.org/10.3390/ijms21124300>
54. Venkatesh NR, Vijayakumar C, Balasubramanian G, Kandhasami SC, Sundaramurthi S, Sreenath SS, et al. Comparison of different scoring systems in predicting the severity of acute pancreatitis: a prospective observational study. *Cureus*. 2020;12(2):e6943. <https://doi.org/10.7759/cureus.6943>
55. Zhou T, Chen Y, Wu JL, Deng Y, Zhang J, Sun H, et al. Extrap-creatic inflammation on magnetic resonance imaging for the early prediction of acute pancreatitis severity. *Pancreas*. 2020;49(1):46–52. <https://doi.org/10.1097/MPA.0000000000001425>
56. Sun HW, Lu JY, Weng YX, Chen H, He QY, Liu R, et al. Accurate prediction of acute pancreatitis severity with integrative blood molecular measurements. *Aging*. 2021;13(6):8817–34. <https://doi.org/10.18632/aging.202689>
57. Pando E, Alberti P, Mata R, Gomez MJ, Vidal L, Cirera A, et al. Early changes in Blood Urea Nitrogen (BUN) can predict mortality in acute pancreatitis: comparative study between BISAP score, APACHE-II, and other laboratory markers-A prospective observational study. *Can J Gastroenterol Hepatol*. 2021;2021:6643595–8. <https://doi.org/10.1155/2021/6643595>
58. Wu Q, Wang J, Qin M, Yang H, Liang Z, Tang G. Accuracy of conventional and novel scoring systems in predicting severity and outcomes of acute pancreatitis: a retrospective study. *Lipids Health Dis*. 2021;20(1):41. <https://doi.org/10.1186/s12944-021-01470-4>
59. Shen D, Tang C, Zhu S, Huang G. Macrophage migration inhibitory factor is an early marker of severe acute pancreatitis based on the revised Atlanta classification. *BMC Gastroenterol*. 2021;21(1):34. <https://doi.org/10.1186/s12876-020-01598-0>
60. Teng TZJ, Tan JKT, Baey S, Gunasekaran SK, Junnarkar SP, Low JK, et al. Sequential organ failure assessment score is superior to other prognostic indices in acute pancreatitis. *World J Crit Care Med*. 2021;10(6):355–68. <https://doi.org/10.5492/wjccm.v10.i6.355>

61. Wang Y, Xu Z, Zhou Y, Xie M, Qi X, Xu Z, et al. Leukocyte cell population data from the blood cell analyzer as a predictive marker for severity of acute pancreatitis. *J Clin Lab Anal.* 2021;35(7): e23863. <https://doi.org/10.1002/jcla.23863>
62. Yan G, Li H, Bhetuwal A, McClure MA, Li Y, Yang G, et al. Pleural effusion volume in patients with acute pancreatitis: a retrospective study from three acute pancreatitis centers. *Ann Med.* 2021; 53(1):2003–18. <https://doi.org/10.1080/07853890.2021.1998594>
63. Dancu GM, Popescu A, Sirli R, Danila M, Bende F, Tarta C, et al. The BISAP score, NLR, CRP, or BUN: which marker best predicts the outcome of acute pancreatitis? *Med (United States).* 2021;100(51): E28121. <https://doi.org/10.1097/MD.00000000000028121>
64. Bardakçı O, Akdur G, Das M, Siddikoğlu D, Akdur O, Beyazit Y. Comparison of different risk stratification systems for prediction of acute pancreatitis severity in patients referred to the emergency department of a tertiary care hospital. *Ulusal Travma ve Acil Cerrahi Dergisi.* 2022;28(7):967–73. <https://doi.org/10.14744/tjtes.2021.51892>
65. Wu B, Yang J, Dai Y, Xiong L. Combination of the BISAP score and miR-155 is applied in predicting the severity of acute pancreatitis. *Int J Gen Med.* 2022;15:7467–74. <https://doi.org/10.2147/IJGM.S384068>
66. Windsor JA. Assessment of the severity of acute pancreatitis: No room for complacency. *Pancreatology.* 2008;8(2):105–9. <https://doi.org/10.1159/000123604>
67. Kamal A, Akshintala VS, Kamal MM, El Zein M, Besharati S, Kumbhari V, et al. Does etiology of pancreatitis matter? Differences in outcomes among patients with post-endoscopic retrograde cholangiopancreatography, acute biliary, and alcoholic pancreatitis. *Pancreas.* 2019;48(4):574–8. <https://doi.org/10.1097/MPA.0000000000001283>
68. Szakács Z, Gede N, Pécsi D, Izbéki F, Papp M, Kovács G, et al. Aging and comorbidities in acute pancreatitis II.: a cohort-analysis of 1203 prospectively collected cases. *Front Physiol.* 2019;9:1776. <https://doi.org/10.3389/fphys.2018.01776>
69. Nagy A, Juhász M, Gorbe A, Váradi A, Izbéki F, Vincze A, et al. Glucose levels show independent and dose-dependent association with worsening acute pancreatitis outcomes: post-hoc analysis of a prospective, international cohort of 2250 acute pancreatitis cases. *Pancreatology.* 2021;21(7):1237–46. <https://doi.org/10.1016/j.pan.2021.06.003>
70. Murata A, Matsuda S, Mayumi T, Yokoe M, Kuwabara K, Ichimiya Y, et al. Effect of hospital volume on clinical outcome in patients with acute pancreatitis, based on a National Administrative Database. *Pancreas.* 2011;40(7):1018–23. <https://doi.org/10.1097/MPA.0b013e31821bd233>
71. Buxbaum J, Quezada M, Chong B, Gupta N, Yao YC, Lane C, et al. The Pancreatitis Activity Scoring System predicts clinical outcomes in acute pancreatitis: findings from a prospective cohort study. *Am J Gastroenterol.* 2018;113(5):755–64. <https://doi.org/10.1038/s41395-018-0048-1>
72. Langmead C, Lee P, Paragomi P, Greer P, Stello K, Hart PA, et al. A novel 5-cytokine panel outperforms conventional predictive markers of persistent organ failure in acute pancreatitis. *Clin Transl Gastroenterol.* 2021;12(5):e00351. <https://doi.org/10.14309/ctg.0000000000000351>
73. Kui B, Pintér J, Molontay R, Nagy M, Farkas N, Gede N, et al. EASY-APP: an artificial intelligence model and application for early and easy prediction of severity in acute pancreatitis. *Clin Transl Med.* 2022;12(6):e842. <https://doi.org/10.1002/ctm2.842>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Capurso G, Ponz de Leon Pisani R, Lauri G, Archibugi L, Hegyi P, Papachristou GI, et al. Clinical usefulness of scoring systems to predict severe acute pancreatitis: a systematic review and meta-analysis with pre and post-test probability assessment. *United European Gastroenterol J.* 2023;11(9):825–36. <https://doi.org/10.1002/ueg2.12464>