

Early laboratory biomarkers for severity in acute pancreatitis; A systematic review and meta-analysis

Fons F. van den Berg^{a, *}, Anna C. de Bruijn^a, Hjalmar C. van Santvoort^{b, c}, Yama Issa^a, Marja A. Boermeester^a

^a Department of Surgery, Amsterdam UMC, University of Amsterdam, Amsterdam Gastroenterology Endocrinology Metabolism, the Netherlands

^b Department of Surgery, St. Antonius Hospital, the Netherlands

^c Department of Surgery, University Medical Center Utrecht, the Netherlands

ARTICLE INFO

Article history:

Received 20 June 2020

Received in revised form

25 August 2020

Accepted 5 September 2020

Available online 8 September 2020

Keywords:

Prediction studies

Acute pancreatitis

Biomarkers

ABSTRACT

Background/Objectives: Acute pancreatitis is complicated by local and systemic complications in 20–30% of the patients. Accurate prediction of severity may be important for clinical decision making. Our aim is to identify and compare the accuracy of laboratory biomarkers that predict severity and complications in adult patients.

Methods: Medline, EMBASE, Web of Science and Cochrane Library (1993 to August 2020) were searched for studies with an unselected population of patients with acute pancreatitis, that contains accuracy data for ≥ 1 laboratory biomarker(s) and/or APACHE-II score for the prediction of a patient outcomes of interest during the first 48 h of admission. The primary outcome is moderate severe or severe acute pancreatitis (MSAP/SAP). Secondary outcomes are severe acute pancreatitis, pancreatic necrosis and organ failure. Risk of bias was assessed using QUADAS-2. Biomarkers extracted from ≥ 3 unique sources, were analyzed using hierarchical summary receiver operating characteristic (HSROC) and bivariate model analysis.

Results: In total, 181 studies were included in the qualitative analysis reporting on 29 biomarkers. For the primary outcome at admission, summary sensitivities and specificities were, respectively, 87% (95% CI 69–95%) and 88% (95% CI 80–93%) for IL-6 at a threshold of >50 pg/ml, 72% (95% CI 64–79%) and 76% (95% CI 67–84%) for an APACHE-II score of ≥ 8 , and 53% (95% CI 35–71%) and 82% (95% CI 74–88%) for CRP >150 mg/l. HSROC curve analysis confirmed these results.

Conclusion: This study indicates superiority of IL-6 for the early prediction of MSAP/SAP and may be used for to guide clinical decision making.

© 2020 IAP and EPC. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Introduction

Acute pancreatitis is a disease that usually has a mild disease course. However, around 20–30% of the patients develop severe complications [1]. Persistent organ failure, with or without the presence of local complications such as (peri-) pancreatic necrosis and secondary infections, is the main determinant for mortality, and these patients are classified as having severe acute pancreatitis according to the revised Atlanta classification [2–4].

Current guidelines on the management of acute pancreatitis

report different score systems for risk stratification [5,6]. For example, guidelines from the International Association of Pancreatology/American Pancreatic Association (IAP/APA) currently recommends using the systemic inflammatory response syndrome (SIRS) criteria to predict severe acute pancreatitis at admission and at 48 h, but it is recognizes that it is unclear which predictor or scoring system is superior for the prediction of severity, complications or mortality [5].

The goal of our systematic review is to systematically identify and analyze predictive laboratory biomarkers to compare prognostic accuracy to predictors that are commonly used in daily practice, such as serum C-reactive protein and the Acute Physiology and Chronic Health Evaluation (APACHE)-II score.

* Corresponding author. Meibergdreef 9, 1105AZ, Amsterdam, PO 22660, 1100, DD Amsterdam, the Netherlands.

E-mail address: f.f.vandenberg@amsterdamumc.nl (F.F. van den Berg).

Methods

The protocol has been registered at the PROSPERO database (CRD42018087157). This review is concordant with the Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) statement [7].

Data sources and searches

We performed literature searches in the online databases MEDLINE (Pubmed), EMBASE (Ovid), Web of Science (Ovid) and Cochrane Library on August 12, 2020. The search strategy was based on the disease of interest (acute pancreatitis), keywords for biomarkers, predictors or prognosis, and outcomes of interest (severity, organ failure and necrosis). No restriction on language was applied. The full search strategies are presented in **Methods** in [Supporting information](#). Reference lists of included studies were manually searched to identify additional studies.

Study selection

The target population consists of patients with acute pancreatitis that were admitted to a hospital. Studies that enrolled a selected population of patients, for example only patients with necrotizing or predicted severe acute pancreatitis, were not eligible for inclusion. Retrospective, prospective and cross-sectional studies, as well as clinical trials that evaluated the predictive value of a biomarker for the clinical outcome measures were eligible for inclusion. Accuracy data (true positives, false negative, false positive, true negatives) had to be available, or be able to be calculated based on the reported sensitivities and specificities. Severity of the episode of acute pancreatitis had to be defined by (or in line with) the original [1] or revised Atlanta classification [4]. The primary outcome of interest was the prediction of moderate severe (MSAP) and severe acute pancreatitis (SAP) as described by the revised Atlanta classification. It is defined as the presence of local or systematic complications, which is consistent with the definition of severe acute pancreatitis as described by the original Atlanta classification. Secondary outcomes were SAP, pancreatic necrosis and organ failure. Persistent organ failure is defined as renal, cardiovascular or pulmonary failure that exists longer than 48 h. We deemed any definition of renal, pulmonary or cardiovascular failure described by the original authors appropriate. Pancreatic necrosis had to be assessed by either imaging, biopsy, fine needle aspiration, peri-operative assessment or autopsy. Case reports, case series, book chapters, letters, abstracts and reviews were excluded, as were animal and in-vitro studies. Studies that preceded the publication of the Atlanta classification (i.e. before the year 1993) were excluded. Studies that reported on imaging-based markers were outside the scope of this review. Genetic biomarkers are covered elsewhere [8].

Studies were first screened in duplicate (F.B. and A.B.) for eligibility based on the title and abstract, and subsequently based on full-text using Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia. www.covidence.org).

Data extraction and quality assessment

Data was extracted by two reviewers (F.B. and A.B.) independently with a data extraction sheet. Disagreements were resolved by discussion between the reviewers. The quality of the studies and markers included in the qualitative meta-analysis were assessed using piloted criteria adapted from the QUADAS-2 tool ([Table S1](#) in [Supporting information](#)).

Data synthesis and analysis

For all studies that reported on total participants, number of participants with the outcome of interest, sensitivity and specificity, a 2×2 contingency table was constructed. Data for timepoints of sample collection were pooled; day 1 represents timepoints from admission until 24 h after admission, or when the articles state “day 0” or “day 1”. Day 2 represents timepoints from 24 until 48 after admission, or when described as “day 2” by the authors. When there are multiple timepoints within a time period, for day 1, the timepoint closest to admission was selected for analysis. For day 2, the timepoint closest to 48 h after admission was used. A meta-analysis was done only when there were at least three unique published sources for a biomarker at a pre-specified timepoint (day 1 or day 2).

We first constructed hierarchical summary receiver-operating characteristic (HSROC) curves that allows summarization of studies that use different cut-off values [9]. When multiple cut-offs for the same study were given, the most commonly used cut-off value was used, or otherwise the most median cut-off value. Descriptive forest plots were also constructed. Next, summary accuracy estimates (summary sensitivity, specificity and accuracy) were calculated using the bivariate random effects model for a common cut-off value [10]. We pre-specified the thresholds for CRP and APACHE at >150 mg/l and ≥ 8 , respectively. These are thresholds that are most often used in the clinical settings [11]. For the other biomarkers, we aimed to identify a common threshold; studies that deviated less than 10% of the identified threshold value were pooled. Summary likelihood ratios and diagnostic odds ratios were calculated from the bivariate model summary estimates using the Markov Chain Monte Carlo method [12]. Heterogeneity analysis for study design (prospective studies), and hospital setting (tertiary and multicenter studies) was done using the HSROC model; sensitivity analyses for risk of bias assessment, and for studies that included patients within 72 h of disease onset were planned but not performed, because of insufficient remaining studies. Analyses were performed using the package “MADA” [13] (<https://CRAN.R-project.org/package=mada>), which enables both the HSROC and bivariate analysis, with R statistical software (version 3.5.1. R Core team, 2014). HSROC curves and forest plots were visualized using Review Manager (Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).

Results

Study selection

In total, 8471 unique records were screened, of which 950 were evaluated for eligibility in full-text ([Fig. 1](#)). We extracted data from 309 studies; 128 studies did not provide sufficient accuracy data to construct 2×2 contingency tables, or did not report on a biomarker by at least three unique data sources. So finally, we included 181 studies in our meta-analysis.

Study characteristics

The characteristics of the included studies are shown in [Table S2](#) in [Supporting information](#). The 181 included studies contained 33,640 patients with a median of 117 patients per study (IQR 135). Consisting of 134 prospective, 45 retrospective and 2 case-control studies. Most studies ($n = 164$) were single-center studies performed in secondary (9.0%) or tertiary (83.1%) centers, while 7.9% were performed in a multi-center setting. 73 studies reported on the recruitment department, in which patients were most frequently recruited from departments of surgery and

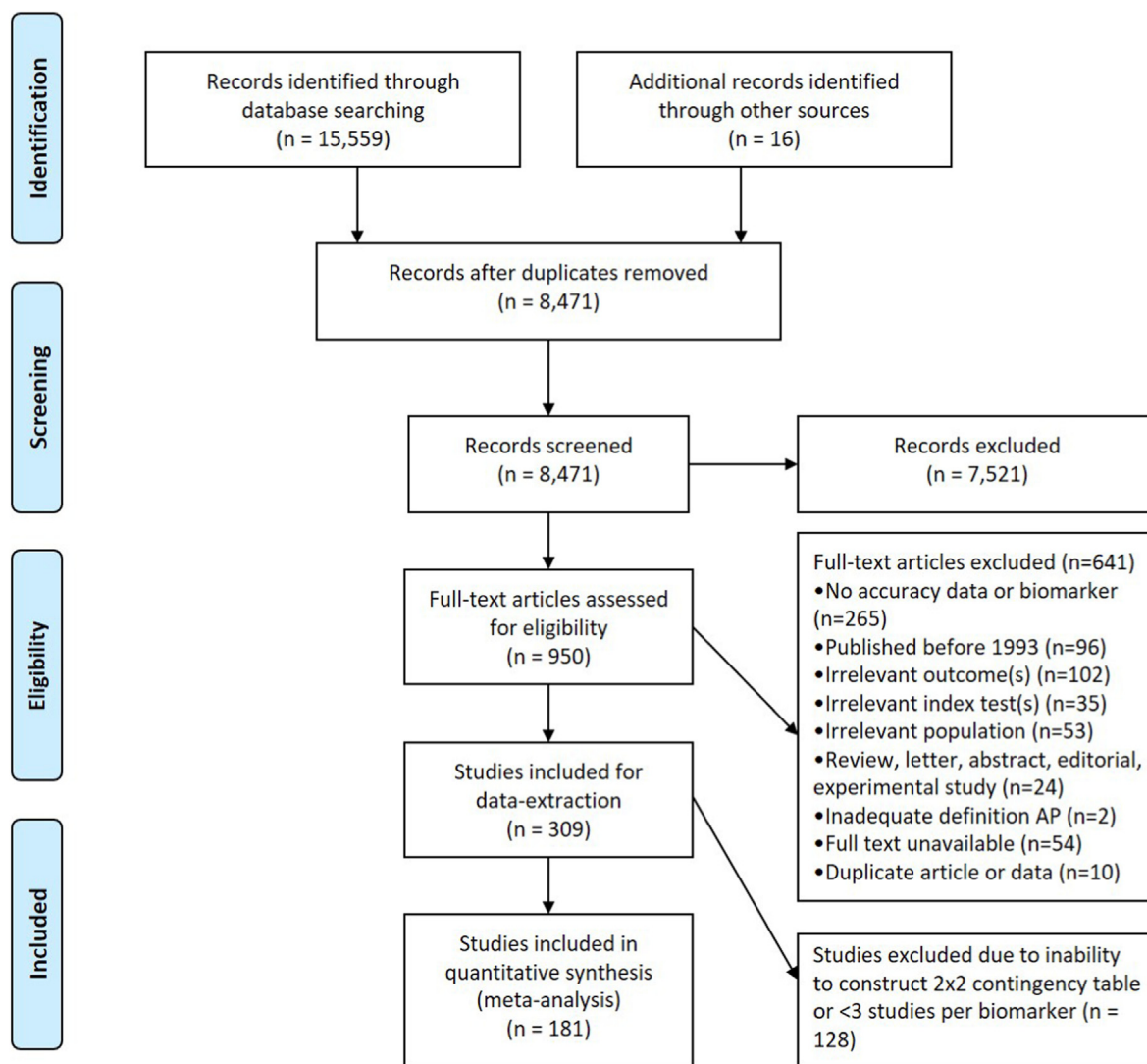


Fig. 1. PRISMA flow chart.

gastroenterology (63%). Accuracy data was reported in 111 studies for the primary outcome, moderate severe or severe acute pancreatitis (MSAP/SAP), in 61 studies for severe acute pancreatitis (SAP), in 25 studies for pancreatic necrosis (PNEC) and in 17 studies for organ failure (OF). Some heterogeneity existed in the prevalence of the outcomes; median prevalence of MSAP/SAP, SAP, PNEC and OF were 30% (IQR 21), 17% (IQR 14), 20% (IQR 16) and 25% (IQR 16), respectively. A summary of the included biomarkers (twenty-six serum and three urine markers) for meta-analysis are shown in Table S3 in Supporting information.

Quality/risk of bias assessment

None of the 181 studies that were included in the pooled analysis had a low risk of bias in all four domains of the QUADAS-2 assessment (Fig. S1 in Supporting information). Although the majority of studies did not report sufficient details (107 studies, 59.1%), only 31 studies (17.1%) included consecutive or random subjects, and avoided a case-control design and inappropriate exclusions. Although recommended by the guidelines [14,15], only 38.1% (69 studies) of the studies made use of pre-defined thresholds to calculate prognostic accuracy (sensitivity and specificity).

Also, few studies reported sufficient information on blinding of the outcome during interpretation of the index test and vice versa.

Moderate severe or severe AP

Primary analysis

Twenty-one biomarkers were appropriate for HSROC analysis, of which C-reactive protein (CRP) and APACHE-II were the most described predictors (41 and 42 studies, respectively), followed by interleukin-6 (IL-6, 18 studies) and procalcitonin (PCT, 16 studies). Polymorphonuclear elastase (PMN-E) was described in 5 studies. Additionally, eight biomarkers with a comparable cut-off were appropriate for bivariate model analysis (Table 1).

Summary sensitivity and specificity of IL-6 at a cut-off of 50 ng/ml (4 studies), at day 1, were 87% (95% CI, 69–95%) and 88% (95% CI, 80–93%), respectively, with a diagnostic odds ratio (DOR) of 66 (95% CI, 11.4 to 234) (Fig. 2, Fig. S2 in Supporting information). Summary accuracy measures of PMN-E at a threshold of 100 µg/l were 85% (95% CI, 65–94%), 81% (95% CI, 60–93%), and 43.9 (95% CI 3.01 to 187), for sensitivity, specificity, and DOR respectively. In comparison, summary sensitivity and specificity of commonly used markers were 53% (95% CI, 35–71%) and 82% (95% CI, 74–88%),

Table 1
Bivariate model for prognostic values.

Biomarker	No. studies	Weighted mean prevalence, %	AUC Summary sensitivity % (95% CI)	Summary specificity % (95% CI)	Summary accuracy % (95% CI)	Mean (95% CI) LR+	Mean (95% CI) LR-	Mean (95% CI) DOR
<i>severity (MSAP/SAP)</i>								
IL-6 day 1 > 50 mg/l	4	30.2	0.93 87 (69–95)	88 (80–93)	88 (77–94)	7.51 (3.9–13.6)	0.17 (0.05–0.37)	66 (11.4–234)
PMN-E day 1 > 100 µg/l	3	30.5	0.90 85 (65–94)	81 (60–93)	82 (62–93)	5.24 (1.7–12.3)	0.22 (0.06–0.56)	43.9 (3.01–187)
PCT day 1 > 0.5 ng/ml	9	37.7	0.81 75 (50–90)	76 (60–86)	76 (56–88)	3.11 (1.79–5.28)	0.35 (0.14–0.68)	10.8 (2.85–27.9)
APACHE-II day 1 ≥ 8	20	29.6	0.80 72 (64–79)	77 (67–84)	76 (66–83)	3.14 (2.13–4.51)	0.37 (0.27–0.48)	8.9 (4.63–16)
CRP day 2 > 150 mg/l	13	25.5	0.80 74 (66–80)	74 (68–79)	74 (67–79)	2.83 (2.24–3.61)	0.36 (0.27–0.47)	8.1 (4.93–13.2)
APACHE-II day 2 ≥ 8	5	37.3	0.79 76 (66–84)	70 (54–82)	72 (58–83)	2.62 (1.57–4.31)	0.36 (0.22–0.54)	8.0 (3.10–16.9)
uTAP day 1 > 35 mmol/l	4	19.9	0.79 64 (48–77)	77 (73–82)	74 (68–81)	2.83 (2.11–3.54)	0.47 (0.31–0.67)	6.41 (3.17–10.7)
NLR day 1 > 10	5	37.2	0.79 79 (73–83)	60 (46–72)	67 (56–76)	1.99 (1.48–2.73)	0.36 (0.28–0.47)	5.63 (3.22–8.81)
CRP day 1 > 150 mg/l	8	24.3	0.79 53 (35–71)	82 (74–88)	75 (65–84)	2.89 (2.29–3.63)	0.57 (0.38–0.75)	5.25 (3.27–8.36)
HCT day 1 > 40%–44%	6	27.3	0.52 47 (38–56)	63 (49–75)	59 (46–70)	1.3 (0.99–1.76)	0.85 (0.71–1.02)	1.57 (0.97–2.4)
<i>severity (SAP)</i>								
CAL day 1 > 1.9 mg/dl	3	21.4	0.87 74 (52–88)	85 (71–93)	83 (67–92)	5.02 (2.97–7.69)	0.31 (0.16–0.5)	16.5 (11.5–22.5)
ALB day 1 < 30–33 mg/l	5	18.1	0.83 71 (55–83)	81 (71–88)	79 (68–87)	3.79 (2.8–4.87)	0.36 (0.24–0.5)	10.7 (8–13.8)
CRP day 1 > 150 mg/l	4	26.5	0.81 69 (51–83)	81 (64–91)	78 (61–89)	3.71 (1.97–6.82)	0.4 (0.22–0.6)	10.3 (3.68–22.1)
CRP day 2 > 150 mg/l	3	10.2	0.86 87 (78–93)	60 (49–70)	63 (52–72)	2.18 (1.69–2.85)	0.23 (0.13–0.38)	10.1 (4.92–19.1)
WBC day 1 14	3	9.7	0.8 64 (32–87)	79 (75–83)	78 (71–83)	2.96 (1.67–3.9)	0.47 (0.18–0.85)	7.96 (1.97–21.5)
APACHE-II day 1 ≥ 8	12	15.2	0.79 74 (61–84)	72 (63–79)	72 (63–80)	2.63 (2.19–3.08)	0.37 (0.25–0.51)	7.4 (5.18–9.96)
CRT day 1 > 70 µmol/l	3	19.2	0.65 66 (60–71)	75 (61–85)	73 (61–85)	2.75 (1.69–4.48)	0.46 (0.37–0.58)	6.22 (2.9–11.7)
HCT day 1 > 40%–44%	6	16.2	0.65 48 (38–58)	74 (66–81)	70 (61–77)	1.87 (1.11–2.96)	0.72 (0.53–0.94)	2.77 (1.18–5.61)
<i>organ failure</i>								
APACHE-II day 1 ≥ 8	9	18.3	0.86 80 (65–89)	79 (65–89)	79 (68–87)	4.03 (1.98–7.2)	0.27 (0.13–0.49)	18 (14.8–49.5)
<i>pancreatic necrosis</i>								
HCT day 1 > 44%	6	18.6	0.65 63 (55–69)	69 (51–83)	68 (52–80)	2.11 (1.3–3.52)	0.55 (0.44–0.73)	3.97 (1.8–7.72)
APACHE-II day 1 ≥ 8	8	18.0	0.62 55 (37–72)	63 (49–76)	62 (47–75)	1.51 (1.21–1.9)	0.71 (0.54–0.89)	2.18 (1.38–3.30)

AUC = area under curve; CI = confidence interval; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; DOR = diagnostic odds ratio; IL-6 = Interleukin-6; CAL = calcium; CRP = C-reactive protein; PCT = procalcitonin; HCT = hematocrit; BUN = blood urea nitrogen; ALB = albumin; WBC = white blood cell count; CRT = creatinine.

respectively, for CRP at a threshold of 150 mg/l (8 studies), and 72% (95% 64–79%) and 77% (95% CI, 67–84%), respectively, for APACHE-II (8 or higher) at day 1 (20 studies). We compared predictive accuracy by meta-regression and found that summary sensitivity for IL-6 was significantly higher than for CRP at day 1 ($P = .03$), with comparable specificity ($P = .71$) (Fig. 3A). Also, summary specificity for IL-6 was significantly higher than for CRP at day 2 ($P = .02$), while sensitivity was not significantly different ($P = .14$) (Fig. 3B). Meta-regression did not show significant differences in sensitivity or specificity between IL-6 at day 1 and APACHE-II on day 1, although specificity was significantly higher compared to APACHE-II on day 2 ($P = .02$) with comparable sensitivity ($P = 0.14$, Fig. 3C and D). There were no differences for PMN-E compared to APACHE-II (data not shown).

HSROC analysis shows higher curves of IL-6 and PMN-E for the prediction of MSAP/SAP at day 1 (Fig. 4). The urinary marker trypsinogen activation peptide (TAP) was studied in three studies and the HSROC curve at day 2 was comparable with serum IL-6 at day 1. Other studied biomarkers showed comparable or lower

HSROC curves than the currently used predictors CRP and APACHE-II (Fig. S3 in Supporting information).

Head-to-head comparisons

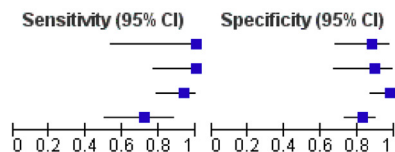
We intended to perform bivariate analysis for studies that made head-to-head comparisons of biomarkers, but there were sufficient data sources only for the comparison IL-6 at day 1 versus CRP at day 2. Meta-regression of 3 studies showed a significant difference in specificity (90% for day 1 IL-6 versus 74% for day 2 CRP, $P < .05$); the difference in sensitivity was not statistically significant (92% versus 84%, respectively) (Fig. 5).

Severe AP

For the prediction of specifically severe acute pancreatitis (persistent organ failure), seventeen biomarkers were analyzed using the HSROC method (Table S3, Fig. S4 in Supporting information), and eight markers (calcium, albumin, CRP, white blood cell count, creatinine, neutrophil-lymphocyte ratio, hematocrit and

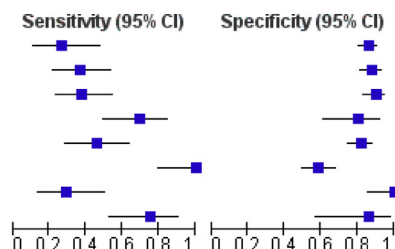
IL-6 24 hr

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Gurleyik 2004	6	3	0	21	1.00 [0.54, 1.00]	0.88 [0.68, 0.97]
Jiang 2004	14	2	0	17	1.00 [0.77, 1.00]	0.89 [0.67, 0.99]
Khanna 2013	29	1	2	40	0.94 [0.79, 0.99]	0.98 [0.87, 1.00]
Rodriguez-Nicolas 2018	18	16	7	76	0.72 [0.51, 0.88]	0.83 [0.73, 0.90]



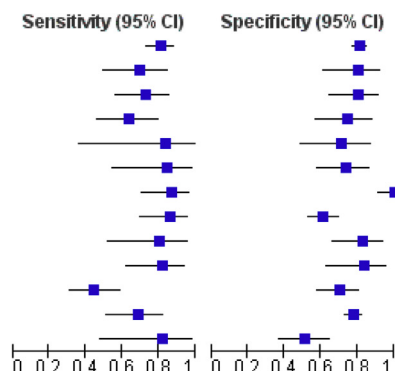
CRP 24 hr

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Johnson 2004	7	25	19	154	0.27 [0.12, 0.48]	0.86 [0.80, 0.91]
Kylanpaa-Back 2001	14	15	24	109	0.37 [0.22, 0.54]	0.88 [0.81, 0.93]
Lempinen 2001	16	11	26	97	0.38 [0.24, 0.54]	0.90 [0.83, 0.95]
Ma 2019	20	6	9	24	0.69 [0.49, 0.85]	0.80 [0.61, 0.92]
Neoptolemos 2000	16	25	19	112	0.46 [0.29, 0.63]	0.82 [0.74, 0.88]
Regner 2008	16	51	0	73	1.00 [0.79, 1.00]	0.59 [0.50, 0.68]
Shokuhi 2002	8	0	19	24	0.30 [0.14, 0.50]	1.00 [0.86, 1.00]
Ueda 1997	18	2	6	12	0.75 [0.53, 0.90]	0.86 [0.57, 0.98]



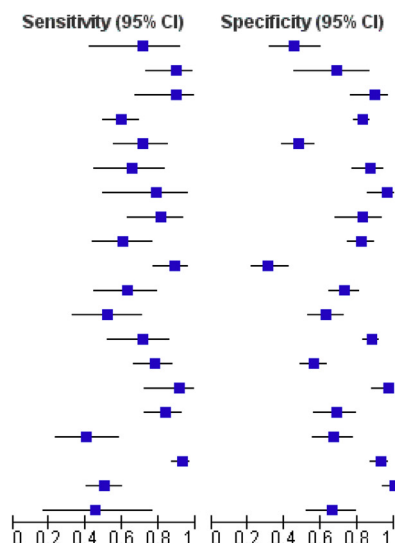
CRP 48 hr

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Basak 2017	100	98	23	417	0.81 [0.73, 0.88]	0.81 [0.77, 0.84]
de-Madaria 2018	20	6	9	24	0.69 [0.49, 0.85]	0.80 [0.61, 0.92]
Frossard 2006	29	8	11	32	0.72 [0.56, 0.85]	0.80 [0.64, 0.91]
Gomercic 2016	23	9	13	26	0.64 [0.46, 0.79]	0.74 [0.57, 0.88]
Gurleyik 2004	5	7	1	17	0.83 [0.36, 1.00]	0.71 [0.49, 0.87]
Gurleyik 2005	11	11	2	31	0.85 [0.55, 0.98]	0.74 [0.58, 0.86]
Khanna 2013	27	0	4	41	0.87 [0.70, 0.96]	1.00 [0.91, 1.00]
Neoptolemos 2000	30	53	5	84	0.86 [0.70, 0.95]	0.61 [0.53, 0.70]
Pongprasobchai 2010	12	6	3	29	0.80 [0.52, 0.96]	0.83 [0.66, 0.93]
Shokuhi 2002	22	4	5	20	0.81 [0.62, 0.94]	0.83 [0.63, 0.95]
Simoes 2011	25	21	31	49	0.45 [0.31, 0.59]	0.70 [0.58, 0.80]
Stirling 2017	26	66	12	233	0.68 [0.51, 0.82]	0.78 [0.73, 0.82]
Zuidema 2014	9	26	2	27	0.82 [0.48, 0.98]	0.51 [0.37, 0.65]



APACHE-II 24 hr

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ammori 2003a	10	30	4	25	0.71 [0.42, 0.92]	0.45 [0.32, 0.59]
Bezmarevic 2012	26	7	3	15	0.90 [0.73, 0.98]	0.68 [0.45, 0.86]
Bulboller 2006	17	5	2	41	0.89 [0.67, 0.99]	0.89 [0.76, 0.96]
Chen 2013	60	70	41	326	0.59 [0.49, 0.69]	0.82 [0.78, 0.86]
Fan 1993	30	70	12	64	0.71 [0.55, 0.84]	0.48 [0.39, 0.57]
Gutierrez-Jimenez 2014	17	9	9	61	0.65 [0.44, 0.83]	0.87 [0.77, 0.94]
Hagjer 2018	11	2	3	44	0.79 [0.49, 0.95]	0.96 [0.85, 0.99]
Khanna 2013	25	7	6	34	0.81 [0.63, 0.93]	0.83 [0.68, 0.93]
Kylanpaa-Back 2001	23	22	15	102	0.61 [0.43, 0.76]	0.82 [0.74, 0.89]
Lee 2016	53	59	7	27	0.88 [0.77, 0.95]	0.31 [0.22, 0.42]
Neoptolemos 2000	22	37	13	100	0.63 [0.45, 0.79]	0.73 [0.65, 0.80]
Pallisera 2014	15	38	14	64	0.52 [0.33, 0.71]	0.63 [0.53, 0.72]
Park 2013	22	34	9	238	0.71 [0.52, 0.86]	0.88 [0.83, 0.91]
Pearce 2006	53	87	15	110	0.78 [0.66, 0.87]	0.56 [0.49, 0.63]
Rathnakar 2017	21	2	2	57	0.91 [0.72, 0.99]	0.97 [0.88, 1.00]
Simoes 2011	47	22	9	48	0.84 [0.72, 0.92]	0.69 [0.56, 0.79]
Stimac 2006	14	27	21	55	0.40 [0.24, 0.58]	0.67 [0.56, 0.77]
Vasudevan 2018	157	13	13	160	0.92 [0.87, 0.96]	0.92 [0.87, 0.96]
Venkatesh 2020	52	0	52	60	0.50 [0.40, 0.60]	1.00 [0.94, 1.00]
Zuidema 2014	5	18	6	35	0.45 [0.17, 0.77]	0.66 [0.52, 0.78]



APACHE-II 48 hr

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Fan 1993	33	64	9	70	0.79 [0.63, 0.90]	0.52 [0.43, 0.61]
Gurleyik 2004	5	2	1	22	0.83 [0.36, 1.00]	0.92 [0.73, 0.99]
Lin 2020	95	56	22	86	0.81 [0.73, 0.88]	0.61 [0.52, 0.69]
Neoptolemos 2000	20	49	15	88	0.57 [0.39, 0.74]	0.64 [0.56, 0.72]
Simoes 2011	44	12	12	58	0.79 [0.66, 0.88]	0.83 [0.72, 0.91]

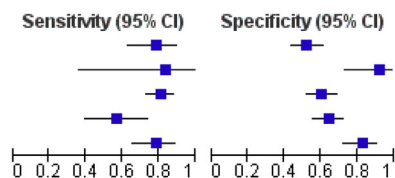


Fig. 2. Forest plots for IL-6, CRP and APACHE-II included in the bivariate analysis of MSAP/SAP.

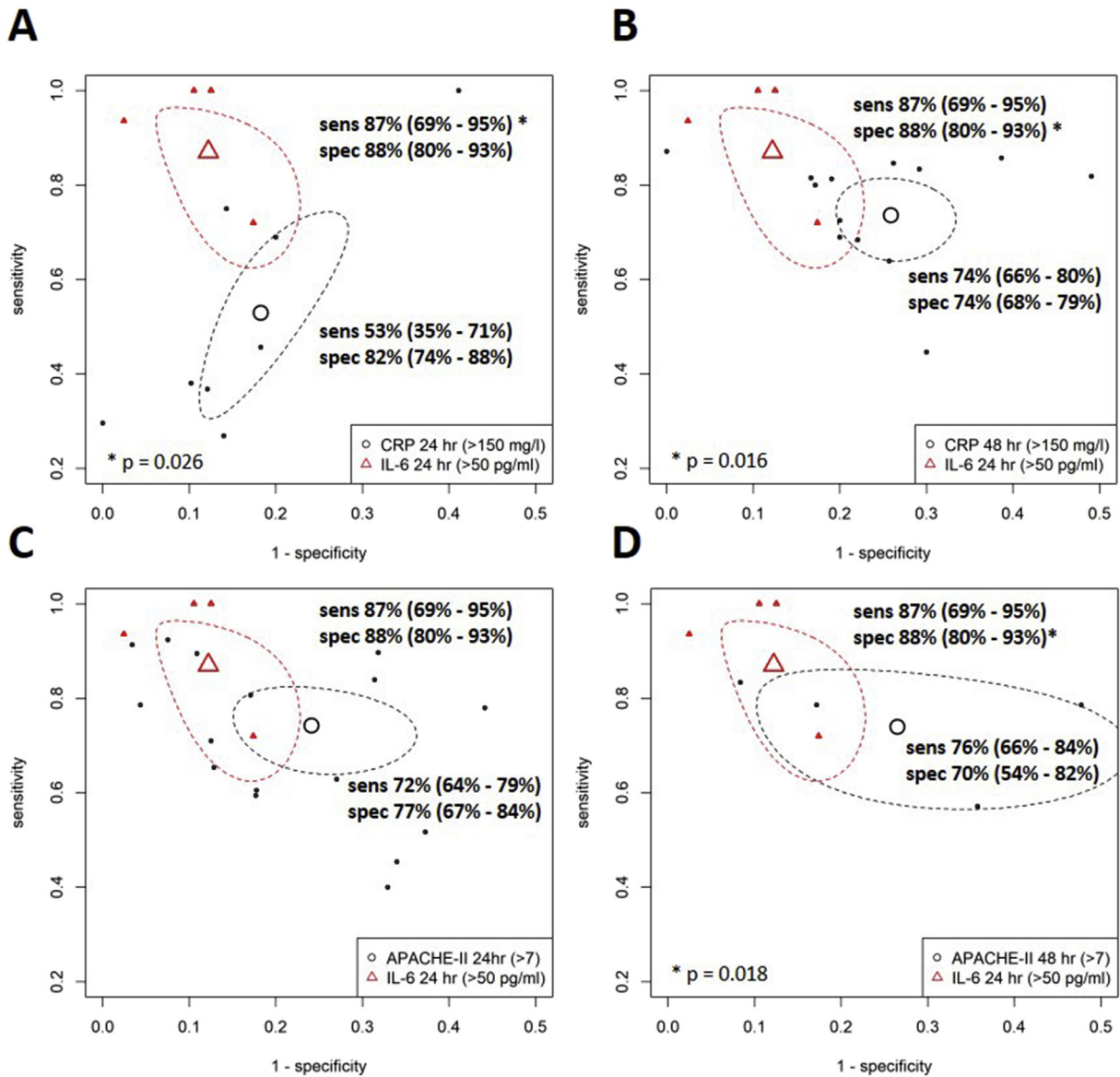


Fig. 3. Summary point estimates of IL-6, APACHE-II and CRP using the bivariate model and meta-regression analysis for the prediction of MSAP/SAP (A) unpaired analysis for CRP and IL-6 at 0–24 h (B) unpaired analysis for IL-6 at 0–24 h and CRP 24–48 h (C) unpaired analysis for IL-6 and APACHE-II at 0–24 h (D) unpaired analysis for IL-6 at 0–24 h and APACHE-II at 24–48 h. Summary point estimates of sensitivity (sens) and specificity (spec) with 95% confidence interval are shown and compared with meta-regression. CRP = C-reactive protein; IL-6 = interleukin-6; PCT = procalcitonin; HCT = hematocrit; * = p-value < 0.05.

APACHE-II) using the bivariate model (Table 1, Fig. S5 in Supporting information).

Summary specificity and sensitivity were highest for serum calcium, with sensitivity, specificity and DOR of 74% (95% CI 52–88%), 85% (95% CI 71–93%) and 16.5 (11.5–22.5), respectively, following by albumin at day 1 (DOR 10.7, 95% CI 8 to 13.8) and CRP at day 1 (DOR 10.3, 95% CI 3.68 to 22.1). Meta-regression showed a significant higher specificity of albumin at day 1 compared to CRP at day 2 ($P = 0.01$), but not to other predictors. Neither did serum calcium demonstrate significant differences in terms of sensitivity or specificity. The HSROC curve of angiotensin-2 (3 studies, 374 subjects) is entirely above the curves of CRP at both day 1 and day 2, but crosses the HSROC curves of other predictors such as APACHE-II, blood urea nitrogen (BUN) and albumin (Fig. S4 in Supporting information).

Organ failure

Three predictors of organ failure, CRP, procalcitonin (PCT) and APACHE-II at day 1, were suitable for meta-analysis (HSROC), of which APACHE-II with a threshold of 8 or higher being also appropriate for the bivariate model analysis.

Summary sensitivity and specificity of APACHE-II were 80% (95% CI, 65–89%) and 79% (95% CI 65–89%), with a DOR of 18.0 (95% CI 14.8 to 49.5) (Table 1, Fig. S6 in Supporting information). At day 1, the HSROC curve of PCT was entirely above the curves of CRP and APACHE-II. This indicates that PCT has superior predictive accuracy for organ failure compared to CRP. However, visual inspection of the HSROC and forest plots shows relatively low heterogeneity among studies for both APACHE and CRP (Fig. S7 in Supporting information), in contrast to PCT.

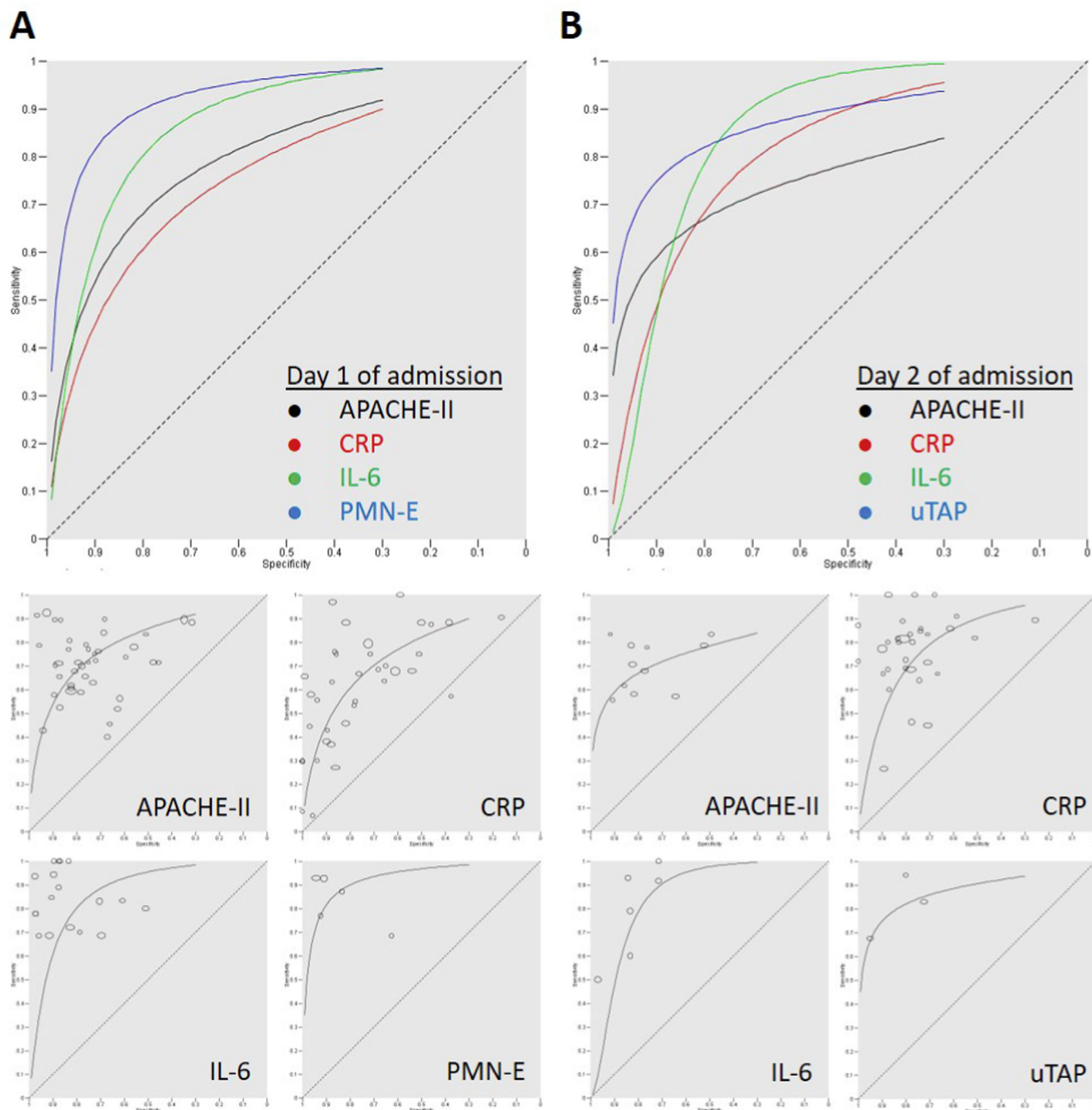


Fig. 4. HSROC model for prediction of severity (MSAP/SAP) at A) 0–24 h after admission, B) 24–48 h after admission. The ovals are representations of the study sample size. HSROC = hierarchical summary receiver operating characteristic; APACHE = Acute Physiology and Chronic Health Evaluation; CRP = C-reactive protein; IL-6 = interleukin-6; PMN-E = polymorphonuclear leukocyte elastase; uTAP = urine trypsinogen activation peptide.

Pancreatic necrosis

Four biomarkers (APACHE-II, CRP, IL-6 and hematocrit) were appropriate for HSROC analysis, of which APACHE-II at a (at threshold 8 or higher) and hematocrit (at threshold 44%) at day 1 were suitable for the bivariate model analysis.

The predictive value of hematocrit (HCT) and APACHE-II were low, with sensitivities and specificity of 63% (95% CI, 55–69%), and 69% (95% CI, 51–83%), respectively for HCT and 55% (95% CI,

37–72%) and 63% (95% CI, 49–76%) for APACHE-II (Table 1, Fig. S8 in Supporting **information**). HSROC analysis shows comparable curves of IL-6 and CRP within 24 h of admission or at 48 h (Fig. S9 in Supporting **information**). The forest plots and HSROC curves showed great heterogeneity for especially APACHE-II and HCT.

Heterogeneity analyses

Two subgroup analyses for the bivariate and HSROC model were

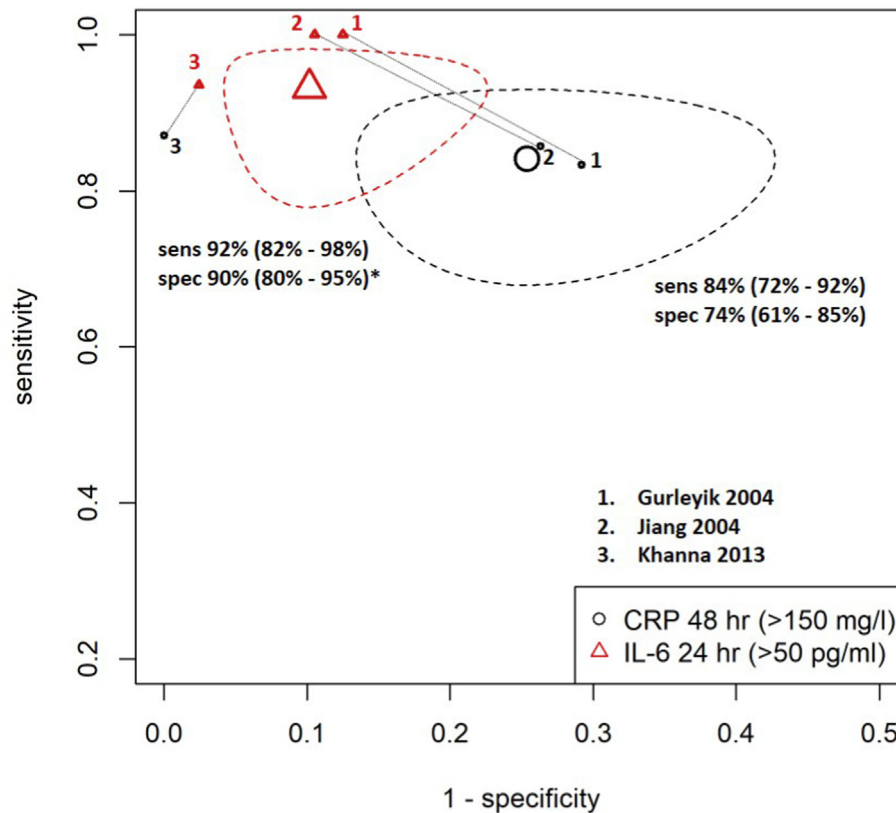


Fig. 5. Summary point estimates for head-to-head comparison using the bivariate model and meta-regression analysis for the prediction of MSAP/SAP Paired analysis for IL-6 at 0–24 h and CRP 24–48 h. Summary point estimates of sensitivity (sens) and specificity (spec) with 95% confidence interval are shown and compared with meta-regression. CRP = C-reactive protein; IL-6 = interleukin-6; * = p -value < 0.05.

performed, first excluding studies with a retrospective or case-control study design (Table S4 in Supporting **information**), and second in studies performed in tertiary centers (Table S4 in Supporting **information**). No major differences in the outcomes were observed.

Discussion

The main finding of this study is that interleukin-6 (IL-6) at admission is superior in predicting moderate severe or severe acute pancreatitis (MSAP/SAP) in comparison to commonly used markers such as serum CRP and APACHE-II. Previous studies indicate that serum IL-6 levels peak at around 36 h from onset of symptoms, and remains elevated for at least 5 days in patients with MSAP/SAP, and in contrast to patient with mild acute pancreatitis (MAP) where levels remain low [16,17]. Our bivariate model analysis shows that summary sensitivity (87%) and specificity (88%) of IL-6 at a threshold of 50 ng/ml are higher than CRP (threshold >150 mg/l) at day 1 (53% and 82%, respectively) or day 2 (74% and 74%, respectively) and APACHE II (threshold 8 or higher) at day 1 (72% and 77%, respectively) or day 2 (76% and 70%, respectively). Meta-regression showed significantly better predictive accuracy over all biomarkers except the APACHE-II score at day 1. Our HSROC model analysis further underlines the superior predictive capacity of IL-6 in respect to the APACHE II score. Although the HSROC curve of polymononuclear elastase (PMN-E) is above the other predictors, the meta-analyses contained only 5 studies with limited samples size (483 patients), and a small study that indicated poor predictive accuracy [18]. This was confirmed by the bivariate analysis that did not show superior accuracy of PMN at a threshold of 100 μ g/l over

other predictors.

This is the first comprehensive systematic review that systematically analyzed laboratory biomarkers for early severity prediction in acute pancreatitis without restriction for specific biomarkers. Previous comprehensive studies focused on mortality [19,20] or infected pancreatic necrosis and persistent organ failure [21] as primary outcomes. Meta-analyses have been published that focused on a single predictors, including on cytokines such as IL-6 [22,23]. Using the Moses-Littenberg random-effects model, they reported pooled sensitivities and specificities for IL-6 in the range of 75%–91% and 75%–86%, respectively. Aoun and colleagues also performed a pooled analysis of diagnostic odds ratios (DOR) and reported a mean DOR of 23.8 at the second day of admission [22]. Our meta-analysis has some advantages. Most importantly, as recommended we applied hierarchical statistical models, instead of the Moses-Littenberg model which is considered to be outdated and should be avoided [24,25]. Pooling of likelihood and diagnostic odds ratio can produce erroneous results [12], therefore we calculated summary likelihood and diagnostic odds ratios using a MCMC model [24]. Secondly, we applied the bivariate model for studies that used a common threshold, and compared them to predictors that are currently used for clinical decision making, enabling clinicians to better interpret the results.

A previous study on clinical prediction scores for mortality in acute pancreatitis did not include a meta-analysis due to heterogeneity [19]. We anticipated that variation in study population, definitions of clinical outcomes and sampling time points were the major sources of heterogeneity in our study. Therefore, the strict eligibility criteria included the exclusion of studies that were done in a subpopulation (i.e. predicted severe pancreatitis) in an attempt

to perform the meta-analyses in an unselected patient population, and studies that used definitions for severity other than the (original or revised) Atlanta classification were excluded. Also, timing of sample collection was accounted for by performing separate analyses for day 1 and 2 of admission. Finally, the heterogeneity analysis indicated that study design (prospective studies) and hospital setting (tertiary centers) did not impact the results.

Despite our efforts to optimize our study, there are some limitations that are applicable to most diagnostic test accuracy reviews, and limitations specific for our study. Well-known difficulties are the lack of data reporting by the original authors, and between-study heterogeneity [25]. Ideally, studies should limit inclusion to patients with onset of symptoms less than 48–72 h before hospital admission, to better reflect the actual disease state. However, since the majority of studies do not report on the duration of symptoms before hospitalization or recruitment of the subjects, a sensitivity analysis with these studies was not feasible. Laboratory methods and assays, which have variable properties in terms of accuracy and quality, and treatment regimens, which might vary by country, hospital setting and time period of inclusion, may be sources of heterogeneity. Risk of bias assessment demonstrates poor quality for most included studies, and a lack of remaining studies precluded a sensitivity analysis for low risk of bias studies in this study. In light of this, there is a clear need for an adequately powered, well-designed study that ideally consists of multiple large international cohorts comparing the best performing (laboratory, imaging and clinical) predictors from the literature, including IL-6.

Early and accurate identification of patients with acute pancreatitis at risk of developing severe complications is needed to guide clinical decision making, triage and to select patients for novel prophylactic treatments in clinical trials. Very accurately performing diagnostic tests tend to have LR + above 10 or LR- below 0.1 [26]. For a diagnostic test to be potentially interesting, it needs to have a LR+ of at least 7 or a diagnostic odds ratio over 30, a requirement that is only met by IL-6 according to our meta-analysis. Recent meta-analyses that focused on clinical and imaging scores also demonstrated a maximum accuracy of around 80% for the prediction of severity [21], of which the modified computed tomography severity index (mCTSI) appears to be the best predictor with a DOR of around 29 (sensitivity 88%, specificity 80%) [27]. Predictive accuracy of biomarkers for (persistent) organ failure and pancreatic necrosis is inadequate and not suitable for individual patient prediction. Multiple predictor scoring systems, such as the APACHE-II are producing inconsistent results, are cumbersome in use and are therefore often not routinely performed by clinicians. The inconsistent use and performance of the APACHE-II score for the prediction of severity is demonstrated by the study of Mounzer et al. where the accuracy parameters in the training cohort (sensitivity 84%, specificity 71%) differed greatly from the validation cohort (sensitivity 97%, specificity 44%) using the same methods, in contrast to the other scoring systems that were under investigation [28]. The distribution of phenotypes that determine MSAP/SAP, organ failure and pancreatic necrosis, might explain this, since the APACHE-II score mainly includes parameters for organ failure. Although future predictors should ideally be single, pancreatitis-specific biomarkers with high accuracy and rapid turnover time, combining multiple high-accuracy biomarkers (such as IL-6) or applying artificial intelligence may also represent suitable strategies.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.pan.2020.09.007>.

References

- [1] Bradley EL. A clinically based classification-system for acute-pancreatitis - summary of the international-symposium on acute-pancreatitis, atlanta, ga, september 11 through 13, 1992. *Arch Surg* 1993;128:586–90.
- [2] Johnson CD, Abu-Hilal M. Persistent organ failure during the first week as a marker of fatal outcome in acute pancreatitis. *Gut* 2004;53:1340–4.
- [3] Vege SS, Gardner TB, Chari ST, Munukuti P, Pearson RK, Clain JE, et al. Low mortality and high morbidity in severe acute pancreatitis without organ failure: a case for revising the atlanta classification to include “moderately severe acute pancreatitis”. *Am J Gastroenterol* 2009;104:710–5.
- [4] 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;62:102–11.
- [5] Besselink M, van Santvoort H, Freeman M, Gardner T, Mayerle J, Vege SS, et al. IAP/AGA evidence-based guidelines for the management of acute pancreatitis. *Pancreatology* : official journal of the International Association of Pancreatology (IAP) 2013;13:E1–15.
- [6] Tenner S, Baillie J, DeWitt J, Vege SS, American College of G. American college of gastroenterology guideline: management of acute pancreatitis. *Am J Gastroenterol* 2013;108:1400–15. 1416.
- [7] McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the P-DTAG, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the prisma-dta statement. *J Am Med Assoc* 2018;319:388–96.
- [8] van den Berg FF, Kempeneers MA, van Santvoort HC, Zwinderman AH, Issa Y, Boermeester MA. Meta-analysis and field synopsis of genetic variants associated with the risk and severity of acute pancreatitis. *BJs open* 2020;4:3–15.
- [9] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865–84.
- [10] Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982–90.
- [11] Banks PA, Freeman ML. Practice parameters committee of the American college of G: practice guidelines in acute pancreatitis. *Am J Gastroenterol* 2006;101:2379–400.
- [12] Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med* 2008;27:687–97.
- [13] Doebler P. Mada: meta-analysis of diagnostic accuracy. 2019. R package version 0.5.9.
- [14] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. Stard 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open* 2016;6:e012799.
- [15] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.
- [16] Heath DI, Cruickshank A, Gudgeon M, Jehanli A, Shenkin A, Imrie CW. Role of interleukin-6 in mediating the acute phase protein response and potential as an early means of severity assessment in acute pancreatitis. *Gut* 1993;34:41–5.
- [17] Naskalski JW, Kusnierz-Cabala B, Panek J, Kedra B. Poly-c specific ribonuclease activity correlates with increased concentrations of il-6, il-8 and tnfr55/tnfr75 in plasma of patients with acute pancreatitis. *J Physiol Pharmacol* 2003;54:439–48.
- [18] Berney T, Gasche Y, Robert J, Jenny A, Mensi N, Grau G, et al. Serum profiles of interleukin-6, interleukin-8, and interleukin-10 in patients with severe and mild acute pancreatitis. *Pancreas* 1999;18:371–7.
- [19] Di MY, Liu H, Yang ZY, Bonis PA, Tang JL, Lau J. Prediction models of mortality in acute pancreatitis in adults: a systematic review. *Ann Intern Med* 2016;165:482–90.
- [20] Gravante G, Garcea G, Ong SL, Metcalfe MS, Berry DP, Lloyd DM, et al. Prediction of mortality in acute pancreatitis: a systematic review of the published evidence. *Pancreatology* : official journal of the International Association of Pancreatology (IAP) [et al] 2009;9:601–14.
- [21] Yang CJ, Chen J, Phillips ARJ, Windsor JA, Petrov MS. Predictors of severe and critical acute pancreatitis: a systematic review. *Dig Liver Dis* 2014;46:446–51.
- [22] Aoun E, Chen J, Reighard D, Gleeson FC, Whitcomb DC, Papachristou GI. Diagnostic accuracy of interleukin-6 and interleukin-8 in predicting severe acute pancreatitis: a meta-analysis. *Pancreatology* : official journal of the International Association of Pancreatology (IAP) [et al] 2009;9:777–85.
- [23] Zhang J, Niu J, Yang J. Interleukin-6, interleukin-8 and interleukin-10 in estimating the severity of acute pancreatitis: an updated meta-analysis. *Hepato-Gastroenterology* 2014;61:215–20.
- [24] Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy chapter 10 analysing and presenting results. *Cochrane Handb Syst Rev Diagnostic Test Accuracy* 2010:1–44.
- [25] Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Cochrane diagnostic test

- accuracy working G: systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;149:889–97.
- [26] Jaeschke R, Guyatt GH, Sackett DL, Guyatt G, Bass E, Brill-Edwards P, et al. Users' guides to the medical literature: lii. How to use an article about a diagnostic test b. What are the results and will they help me in caring for my patients? *J Am Med Assoc* 1994;271:703–7.
- [27] Miko A, Vigh E, Matrai P, Soos A, Garami A, Balasko 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.
- [28] 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: 1476–82.