Published online: 2020-05-26

Clinical Prediction Scores for Pediatric Appendicitis

Johanna Gudjonsdottir[1] Emma Marklund[1] Lars Hagander[1,2] Martin Salö[1,2]

[1] Department of Clinical Sciences, Pediatrics, Lund University, Lund, Sweden
[2] Department of Pediatric Surgery, Skane University Hospital, Lund, Sweden

Address for correspondence Martin Salö, MD, PhD, Department of Pediatric Surgery, Skane University Hospital, Lund, Sweden (e-mail: martin.salo@med.lu.se; martin.salo@skane.se).

Eur J Pediatr Surg 2021;31:252–260.

Abstract

Introduction The rate of misdiagnosis of appendicitis in children is a challenge and clinical prediction scores could be part of the solution. However, the pediatric appendicitis score (PAS) and the Alvarado score have shown disappointing diagnostic accuracy in pediatric validation studies, while the appendicitis inflammatory response (AIR) score and the novel pediatric appendicitis risk calculator (pARC) have not yet been validated thoroughly. Therefore, the aim of the present study was to evaluate these four prediction scores prospectively in children with suspected appendicitis.

Materials and Methods A prospective study was conducted over a 2-year period. All patients <15 years with suspected appendicitis were eligible for inclusion. The four prediction scores were compared regarding predictive values, receiver operating characteristics (ROC) curves, decision curve analysis, and clinical outcome.

Results Of the 318 patients included, 151 (47 %) patients had appendicitis. The AIR score and the pARC had substantially higher specificity and positive predictive value, and lower rate of false positives (7% and 2%), than the PAS and Alvarado score (36 and 28%, $p < 0.001$). Across the different gender and age groups, the AIR score and the pARC generally had fewer false positives than the PAS and Alvarado score. There were no significant differences in sensitivity, negative predictive values, rates of missed appendicitis, or ROC curve analysis. In decision curve analysis, the AIR score and the pARC outperformed the PAS and Alvarado score at most threshold probabilities.

Conclusion The AIR score and the pARC are superior to the PAS and Alvarado score in diagnosing children with suspected appendicitis.

Keywords
- appendicitis
- appendicitis inflammatory response score
- pediatric appendicitis risk calculator
- clinical prediction scores

Introduction

Appendicitis is the most common abdominal emergency in children[1,2] and appendectomy is the most common acute abdominal operation performed worldwide.[3] Yet, a significant proportion of children are initially misdiagnosed, especially younger children and girls.[1,4–6] Misdiagnosis leads to prolonged observation periods, increased risk of negative appendectomies, and adverse effects such as perforations and pelvic abscesses.[7] These, in turn, cause morbidity, increased costs, and inadequate use of health care resources.

As a result, clinical prediction scores have been developed, based on combinations of history, symptoms and basic laboratory results.[8] They should be used to determine which patients can be sent home, further evaluated with ultrasonography (US) or computed thomography (CT), or taken straight to surgery. Three of the most well-established scoring systems are the Alvarado score,[9] the pediatric appendicitis score (PAS),[10] and the appendicitis inflammatory response (AIR) score[11] (►**Supplementary Table S1** [available in the online version]). Several studies have evaluated the use of the Alvarado score and PAS in children with suspected appendicitis, with varying results.[12–16] The AIR score outperforms other scoring systems in adult populations[11,17,18] and was superior in a retrospective pediatric cohort study[19];

received
November 25, 2019
accepted
April 7, 2020
published online
May 26, 2020

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0040-1710534.
ISSN 0939-7248.

however, it has not been yet been evaluated exclusively in a prospective pediatric cohort. Recently, the pediatric appendicitis risk calculator (pARC) was introduced as a result of a prospective multicenter study.[20] This new instrument uses a multivariable prediction model to quantify the risk of appendicitis on a continuous scale and has shown promising diagnostic accuracy; however, it remains to undergo external validation. Therefore, the aim of the present study was to prospectively evaluate these four clinical prediction scores for suspected appendicitis in children, regarding diagnostic values and receiver operating characteristics (ROC) as well as impact on clinical decision. We hypothesized that the AIR score would outperform the other scores overall and across different age and gender groups.

Materials and Methods

This was a prospective study of a 2-year consecutive cohort of children with suspected appendicitis at a tertiary center of pediatric surgery with a catchment area of 350,000 inhabitants for primary surgical care. The study was approved by the regional ethical committee (DNR 2010/49 and 2013/614) and by the hospital review board. The included subjects all agreed to participation through parental informed consent.

Inclusion and Exclusion Criteria

All children <15 years of age presenting to a pediatric surgeon with suspected appendicitis were eligible for inclusion in the study. The exclusion criteria were previous episode of suspected appendicitis, or current treatment with anti-inflammatory drugs, or severe chronic illness. Two patients were excluded: one because of a previous episode of appendicitis and the other one due to ongoing treatment with corticosteroids. Another 10 patients were excluded as a result of missing laboratory results (►Fig. 1). Hence, a total of 318 patients remained for further analyses of the PAS, AIR score, and the Alvarado score. The pARC could be calculated for 200 patients, with exclusions as a result of

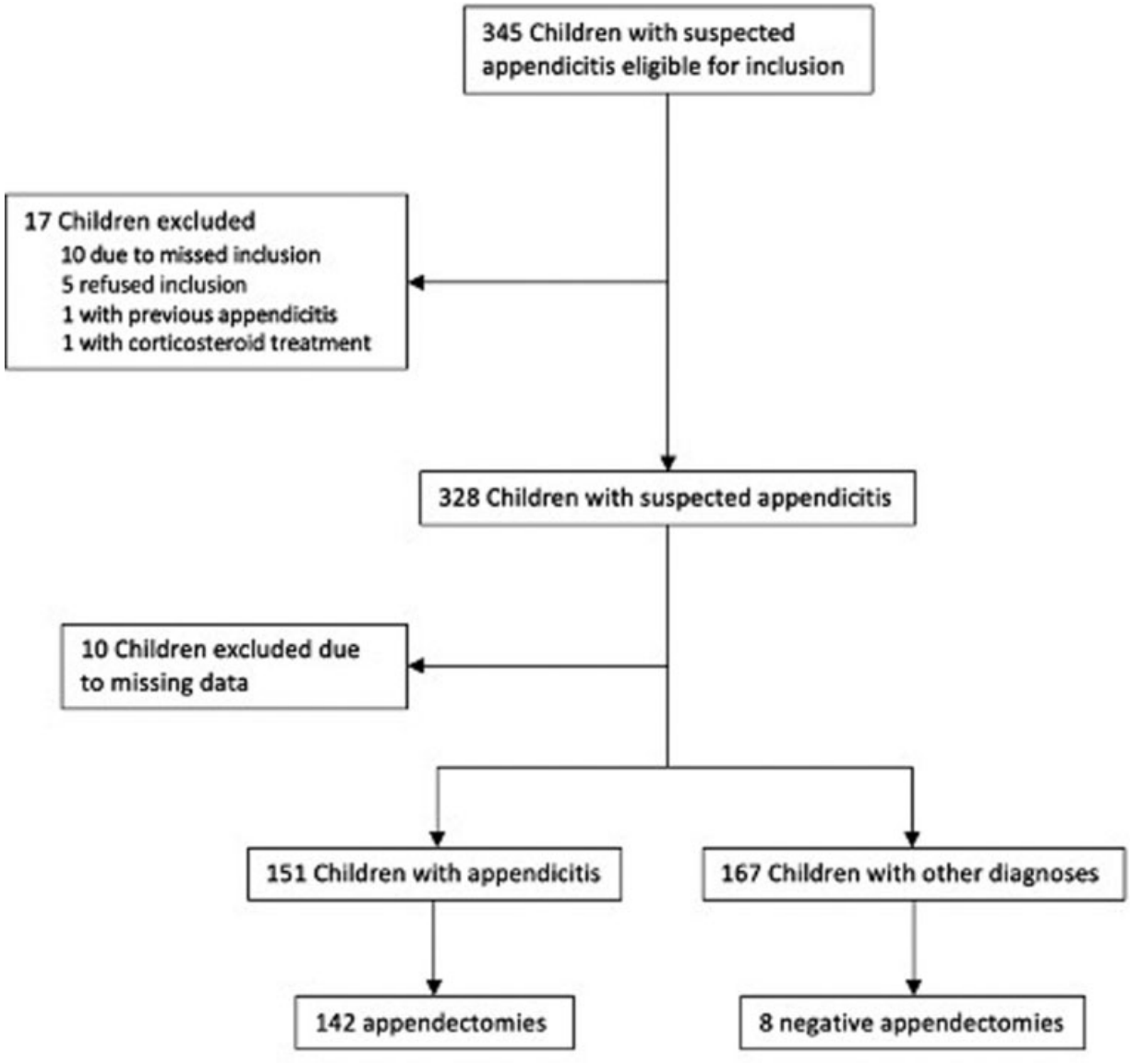

Fig. 1 Flow chart of inclusion and exclusion of the study subjects.

its 5-year age-cutoff ($n = 40$), and missing data regarding symptom duration ($n = 78$).

Data Collection

During two consecutive years (March 1, 2016–February 28, 2018), data were collected prospectively at the pediatric emergency department. The pediatric surgeon on call—who examined the patients—registered the clinical data including laboratory values in a study form, from which the clinical prediction scores were later derived. Medical records were reviewed to determine the patients' final diagnoses, whether they had undergone surgery, and the results from the histopathologic examination. Since the Swedish health care system provides state-wide electronical medical records, patients without readmission to any pediatric ED in the state were assumed not to have appendicitis.

Primary and Secondary Outcome

Primary outcomes were appendicitis and complicated appendicitis. The diagnosis of appendicitis was based on the intraoperative findings and histopathologic diagnosis.[21] Phlegmonous appendicitis was considered uncomplicated, whereas gangrenous—perforated and appendiceal— abscess was considered complicated. Phlegmonous appendicitis was defined as infiltration of neutrophil granulocytes in the muscularis propria layer.[21–24] Gangrenous appendicitis was defined as an inflamed appendix with significant gray or black discoloration with clear histological evidence of transmural necrosis and absence of the criteria for perforation.[21,22] Perforation was defined as a visual hole in the appendix, perioperative finding of an appendicolith in the abdomen, or the spread of pus in the abdominal cavity.[23] The diagnosis of appendiceal abscess was based on US or CT examinations.

Secondary outcomes were missed appendicitis and no appendicitis. Missed appendicitis was defined as a patient with appendicitis who was classified as low risk. No appendicitis was defined as patient discharge without primary surgery or subsequent readmission, or the finding of anon-inflamed appendix during laparoscopy or histopathologic examination (negative appendectomy). The histological definition of no appendicitis was absence of any signs of inflammation, or inflammatory changes limited to the mucosa.[25,26] Noninflamed appendices were left *in situ*.[27]

Definitions

The parameter "fever" in PAS was not specified in the original study,[12,13,28] and the cut-off temperature was set at ≥38.0°C. A "leucocyte shift" in the Alvarado score was equated with neutrophilia, with different cut-off values depending on age. The normal reference intervals for leucocytes at different ages were: 6.0 to 16.0 × 10^9/L (3 months–3 years), 5.0 to 15.0 × 10^9/L (3–6 years), and 5.0 to 13.0 × 10^9/L (6–18 years). The normal reference intervals for neutrophils at different ages were: 1.6 to 5.3 × 10^9/L (3 months–1 year), 1.6 to 6.5 × 10^9/L (1–5 years), 2.4 to 6.5 × 10^9/L (5–10 years), and 1.7 to 7.0 × 10^9/L (10–15 years).

Statistical Analyses

A power analysis was performed and 262 patients were required to show a difference of 10% between scores with a power of 80% and a p-value of <0.05. Continuous normal distributed and nonnormal distributed variables were reported as mean ± standard deviation (SD) and median (minimum–maximum), respectively, with differences between groups assessed using the Student's t-test and Mann–Whitney U test. Dichotomous variables were presented as frequencies and percentages, with differences between groups assessed using Fisher's exact test or Chi-squared test. A *post hoc* test (Bonferroni) was performed for comparisons between >2 groups. Sensitivity, specificity, predictive values, rates of missed appendicitis (false negatives), and rates of no appendicitis (false positives) were calculated for each score's respective cut-off levels for low, intermediate, and high risk of appendicitis. Cut-off levels according to the original publications of each scoring system[9–11,20] were used.

A ROC curve was performed with analysis of the area under the curve (AUC) for comparisons of the scores in the total study populations, as well as in different age (0–4, 5–9, and 10–14 years) and gender groups. Since the pARC could not be calculated for children aged <5 years, this scoring system was not included in these analyses. In addition, decision curves were created by calculating net benefit using the formula: (true positives/n)-((false positives/n)*(threshold probability/(1-threshold probability))), and plotted on different threshold probabilities.[29] Every patient's risk of appendicitis (predicted probability) according to each scoring system was calculated through logistic regressions. There is currently no established equivalent of the net benefit formula for true and false negatives. Hence, the decision curves' threshold probabilities spanned from 70 to 100%. Since the pARC could not be calculated for patients younger than 5 years, two curves were created: one including all patients and one with only 5 to 14 years old including the pARC. The net benefit for treating all patients was negative at every threshold (values ranging from –0.8 to –51), and thus not included in the graph. R software, version 3.5.1 (R foundation for statistical computing, Vienna, Austria) was used to generate the decision curves and the ROC plots. All other statistical analyses were performed in IBM SPSS Statistics for Macintosh, version 24.0 (IBM Corp, Armonk, New York, United States).

Results

Study Population

Of the 318 patients included, 176 (55%) were boys and the mean age was 9 years. Of these, 151 (47%) patients were diagnosed with appendicitis, 84 (56%) patients with phlegmonous appendicitis, and 67 (44%) patients with complicated appendicitis (►Table 1). Among the 167 patients without appendicitis, the most frequent diagnoses were nonspecific abdominal pain (22%) and mesenteric lymphadenitis (12%). The PAS, AIR score, Alvarado score, and the Parc were all significantly higher in patients with appendicitis ($p < 0.001$; ►Table 1).

Table 1 Parameters, total points of the pediatric appendicitis score, appendicitis inflammatory response score, Alvarado score and the pediatric appendicitis risk calculator, and final diagnoses of 318 patients with suspected appendicitis

	Appendicitis ($n = 151$)	No appendicitis ($n = 167$)	*p*-Value
Age (y)	11 (2–14)	9 (2–14)	0.034[c]
Sex (male/female)	102/49	74/93	<0.001[a]
Nausea	125 (83%)	106 (63%)	<0.001[a]
Vomiting	103 (68%)	54 (32%)	<0.001[a]
Anorexia	121 (80%)	111 (66%)	0.006[a]
Pain migration	80 (52%)	49 (29%)	<0.001[a]
Temperature (°C)	37.8 (±0.9)	37.7 (±0.9)	0.27
Pain RLQ	149 (99%)	132 (79%)	<0.001[a]
Hopping/coughing/ percussion tenderness in RLQ	139 (92%)	65 (39%)	<0.001[a]
Rebound tenderness/ involuntary defense	125 (83%)	30 (18%)	<0.001[a]
Light	48 (32%)	22 (13%)	<0.001[a]
Medium	47 (31%)	6 (4%)	<0.001[a]
Strong	30 (20%)	2 (1%)	<0.001[a]
Leucocytes	16.3 (4.4–34.5)	9.5 (3.3–25.5)	<0.001[c]
10.0–14.9 × 10^9/L	41 (27%)	48 (29%)	0.752[a]
≥15.0 × 10^9/L	93 (62%)	29 (17%)	<0.001[a]
Neutrophils	13.2 (2.8–30.5)	6.0 (1.1–22.3)	<0.001[c]
70–84%	108 (72%)	60 (36%)	<0.001[a]
≥85%	30 (20%)	15 (9%)	0.005[a]
CRP	37.5 (<5–328)	19.5 (<0.6–186)	
10–49 mg/L	53 (35%)	64 (38%)	0.552[a]
≥50 mg/L	54 (36%)	13 (8%)	<0.001[a]
PAS	8.1 (±1.5)	5.3 (±2.1)	<0.001[b]
AIR	7.1 (±2.4)	3.2 (±2.1)	<0.001[b]
Alvarado	8.3 (±1.7)	5.3 (±2.1)	<0.001[b]
pARC[d]	80.0 (4.0–100)	23.5 (0–88.0)	<0.001[c]
Final diagnosis (*n*)	Phlegmonous (84) Gangrenous (29) Perforated (28) Abscess (10)	Unspecified abdominal pain (71) mesenteric lymphadenitis (39), Viral infection (22), constipation (14), pneumonia (6), ovulation (3), meckel's diverticulum (2), ruptured ovarian cyst (2) UTI (2), tonsillitis (2), pancreatitis (1), retrograde menstruation (1), salmonella infection (1), URTI (1)	

Abbreviations: AIR, appendicitis inflammatory response; CRP, C-reactive protein; PAS, pediatric appendicitis score; pARC, pediatric appendicitis risk calculator; RLQ, right lower quadrant; UTI, urinary tract infection; URTI, upper respiratory tract infection.
Note: Values presented as *n* (%), mean (±standard deviation) and median (minimum–maximum).
[a]Chi-square test.
[b]Student's *t*-test.
[c]Mann–Whitney U test.
[d]pARC only calculated for 200 patients (108 with appendicitis).

Score Comparison

In the high-risk group, the AIR score and the pARC had substantially higher specificity and PPV than the PAS and Alvarado score (►**Table 2**). The AIR score and the pARC also had significantly fewer cases of false positives (7 and 2%) than the PAS and Alvarado score (36 and 28%; $p < 0.001$). In the low-risk group, all scoring systems displayed similar sensitivity and NPV, and there were no differences in rates of missed appendicitis (e.g., patients with appendicitis but with a low risk according to the scores' cut-off values) that ranged from 7 to

Table 2 Diagnostic values and clinical outcome of prediction scores for pediatric appendicitis according to the published cut-off points for all cases of appendicitis

	PAS	AIR		Alvarado		pARC	
		Low	High	Low	High	Low	High
Sensitivity (%)	95.3 (90.3–98.0)	88.1 (81.6–92.6)	27.8 (21.0–35.8)	96.7 (92.0–98.8)	84.1 (77.1–89.4)	97.2 (91.5–99.3)	39.8 (30.7–49.7)
Specificity (%)	51.5 (43.7–59.2)	77.8 (70.6–83.7)	98.2 (94.4–99.5)	33.5 (26.5–41.3)	70.1 (62.4–76.8)	41.3 (31.3–52.1)	98.9 (93.2–99.9)
PPV (%)	64.0 (57.3–70.2)	78.2 (71.1–84.0)	93.3 (80.7–98.3)	56.8 (50.5–62.9)	71.8 (64.4–78.1)	66.0 (58.1–73.2)	97.7 (86.5–99.9)
NPV (%)	92.5 (84.6–96.7)	87.8 (81.2–92.4)	60.0 (54.0–65.9)	91.8 (81.2–96.9)	83.0 (75.5–88.6)	92.7 (79.0–98.1)	58.3 (50.2–66.1)
Missed appendicitis (*n* %)	7 (8)	18 (12)		5 (8)		3 (7)	
No appendicitis (*n* %)	81 (36)		3 (7)[a]		50 (28)		1 (2)[b]
Negative appendectomy (*n* %)	8 (5)		0 (0)		3 (2)		1 (2)

Abbreviations: AIR, appendicitis inflammatory response; NPV, negative predictive value; PAS, pediatric appendicitis score; pARC, pediatric appendicitis risk calculator; PPV, positive predictive value.
Note: Sensitivity, specificity, PPV, and NPV presented as % (95% confidence interval), missed appendicitis and no appendicitis presented as *n* (%).
[a]$p < 0.05$ when comparing the AIR score to the PAS and Alvarado score through Chi-square test.
[b]$p < 0.05$ when comparing the pARC to the PAS and Alvarado score.
PAS: low = 0–5 and high = 6–10; AIR score: low = 0–4 and high = 9–12; Alvarado score low = 0–4 and high = 7–10; pARC: low = 0–14% and high = 85–100%.

12% (►Table 2). Very few of these patients had a complicated appendicitis or suffered any complications after surgery. When evaluating the scoring systems' performance in cases of only complicated appendicitis, the sensitivity and NPV of all scoring systems increased, the specificity remained unchanged, and the PPV decreased (►Supplementary Table S2 [available in the online version]).

The pARC assigned a higher proportion of patients to the intermediate-risk group (57%) than the AIR score (39%) and Alvarado score (25%; $p < 0.001$). The AIR score assigned a higher proportion of patients to the low-risk group (47%) than the other three scoring systems ($p < 0.001$). The PAS and Alvarado score assigned a greater proportion of patients to the high-risk group (71 and 56%, respectively) than the AIR score and the pARC (14 and 22%; $p < 0.001$; ►Table 3).

Gender and Age Analysis

Among boys, the AIR score had significantly lower false positive rate compared to the PAS, and the pARC had lower false positive rate compared with the PAS ($p = 0.005$) and Alvarado score ($p = 0.01$). Among girls, the AIR score had significantly lower false positive rate than the PAS ($p = 0.002$) and Alvarado score ($p = 0.02$), and the pARC had lower false positive rate than the PAS ($p = 0.011$). Overall, there was no difference in NPV or missed appendicitis rates between girls and boys. In the group of 10 to 14 years old, the pARC outperformed both the PAS and Alvarado score in terms of false positive rate ($p = 0.001$ and $p = 0.03$, respectively). The AIR score had lower false positive rate than the PAS ($p = 0.026$), but the differences were not significant when compared with the pARC ($p = 0.65$) and Alvarado score ($p = 0.19$). The PAS had a significantly higher false positive rate than the Alvarado score ($p = 0.028$). Among the 0 to 4 years old, no significant differences in false positive rate were shown. Neither of the groups displayed a difference in rates of missed appendicitis. However, the rates of missed appendicitis generally seemed to increase with age. The NPV seemed to decrease with increasing age. (►Supplementary Table S3 [available in online version only]).

Receiver Operating Characteristics Curve and Decision Curve Analysis

In the total cohort, AUC values from the ROC curves of the different scoring systems were similar, ranging from 0.90 for the pARC and 0.86 for the Alvarado score. In the group with complicated appendicitis, the PAS and Alvarado score had the lowest AUCs (0.91) and the AIR score the highest (0.94). In the different age and gender groups, the AUC did not differ strongly between the four prediction scores (►Fig. 2).

In the decision curve analysis, the AIR score had a better net benefit than the PAS and Alvarado score at most threshold probabilities, except around 0.90, where the net benefit of the PAS was higher, and above 0.95, where the Alvarado score was higher (►Fig. 3). When including the pARC, and thus looking only at the patients aged 4 to 15 years, the pARC displayed the highest net benefit almost throughout the entire span of threshold probabilities. Between thresholds of 0.86 and 0.89, the net benefit of the AIR score was higher (►Fig. 3).

Discussion

This is the first prospective comparison and validation of the PAS, AIR score, Alvarado score, and the pARC in a pediatric population focusing on the scores' overall performance, as

Table 3 Distribution of outcomes in different risk categories according to the pediatric appendicitis score, appendicitis inflammatory response score, Alvarado score and pediatric appendicitis risk calculator in pediatric patients with and without appendicitis

	PAS	AIR	Alvarado	pARC	*p*-Value
Low risk					
Total cohort	93 (29)	148 (47)	61 (19)	41 (21)	<0.001
No appendicitis	86 (51)	130 (78)	56 (33.5)	38 (41)	<0.001
Appendicitis	7 (5)	18 (12)	5 (3)	3 (3)	0.02
Complicated appendicitis	1 (1)	4 (6)	2 (3)	1 (2)	0.499
Intermediate risk					
Total cohort		125 (39)	80 (25)	115 (57)	<0.001
No appendicitis		34 (20)	61 (36.5)	53 (58)	<0.001
Appendicitis		91 (60)	19 (13)	62 (57)	<0.001
Complicated appendicitis		30 (45)	4 (6)	22 (50)	<0.001
High risk					
Total cohort	225 (71)	45 (14)	177 (56)	44 (22)	<0.001
No appendicitis	81 (49)	3 (2)	50 (30)	1 (1)	<0.001
Appendicitis	144 (95)	42 (28)	127 (84)	43 (40)	<0.001
Complicated appendicitis	66 (99)	33 (49)	61 (91)	21 (48)	<0.001

Abbreviations: AIR, appendicitis inflammatory response; PAS, pediatric appendicitis score; pARC, pediatric appendicitis risk calculator.
Note: Values presented as *n* (%).
PAS: low = 0–5 and high 6–10; AIR score: low = 0–4 and high = 9–12; Alvarado score low = 0–4 and high = 7–10; pARC: low = 0–14% and high = 85–100%.

well as in subgroups of gender and different ages. Overall, the AIR score and the pARC had a higher diagnostic accuracy compared with the PAS and Alvarado score.

The aim of clinical prediction scores is to predict clinical outcome. In addition to identifying patients with appendicitis, it is important to evaluate the prediction scores' ability to rule out appendicitis to avoid unnecessary investigations or surgery. Hence, the false positive rate and the missed appendicitis rate are valid measurements of the scoring systems' diagnostic performance. Even though the AUROC of all clinical prediction scores were high, our results demonstrated considerable differences in clinical outcome if the scores were used as proposed by their original authors, where the AIR score's and the pARC's ability to diagnose appendicitis accurately were significantly greater than that of the Alvarado score and the PAS. No differences in the scoring systems' ability to exclude appendicitis were found, and all scoring systems displayed more or less unsatisfactory numbers of missed appendicitis. One possible explanation for this is that the prediction scores might have been calculated at too early a point in time and the patients' scores would have progressed along with the course of the disease had the clinical examination and laboratory tests been repeated. It has also been suggested that phlegmonous appendicitis can be a self-limiting disease that can sometimes resolve spontaneously.[29] One could, therefore, claim that a high rate of missed diagnosis does not necessarily mean that the patient will suffer from complications of no or delayed diagnosis.

The PAS was constructed through a prospective study of 1,170 patients aged 4 to 15 years and showed excellent diagnostic accuracy in the original study.[10] However, several studies have failed to reproduce these results.[12,13,30,31] The AIR score was developed through a prospective study of 545 patients of all ages and focused mainly on identifying patients with complicated appendicitis.[11] Although developed for all ages, it showed a high discriminating power—exceeding the ones of the PAS and Alvarado score—when evaluated retrospectively in children[19] and in a prospective randomized controlled trial in all age groups.[32] The Alvarado score was developed through a retrospective study of 305 hospitalized patients, both children and adults, with abdominal pain suggestive of appendicitis. The original report did not explicitly present the test's performance,[9] but an evaluating study calculated the diagnostic values according to available data from the original study.[19] Further evaluating studies have shown varying results.[14,31,33] When comparing the parameters of the different scoring systems, the AIR score and the pARC put more emphasis on objective findings in the clinical examination and in laboratory results, while the PAS and Alvarado score focus more on medical history. One might hypothesize that this is part of the explanation as to why the AIR score and the pARC perform better in children since they do not rely on the patient's ability to narrate the course of their illness—which can be a challenging task for a pediatric (and sometimes nonverbal) patient.

This study is the first of its kind to incorporate the method of net benefit and decision curves in children with appendicitis. Net benefit is an analytical measure that puts benefit and harm on the same scale by comparing specified threshold probabilities, that is, comparing the scoring systems' performance under different scenarios by varying the risk of missed appendicitis or negative laparotomy one is willing to accept. It is thereby a way

	N	PAS	AIR	Alvarado	pARC
All appendicitis	318	0.87 (0.83–0.91)	0.88 (0.85–0.92)	0.86 (0.82–0.90)	0.90 (0.86–0.95)
Complicated appendicitis	234	0.91 (0.87–0.95)	0.94 (0.90–0.97)	0.91 (0.86–0.95)	0.92 (0.87–0.97)
Sex					
Boys	176	0.88 (0.83–0.93)	0.88 (0.83–0.93)	0.87 (0.82–0.93)	0.91 (0.85–0.96)
Girls	142	0.84 (0.77–0.91)	0.87 (0.80–0.94)	0.84 (0.76–0.91)	0.88 (0.81–0.95)
Age					
0–4 years	40	0.90 (0.79–1)	0.92 (0.82–1)	0.90 (0.79–1)	–
5–9 years	120	0.88 (0.82–0.94)	0.89 (0.83–0.95)	0.87 (0.81–0.94)	0.91 (0.85–0.98)
10–14 years	158	0.85 (0.78–0.91)	0.89 (0.83–0.94)	0.85 (0.79–0.91)	0.90 (0.85–0.96)

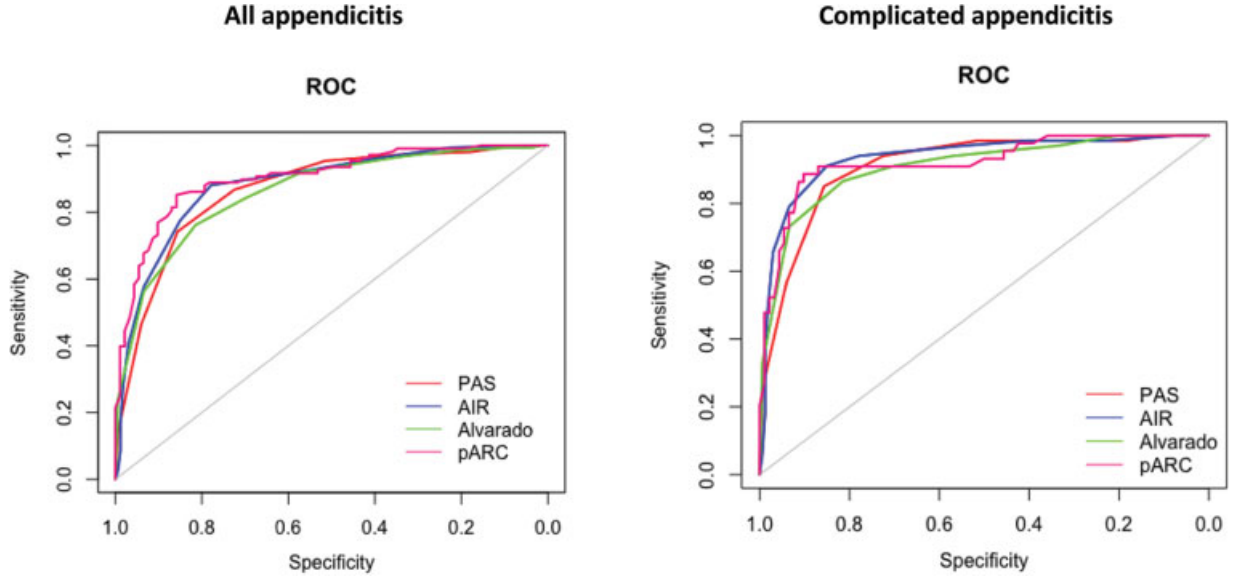

Fig. 2 Receiver operating characteristics curve analyses with area under the curve for different prediction scores in 318 children with suspected appendicitis. Values presented as area under the curve (AUC) (95% Confidence Interval). PAS, Pediatric Appendicitis Score; AIR, appendicitis inflammatory response; pARC, pediatric Appendicitis Risk Calculator.

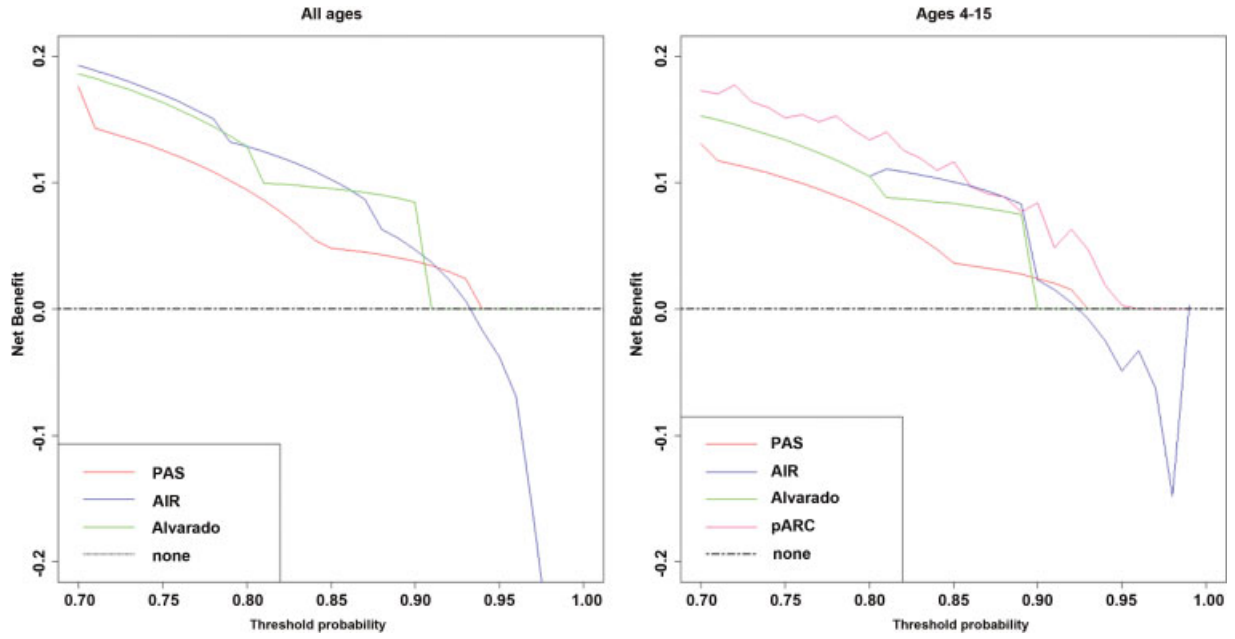

Fig. 3 Decision curves (threshold probabilities 70–100%) for the pediatric appendicitis score, appendicitis inflammatory response score, Alvarado score and the pediatric appendicitis risk calculator for all ages and 4 to 15 years old. In the right graph, the threshold probabilities between 0.7 and 0.8, the lines of AIR score and Alvarado score are overlapping. PAS, Pediatric Appendicitis Score; AIR, appendicitis inflammatory response; pARC, pediatric Appendicitis Risk Calculator.

to integrate and quantify the clinical consequences of the different scoring systems in our analyses (i.e., the benefit of adequately diagnosing a case of appendicitis vs. misdiagnosing a child and wrongfully sending him or her home or to the operating theater). Thus, the unit of net benefit is the number of true positives/patients. Hence, if the difference in net benefit between two scores is 0.05, the better score will result in five (of 100) more patients with appendicitis being correctly identified without an increase in misdiagnosis/negative appendectomies. The method further described by Vickers et al.[29] The decision curve analysis in this study shows a higher net benefit of the pARC compared with the other scoring systems almost throughout the threshold span between 0.7 and 1. When excluding the pARC, and thus evaluating net benefit over the entire age span (0–14 years), the AIR score has a higher net benefit than the PAS and Alvarado score. This further suggests a superiority of the AIR score and the pARC over the other two scoring systems. A weakness of the pARC is that it placed the majority of the patients in the intermediate risk group. The AIR score assigned a large proportion of the patients to the low-risk group. Even so, the rates of missed appendicitis did not differ significantly between the scoring systems. In conclusion, these results strongly suggest that the AIR score and the pARC are superior to the PAS and Alvarado score, supporting the results of previous studies.[19,20,32] Our recommendations are that the PAS and Alvarado score should be used with great caution in a clinical setting and barely in further research in the field.

It has been shown that imaging enhances the performance of scores.[34] We consider imaging to be a crucial part of the clinical work-up in children with suspected appendicitis, and our study findings could help delineating its greatest benefit to patients stratified to the intermediate risk group. In our center, diagnostic imaging for all patients would result in an excess demand of radiological examinations, possibly with a risk of unnecessarily diagnosing some patients with mild symptoms and possibly self-limiting phlegmonous appendicitis while delaying necessary surgical care for those with unequivocal clinical findings. US is the first choice and is a reliable tool, especially in experienced hands,[34] but even under such circumstances, it is not always conclusive due to difficulties in visualizing appendix.[33] CT could certainly be an alternative for selected patients with intermediate risk of appendicitis[34] but should be used with caution in children, considering the radiation-associated long-term risk of cancer.[35] In children with nonconclusive US and intermediate risk, the prevalence of appendicitis is often low.[36] However, under uncertain circumstances, it should be remembered that MRI is a noninvasive modality with high diagnostic accuracy for appendicitis in children[37,38] even if for many centers, such as ours, the lack of availability remains a practical limitation. Children with lower risk categories and nonconclusive US often have negative or unequivocal results also on their MRI.[39]

Another alternative or complement to radiologic imaging in the children with intermediate risk of appendicitis is active observation and repeated scoring. In hospital delay to appendectomy does not increase the rates of perforation or complications.[40–42] Children stratified to the low- and high-risk groups according to pARC and AIR score should be sent home or taken to the operating theater, respectively.

Limitations

The current study had a relatively smaller number of patients compared with other studies. However, unlike many other studies, it is a prospective evaluation of the scoring systems. Another limitation is the lack of data for symptom duration in a substantial part of the cohort, reducing the pARC cohort. A power calculation was not performed regarding the subgroup analyses, but considering the relatively small sample size, these are probably underpowered.

Only children under 15 years were included due to the cut-off limit at all Swedish pediatric surgery centers. Only children referred to the pediatric surgeon on call were included. The referral could come from a pediatrician at the ED, a family practitioner or a nurse at the pediatric ED. The study, therefore, focuses mainly on patients whose original risk of appendicitis was regarded as high. Patients with low suspicion of appendicitis, who were sent home and did not return to the hospital, were assumed not suffering from appendicitis. These patients might have a spontaneously resolving appendicitis and therefore misclassified. Further, the study was confined to the ED, and no data on repeated scoring were gathered. This could be an interesting topic for future studies with comparison between different scores. A strength of the study was that the cohort was stratified according to sex and age of the patients. However, obesity was not registered in our database, yet one could hypothesize that clinical prediction scores with emphasis on findings from the abdominal examination (clinical signs of peritonitis) might bias obese children to a lower risk group, or at least to a lower score, due to masked symptoms. Future studies should elucidate if obesity results in higher rates of false negative scoring and unnecessary appendectomies.[43]

Further, the study was confined to the ED, and no data on repeated scoring were gathered. This could be an interesting topic for future studies with comparison between different scores.

Conclusion

The AIR score and the pARC have an overall higher diagnostic accuracy in children with suspected appendicitis compared with the PAS and Alvarado score. Therefore, we recommend these scores when evaluating a child with suspicion of appendicitis in the ED. Safely ruling out a diagnosis of appendicitis through clinical prediction scores remains a challenge.

Funding

This study received its financial support from Bengt Ihres Foundation.

Conflict of Interest

None declared.

Acknowledgment

We would like to thank statistician Anna Åkesson, who helped us with the decision curve analysis.

References

1 Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. Am J Epidemiol 1990;132(05):910–925
2 Lee JH, Park YS, Choi JS. The epidemiology of appendicitis and appendectomy in South Korea: national registry data. J Epidemiol 2010;20(02):97–105
3 GlobalSurg Collaborative. Mortality of emergency abdominal surgery in high-, middle- and low-income countries. Br J Surg 2016;103(08):971–988
4 Marzuillo P, Germani C, Krauss BS, Barbi E. Appendicitis in children less than five years old: a challenge for the general practitioner. World J Clin Pediatr 2015;4(02):19–24
5 Horwitz JR, Gursoy M, Jaksic T, Lally KP. Importance of diarrhea as a presenting symptom of appendicitis in very young children. Am J Surg 1997;173(02):80–82
6 Salö M, Ohlsson B, Arnbjörnsson E, Stenström P. Appendicitis in children from a gender perspective. Pediatr Surg Int 2015;31(09):845–853
7 Graff L, Russell J, Seashore J, et al. False-negative and false-positive errors in abdominal pain evaluation: failure to diagnose acute appendicitis and unnecessary surgery. Acad Emerg Med 2000;7(11):1244–1255
8 Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. Ann Emerg Med 1999;33(04):437–447
9 Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 1986;15(05):557–564
10 Samuel M. Pediatric appendicitis score. J Pediatr Surg 2002;37(06):877–881
11 Andersson M, Andersson RE. The appendicitis inflammatory response score: a tool for the diagnosis of acute appendicitis that outperforms the Alvarado score. World J Surg 2008;32(08):1843–1849
12 Goldman RD, Carter S, Stephens D, Antoon R, Mounstephen W, Langer JC. Prospective validation of the pediatric appendicitis score. J Pediatr 2008;153(02):278–282
13 Bhatt M, Joseph L, Ducharme FM, Dougherty G, McGillivray D. Prospective validation of the pediatric appendicitis score in a Canadian pediatric emergency department. Acad Emerg Med 2009;16(07):591–596
14 Bond GR, Tully SB, Chan LS, Bradley RL. Use of the MANTRELS score in childhood appendicitis: a prospective study of 187 children with abdominal pain. Ann Emerg Med 1990;19(09):1014–1018
15 Macklin CP, Radcliffe GS, Merei JM, Stringer MD. A prospective evaluation of the modified Alvarado score for acute appendicitis in children. Ann R Coll Surg Engl 1997;79(03):203–205
16 van Amstel P, Gorter RR, van der Lee JH, Cense HA, Bakx R, Heij HA. Ruling out appendicitis in children: can we use clinical prediction rules? J Gastrointest Surg 2019;23(10):2027–2048
17 de Castro SM, Ünlü C, Steller EP, van Wagensveld BA, Vrouenraets BC. Evaluation of the appendicitis inflammatory response score for patients with acute appendicitis. World J Surg 2012;36(07):1540–1545
18 Kollár D, McCartan DP, Bourke M, Cross KS, Dowdall J. Predicting acute appendicitis? A comparison of the Alvarado score, the appendicitis inflammatory response score and clinical assessment. World J Surg 2015;39(01):104–109
19 Macco S, Vrouenraets BC, de Castro SM. Evaluation of scoring systems in predicting acute appendicitis in children. Surgery 2016;160(06):1599–1604
20 Kharbanda AB, Vazquez-Benitez G, Ballard DW, et al. Development and validation of a novel pediatric appendicitis risk calculator (pARC). Pediatrics 2018;141(04):e20172699
21 Carr NJ. The pathology of acute appendicitis. Ann Diagn Pathol 2000;4(01):46–58
22 Emil S, Gaied F, Lo A, et al. Gangrenous appendicitis in children: a prospective evaluation of definition, bacteriology, histopathology, and outcomes. J Surg Res 2012;177(01):123–126
23 St Peter SD, Sharp SW, Holcomb GW III, Ostlie DJ. An evidence-based definition for perforated appendicitis derived from a prospective randomized trial. J Pediatr Surg 2008;43(12):2242–2245
24 Herd ME, Cross PA, Dutt S. Histological audit of acute appendicitis. J Clin Pathol 1992;45(05):456–458
25 Pieper R, Kager L, Näsman P. Clinical significance of mucosal inflammation of the vermiform appendix. Ann Surg 1983;197(03):368–374
26 Campbell JS, Fournier P, Dasilva T. When is the appendix normal? A study of acute inflammations of the appendix apparent only upon histologic examination. Can Med Assoc J 1961;85:1155–1157
27 Dingemann J, Metzelder M, Kuebler JF, Ure B. Laparoscopy for suspected appendicitis in children: May a macroscopically normal appendix be left in situ? Eur J Pediatr Surg 2009;19(03):153–156
28 Pogorelić Z, Rak S, Mrklić I, Jurić I. Prospective validation of Alvarado score and Pediatric Appendicitis Score for the diagnosis of acute appendicitis in children. Pediatr Emerg Care 2015;31(03):164–168
29 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6
30 Schneider C, Kharbanda A, Bachur R. Evaluating appendicitis scoring systems using a prospective pediatric cohort. Ann Emerg Med 2007;49(06):778–784, 784.e1
31 Mandeville K, Pottker T, Bulloch B, Liu J. Using appendicitis scores in the pediatric ED. Am J Emerg Med 2011;29(09):972–977
32 Andersson M, Kolodziej B, Andersson RE, Group SS; STRAPPSCORE Study Group. Randomized clinical trial of appendicitis inflammatory response score-based management of patients with suspected appendicitis. Br J Surg 2017;104(11):1451–1461
33 Escribá A, Gamell AM, Fernández Y, Quintillá JM, Cubells CL. Prospective validation of two systems of classification for the diagnosis of acute appendicitis. Pediatr Emerg Care 2011;27(03):165–169
34 Dingemann J, Ure B. Imaging and the use of scores for the diagnosis of appendicitis in children. Eur J Pediatr Surg 2012;22(03):195–200
35 Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet 2012;380(9840):499–505
36 Löfvenberg F, Salö M. Ultrasound for appendicitis: performance and integration with clinical parameters. BioMed Res Int 2016;2016:5697692
37 Kulaylat AN, Moore MM, Engbrecht BW, et al. An implemented MRI program to eliminate radiation from the evaluation of pediatric appendicitis. J Pediatr Surg 2015;50(08):1359–1363
38 Aspelund G, Fingeret A, Gross E, et al. Ultrasonography/MRI versus CT for diagnosing appendicitis. Pediatrics 2014;133(04):586–593
39 Sincavage J, Buonpane C, Benyamen B, et al. Alvarado scores predict additive value of magnetic resonance imaging in workup of suspected appendicitis in children. J Surg Res 2019;244:42–49
40 Yardeni D, Hirschl RB, Drongowski RA, Teitelbaum DH, Geiger JD, Coran AG. Delayed versus immediate surgery in acute appendicitis: do we need to operate during the night? J Pediatr Surg 2004;39(03):464–469
41 Schnüriger B, Laue J, Kröll D, Inderbitzin D, Seiler CA, Candinas D. Introduction of a new policy of no nighttime appendectomies: impact on appendiceal perforation rates and postoperative morbidity. World J Surg 2014;38(01):18–24
42 Andersson RE. Does delay of diagnosis and treatment in appendicitis cause perforation? World J Surg 2016;40(06):1315–1317
43 Kutasy B, Hunziker M, Laxamanadass G, Puri P. Increased incidence of negative appendectomy in childhood obesity. Pediatr Surg Int 2010;26(10):959–962