World J Surg (2017) 41:1769–1781
DOI 10.1007/s00268-017-3926-6

SCIENTIFIC REVIEW

Clinical Prediction Rules for Appendicitis in Adults: Which Is Best?

Malsha Kularatna[1] · Melanie Lauti[2] · Cheyaanthan Haran[2] · Wiremu MacFater[2] · Laila Sheikh[1] · Ying Huang[3] · John McCall[4] · Andrew D. MacCormick[1,2]

Published online: 3 March 2017

Abstract

Background Clinical prediction rules (CPRs) provide an objective method of assessment in the diagnosis of acute appendicitis. There are a number of available CPRs for the diagnosis of appendicitis, but it is unknown which performs best.

Aim The aim of this study was to identify what CPRs are available and how they perform when diagnosing appendicitis in adults.

Method A systematic review was performed in accordance with the PRISMA guidelines. Studies that derived or validated a CPR were included. Their performance was assessed on sensitivity, specificity and area under curve (AUC) values.

Results Thirty-four articles were included in this review. Of these 12 derived a CPR and 22 validated these CPRs. A narrative analysis was performed as meta-analysis was precluded due to study heterogeneity and quality of included studies. The results from validation studies showed that the overall best performer in terms of sensitivity (92%), specificity (63%) and AUC values (0.84–0.97) was the AIR score but only a limited number of studies investigated at this score. Although the Alvarado and Modified Alvarado scores were the most commonly validated, results from these studies were variable. The Alvarado score outperformed the modified Alvarado score in terms of sensitivity, specificity and AUC values.

Conclusion There are 12 CPRs available for diagnosis of appendicitis in adults. The AIR score appeared to be the best performer and most pragmatic CPR.

✉ Malsha Kularatna
malsh87@hotmail.com;
Malsha.Kularatna@middlemore.co.nz

✉ Andrew D. MacCormick
andrew.maccormick@middlemore.co.nz

[1] Department of Surgery, Middlemore Hospital, Counties Manukau Health, Auckland, New Zealand

[2] Department of Surgery, South Auckland Clinical Campus, University of Auckland, Auckland, New Zealand

[3] Section of Epidemiology and Biostatistics, School of Population Health, University of Auckland, Auckland, New Zealand

[4] Department of Surgery, University of Otago, Dunedin, New Zealand

Introduction

Appendicitis is one of the most common acute surgical illnesses with a lifetime prevalence of one in seven [1]. It continues to be clinically challenging to diagnose as it mimics a variety of other pathologies, especially in females [1]. Diagnosis is usually based on the clinical history, examination, correlated with laboratory and imaging investigations. The final diagnosis may require diagnostic laparoscopy, which itself is not without risk.

Clinical prediction rules (CPRs) are one of the most commonly described tools used to aid the diagnosis of appendicitis. CPRs are derived from systematic clinical observations and aim to reduce uncertainty by standardising

the collection and interpretation of clinical data [2, 3]. They have been shown to provide a more objective method of assessment and standardisation of care for patients with suspected appendicitis, thereby reducing the number of unnecessary operations and patient exposure to radiation [4]. Although a plethora of CPRs exist for the diagnosis of appendicitis, it is unclear which of these performs most reliably.

The aim of this systematic review was to identify all current CPRs for the diagnosis of appendicitis in adults and assess their performance.

Methods

Search strategy

This study was completed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [5]. A comprehensive literature search was performed in MEDLINE, EMBASE, Pubmed and Cochrane Central Register of Controlled Trials databases from inception to February 2016. The search strategy is outlined in Table 1. Studies were restricted to English language and humans only. The reference list of all included and relevant review articles were also searched to identify further potentially eligible manuscripts.

Inclusion and exclusion criteria

Only studies that derived or validated the impact of a CPR for use in adults presenting with right lower quadrant (RLQ) pain, right iliac fossa (RIF) pain or abdominal pain suspicious of appendicitis were included. For the purposes of this study, a CPR was defined as one that [2, 3, 6];

- Had three or more predictive variables obtained from the history, physical exam and simple diagnostic tests
- Provided a probability of an outcome or suggested a diagnostic/therapeutic course of action.
- Was not a decision analysis, decision tree or practice guideline.

Table 1 Search terms used

Search terms
(appendicitis or appendicectom$ or appendectom$ or right iliac fossa pain or rif pain or right lower quadrant pain or rlq pain)
AND
(nomogram$ or algorithm$ or guideline$ or decision or checklist$ or score or scores or scoring or probabilit$ or protocol$ or pathway$ or rule or rules or predictive)

Both CPR derivation and validation studies were included. A derivation study was defined as a study that described the method of how a new CPR was formed and explained how it should be applied in a clinical setting. A validation study assessed performance of an existing CPR by ascertaining the sensitivity, specificity and/or AUC. If derivation studies included an internal validation component, the validation component was excluded from the validation study analysis due to a high risk of potential bias [2].

Exclusion criteria for derivation studies

When assessing articles which derived a CPR, studies that modified an existing scoring system in order to generate a new scoring system were included if the new parameters and cut-off values were clearly defined. There was no restriction on study design. Scores which were derived for use solely in paediatric, elderly, pregnant or single gender populations and those that did not assess the primary outcomes of appendicitis versus non-appendicitis and/or required the use of neural networks were excluded.

Exclusion criteria for validation studies

Studies that validated CPRs in elderly populations, a single gender only or included patients younger than 14 years, were excluded. Studies that looked at a subset of the scoring system or only patients that had imaging were also excluded. Three studies that did not state the age of the participants were also excluded. Studies that included patients younger than 14 years of age with a separate analysis for adults were included.

Selection of studies

The initial search, title and abstract screen were performed independently by MK and CH. Any discrepancy between the two reviewers was discussed with senior author AM. A total of 224 articles were identified as relevant and underwent full text review by authors MK, CH, ML, WM and LS.

Data extraction and statistical analysis

Derivation studies

Data from studies describing derivation of a CPR were extracted using a standard pro forma. Study characteristics, derivation methodology, scoring systems characteristics (e.g. use of weighting, positive versus negative scoring)

and variables comprising the CPR were recorded for each study.

Validation studies

Extracted data from CPR validation studies were also extracted using a standard proforma. These included study design, results obtained for sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), likelihood ratios and AUCs values from receiver operating curve (ROC) analysis.

When more than two cut-off values were evaluated for the prediction of high risk of having appendicitis, only the cut-off recommended in the original derivation paper was used for analysis. When sensitivity and specificity were not calculated in the validation studies, these were calculated from the data available using a two-by-two table by author MK and confirmed by YT. Forrest plot confidence intervals (CI) were calculated using the variance method for all studies to minimise bias [7].

Assessment of methodological quality of validation studies

The quality of included validation studies was assessed and scored using 15 pre-defined criteria by Wasson et al. (Table 2) [3]. These criteria were specifically designed to assess articles describing clinical prediction rules.

Table 2 Quality assessment criteria for validation studies based on previously defined criteria by Wasson et al. [3]

Data collected prospectively	1
Study site well described (place, department, number of centres (2/3 have to be present)	1
Rule derived/validated on all patients at risk (i.e. all patients with abdominal pain at risk of appendicitis	1
Study population well described (age, sex)	1
Outcomes studied well defined (appendicitis vs no appendicitis), confirmation of appendicitis based on histology.	1
Blinding of those assessing predictors to outcome and/or vice versa	1
Adequate follow-up of outcomes (phone call or evaluation of readmission from clinical notes for at least 30 days' post-admission)	1
Predictors evaluated defined well—uniform definition of signs and symptoms of the scoring system and agreement regarding clinical assessment methods.	1
Mathematical techniques used reported;	1–3
descriptive 1	
logistic regression 2	
cross-validation 3	
95% confidence interval reported	1
Adequate precision estimation/power calculation	1
Adequate reporting of results (report four out of six of; sensitivity, specificity, negative predictive value, positive predictive value, likelihood ratio or area under the curve)	1
Reproducible methodology	1

Results

Study selection

The initial database search identified 7696 titles, and a further 56 identified through the manual search. Of these, 4398 were potentially relevant after removal of duplicates and further screening. Following abstract review 257 papers met criteria for full text review. Of these, 12 papers describing derivation of CRPs and 22 describing validation were included. The PRISMA flow diagram is presented in Fig. 1 [5].

Derivation studies

Characteristics of CPRs derived for use in adults with suspected appendicitis demonstrated significant heterogeneity in both study population and methodology (Table 3). Among the discrepancies in methodology was the variation in statistical analyses. Three studies used univariate analysis, while seven studies used multivariate analysis (Table 3) [8–16].

The most commonly incorporated variable was the white cell count, which appeared in all 12 studies (Table 4) [8–18]. Temperature, rebound tenderness and migratory pain were the next most common across all studies (Table 4) [8–11, 16–18]. Studies that used multivariate analysis identified gender, elevated C-reactive protein, RIF pain, neutrophilia, vomiting and signs of peritonism (guarding, rigidity) as likely variables [9, 10, 13–16]. Rectal tenderness, diarrhoea and Rovsing's sign were the least commonly used variables and appeared only in CPRs that used univariate analysis [11, 12, 18].

Validation studies

The 22 included validation studies demonstrated heterogeneity with respect to study population, study design and cut-off values evaluated (Table 5). Two of the 22 studies only had AUC values available for the adult population. A scatter plot of all sensitivity and specificity values adjusted for sample size is shown in Fig. 2. A Forrest plot could only be generated for sensitivity as the number of true negatives was unable to be calculated from the majority of the studies due to incomplete follow-up of discharged patients (Fig. 3). As CIs displayed in the Forrest plot were

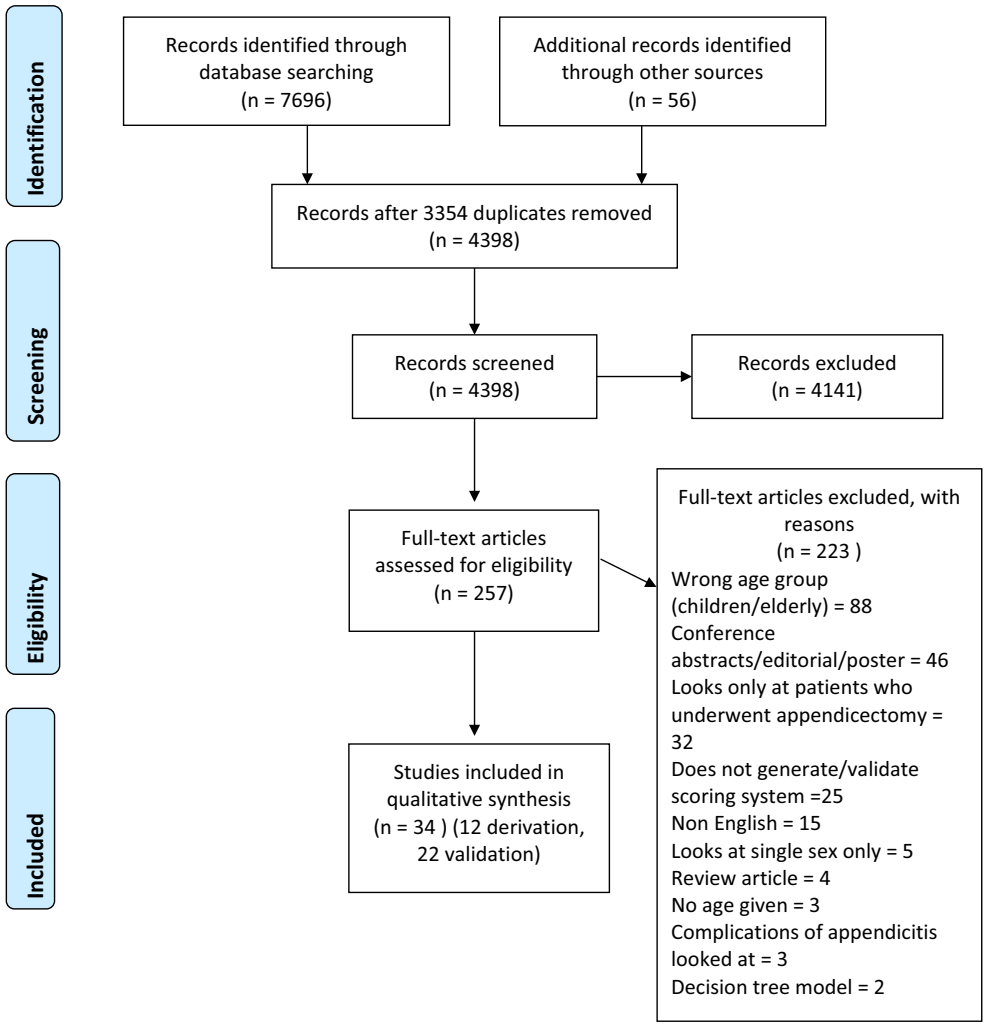

Fig. 1 Flow diagram of showing systematic inclusion of papers

calculated using the variance method, the values presented in Fig. 3 may differ to those published in the original studies due to different calculation methods. The studies published by Scott et al. (year), Erdem et al. (year) do not have CIs calculated as the sensitivity and sample size values were too similar.

The majority of studies had a quality score between six and eight, while only six studies scored ten or more out of fifteen (Table 5). Of these, the two highest quality studies validated the acute inflammatory response (AIR) and Lintula scores [19, 20].

A general trend demonstrated that at higher cut-off values, the specificity of scoring systems improved but at the expense of the sensitivity. Clinically, this means CPRs with high cut-off values are better for ruling out a diagnosis of appendicitis due to the good positive predictive value (Table 5; Figs. 2, 3). This was especially apparent in the Alvarado and AIR scores.

The most commonly validated CPR was the Alvarado score, followed by the Kalan's modified Alvarado score (Figs. 2, 3) [21–37]. The average AUC value for the Alvarado score that ranged between 0.74 and 0.88 was higher than the modified Alvarado score which had an AUC of 0.69 from a single study (Table 5).

The sensitivity of the Alvarado score ranged from 67.65 to 96.3%, while specificity ranged from 58.18 to 89.39% when the originally recommended cut-off of seven was used. This variability was also seen in Kalan's modified Alvarado score where the sensitivity ranged from 53.8 to 97.6%, and specificity ranged from 28.57 to 80% for the same cut off value. This variability remained regardless of the quality of the studies (Table 5).

The AIR, Raja Isteri Pengiran Anak Saleha Appendicitis (RIPASA), Ohmman, Lintula and Eskelinen scores each had only a single validation study from which sensitivity and specificity could be obtained [19, 20, 38].

Table 3 Characteristics of studies and the clinical prediction rules from derivation studies

Author, study dates, name of CPR	Number of patients	Study design	Mean age (years)	Patient population	Derivation technique	Number of variables	Range of score	Weighting	Application of score
Alvarado [8] 1975–1976	305	Retrospective	25.3	Abdominal pain suggestive of appendicitis	Univariate statistical analysis	8	0–10	Based on diagnostic weight	0–3 = discharge 4–6 = observe 7–10 = operation
Andersson [9] 1992–1993, 1997 (AIR)	316	Retrospective	25.9	Suspected acute appendicitis	Multivariate logistic regression	7	0–12	Based on regression coefficient	>8 = high probability 5–8 = indeterminate <5 = low probability
Andersson [10] Dec 2003–Aug 2005 (modified AIR score)	432	Prospective	21	Presenting to ED with lower abdominal pain <5 days	Multivariate logistic regression	13	0–14	Based on regression coefficient	≤5 = discharge 6–9 = observe 10–14 = theatre
Christian [17] 1988–1989	58 cases, 59 controls	Prospective case–control study	24	Suspected acute appendicitis	Derivation not described	5	0 to 5	None	>= 4 = Appendicitis <4 = not appendicitis
Fenyo, [11] 1982	259	Prospective	Can be used in adults but no mean age given	RLQ pain	Univariate statistical analysis	19	−70 to 70	Based on likelihood ratio of each variable	>12 = laparotomy −16 to 12 = re-evaluate <−16 = non-operative
Eskelinen, [49] 1978–1984	1333	Prospective	38	Acute abdominal pain	Multivariate logistic regression	6	33.8–67.6	Based on regression coefficient	>57 = appendicectomy 50–57 = non-defined <50 = no appendicectomy
Goh, [18] 2006	238	Prospective review of retrospective data	Can be used in adults but no mean age given	Suspected acute appendicitis	Two variables removed from "Modified Alvarado score"	5	0 to 7	As per Alvarado score	As per Alvarado
Jahn, [12] 1990–1992	222	Prospective study	27	Suspected acute appendicitis from 0800- midnight	Univariate statistical analysis	11	−75 to 54	Based on likelihood ratio of each variable	>16 = High risk -20 – 16 = Intermediate risk <-20 = Low risk
Lindberg, [13] dates not specified	746	Prospective	Can be used in adults but no mean age given	Acute abdominal pain	Multivariate logistic regression	10	−84 to 66	Based on likelihood ratio of each variable	>0 = High risk -26 – 0 = Grey zone <-26 = Low risk

Table 3 continued

Author, study dates, name of CPR	Number of patients	Study design	Mean age (years)	Patient population	Derivation technique	Number of variables	Range of score	Weighting	Application of score
Sammalkorpi, [14] 2014	829	Prospective	Median 32 Range 25–47	Suspected acute appendicitis or pain in RLQ	Multivariate logistic regression	7	4–23	Based on regression coefficient	≥16 = high 11–15 = intermediate 0–10 = low
Tzanakis, [15] 1998 - 2001	303	Prospective	28.3	Suspected acute appendicitis	Multivariate logistic regression	4	0–15	Based on regression coefficient	>8 = Appendicitis <8 = Not appendicitis
Van den Broek, [16] 1994–1995	577	Prospective	>11	Suspected acute appendicitis	Multivariate logistic regression	5	0–9	Based on regression coefficient	0–3 = Observe 4–6 = Diagnostic laparoscopy

AIR acute inflammatory response, *RLQ* right lower quadrant

The AIR score showed a high sensitivity (92%) and moderate specificity (63%) at a cut-off value above five. This reverted to 20 and 97%, respectively, for a cut-off value above eight, which was the original cut-off recommended by the authors. The AUC values generated for this CPR ranged from 0.805 to 0.97, with an average value of 0.872 [20, 39, 40].

The Lintula score which was originally derived for use in paediatrics showed high performance in adults with a sensitivity of 87% and specificity of 96% [19]. The final score looked at in this study was based on repeated calculations for patients who were observed as inpatients. This is in comparison with other studies which only reported diagnostic indices based on scores at admission. There were no AUC values available for this CPR.

Erdem et al. validated the Alvarado, RIPASA, Eskelinen and Ohmann CPRs in a single study with a quality score of ten [38]. While the RIPASA, Eskelinen and Ohmann scores showed superior sensitivity and AUC values to the Alvarado scoring system, they showed poor specificity.

The pragmatic utility of these scoring systems (Table 5) demonstrated that the modified Alvarado score, Alvarado and AIR score are the most user-friendly CPRs. The use of decimal points and multiple weightings make the other scores difficult to calculate in a busy clinical setting.

Discussion

There are currently 12 published CPRs available to aid diagnosis of adults presenting with suspected appendicitis. These have been validated in 22 separate studies. The aim of this systematic review was to ascertain which of these available scores performed the best. The heterogeneity of included studies precluded the possibility of performing a meta-analysis. Based on a narrative review, however, it appears the AIR score performs the best.

Assessing the best performing CPR without meta-analysis meant narratively assessing sensitivity, specificity, AUC values, usability and the quality of available studies. Although the Lintula score performed highly in terms sensitivity and specificity, this score is difficult to use in a busy clinical setting and the comparability of the results obtained remains in question as the final score was based on repeated calculations as opposed to calculation at a single point in time. While the Eskelinen, RIPASA and Ohmann scores had good sensitivity and AUC values, they are difficult to calculate given the number of variables and range of weightings used. Thus, the overall best performer in terms of the quality of studies, results and usability was the AIR score. It is easy to calculate manually, and all parameters are easy to interpret except perhaps for the recommended subjective grading of rebound tenderness (as

Table 4 Variables incorporated within CPRs

	Alvarado [8]	Andersson [9]	Andersson [10]	Christian [17]	Eskelinen [49]	Fenyo [11]	Goh [18]	Jahn [12]	Lindberg [13]	Sammalkorpi [14]	Tzanakis [15]	Van-den Boek [16]	Total
Gender						●		●	●	●		●	5
Right iliac fossa pain		●	●		●					●			4
Other abdominal pain				●	●				●				3
Onset of pain						●		●					2
Migratory pain	●					●	●	●	●	●		●	7
Duration of pain					●	●			●				3
Intensity of pain						●		●					2
Worse with movement						●		●					2
Worse with cough						●		●	●				3
Nausea	●					●							2
Vomiting		●	●	●		●		●	●				6
Anorexia	●					●		●					3
Diarrhoea						●							1
RIF tenderness	●			●	●		●				●		5
Other abdo tenderness					●	●							2
Rebound tenderness		●	●		●	●	●		●		●	●	8
Rigidity						●	●	●	●				4
Guarding		●	●		●			●		●			5
Rectal tenderness						●							1
Rovsing's sign								●					1
Temperature	●	●	●	●		●	●					●	7
White cell count	●	●	●	●	●	●	●	●	●	●	●	●	12
Neutrophils	●	●	●							●			4
C-reactive protein (CRP)		●	●							●			3
others	Left shift		*			Subjective fever			Progression of pain				

* Chemokine C–C motif, ligand, chemokine ligand, interleukin, matrix metalloproteinase, myeloperoxidase, SAA serum amyloid

Table 5 Results for sensitivity, specificity, quality of external validation studies (ordered by score) and pragmatic utility of score

CPR validated	Cut off value evaluated	Sensitivity	Specificity	AUC values	Quality assessment score [reference]	Usability in a clinical setting
Alvarado	≥7	67.7 (79–91)	89.4		5 [36]	Pros-
		96.3 (92–100)	81.3		7 [32]	8 variables only
		Not available		18–49 = 0.88, ≥ 50 = 0.75	7 [39]	Easy to calculate with score of 1 or 2 per variable
		Not available		0.853	8 [40]	
		80 (73–87)	71.4		8 [24]	
		87.4 (85–90)	74.4	0.74	8 [28]	
		92.8 (89–97)	58.2		8 [33]	Contra;
		79 (74–84)	68.9	0.81	8 [34]	'Neutrophil left shift parameter' difficult to obtain and interpret
		85.3 (79–91)	87.2		8 [35]	
		82	75	0.818	10 [38]	
	≥ 8	44.04 (38–50)	76.8		6 [37]	
		42.2 (36–49)	79.1 (67.9–87.1		10 [27]	
		37.73 (32–43)	98.4 (96–99)		10 [26]	
AIR	≥ 5	92 (84–97)	63 (56–69)	0.84	11 [20]	Pros:
	≥ 9	20 (11–30)	97 (94–99)			7 variables Easy to calculate with score of 1,2 or 3 per variable and a total out of twelve
		Not available for adults		18–49 = 0.97, ≥50 = 0.92,	7 [39]	Cons;
		Not available for adults		0.805	8 [40]	Rebound tenderness graded as light, medium or strong which is subjective and dependent on clinical experience
RIPASA	≥ 7.5	100	28	0.857	10 [38]	Pros; Clearly defined variables Cons; 15 variables with decimal points. Difficult to calculate
Ohmann	≥ 12	96	42	0.899	10 [38]	Pros; Only 8 variables Clearly defined Cons; Difficult to calculate due different weightings with decimal points 1, 1.5, 2, 2.5 and 4.

Table 5 continued

CPR validated	Cut off value evaluated	Sensitivity	Specificity	AUC values	Quality assessment score [reference]	Usability in a clinical setting
Eskelinen	> 57	100	44	0.867	10 [38]	Pros: Only 6 variables Cons Difficult to calculate with different weightings and decimal points Total score between 33.8 and 67.6
Lintula	≥21	87 (81–94)	95		11 [19]	Pros; Only 9 variables Cons: Multiple weights placed on variables ranging from 0 to 7 with a total possible score of 32
Modified Alvarado score	> 4	72 (61–82)	54	0.69	9 [51]	Pros-
	≥7	79.7 (71–88)	56		6 [22]	7 variables only
		74 (67–81)	39		6 [23]	Easy to calculate with score of 1 or 2 per variable
		82.2 (74–90)	28.57		7 [29]	Neutrophil left shift removed
		97.6 (94–100)	66.7		7 [52]	Contra-Nil
		53.8 (46–62)	80		8 [21]	

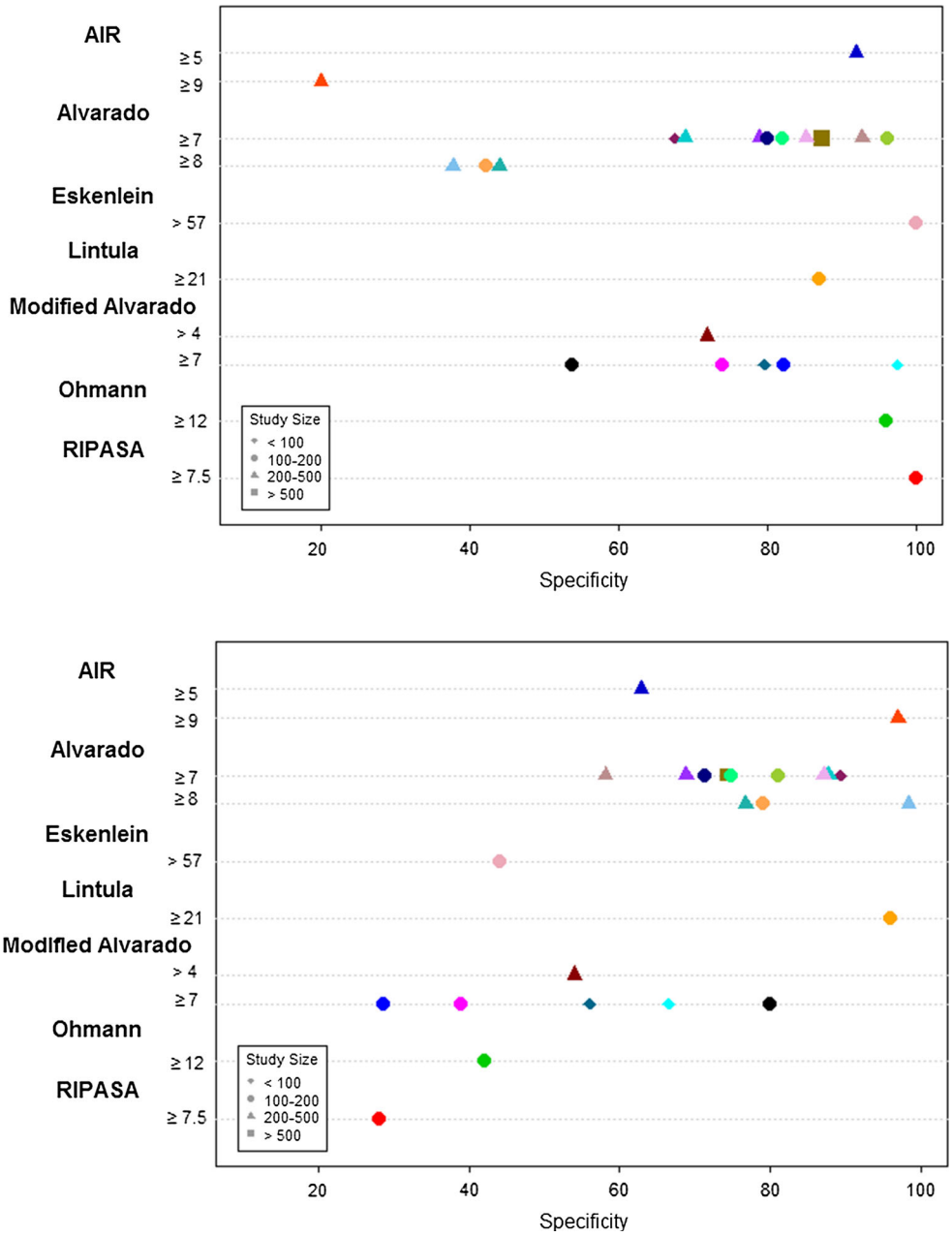

Fig. 2 Dot plot of sensitivity and specificity adjusted for sample size of each population. *Different colours* and *shapes* have been used to differentiate populations

this requires clinical experience which may be limited in junior doctors). A score of ≥five appears to be better than the originally recommended cut-off of nine as there is lower number of missed diagnoses without a significant reduction in specificity.

The majority of published validation studies evaluated the Alvarado score and Kalan's modified Alvarado score. This is probably because Alvarado was among the pioneers to generate a CPR as a diagnostic aid for appendicitis [8]. Although the Alvarado score is simple to calculate, the interpretation of left shift in neutrophils is time consuming. The results from the available studies demonstrated wide variation for both sensitivity and specificity. This variation was further emphasised as cut-off value increased and was also attributable to study design (e.g. prospective verses retrospective), variations in the characteristics of the evaluated patients, interpretation of variables of the CPR by different clinicians in different settings as well as the clinical expertise of the clinicians. While the overall sensitivity did not appear to show much variation between the Alvarado and modified Alvarado scores, the specificity appeared to be lower for the modified Alvarado score [8, 41]. Thus although the modified Alvarado score provides a more user friendly CPR, the removal of the left

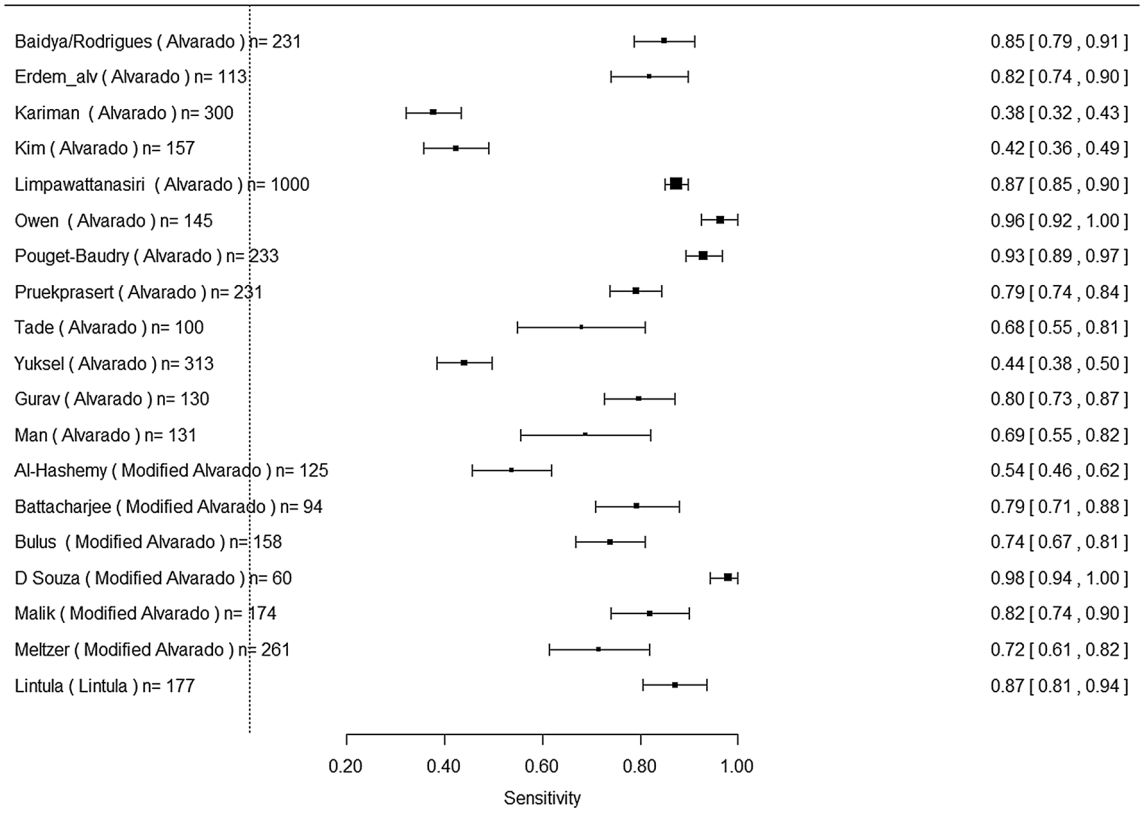

Fig. 3 Forrest plot for sensitivity. A variance calculation has been used for unbiased estimation of the confidence interval for each study. This may not be the same as those published in the original article as they may have used a different method. For two of the included studies a confidence interval could not be determined due to the sample size and sensitivity being equivalent

shift in neutrophils appeared to increase the number of false positives and was less accurate than the original CPR [8, 41].

Among derivation studies, there was wide discrepancy in the derivation methodology used. Multivariate logistic regression is known to be more reliable than using univariate analysis. This is highlighted by those CPRs derived with the multivariate method consistently identifying variables used in clinical practice [7, 42, 43] [44]. Variables such as rectal tenderness and diarrhoea that were identified in studies employing univariate analysis are seldom used clinically in the diagnosis of appendicitis [44–47]. The reliability of multivariate logistic regression analysis is further emphasised by CPRs which used this methodology such as the Lintula, AIR and Eskelinen scores showing better sensitivity and AUC values compared to the Alvarado score which was derived using univariate analysis.

Several studies investigating CPRs for appendicitis conclude that clinical judgement is comparable to CPR stratification, especially when performed by a senior surgeon [21, 27, 30, 34, 48]. While this could imply that CPRs do not improve diagnostic accuracy compared to a senior surgeon, it provides evidence that CRPs can improve diagnostic accuracy to the level of an experienced surgeon when used by less experienced staff [21, 30, 48, 49]. Given that junior staff usually undertake initial evaluation of patients with suspected appendicitis, the use of a CPR is valuable in this context. Patient care is likely to be more standardised and unnecessary exposure to radiation and invasive investigations, including laparoscopy, minimised.

The heterogeneity and quality of included studies precluded meta-analysis of available data. A further limitation was the pre-defined age criteria as many of the studies included children were excluded because the finding for children and adults could not be separated. The exclusion of non-English publications may also have excluded important validation studies done in other populations.

Conclusion

There are currently 12 CPRs available for use in adults with suspicion of appendicitis. Heterogeneity in methodology and quality of available studies precluded a meta-analysis. The AIR score performed best in terms of

sensitivity, specificity AUC values and usability but has been validated in only a small number of studies. The Alvarado and modified Alvarado were the most commonly validated CPRs, but their performance was variable. The original Alvarado score outperformed the modified Alvarado score across all three criteria (sensitivity, specificity and AUC values).

Acknowledgements Irene Zeng—Research Biostatistician, Department of knowledge and information management, Middlemore Hospital.

Authors' contribution MK designed the study, performed the initial screen and review of all articles included and composed the manuscript. ML was involved in the extraction of data from articles included and assisted with preparing the final manuscript. CH was involved in initial article screen and extraction of data from articles included. WM was involved in extraction of data from articles included. LS was involved in extraction of data from articles included. YH provided statistical expertise used to develop the dot plot. JM is a senior author and primary supervisor of masters student who completed this project and also assisted in preparing the final manuscript. AM is a senior author and principal investigator provided supervision to the co-authors. All authors have approved the manuscript as representing honest work, and each author meets the requirements for authorship.

References

1. Stephens PL, Mazzucco JJ (1999) Comparison of ultrasound and the Alvarado score for the diagnosis of acute appendicitis. Conn Med 63:137–140
2. Laupacis A, Sekar N (1997) Stiell l G. Clinical prediction rules: a review and suggested modifications of methodological standards JAMA 277:488–494
3. Wasson JH, Sox HC, Neff RK et al (1985) Clinical prediction rules. N Engl J Med 313:793–799
4. Ohle R, O'Reilly F, O'Brien KK et al (2011) The Alvarado score for predicting acute appendicitis: a systematic review. BMC Med 9:139
5. Moher D, Liberati A, Tetzlaff J et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 151:264–269
6. McGinn TG, Guyatt GH, Wyer PC et al (2000) Users' guides to the medical literature: Xxii: how to use articles about clinical decision rules. JAMA 284:79–84
7. Zhou X-H, Obuchowski NA, McClish DK (2011) Statistical analysis for meta-analysis statistical methods in diagnostic medicine. Wiley, Hoboken, pp 435–448
8. Alvarado A (1986) A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 15:557–564
9. Andersson M, Andersson RE (2008) The appendicitis inflammatory response score: a tool for the diagnosis of acute appendicitis that outperforms the Alvarado score. World J Surg 32:1843–1849. doi:10.1007/s00268-008-9649-y
10. Andersson M, Ruber M, Ekerfelt C et al (2014) Can new inflammatory markers improve the diagnosis of acute appendicitis? World J Surg 38:2777–2783
11. Fenyo G (1987) Routine use of a scoring system for decision-making in suspected acute appendicitis in adults. Acta Chirurgica Scandinavica 153:545–551
12. Jahn H, Mathiesen FK, Neckelmann K et al (1997) Comparison of clinical judgment and diagnostic ultrasonography in the diagnosis of acute appendicitis: experience with a score-aided diagnosis. Eur J Surg 163:433–443
13. Lindberg G, Fenyö G (1988) Algorithmic diagnosis of appendicitis using Bayes' theorem and logistic regression Bayesian statistics 3:665–668
14. Sammalkorpi HE, Mentula P, Leppaniemi A (2014) A new adult appendicitis score improves diagnostic accuracy of acute appendicitis—a prospective study. BMC Gastroenterol 14(1):114. doi:10.1186/1471-230X-14-114
15. Tzanakis NE, Efstathiou SP, Danulidis K et al (2005) A new approach to accurate diagnosis of acute appendicitis. World J Surg 29:1151–1156 **(discussion 1157)**
16. van den Broek WT, Bijnen BB, Rijbroek B et al (2002) Scoring and diagnostic laparoscopy for suspected appendicitis. Eur J Surg 168:349–354
17. Christian F, Christian GP (1992) A simple scoring system to reduce the negative appendicectomy rate. Ann R Coll Surg Engl 74:281–285
18. Goh PL (2010) A simplified appendicitis score in the diagnosis of acute appendicitis. Hong Kong J Emerg Med 17:230–235
19. Lintula H, Kokki H, Pulkkinen J et al (2010) Diagnostic score in acute appendicitis. validation of a diagnostic score (Lintula score) for adults with suspected appendicitis. Langenbecks Arch Surg 395:495–500
20. Scott AJ, Mason SE, Arunakirinathan M et al (2015) Risk stratification by the appendicitis inflammatory response score to guide decision-making in patients with suspected appendicitis. Br J Surg 102:563–572
21. Al-Hashemy AM, Seleem MI (2004) Appraisal of the modified Alvarado score for acute appendicitis in adults. Saudi Med J 25:1229–1231
22. Bhattacharjee PK, Chowdhury T, Roy D (2002) Prospective evaluation of modified Alvarado score for diagnosis of acute appendicitis. J Indian Med Assoc 100(310–311):314
23. Bulus H, Tas A, Morkavuk B et al (2013) Can the efficiency of modified Alvarado scoring system in the diagnosis acute appendicitis be increased with tenesmus? Wien Klin Wochenschr 125:16–20
24. Gurav P, Hombalkar N, Dhandore P et al (2013) Evaluation of right iliac fossa pain with reference to alvarado score can we prevent unnecessary appendicectomies? JKIMSU 2(2):24–29
25. Huang TH, Huang YC, Tu CW (2013) Acute appendicitis or not: facts and suggestions to reduce valueless surgery. J Acute Med 3:142–147
26. Kariman H, Shojaee M, Sabzghabaei A et al (2014) Evaluation of the Alvarado score in acute abdominal pain. Ulusal Travma ve Acil Cerrahi Dergisi 20:86–90
27. Kim K, Rhee JE, Lee CC et al (2008) Impact of helical computed tomography in clinically evident appendicitis. Emerg Med J 25:477–481
28. Limpawattanasiri C (2011) Alvarado score for the acute appendicitis in a provincial hospital. J Med Assoc Thai 94:441–449
29. Malik AA, Wani NA (1998) Continuing diagnostic challenge of acute appendicitis: evaluation through modified Alvarado score. Aust N Z J Surg 68:504–505
30. Man E, Simonka Z, Varga A et al (2014) Impact of the Alvarado score on the diagnosis of acute appendicitis: comparing clinical judgment, Alvarado score, and a new modified score in suspected appendicitis: a prospective, randomized clinical trial. Surg Endosc 28:2398–2405
31. Meltzer AC, Baumann BM, Chen EH et al (2013) Poor sensitivity of a modified Alvarado score in adults with suspected appendicitis. Ann Emerg Med 62:126–131

32. Owen TD, Williams H, Stiff G et al (1992) Evaluation of the Alvarado score in acute appendicitis. J R Soc Med 85:87–88
33. Pouget-Baudry Y, Mucci S, Eyssartier E et al (2010) The use of the Alvarado score in the management of right lower quadrant abdominal pain in the adult. J Visc Surg 147:e40–44
34. Pruekprasert P, Maipang T, Geater A et al (2004) Accuracy in diagnosis of acute appendicitis by comparing serum C-reactive protein measurements, Alvarado score and clinical impression of surgeons. J Med Assoc Thail 87:296–303
35. Rodrigues G, Rao A, Khan SA (2006) Evaluation of Alvarado score in acute appendicitis: a prospective study. Internet J Surg 9:1–5
36. Tade AO (2007) Evaluation of Alvarado score as an admission criterion in patients with suspected diagnosis of acute appendicitis. West Afr J Med 26:210–212
37. Yuksel Y, Dinc B, Yuksel D et al (2014) How reliable is the Alvarado score in acute appendicitis? Ulusal Travma ve Acil Cerrahi Dergisi 20:12–18
38. Erdem H, Cetinkunar S, Das K et al (2013) Alvarado, Eskelinen, Ohhmann and Raja Isteri Pengiran Anak Saleha Appendicitis scores for diagnosis of acute appendicitis. World J Gastroenterol 19:9057–9062
39. de Castro SM, Unlu C, Steller EP et al (2012) Evaluation of the appendicitis inflammatory response score for patients with acute appendicitis [Erratum appears in World J Surg. 2012 Sep; 36(9):2271]. World J Surg 36:1540–1545
40. Kollar D, McCartan DP, Bourke M et al (2015) Predicting acute appendicitis? A comparison of the Alvarado score, the Appendicitis Inflammatory Response Score and clinical assessment [Erratum appears in World J Surg. 2015 Jan; 39(1):112; PMID: 25315090]. World J Surg 39:104–109
41. Kalan M, Talbot D, Cunliffe WJ et al (1994) Evaluation of the modified Alvarado score in the diagnosis of acute appendicitis: a prospective study. Ann R Coll Surg Engl 76:418–419
42. Zhou X-H, Obuchowski NA, McClish DK (2011) Regression analysis for independent ROC data statistical methods in diagnostic medicine. Wiley, Hoboken, pp 261–296
43. Zhou X-H, Obuchowski NA, McClish DK (2011) Analysis of multiple reader and/or multiple test studies statistical methods in diagnostic medicine. Wiley, Hoboken, pp 297–328
44. Beasly SW (2000) Can we improve diagnosis of acute appendicitis? BMJ 321:907–908
45. Abou Merhi B, Khalil M, Daoud N (2014) Comparison of Alvarado score evaluation and clinical judgment in acute appendicitis. Med Arh 68:10–13
46. Kasimov RR, Mukhin AS (2013) Current state of acute appendicitis diagnosis. Sovremennye Tehnologii v Medicine 5:112–116
47. Wagner JM, McKinney WP, Carpenter JL (1996) Does this patient have appendicitis? JAMA 276:1589–1594
48. Yegane R, Peyvandi H, Hajinasrollah E et al (2008) Evaluation of the modified Alvarado score in acute appendicitis among iranian patients. Acta Medica Iranica 46:501–506
49. D'Souza C, Martis J, Vaidyanathan V (2013) Diagnostic efficacy of modified alvarado score over graded compression ultrasonography Nitte University. J Health Sci 3:105–108