British Journal of Medicine & Medical Research
4(2): 711-730, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Diagnostic Scores for Appendicitis: A Systematic Review of Scores' Performance

Chumpon Wilasrusmee[1,3], Thunyarat Anothaisintawee[2,3], Napaphat Poprom[1], Mark McEvoy[4], John Attia[4] and Ammarin Thakkinstian[3*]

[1]Department of Surgery, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand.
[2]Department of Family Medicine, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand.
[3]Section for Clinical Epidemiology and Biostatistics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand.
[4]Centre for Clinical Epidemiology and Biostatistics, Newcastle University, Newcastle, NSW, UK.

Authors' contributions

This work was carried out in collaboration between all authors. Author CW: Study design, study selection, data extraction, data analysis and interpretation, and draft the manuscript. Author TA: Quality assessment and study design. Author NP: Study selection and data extraction. Author MM: Study design, interpretation, and critical revision of the manuscript. Author JA: Study design, interpretation, and critical revision of the manuscript. Author AT: Study design, study selection, data extraction, data analysis and interpretation, and critical revision of the manuscript. All authors read and approved the final manuscript.

Original Research Article

Received 12th June 2013
Accepted 10th September 2013
Published 19th October 2013

ABSTRACT

Aims: Several scoring systems have been developed for diagnosis of appendicitis. This study aims to systematically explore how those scores were derived and validated, and to compare their performance.
Study Design: Systematic review.
Place and Duration of Study: We searched Medline from 1949 and EMBASE from 1974 to March 2012 to identify relevant articles published in English.
Methodology: Information about model development and performance was extracted.

**Corresponding author: Email: ammarin.tha@mahidol.ac.th;*

The "risk of bias" assessment tool was developed based on a critical appraisal guide for clinical prediction rules. Calibration (O/E ratio) and discrimination (C-statistic) coefficients were estimated. A meta-analysis was applied to pool calibration coefficients and C-statistics.

Results: Forty-four out of 468 studies were eligible. Of these, 14 developed or modified diagnostic scoring systems and 30 validated existing models. Four scores had been most frequently validated, i.e., Alvarado, modified Alvarado, Fenyo, and Eskelinen. Among them, only the Eskelinen model was derived based on a multivariate regression whereas the rest used univariate or non-statistical methodology. All studies reported very good but imprecise calibration. For discrimination, the pooled C-statistics for these corresponding scores were 0.77, 0.86, 0.81, and 0.84 respectively. In the external validation, the discriminative performance decreased about 25.3% and 10.1% for the Alvarado and Fenyo scores respectively.

Conclusion: The research methods for scoring systems of appendicitis were inconsistent. More efficient scoring systems which have been internally and externally validated are required.

Keywords: Appendicitis; prediction score; systematic review; C-statistic; calibration.

ABBREVIATIONS

***CI:** confidence interval; **E/O ratio:** expected/observed ratio; **PMN:** polymorphonuclear; **RLQ:** right lower quadrant; **ROC:** Receiver Operating Characteristic; **WBC:** white blood cell.*

1. INTRODUCTION

Appendicitis is one of the most important clinical causes of acute abdominal pain, with an incidence of 110/100,000 [1] over a life time period. Although many attempts have been made to improve diagnostic accuracy, false positives and false negatives remain common with rates of negative appendectomy ranging from 15% to 26% [2,3] and perforated appendectomy ranging from 10% to 30%. [4] Several scoring systems including computer-based models and algorithms have been developed with good initial performance (in the derivation dataset), but this usually falls when validated in the general populations. Nevertheless, these scoring systems have been occasionally applied in general routine practice because of a lack of accuracy of physical examination [5]. Outcomes following a negative appendectomy (i.e., false positive) are less life threatening than the outcome of a false negative diagnosis, in which its mortality rate was high from appendiceal perforation and peritonitis. As a result, an aggressive surgical approach is frequently applied when the diagnosis is in doubt and this sometimes results in the removal of normal appendices. In order to reduce the aggressive management, an accurate and reliable diagnostic test for appendectomy is required to effectively discriminate between patients who require prompt surgical intervention from the patients who need more conservative management.

Imaging modalities have been used to improve diagnostic accuracy. However, disadvantages include excessive cost, lack of accessibility (particularly in developing countries), lack of radiologists, examiner-dependent efficacy (e.g., ultrasound), potential harmful ionization (e.g., computerized tomography), and low performance in low or high prevalence populations. Clinical scoring systems that synthesize clinical information have been developed and may be useful for those countries where imaging is less accessible.

These scores have been derived by incorporating clinical signs and symptoms into a mathematical equation which predicts the probability of appendicitis. Various statistical methods have been used to construct a number of diagnostic scoring algorithims [1,6-40], some of which have been validated [18,33] either internally [8,39] or externally [7-9,11,15,16, 21,27,28,30,37-39], while others have been applied without validation. The performance of those scores that have been validated varies from fair to good. We therefore conducted a systematic review to explore score performance in both the development and validation phases of these studies. The strengths and limitations of previous diagnostic scoring algorithms have been critically appraised. Findings from this review will help to identify the most valid and appropriate model to use across settings or may highlight the need to create new model/s with higher diagnostic accuracy.

2. MATERIALS AND METHODS

2.1 Search Strategy

We searched Medline from 1949 and EMBASE from 1974 to March 2012 to identify relevant studies published in English. Search terms were as follows: appendicitis, gangrenous appendicitis, phlegmon, perforated appendicitis, abdominal pain, score, scoring system, prediction score, prediction model, diagnostic score, assessment tool, ultrasonogram, ultrasonography, computer tomography, accuracy, negative appendectomy, sensitivity, specificity, likelihood ratio, false positive, false negative, true positive, true negative, receive operating characteristic (ROC), area under curve (AUC). Search strategies have been described in the appendix. We contacted authors for studies where data were insufficient.

2.2 Study Selection

Studies were reviewed based on titles and abstracts. If a decision could not be made, full articles were retrieved. Observational studies (cohort, case-control, or cross-sectional) published in English were selected if they met with the following criteria: suspected adult appendicitis, considered more than one risk factor in the prediction score, had the outcome as appendicitis versus non-appendicitis, applied any equation (e.g., Logistic regression, Bayesian method, or non-mathematical-investigator opinion base) to develop the prediction model, and reported the model's performance (i.e., calibration and discrimination parameters).

2.3 Data Extraction

The general characteristics of studies (i.e., author, journal, publication year, type of participants, ethnicity, study design, number of subjects, rate of negative appendectomy, percent of complicated appendicitis, and specific objective/s (i.e., to develop or validate score, or both)) were extracted. If the diagnostic model described its development then specific information about model building (i.e., type of statistical model, predictive factors, creating scores using coefficients or exponential of coefficients) was extracted. Calibration (a ratio of expected versus observed value (E/O ratio)), and discrimination parameters (i.e., the concordance statistic (C-statistic)) along with 95% confidence interval (CI) were also extracted. These parameters were calculated if the study did not directly report them, but did provide summary data allowing for calculations. For studies describing model validation, the type of validation (internal, external, or both) and results were also recorded. If authors had modified a previous prediction model, the following aspects were recorded: whether any of

the original included variables were removed or modified; and whether new predictive factors were added.

2.4 Methodological Assessment

The methodological assessment tool used in this review was developed based on a critical appraisal guide for clinical prediction rules [41] which considered both derivation and validation phases. Four domains were considered for the derivative phase, i.e., selection bias (representative of spectrum), information bias (ascertainment of outcome measurements, blinding outcome assessment, number of predictors, assessment of predictors without knowledge of outcome, proportion of important predictors), confounding bias (properly used multi-variate regression analysis to create score), and other issues (sample size, clinically sensible). For the validation phase, only 3 domains were considered, i.e., selection bias (representative of spectrum), information bias (ascertainment of outcome measurement, blinded outcome assessment, accurate interpretation), and other issues (i.e., follow up). Each item was classified as yes (low risk of bias), no (high risk of bias), or unclear if there was insufficient information to judge. Two reviewers (CW and TA) independently extracted data and assessed risk of bias for all included studies. Any disagreement was discussed with a third party (AT).

2.5 Statistical Analysis

Model performance was described separately for the derivation and validation phases. Calibration (O/E ratio) and discrimination (C-statistic) coefficients along with their 95% confidence intervals (CI) were estimated for each study. A meta-analysis was applied to pool the O/E and C-statistic using the equations as described in the appendix. Heterogeneity was assessed using the Q statistic and a degree of heterogeneity I^2 was estimated. If heterogeneity was present (p value <0.10 or I^2 > 25%), a random-effect model was used to pool data, otherwise a fixed-effect model was applied. All analyses were performed using STATA version 12.0.

3. RESULTS

3.1 Selection of Studies

We identified 468 studies of which 44 studies met our inclusion criteria and thus were eligible for the review (Fig. 1). The characteristics of these studies are described in Table 1 and Supplement Table 1. Of the 44 included studies, 9 studies [7,10,15,16,18,21,27,28,37] exclusively derived new prediction scores or modified previous prediction models (hereafter called derivation studies), 5 studies [8,9,33,38,39] derived and internally or externally validated their models in the same studies, whereas 30 studies exclusively examined internal [30,42] and external [1,6,10,12-14,17,19,20,22-26,29,31,32,34-36,40,43-49] model validations.

Among the 14 derivation studies, [7-9,11,15,16,18,21,27,28,33,37-39] all focused on adult patients, and most included patients with suspected appendicitis who received surgery or observation only, although 3 studies [33,37,39] included only patients who received surgery. Ten models [7,8,11,15,16,18,27,33,37-39,42] were derived in Caucasian populations while three models [11,21,28] were in Asian populations. The models were mainly constructed within cohorts, either retrospectively [7,33,39] or prospectively. [8,15,16,18,21,27,28,38,42].

Among 30 studies that exclusively performed validation, 27 studies had validated models on patients with suspected appendicitis whereas 3 studies focused on surgical patients. Most of these studies were conducted within prospective cohorts. Eighteen and twelve studies were conducted within Caucasian and Asian populations, respectively.

3.2 Methodological Assessments

Results of methodological assessments are described in Table 2. Among the 14 derivation studies, 8/14 (57.1%) studies had recruited consecutive patients with the chief complaint of abdominal pain, or randomly selected patients from a well-defined, population-based sampling frame of abdominal pain for retrospective studies; whereas the rest of the studies had recruited a specific group of patients presenting with at least a few clinical signs and symptoms of appendicitis.

Most studies (92.9%) confirmed the diagnosis of appendicitis histologically without mentioning whether the histology was performed with blinding to clinical information. The predictor variables used in the derivation models were considered complete and appropriate (i.e., low risk of bias) if authors used predictors from all clinically relevant predictive categories (i.e. demographic, clinical signs, symptoms, laboratory, and imaging data); otherwise this item was graded as high risk of bias. Ten out of fourteen (71.4%) studies clearly listed all clinically relevant predictor categories, whereas the remaining studies considered only a few predictor categories. Only 5/14 (35.71%) studies measured or collected predictors in which assessors were blinded to the final diagnosis of appendicitis, laboratory, and imaging findings, and 57.14% of studies measured or collected predictors where the assessor was unblinded to the possible diagnosis of appendicitis.

Eleven out of fourteen studies (78.7%) performed statistical estimations or tests for all predictors, whereas 3/14 (21.3%) studies did not apply any statistical method. However, only 5/14 (35.7%) studies applied multivariate regression models which simultaneously included significant predictors in the models, and used coefficients or relative risks from these regression models to create scores. The rest of the studies created prediction scores based on univariate results or non-statistical models.

Twelve (85.7%) studies had sufficient numbers of subjects for either appendicitis or total subjects based on the rule of thumb of 1 predictor per 10-30 appendicitis cases). Some studies (71.4%) included predictors that seemed to be clinically sensible, the scores were easy to apply and also suggested a course of clinical action.

For validation studies, 28/35 (80%) studies were less likely to be affected by selection bias. Ascertainment of diagnosis of appendicitis was clearly defined in 34/35 (97%) studies. Twenty four out of thirty five (68%) studies clearly described that their interpretation of the prediction rule was not influenced by information of the final diagnosis of appendicitis while 25% were potentially influenced by the diagnosis, and these 25% did not mention if blinding of clinical information was applied or not. Only 10 (28%) of studies followed up all non-operative subjects.

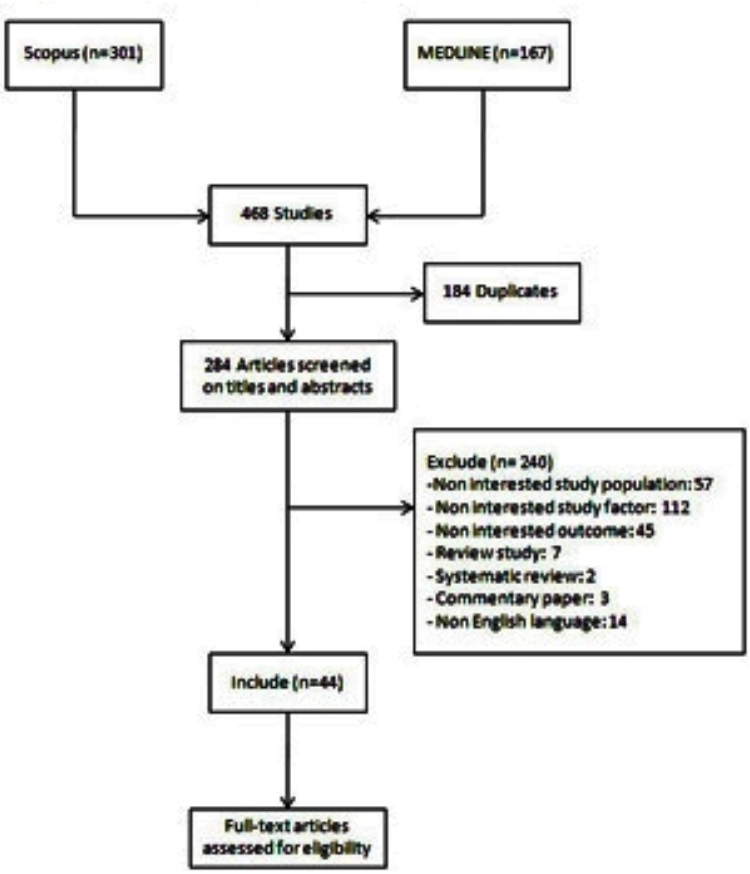

Fig. 1. Identification of studies for inclusion

Table 1. Characteristics of studies that had derived prediction scores for appendicitis

Author	Study phase	Year	Model	Study design	Type of subjects	%Male	Ethnicity	No. Appendicitis	No. Non-appendicitis	Statistical method	%Negative appendicitis	%Complicated appendicitis
Van way [39]	D/I	1982	New	Retro-cohort	Operated patients	NA	Caucasian	360	116	Discrimination analysis	29.83	25.30
Teicher [37]	D	1983	New	Case-control	Operated patients	45.5	Caucasian	100	100	Diagnostic analysis	40	
Alvarado [7]	D	1986	Alvarado score	Retro-cohort	in-patients	NA	Caucasian	227	50	Diagnostic analysis	7	18.77
Fenyo [16]	D	1987	New	Pro-cohort	Suspected appendicitis	NA	Caucasian	365	833	Diagnostic analysis	18	14.00
Christian [11]	D	1992	New	Quasi-experimental design	Suspected appendicitis	77.6	Asian	43	15	non-statistical base	6.5	6.50
Eskelinen [15]	D	1992	New	Pro-cohort	in-patients	NA	Caucasian	270	1333	Multiple logistic regression	21.6	6.74
Kalan [21]	D	1994	Modified Alvarado	Pro-cohort	in-patients	55.3	Asian	40	9	non-statistical base	23.68	NA
Ramirez [33]	D/I	1994	New	Retro-cohort	Operated patients	63.0	Caucasian	293	67	Bayesian, Likelihood ratio weight	18.61	NA
Gallego [18]	D	1998	New	Prospective Cohort	Suspected appendicitis	NA	Caucasian	101	91	Bayesian, Likelihood ratio weight	8.85	18.23
Tzanakis [38]	D/I/E	2005	New	Prospective Cohort	in-patients	56.1	Caucasian	217	504	Logistic regression	19.20	10.23
Lintula [27]	D	2005	New	Pro-cohort	Suspected appendicitis	100	Caucasian	43	84	Logistic regression	13	NA
Malik [28]	D	2007	Modified Alvarado	Pro-cohort	in-patients	55.1	Asian	174	80	non-statistical base	11.49	12.07
Andersson [8]	D/I/E	2008	New	Pro-cohort	in-patients	46.0	Caucasian	191	254	Logistic regression	11.00	14.00
Chong [9]	D/I	2010	RIPASA score	Retro-cohort	Emergency appendectomy	57.7	Asian	261	51	Univariate analysis	16.30	NA

Table 2. Describe methodological assessments

Study	Selection bias	Information bias						Confounding bias		Other issue		
	Representative spectrum	Ascertainment of outcome measurement	Blinded assess outcome	No predictors	Predictors blinded outcome	Significant predictors	Accurate interpretation	Multivariate regression analysis	Created score properly	Sample size	Clinical sensible	Other issues
Van way,1982 [39]	N	NA	NA	N	N	Y	-	Y	Y	Y	Y	-
Teicher,1983 [37]	Y	Y	NA	Y	NA	Y	-	N	N	Y	Y	-
Alvarado,1986 [7]	Y	Y	NA	Y	NA	Y	-	N	N	Y	Y	-
Fenyo,1987 [16]	Y	Y	NA	Y	NA	Y	-	N	N	Y	N	-
Christian,1992 [11]	N	Y	NA	N	Y	N	-	N	N	N	Y	-
Eskelinen,1992 [15]	Y	Y	NA	Y	Y	Y	-	Y	Y	Y	Y	-
Kalan,1994 [21]	N	Y	NA	N	NA	Y	-	N	N	N	Y	-
Ramirez,1994 [33]	N	Y	NA	Y	NA	Y	-	N	N	Y	N	-
Gallego,1998 [18]	Y	Y	NA	Y	NA	Y	-	N	N	Y	N	-
Tzanakis,2005 [38]	Y	Y	NA	Y	Y	Y	-	Y	Y	Y	Y	-
Lintula, 2005 [27]	Y	Y	NA	Y	Y	Y	-	Y	Y	Y	Y	-
Malik,2007 [28]	N	Y	NA	N	NA	Y	-	N	N	Y	Y	-
Andersson, 2008 [8]	Y	Y	NA	Y	Y	Y	-	Y	Y	Y	Y	-
Chong,2010 [9]	N	Y	NA	Y	NA	Y	-	N	N	Y	N	-
Van way,1982 [39]	N	NA	NA	-	-	-	N	-	-	-	-	N
Fenyo,1987 [16]	Y	Y	NA	-	-	-	Y	-	-	-	-	N
Ramirez,1994 [33]	N	Y	NA	-	-	-	N	-	-	-	-	N
Tzanakis,2005 [38]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Andersson, 2008 [8]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Lintula,2010 [42]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Chong,2010 [9]	N	Y	NA	-	-	-	N	-	-	-	-	Y
Fenyo, 1997 [17]	Y	Y	NA	-	-	-	N	-	-	-	-	N
Denizbasi,2003 [13]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Chan,2003 [43]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Win,2004 [48]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
AlQahtani,2004 [6]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Pruekprasert, 2004 [32]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Enochsson, 2004 [14]	UN	Y	NA	-	-	-	Y	-	-	-	-	NA
Sitter,2004 [34]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Tzanakis,2005 [38]	Y	Y	NA				NA	-	-	-	-	Y
	Y	Y	NA	-	-	-	NA					Y
	Y	Y	NA				NA					Y
	Y	***Y***	***NA***				***NA***					***Y***

Table 2 continues……

	Y	Y	NA				NA					Y
	Y	Y	NA				NA					Y
Mckay,2007 [29]	Y	Y	NA	-	-	-	N	-	-	-	-	NA
Andersson,2008 [8]	Y	Y	NA	-	-	-	NA	-	-	-	-	Y
Kurane,2008 24]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Sun, 2008 [35]	Y	Y	NA	-	-	-	N	-	-	-	-	NA
Kim,2008 [44]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Yildrim,2008 [49]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Prabhudesai,2008 [45]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Singh,2008 [46]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Soomro,2008 [47]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Talukder,2009[36]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Hsieh, 2010 [19]	Y	Y	NA	-	-	-	N	-	-	-	-	NA
Pouret-Baudry,2010 [31]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Chong, 2011 [10]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
	Y	Y	NA				Y					NA
Inci,2011 [20]	Y	Y	NA	-	-	-	Y	-	-	-	-	Y
Limpawattan, 2011 [26]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Konan,2011 [23]	N	Y	NA	-	-	-	N	-	-	-	-	NA
Kanumba,2011 [22]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
Yoldas,2011 [40]	N	Y	NA	-	-	-	N	-	-	-	-	NA
Castro,2012 [12]	Y	Y	NA	-	-	-	Y	-	-	-	-	NA
	Y	Y	NA				Y					NA

3.3 Score Development

Among 14 derivation studies, 5 predictive categories were considered in the models including demographic data, clinical signs, clinical symptoms, laboratory, and imaging data, (Table 1). Of the 2 demographic variables, gender was more commonly included in the model compared with age (42.9% vs 14.3%). Ten symptom variables were considered, in which nausea (9/14, 64.3 %) was most commonly included in the model follow by migration of pain, pain at presentation, or duration of pain (all were 46.2%). Nine clinical signs were considered with the most common variable being rebound tenderness (76.9%), followed by right lower quadrant (RLQ) tenderness (61.5%), and RLQ guarding (53.9%) or elevated temperature (53.9%). Among the 10 clinical symptoms, nausea/vomiting (53.9%) followed by migration and duration of pain (46.4%) were most commonly included. Most studies (84.6%) considered at least one laboratory variable. Among these, elevated white blood cell count (76.9%) was most commonly used followed by left shift of polymorphonuclear (PMN) cells (46.2%). Only two studies [18,38] used radiological data (e.g. ultrasonography and/or abdominal radiograph) in their scoring systems.

As described in Table 1 and Supplement Table 1, these prediction scores were developed using statistical modeling in 5 studies [8,15,27,38,39] whereas 9 studies [7,9,11,16,18,21, 28,33,37] did not apply statistical modeling. Among the 5 studies which used statistical modeling, 4 studies [8,15,27,38] applied multivariate logistic regression and 1 study [39] used discriminant analysis. Scoring schemes of these models were created based on coefficients of the logit or discriminant regression models. Among the 9 studies that did not apply statistical models, univariate analysis (e.g., Chi-square test, relative risk) and estimated diagnostic parameters (e.g., likelihood ratio, sensitivity, specificity) were used for assessing associations in 6 studies whereas 3 studies did not apply any statistical test.

3.4 Model Performance

Model performances using C-statistics and O/E calibration coefficients were extracted from individual studies if reported, otherwise they were estimated using summary data reported in the articles (Table 3). The Alvarado model [7] was the most frequent externally validated model without internal validation by 50 studies. [6,8,10,12,13,19,20,23,26,29,31,32,35,38,43-77] Seventeen studies [50-62,65,72,74,77] included pediatric population, 2 studies [66,71] were published in non English and 10 studies [44,63,64,67-70,73,75,76] presented non-interested outcomes. The model was originally derived in 277 Caucasians using diagnostic accuracy for weighting each of 8 predictors, i.e., migration of pain, anorexia, nausea/vomiting, elevated temperature, rebound tenderness, RLQ tenderness, elevated WBC, and PMN cell left shift. The point estimate of the O/E calibration coefficient was very good (1.0), although the confidence interval varied from 0.75 to 1.25. The C-statistic was 0.78, which indicated reasonably good discrimination. All eight variables were included in the externally validated models. Combining E/O calibration coefficients and C-statistics across 21 studies yielded a pooled O/E and the pooled C statistic of 1 (95%CI, 0.98 to 1.02) and 0.77 (95%CI, 0.73 to 0.81), respectively.

The Alvarado score was also modified by two subsequent studies by excluding the shift to left of PMN cells due to unavailability of routine laboratory data in the study performed by Kalan et al. [21] or it was replaced with other variables (i.e. cough test, Rovsing's sign, rectal tenderness) in the study performed by Malik et al. [28] In these studies the O/E calibration coefficients did not change much when compared with the original model.

Table 3. Describe performances of predictive models of appendicitis

	Derived model			Internal validation			External validation			
	No. variable/ event/N	**O/E**	**C-statistic**	**No. variable/ event/N**	**O/E**	**C-statistic**	**Author, Year**	**No. variable /event/N**	**O/E**	**C-statistic**
Alvarado, 1986[7]	8/227/277	1(0.93,1.07)	0.80(0.73,0.86)	-	-	-	Denizbasi, 2003[13]	8/175/221	0.99(0.91,1.09)	0.71(0.63,0.78)
							Chan, 2003[43]	8/75/175	1(0.83,1.173)	0.51(0.43,0.58)
							Win, 2004[48]	8/57/122	1(0.67,1.33)	0.85(0.79,0.92)
							AlQahtani, 2004[6]	8/121/151	1(0.79,1.21)	0.84(0.76,0.93)
							Pruekprasert, 2004[32]	8/186/231	1(0.93,1.07)	0.74(0.67,0.81)
							Tzanakis, 2005[38]	8/87/201	1(0.69,1.31)	0.83(0.78,0.88)
							McKay, 2007[29]	8/48/144	1(0.70,1.30)	0.74(0.66,0.82)
							Andersson, 2008[8]	8/76/225	0.99(0.79,1.21)	0.63(0.58,0.68)
							Sun, 2008[35]	8/213/372	1(0.90,1.10)	0.65(0.61,0.70)
							Kim, 2008[44]	8/90/157	1(0.86,1.14)	0.61(0.54,0.68)
							Yildrim, 2008[49]	8/55/143	1(0.40,1.59)	0.94(0.89,0.98)
							Prabhudesai, 2008[45]	8/24/60	1(0.55,1.45)	0.86(0.77,0.95)
							Singh, 2008[46]	8/62/100	0.99(0.79,1.21)	0.79(0.71,0.88)
							Soomro, 2008[47]	8/178/227	1(0.93,1.07)	0.67(0.61,0.74)
							Hsieh, 2010[19]	8/115/180	1(0.85,1.15)	0.77(0.70,0.83)
							Pouget-Baudry,2010[31]	8/171/233	1(0.92,1.08)	0.68(0.62,0.73)
							Chong, 2011[10]	8/101/192	0.99(0.82,1.18)	0.78(0.72,0.84)
							Inci, 2011[20]	8/57/66	1(0.89,1.11)	0.75(0.58,0.92)
							Limpawattanasiri, 2011[26]	8/715/1000	1(0.94,1.06)	0.81(0.78,0.84)
							Konan,2011[23]	8/41/82	1(0.69,1.31)	0.84(0.76,0.91)
							Castro, 2012[12]	8/340/935	1(0.91,1.09)	0.62(0.60,0.65)
							Pooled		1(0.98, 1.02)	0.77 (0.73, 0.81)
Kalan, 1994[21] (Modified Alvarado)	7/40/49	1(0.74,1.26)*	0.76(0.60,0.92)*	-	-	-	Lamparelli, 2000[25]	7/56/84	1(0.80,1.20)	0.82(0.74,0.90)
							Kurane, 2008[24]	7/23/60	1(0.49,1.51)	0.81(0.71,0.92)
							Kanumba, 2011[22]	7/85/127	0.99(0.72,1.28)	0.92(0.87,0.97)
							Pooled		0.99 (0.84, 1.15)	0.86 (0.78, 0.94)
Malik, 2007[28] (Modified Alvarado)	8/174/254	1(0.93,1.07)	0.54(0.45,0.63)	-	-	-	Talukder, 2009[36]	8/84/100	1(0.91,1.09)	0.65(0.52,0.78)
							Fenyo,1997[17]	18/392/1167	1(0.88,1.12)	0.80(0.77,0.82)
							Enochsson,2004[14]	18/330/426	1(0.94,1.06)	0.73(0.68,0.78)
							Tzanakis, 2005[38]	18/87/201	1(0.65,1.35)	0.88(0.83,0.92)
							Pooled		1.0 (0.95, 1.05)	0.81 (0.74, 0.87)

Table 3 continues……..

Eskelinen, 1992[15]	6/572/1333	1(0.93,1.07)	0.59(0.58,0.61)	-	-	-	Sitter, 2004[34]	6/662/2359	1(0.90,1.10)	0.82(0.80,0.84)
							Tzanakis, 2005[38]	6/87/201	1(0.71,1.28)	0.87(0.83,0.92)
							Pooled		1.0(0.91, 1.09)	0.84 (0.79, 0.89)
Christian, 1992[11]	5/43/58	1(0.56,1.44)*	0.87(0.77,0.98)*	-	-	-	Tzanakis, 2005[38]	5/87/201	1(0.71,1.29)	0.85 (0.80,0.90)
Tzanakis, 2005[38]	10/130/303	1(0.45,1.55)	0.97(0.95,0.99)	NA	NA	NA	Tzanakis, 2005[38]	10/87/201	-	0.96(0.93,0.99)
Lintula, 2005[27]	NA	NA	NA	9/52/96	1(0.64,1.36)	0.92(0.87,0.97)	Yoldas, 2011[40]	9/132/156	1(0.92,1.08)	0.79(0.73,0.86)
Andersson, 2008[8]	7/115/316	NA	0.87	7/115/316	1(0.43,1.57)	0.89(0.86,0.93)	Castro, 2012[12]	7/348/945	0.99(0.91,1.09)	0.55(0.54,0.57)
Chong, 2010[9]	NA	NA	NA	15/101/192	1(0.45,1.55)	0.90(0.85,0.94)	Chong, 2011[10]	15/101/192	0.99(0.82,1.18)	0.78(0.72,0.84)
Van way, 1982[39]	4/NA/219	-	-	4/169/257	1(0.92,1.08)	0.61(0.54,0.68)	Tzanakis, 2005[38]	4/87/201	1(0.76,1.24)	0.78(0.72,0.84)
Teicher, 1983[37]	7/100/200	1(0.75,1.25)	0.78(0.72,0.83)	-	-	-	Tzanakis, 2005[38]	7/87/201	1(0.69,1.31)	0.86(0.81,0.91)
Ramirez, 1994[33]	7/293/360	1(0.94,1.06)	0.72(0.60,0.84)	NA	NA	NA	-	-	-	-
Gallego, 1998[18]	6/101/192	1(0.78,1.23)	0.89(0.85,0.93)	-	-	-	-	-	-	-
Ohmann, 1999[30]	-	-	-	8/235/1254	1(0.44,1.56)	0.87(0.84,0.90)	Tepel, 2004[1]	8/113/400	1(0.78,1.20)	0.73(0.68-0.78)

Conversely, the C statistics decreased from the original O/E calibration coefficient of 0.80 (95%CI, 0.73 to 0.86) to 0.76 (95%CI, 0.60 to 0.92) when the PMN cell left shift variable was excluded, and performance was even poorer when PMN cell left shift was replaced with the cough test, Rovsing's sign, or rectal tenderness variables (C-statistic= 0.54; 95%CI, 0.45 to 0.63). However, the external performance of the modified Alvarado model by Kalan, [21] which was validated by 3 studies, [22,24,25] performed well in terms of calibration and discrimination.

The Malik model had only fair discrimination (C-statistic = 0.65; 95% CI: 0.52 to 0.78).

The Fenyo et al. [16] model, which was the second most externally validated after Alvarado, was developed in 1987 and included 18 variables. A positive likelihood ratio was used to create a score for each variable. The score performance in the derivation phase was outstanding for both calibration and discrimination with an O/E calibration coefficient and C statistic of 1(0.75 to 1.25) and 0.91(0.89 to 0.93), respectively. The Fenyo model was tested in 3 other studies, [14,17,38] which yielded a 10% decrease of the pooled C statistic (0.81, 95% CI: 0.74 to 0.87) compared with the original study.

The Eskelinen model was developed in 1992 and included the largest sample size of cases of appendicitis (n= 572/1333). The logistic regression model was constructed by including 6 variables in the equation and the score was derived using coefficients from the logit equation. Among the 6 included variables, 4 variables (i.e., rebound tenderness, rigidity, RLQ tenderness, and increase of WBC) were similar to those used in the derivation of the Alvarado score with the exception of 2 variables (i.e., duration of pain and pain at presentation). The estimated O/E calibration coefficient for the derived model was high and precise (1.0, 95% CI: 0.93 to 1.07) whereas the discriminative performance was only fair with a pooled C statistic of 0.59 (95% CI: 0.58 to 0.61). In contrast, external validation of this score suggested good discrimination with the pooled C statistic of 0.84 (95% CI: 0.79 to 0.89) in different groups of population in 2004 and 2005.

4. DISCUSSION

We have reviewed the performance of 14 diagnostic scores for appendicitis. Diagnostic scores were judged based on their discrimination, i.e. distinguishing cases from non-cases, and their calibration, i.e. the percentage of observed cases was similar to predicted or expected cases. Discrimination was judged by looking at the area under the ROC curve (or C-statistic) which ranges from 0.5 (consistent with chance) to 1.0 (perfect diagnostic ability). [78] Calibration is judged by looking at observed/expected ratios, with 1.0 indicating perfect calibration. [79,80]

Only 35.7% of studies derived scores using statistical modeling whereas the rest used diagnostic parameters (i.e., accuracy or likelihood ratio positive) or univariate analysis (i.e., Chi-square test) without a proper rationale for weighting in prediction scores. Although the Alvarado [7], modified Alvarado [21], and Fenyo [16] scores were not derived using statistical modeling, they were the most frequently applied with externally validated C-statistics of 0.77, 0.86, and 0.81, respectively. The performance of these models did not differ significantly to the Eskelinen [15] scoring model (C-statistic = 0.84; 95% CI; 0.79 to 0.89) which properly applied statistical modeling. All the models seemed to have reasonable calibration (O/E of 1.0) although the Eskelinen and Alvarado scores had the smallest confidence interval. Many factors may influence the performance of a diagnostic model. The association between predictive factors and appendicitis using derived data may occur by chance and thus will

result in poorer performance in a general population (i.e., external validation). Such an event is more likely to occur if a sample size is relatively small compared with the number of diagnostic factors included in the model. [81] With a small sample size, unimportant variables may be selected and some important variables may be omitted from the model. Conversely, a very large sample size is more likely to include statistically significant variables that are not clinically meaningful. At least 10 – 30 cases per one predictive factor is necessary to derive a valid model as suggested by simulation studies. [82] As per our review, the number of predictor variables included in the Alvarado [7], modified Alvarado [21], Fenyo [16], and Eskelinen [15] were 8, 7, 18, and 6 variables respectively, so the required number of appendicitis cases in each study should have been at least 80, 70, 180, and 60 subjects respectively; and 240, 210, 540, and 180 subjects for greater precision. Among these 3 models, the Fenyo [16] model with a sample size of appendicitis cases of 109, was far below the minimum required number of 180 cases. As a result, an over-fitted or an overoptimistic model may be applied if the Fenyo [16] model is applied in a general population. This was confirmed by the fact that the C-statistic dropped by 10% in the validation set compared with the derivation set. In an appropriate sample size model, the C-statistic dropped by only 3.75% and increased from 0.59 to 0.84 in the Alvarado and Eskelinen scores, respectively.

It is generally recommended that derivation of prognostic models [81,83,84] should be developed using a multi-variate regression or Bayesian model rather than developed using a univariate or non-statistical modeling approach. The multivariate model allows for the simultaneous inclusion of multiple variables and adjustment for confounding variables. The use of statistical models should be clearly described and the model's assumptions or goodness of fit should be checked. Although the original and modified Alvarado and Fenyo scores were derived based on non-statistical models, their external predictive performances still provided good discrimination. However, applying these scores to a general population may be problematic due to inappropriately derived scores. The model itself should be simple, easy to apply, and interpret to encourage general surgeons to apply these models in clinical practice. The number of included variables should be limited to a few and they should be easy to measure or examine in a routine clinical practice. Applying the original and modified-Alvarado and Fenyo scores requires 7 and 18 variables respectively, whereas only 6 variables are required when applying the Eskelinen score. All of these predictor variables are signs and symptoms with only 1 laboratory predictor (i.e., WBC count).

5. CONCLUSION

In summary, it is recommended that clinical decision rules should be developed using rigorous statistical approaches, they should be derived and validated in independent populations, they should exhibit good discrimination, i.e. high C-statistic, and exhibit good calibration, i.e. O/E close to 1, and be tested in large samples with sufficient power to accommodate the number of predictors being tested. The rule that comes closest to meeting all these criteria is that by Eskelinen, although there is still much room for improvement and validation.

CONSENT

Consent is not applicable in this study.

ETHICAL APPROVAL

Ethical approval is not applicable in this study.

ACKNOWLEDGEMENTS

This study had no funding.

COMPETING INTERESTS

CW is Ph.D. student in Clinical Epidemiology program, the Faculty of Medicine Ramathibodi Hospital and Faculty of Graduate Studies, Mahidol University, Bangkok, Thailand. This study is a part of his dissertation.

TA, MM, JA, and AT do not have any conflict of interest. All authors have read this version and are agreeable for its publication.

REFERENCES

1. Tepel J, Sommerfeld A, Klomp HJ, Kapischke M, Eggert A, Kremer B. Prospective evaluation of diagnostic modalities in suspected acute appendicitis. Langenbecks Arch Surg. 2004;389(3):219-24.
2. Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. Am J Epidemiol. 1990;132(5):910-25.
3. Horntrich J, Schneider W. [Appendicitis from an epidemiological viewpoint]. Zentralbl Chir. 1990;115(23):1521-9.
4. Temple CL, Huchcroft SA, Temple WJ. The natural history of appendicitis in adults. A prospective study. Ann Surg. 1995;221(3):278-81.
5. Andersson RE. Meta-analysis of the clinical and laboratory diagnosis of appendicitis. Br J Surg. 2004;91(1):28-37.
6. Al Qahtani HH, Muhammad AA. Alvarado score as an admission criterion for suspected appendicitis in adults. Saudi J Gastroenterol. 2004;10(2):86-91.
7. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med. 1986;15(5):557-64.
8. 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(8):1843-9.
9. Chong CF, Adi MIW, Thien A, Suyoi A, Mackie AJ, Tin AS, et al. Development of the RIPASA score: A new appendicitis scoring system for the diagnosis of acute appendicitis. Singapore Medical Journal. 2010;51(3):220-5.
10. Chong CF, Thien A, Mackie AJA, Tin AS, Tripathi S, Ahmad MA, et al. Comparison of RIPASA and Alvarado scores for the diagnosis of acute appendicitis. Singapore Medical Journal. 2011;52(5):340-5.
11. Christian F, Christian GP. A simple scoring system to reduce the negative appendicectomy rate. Ann R Coll Surg Engl. 1992;74(4):281-5.
12. de Castro SM, Unlu C, Steller EP, van Wagensveld BA, Vrouenraets BC. Evaluation of the Appendicitis Inflammatory Response Score for Patients with Acute Appendicitis. World J Surg; 2012.

13. Denizbasi A, Unluer EE. The role of the emergency medicine resident using the Alvarado score in the diagnosis of acute appendicitis compared with the general surgery resident. Eur J Emerg Med. 2003;10(4):296-301.
14. Enochsson L, Gudbjartsson T, Hellberg A, Rudberg C, Wenner J, Ringqvist I, et al. The Fenyö-Lindberg scoring system for appendicitis increases positive predictive value in fertile women: A prospective study in 455 patients randomized to either laparoscopic or open appendectomy. Surgical Endoscopy and Other Interventional Techniques. 2004;18(10):1509-13.
15. Eskelinen M, Ikonen J, Lipponen P. A computer-based diagnostic score to aid in diagnosis of acute appendicitis. A prospective study of 1333 patients with acute abdominal pain. Theoretical Surgery. 1992;7(2):86-90.
16. Fenyo G. Routine use of a scoring system for decision-making in suspected acute appendicitis in adults. Acta Chir Scand. 1987;153(9):545-51.
17. Fenyö G, Lindberg G, Blind P, Enochsson L, Öberg Å. Diagnostic decision support in suspected acute appendicitis: Validation of a simplified scoring system. European Journal of Surgery. 1997;163(11):831-8.
18. Galindo Gallego M, Fadrique B, Nieto MA, Calleja S, Fernández-Aceñero MJ, Ais G, et al. Evaluation of ultrasonography and clinical diagnostic scoring in suspected appendicitis. British Journal of Surgery. 1998;85(1):37-40.
19. Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery. 2011;149(1):87-93.
20. Inci E, Hocaoglu E, Aydin S, Palabiyik F, Cimilli T, Turhan AN, et al. Efficiency of unenhanced MRI in the diagnosis of acute appendicitis: Comparison with Alvarado scoring system and histopathological results. European Journal of Radiology. 2011;80(2):253-8.
21. Kalan M, Talbot D, Cunliffe WJ, Rich AJ. Evaluation of the modified Alvarado score in the diagnosis of acute appendicitis: a prospective study. Ann R Coll Surg Engl. 1994;76(6):418-9.
22. Kanumba ES, Mabula JB, Rambau P, Chalya PL. Modified Alvarado Scoring System as a diagnostic tool for Acute Appendicitis at Bugando Medical Centre, Mwanza, Tanzania. BMC Surgery. 2011;11.
23. Konan A, Hayran M, Kilic YA, Karakoc D, Kaynaroglu V. Scoring systems in the diagnosis of acute appendicitis in the elderly. Ulusal Travma ve Acil Cerrahi Dergisi. 2011;17(5):396-400.
24. Kurane SB, Sangolli MS, Gogate AS. A one year prospective study to compare and evaluate diagnostic accuracy of modified Alvarado score and ultrasonography in acute appendicitis, in adults. Indian Journal of Surgery. 2008;70(3):125-9.
25. Lamparelli MJ, Hoque HMR, Pogson CJ, Ball ABS. A prospective evaluation of the combined use of the modified Alvarado score with selective laparoscopy in adult females in the management of suspected appendicitis. Annals of the Royal College of Surgeons of England 2000;82(3):192-5.
26. Limpawattanasiri C. Alvarado score for the acute appendicitis in a provincial hospital. J Med Assoc Thai. 2011;94(4):441-9.
27. Lintula H, Pesonen E, Kokki H, Vanamo K, Eskelinen M. A diagnostic score for children with suspected appendicitis. Langenbecks Arch Surg. 2005;390(2):164-70.
28. Malik AH, Wani RA, Saima BD, Wani MY. Small lateral access-an alternative approach to appendicitis in paediatric patients: A randomised controlled trial. International Journal of Surgery. 2007;5(4):234-8.

29. McKay R, Shepherd J. The use of the clinical scoring system by Alvarado in the decision to perform computed tomography for acute appendicitis in the ED. Am J Emerg Med. 2007;25(5):489-93.
30. Ohmann C, Franke C, Yang Q. Clinical benefit of a diagnostic score for appendicitis: Results of a prospective interventional study. Archives of Surgery. 1999;134(9):993-6.
31. Pouget-Baudry Y, Mucci S, Eyssartier E, Guesdon-Portes A, Lada P, Casa C, et al. The usefulness of the Alvarado Score in the management of right lower quadrant abdominal pain in the adult. Journal de Chirurgie Viscerale. 2010;147(2):128-32.
32. Pruekprasert P, Geater A, Ksuntigij P, Maipang T, Apakupakul N. Accuracy in diagnosis of acute appendicitis by comparing serum C-reactive protein measurements, Alvarado score and clinical impression of surgeons. Journal of the Medical Association of Thailand. 2004;87(3):296-303.
33. Ramirez JM, Deus J. Practical score to aid decision making in doubtful cases of appendicitis. Br J Surg. 1994;81(5):680-3.
34. Sitter H, Hoffmann S, Hassan I, Zielke A. Diagnostic score in appendicitis. Validation of a diagnostic score (Eskelinen score) in patients in whom acute appendicitis is suspected. Langenbecks Arch Surg. 2004;389(3):213-8.
35. Sun JS, Noh HW, Min YG, Lee JH, Kim JK, Park KJ, et al. Receiver operating characteristic analysis of the diagnostic performance of a computed tomographic examination and the Alvarado score for diagnosing acute appendicitis: emphasis on age and sex of the patients. J Comput Assist Tomogr. 2008;32(3):386-91.
36. Talukder DB. Siddiq AKMZ Modified Alvarado Scoring System in the Diagnosis of Acute Appendicitis. JAFMC Bangladesh. 2009;5(1):18-20.
37. Teicher I, Landa B, Cohen M. Scoring system to aid in diagnoses of appendicitis. Annals of Surgery. 1983;198(6):753-9.
38. Tzanakis NE, Efstathiou SP, Danulidis K, Rallis GE, Tsioulos DI, Chatzivasiliou A, et al. A new approach to accurate diagnosis of acute appendicitis. World J Surg. 2005;29(9):1151-6, discussion 7.
39. Van Way CW, 3rd, Murphy JR, Dunn EL, Elerding SC. A feasibility study of computer aided diagnosis in appendicitis. Surg Gynecol Obstet. 1982;155(5):685-8.
40. Yoldas O, Karaca T, Tez M. External validation of Lintula score in Turkish acute appendicitis patients. International Journal of Surgery. 2012;10(1):25-7.
41. Guyatt GRD. Users' Guides to the Medical Literature: A Menual for Evidence-based Clinical Practice. 2 ed. Chicago: American Medical Association Printed in the United States of America; 2002.
42. Lintula H, Kokki H, Pulkkinen J, Kettunen R, Gröhn O, Eskelinen M. Diagnostic score in acute appendicitis. Validation of a diagnostic score (Lintula score) for adults with suspected appendicitis. Langenbeck's Archives of Surgery. 2010;395(5):495-500.
43. Chan MY, Tan C, Chiu MT, Ng YY. Alvarado score: an admission criterion in patients with right iliac fossa pain. Surgeon. 2003;1(1):39-41.
44. Kim K, Rhee JE, Lee CC, Kim KS, Shin JH, Kwak MJ, et al. Impact of helical computed tomography in clinically evident appendicitis. Emerg Med J. 2008;25(8):477-81.
45. Prabhudesai SG, Gould S, Rekhraj S, Tekkis PP, Glazer G, Ziprin P. Artificial neural networks: useful aid in diagnosing acute appendicitis. World J Surg. 2008;32(2):305-9; discussion 10-1.
46. Singh K, Gupta S, Pargal P. Application of Alvarado scoring system in diagnosis of acute appendicitis. JK Science. 2008;10(2):84-6.

47. Soomro AG, Siddiqui FG, Abro AH, Abro S, Shaikh NA, Memon AS. Diagnostic accuracy of Alvarado Scoring System in acute appendicitis. Journal of the Liaquat University of Medical and Health Sciences. 2008;7(2):93-6.
48. Winn RD, Laura S, Douglas C, Davidson P, Gani JS. Protocol-based approach to suspected appendicitis, incorporating the Alvarado score and outpatient antibiotics. ANZ J Surg. 2004;74(5):324-9.
49. Yildirim E, Karagulle E, Kirbas I, Turk E, Hasdogan B, Teksam M, et al. Alvarado scores and pain onset in relation to multislice CT findings in acute appendicitis. Diagn Interv Radiol. 2008;14(1):14-8.
50. Abdeldaim Y. The Alvarado score as a tool for diagnosis of acute appendicitis. Irish Medical Journal 2007;100(1).
51. 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(9):1014-8.
52. Chan MY, Teo BS, Ng BL. The Alvarado score and acute appendicitis. Ann Acad Med Singapore. 2001;30(5):510-2.
53. Escribá A, Gamell AM, Fernández Y, Quintillá JM, Cubells CL. Prospective validation of two systems of classification for the diagnosis of acute appendicitis. Pediatric Emergency Care. 2011;27(3):165-9.
54. Farahnak M, Talaei-Khoei M, Gorouhi F, Jalali A. The Alvarado score and antibiotics therapy as a corporate protocol versus conventional clinical management: randomized controlled pilot study of approach to acute appendicitis. Am J Emerg Med. 2007;25(7):850-2.
55. Gwynn LK. The diagnosis of acute appendicitis: clinical assessment versus computed tomography evaluation. J Emerg Med. 2001;21(2):119-23.
56. Hsiao KH, Lin LH, Chen DF. Application of the MANTRELS scoring system in the diagnosis of acute appendicitis in children. Acta Paediatr Taiwan. 2005;46(3):128-31.
57. Khan I, ur Rehman A. Application of alvarado scoring system in diagnosis of acute appendicitis. J Ayub Med Coll Abbottabad. 2005;17(3):41-4.
58. Muenzer JT, Jaffe DM, Schwulst SJ, Dixon DJ, Schierding WS, Li Q, et al. Evidence for a novel blood RNA diagnostic for pediatric appendicitis: The riboleukogram. Pediatric Emergency Care. 2010;26(5):333-8.
59. Owen TD, Williams H, Stiff G, Jenkinson LR, Rees BI. Evaluation of the Alvarado score in acute appendicitis. J R Soc Med. 1992;85(2):87-8.
60. Petrosyan M, Estrada J, Chan S, Somers S, Yacoub WN, Kelso RL, et al. CT scan in patients with suspected appendicitis: clinical implications for the acute care surgeon. Eur Surg Res. 2008;40(2):211-9.
61. Rezak A, Abbas HMA, Ajemian MS, Dudrick SJ, Kwasnik EM. Decreased use of computed tomography with a modified clinical scoring system in diagnosis of pediatric acute appendicitis. Archives of Surgery. 2011;146(1):64-7.
62. Schneider C, Kharbanda A, Bachur R. Evaluating appendicitis scoring systems using a prospective pediatric cohort. Ann Emerg Med. 2007;49(6):778-84, 84 e1.
63. Arain G, Sohu, KM., Ahmad, E., Haider, W., Naqi, SA. Role of Alvarado score in diagnosis of acute appendicitis. Pak J Surg. 2001;17:41-6.
64. Baidya N, Rao, A., Khan, SA. Evaluation of Alvarado score in acute appendicitis: a prospective study. Internet J Surg. 2007;9.

65. Borges P, Lima, MdC, Neto, GHF. The Alvarado score validation in diagnosing acute appendicitis in children and teenagers at the Instituto Materno Infantil de Pernambuco, IMIP. Rev Bras Saude Mater Infant. 2003;3:439-45.
66. Canavosso L, Carena P, Carbonell JM, Monjo L, Zuñiga CP, Sánchez M, et al. Right iliac fossa pain and Alvarado score. Cirugia Espanola. 2008;83(5):247-51.
67. Kang WM, Lee CH, Chou YH, Lin HJ, Lo HC, Hu SC, et al. A clinical evaluation of ultrasonography in the diagnosis of acute appendicitis. Surgery 1989;105(2 I):154-9.
68. Lada PE, Ocho S, Rosso F, Ternengo D, Sánchez M, Di Benedetto N, et al. Use of Alvarado's score for the early diagnosis of acute appendicitis. Prensa Medica Argentina. 2005;92(7):447-56.
69. Memon AA, Vohra LM, Khaliq T, Lehri A. Diagnostic accuracy of alvarado score in the diagnosis of acute appendicitis. Pakistan Journal of Medical Sciences. 2009;25(1):118-21.
70. Saidi RF, Ghasemi M. Role of Alvarado score in diagnosis and treatment of suspected acute appendicitis [3]. American Journal of Emergency Medicine. 2000;18(2):230-1.
71. Sanabria A, Domínguez LC, Bermúdez C, Serna A. Evaluation of diagnostic scales for appendicitis in patients with lower abdominal pain. Biomédica : revista del Instituto Nacional de Salud. 2007;27(3):419-28.
72. Shreef KS, Waly AH, Abd-Elrahman S, Abd Elhafez MA. Alvarado score as an admission criterion in children with pain in right iliac fossa. Afr J Paediatr Surg. 2010;7(3):163-5.
73. Shrivastava UK, Gupta A, Sharma D. Evaluation of the Alvarado score in the diagnosis of acute appendicitis. Trop Gastroenterol. 2004;25(4):184-6.
74. Stephens PL, Mazzucco JJ. Comparison of ultrasound and the Alvarado score for the diagnosis of acute appendicitis. Conn Med. 1999;63(3):137-40.
75. Subotić AM, Sijacki AD, Dugalić VD, Antić AA, Vuković GM, Vukojević VS, et al. Evaluation of the Alvarado score in the diagnosis of acute appendicitis. Acta chirurgica Iugoslavica. 2008;55(1):55-61.
76. Tade AO. Evaluation of Alvarado score as an admission criterion in patients with suspected diagnosis of acute appendicitis. West Afr J Med. 2007;26(3):210-2.
77. Wani M, Yousaf, MN., Khan, MA., BabaAbdul, A., Durrani, M., Wani, MM., Shafi, M. Usefulness of the Alvarado scoring system with respect to age, sex and time of presentation, with regression analysis of individual parameters. Internet J Surg. 2007;11.
78. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157-72; discussion 207-12.
79. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
80. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Semin Oncol. 2010;37(1):31-8.
81. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
82. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993-1000.

83. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
84. Sullivan LM, Massaro JM, D'Agostino RB, Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med. 2004;23(10):1631-60.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=298&id=12&aid=2290