Published in final edited form as:
Emerg Med J. 2016 July ; 33(7): 458–464. doi:10.1136/emermed-2015-205301.

Prospective Evaluation of the Ability of Clinical Scoring Systems and Physician-Determined Likelihood of Appendicitis to Obviate the Need for Computed Tomography

Sean K. Golden, BA[1], **John B. Harringa, BS**[1], **Perry J. Pickhardt, MD**[2], **Alexander Ebinger, MD**[3], **James E. Svenson, MD, MS**[1], **Ying-Qi Zhao, PhD**[4], **Zhanhai Li, PhD**[4], **Ryan P. Westergaard, MD, PhD**[5], **William J. Ehlenbach, MD, MSc**[5], and **Michael D. Repplinger, MD, PhD**[1,2]

Sean K. Golden: skgolden@wisc.edu; John B. Harringa: harringa@medicine.wisc.edu; Perry J. Pickhardt: ppickhardt2@uwhealth.org; Alexander Ebinger: alexander.ebinger@ucdenver.edu; James E. Svenson: jes@medicine.wisc.edu; Ying-Qi Zhao: yzhao5@wisc.edu; Zhanhai Li: zhanhaili@wisc.edu; Ryan P. Westergaard: rpw@medicine.wisc.edu; William J. Ehlenbach: wjehlen@medicine.wisc.edu; Michael D. Repplinger: mdreppli@medicine.wisc.edu

[1]University of Wisconsin – Madison, Department of Emergency Medicine, Madison, WI, USA

[2]University of Wisconsin – Madison, Department of Radiology, Madison, WI, USA

[3]University of Colorado School of Medicine, Department of Emergency Medicine, Denver, CO, USA

[4]University of Wisconsin – Madison, Department of Biostatistics and Medical Informatics, Madison, WI, USA

[5]University of Wisconsin – Madison, Department of Medicine, Madison, WI, USA

Abstract

Objective—To determine whether clinical scoring systems or physician gestalt can obviate the need for CT in patients with possible appendicitis.

Methods—Prospective, observational study of patients with abdominal pain at an academic emergency department from 2/2012–2/2014. Patients over 11 years old who had a CT ordered for possible appendicitis were eligible. All parameters needed to calculate the scores were recorded on standardized forms prior to CT. Physicians also estimated the likelihood of appendicitis. Test characteristics were calculated using clinical follow up as the reference standard. ROC curves were drawn.

Corresponding Author: Michael D. Repplinger, MD, PhD, Department of Emergency Medicine, University of Wisconsin School of Medicine & Public Health, 800 University Bay Drive, Suite 310 Mail Code 9123, Madison, WI 53705 (mdreppli@medicine.wisc.edu), Phone 608–890–5963, Fax 608-265-8241.

COMPETING INTERESTS

Perry J. Pickhardt, MD has the following financial disclosures: Co-founder, VirtuoCTC and shareholder, Cellectar Biosciences. The other authors have no financial conflicts.

CONTRIBUTION STATEMENT

Study planning was done by MDR, PJP, AE, and JES. Data collection was done by SKG, AE, and JBH. Data analysis was done by JES, YZ, and ZL. Manuscript preparation and editing was done by MDR, SKG, JBH, PJP, AE, JES, YZ, ZL, RPW, and WJE. MDR assumes responsibility for the overall content of this manuscript.

Results—Of the 287 patients (mean age [range], 31 [12–88] years; 60% women), the prevalence of appendicitis was 33%. The Alvarado score had a positive likelihood ratio [LR(+)] (95% confidence interval) of 2.2 (1.7–3) and a negative likelihood ratio [LR(−)] of 0.6 (0.4–0.7). The modified Alvarado score (MAS) had LR(+) 2.4 (1.6–3.4) and LR(−) 0.7 (0.6–0.8). The RIPASA score had LR(+) 1.3 (1.1–1.5) and LR(−) 0.5 (0.4–0.8). Physician-determined likelihood of appendicitis had LR(+) 1.3 (1.2–1.5) and LR(−) 0.3 (0.2–0.6). When combined with physician likelihoods, LR(+) and LR(−) was 3.67 and 0.48 (Alvarado), 2.33 and 0.45 (RIPASA), and 3.87 and 0.47 (MAS). The AUC was highest for physician-determined likelihood (0.72), but was not statistically significantly different from the clinical scores (RIPASA – 0.67, Alvarado 0.72, MAS 0.7).

Conclusions—Clinical scoring systems performed equally well as physician gestalt in predicting appendicitis. These scores do not obviate the need for imaging for possible appendicitis when a physician deems it necessary.

Keywords

Appendicitis; Clinical decision support; Computed tomography; Appendectomy; sensitivity and specificity

INTRODUCTION

Acute appendicitis is the most common indication for emergent abdominal surgery; over 250,000 appendectomies are performed each year in the United States.[1] In spite of its high incidence, acute appendicitis can often present a diagnostic challenge for the emergency physician since the classic presentation of periumbilical pain followed by nausea, vomiting, and pain migration to the right lower quadrant, occurs in only 50–60% of cases.[2] In an effort to improve the diagnostic accuracy of the clinical diagnosis of appendicitis, several scoring systems have been developed to systematically incorporate laboratory values, symptoms, physical exam findings, and patient characteristics. The Alvarado score was introduced in 1986, and is the most widely reported scoring system used to evaluate for appendicitis.[3] A modified version of the Alvarado score, which uses the same categories but does not require a white blood cell differential, was reported in 1994.[4] More recently, the Raja Isteri Pengiran Anak Saleha Appendicitis (RIPASA) score was created as an alternative to the Alvarado score, aiming to be more suitable to the Southeast Asian population.[5] These scores have been evaluated in a number of studies, but with mixed results regarding test accuracy.[5–9]

Conversely, the use of computed tomography (CT) for the diagnosis of appendicitis has yielded consistently impressive results. In contrast to the performance of clinical assessment alone, the negative laparotomy rate is less than 10% when using multi-detector CT for the evaluation of possible appendicitis and the false negative rate is less than 1%.[10–13] One study found that an observed negative laparotomy rate of 7.5% would have been further reduced to 4.1% if appendectomy had been avoided in patients where CT was interpreted as negative for appendicitis.[13] Another advantage of CT is its ability to evaluate for alternative diagnoses that are responsible for patients' symptoms. Indeed, alternative diagnoses are more common than the underlying frequency of appendicitis.[14] Consequently, there has there has been a 25% absolute increase in CT usage for patients eventually diagnosed with appendicitis.[15] Further, CT use in patients presenting to the emergency department (ED) with abdominal pain was shown to have doubled over a four year period (2001–2005) to 22.5%.[16] Unfortunately, there is an increased risk of developing cancer due to ionizing radiation exposure when undergoing CT scanning, particularly for the pediatric age group.[17]

This has led some researchers to re-evaluate the use of clinical scoring systems, particularly the implementation of a two cut-off system – one threshold below which the diagnosis is excluded and another, higher threshold above which the diagnosis is presumed.[18–20] Such a system could limit patient exposure to ionizing radiation while maintaining a satisfactorily low negative-appendectomy rate by appropriately risk stratifying patients to either very high risk (requiring operative intervention) or very low risk (requiring no intervention). Unlike these structured scoring systems, physicians often order CT scans for patients due to anecdotal experience or, particularly in the United States, malpractice risk intolerance.[21,22] It is possible that applying these scoring systems to patients for whom a physician has ordered a CT scan to evaluate for appendicitis could obviate the use of unnecessary imaging. Similar efforts have been made for patients with chest pain syndromes and have led to decreased cost and radiation exposure without concomitant adverse events.[23]

Therefore, the primary objective of this study is to determine whether a clinical score – the Alvarado score, modified Alvarado score, and RIPASA score – or physician-determined likelihood of appendicitis are accurate enough to obviate the need for CT imaging in a prospectively identified cohort of patients for whom a CT scan was ordered to evaluate for appendicitis. Secondarily, this study aims to ascertain if the accuracy of the scoring systems is enhanced when combined with physician-determined likelihood of appendicitis.

METHODS

Study Design and Setting

This is a HIPAA-compliant, IRB-approved prospective, observational study of a convenience sample of emergency department (ED) patients for whom a CT was ordered to evaluate for appendicitis. The study was conducted at the University of Wisconsin Hospital's ED, which has 48,000 patient encounters annually, between February 2012 and February 2014.

Selection of Participants

Patients were eligible if they were over 11 years old and had a CT ordered to evaluate for appendicitis. Patients who were incarcerated, pregnant, post-appendectomy, unable to have intravenous contrast, unable to speak or read English, or who lacked capacity to provide informed consent/assent were excluded. The age restriction for this study is due to the fact that it was part of a larger study which required that patients be able to undergo MRI with minimal need for sedation. Further, we required the use of CT as an eligibility criterion because our goal was to determine whether a clinical scoring system could obviate the need to order a CT scan in patients for whom the emergency physician, influenced by clinical experience, had determined CT was warranted. While this naturally excluded those at very high and very low risk, those patients were not of interest for this study's purpose. To determine the total number eligible, we queried our medical imaging database for all abdominal CT scans ordered during study hours for patients over 11 years old. We then read through every indication for the scan and only included those which listed appendicitis as the reason for testing.

The decision to order a CT was made by the treating physicians, which included attending and resident emergency physicians, independent of the study protocol. In our center's practice, surgical consultation and ultrasound were not specifically recommended prior to CT, nor were physicians required to calculate a clinical score prior to ordering advanced imaging like CT. Screening and enrollment was completed by emergency physicians or a research assistant who monitored the emergency department's real-time electronic track board. The hours of enrollment varied through the course of the study period, based on MRI availability. All subjects supplied written, informed consent to participate in this study. Written informed assent was obtained for any patient under 18 years old.

Methods and Measurements

Prior to each patient going to CT, the treating physician completed a standardized data collection form, recording all components necessary to calculate each of the three clinical scores. Physicians were then asked to estimate the patient's likelihood of appendicitis by selecting one of four probability intervals: <40%, 40–60%, 60–80%, and >80%. These cut-off values were based on the historically accepted negative laparotomy rate of 20%. A recent meta-analysis has also suggested that the use of cut-off values at the 40% and 60% level were ideal for ruling out appendicitis when using clinical scoring systems.[9]

Outcomes

The primary outcome was the diagnostic accuracy of each clinical score. These were calculated using data recorded on standardized collection forms. Any laboratory findings that were not initially recorded on the form were subsequently collected from the patient's electronic medical record by a research assistant.

As mentioned previously, other studies have proposed using a two-threshold model for applying these clinical scores to clinical practice. The upper threshold defines a level above which the patients is presumed to have appendicitis, and should undergo operative intervention. The lower threshold is the level below which the patient is presumed to not have appendicitis, and therefore should be discharged home without further evaluation/intervention. Intermediate scores (i.e. – between the threshold levels) should undergo further evaluation (e.g. – observation, imaging, etc.). Based on previous literature, the thresholds for the Alvarado and modified Alvarado scores were <4 and ≥7 while 5 and 7.5 were the thresholds used for the RIPASA score.[3–5,24] Since the variable "foreign identify card" was not relevant to our population, we also report results for an upper threshold of 6.5 for the RIPASA score.

In the case of physician-determined likelihood of appendicitis, we defined 60% as the single threshold to determine the "test" to be positive or negative. Notably, the physicians completing the form were unaware of this threshold at the time of prospective evaluation. Though a seemingly low threshold, it was based on a previous meta-analysis suggesting that this was the optimal cut-off threshold for physician gestalt in diagnosing appendicitis.[9]

Surgical findings, pathology reports, CT findings, and clinical follow up were used as the reference standard to determine whether the patient had appendicitis. Clinical follow up was attempted first by phone call, which occurred at least one month from the index emergency department visit to ascertain whether the CT result missed a case of appendicitis. If this was unsuccessful in contacting the patient, a chart review was performed instead. Patients who underwent operative intervention or exploration had surgical and pathological results abstracted from the medical record in an effort to identify falsely positive CT results.

Analysis

Descriptive statistics were used for patient demographics. We used two-tailed t-tests for continuous variables and chi-square tests for categorical variables. The sensitivity, specificity, positive likelihood ratio [LR(+)], negative likelihood ratio [LR(−)], positive predictive value (PPV), and negative predictive value (NPV) for the clinical scoring systems and the physician-determined likelihood scale were calculated using the reference standard defined above. These values are reported as point estimates with 95% confidence intervals. Receiver operating characteristic (ROC) curves were generated for the clinical scoring systems and physician-determined likelihood scale. Areas under the curve (AUC) for each ROC curve were calculated to evaluate and compare the diagnostic potential of each scoring system. All statistical analyses were performed using SAS, version 9.4 (SAS Institute Inc., Cary NC).

RESULTS

Characteristics of study subjects

During the period of enrollment (2/2012–2/2014), 1092 patients met eligibility criteria during study hours on retrospective review of CT order requests. Of the 1092, a total of 287 subjects (26%) were enrolled in this study (see Table 1 for demographics). For those included in the study, symptom duration was <48 hours in 182 (63.4%) patients and 142 (49.4%) had an elevated WBC count.

Sixteen patients did not have a clinic follow-up visit and were not reachable by phone call to verify that they did not have appendicitis. These were considered lost to follow up, but were considered negative for appendicitis in light of a lack of repeat emergency department visit and the negative findings reported in their CT reports. The mean age of these patients was 26 years old (SD 7.1 years), and 11 (68.8%) were female.

Ten patients did not have a urinalysis recorded in their chart, but the other measures were recorded. Since the RIPASA score is the only one to require a urinalysis for its calculation, these patients were not included in the RIPASA score results, but were included for physician gestalt, Alvarado score, and modified Alvarado score.

Though CT was a requirement for study participation, 62 (21.6%) patients also had ultrasound performed. Of these, 23 (37%) were done in patients <18 years old.

Main results

At the higher, “rule-in” cut-off threshold, the RIPASA score had the highest sensitivity (0.78, 95% confidence interval [95% CI] 0.68–0.86), but the lowest specificity (0.36, 95% CI 0.29–0.44). Conversely, the modified Alvarado score had the lowest sensitivity (0.47, 95% CI 0.37–0.57), but the highest specificity (0.81, 95% CI 0.75–0.86). The original Alvarado score had test characteristics between these values. Additionally, we calculated the test characteristics for the clinical scoring systems at a lower, “rule-out” threshold. The NPV for each score varied from 0.75 for the modified Alvarado score to 0.89 for the RIPASA score. Physician-determined likelihood of appendicitis had test characteristics similar to these scores (Table 2).

In table 3, we report the test characteristics of our cohort when stratified by age (pediatric versus adult populations) and gender using the traditional upper level cut-off suggested for each score. The PPV was higher in males versus females when using the Alvarado (0.67 vs 0.4) and RIPASA (0.51 vs 0.3) scores as well as physician-determined likelihood of appendicitis (0.69 vs 0.43) while NPV was higher in females versus males with the modified Alvarado score (0.82 vs 0.63) and physician-determined likelihood of appendicitis (0.89 vs 0.7). In other cases, the scores were not significantly different when stratified by age group or gender (i.e. – point estimates may differ, but the 95% confidence intervals overlap considerably). Additionally, we report the test characteristics of each scoring system at each cut-off level in Appendix 1. Notably, this table presents the numerical data for the results displayed in our Figure, the ROC curve.

We also evaluated the use of these clinical scores when used in series with physician-determined pre-test probability estimates, but this did not substantially affect the results (Table 4).

Receiver operator characteristic (ROC) curves were also generated (Figure). The AUC was greatest for the Alvarado score and physician-determined likelihood of appendicitis (0.72, 95% CI 0.66–0.78), while the RIPASA score had the lowest AUC (0.67, 95% CI 0.60–0.74). None of the values were statistically different from one another, as evidenced by the significant overlap in confidence intervals for the scores’ AUCs.

DISCUSSION

The aim of our study was to compare the ability of three clinical scoring systems – the Alvarado score, modified Alvarado score, and RIPASA score – to diagnose or exclude appendicitis in a prospectively identified cohort of patients for whom a CT scan was ordered to evaluate for appendicitis. We found that the diagnostic accuracy of these scores was not sufficient to obviate the use of CT in this situation. Moreover, none of the scores performed any better than physician-estimated likelihood of appendicitis.

Previous reports have shown that the use of medical imaging, particularly CT, has decreased the negative laparotomy rate for patients who undergo appendectomy.[10] However, CT exposes patients to ionizing radiation, which increases a person’s lifetime risk for developing cancer.[17] Subsequently, the American College of Radiology has made efforts to reduce patients' exposure to radiation through efforts like the "Image Gently" campaign and supporting the ALARA (as low as reasonably achievable) concept, urging physicians to use the minimum amount of radiation necessary to generate images of diagnostic quality. This can be achieved through either decreasing the amount of radiation used when undergoing such imaging or developing clinical scoring systems that obviate the need for these imaging tests by either making or excluding the diagnosis without the need for further testing.

Previous research regarding the utility of clinical scoring systems for the diagnosis of appendicitis has yielded mixed results.[6,7,25] While early data suggested that the Alvarado score had sufficient test characteristics to direct care (discharge versus operative intervention), a meta-analysis performed in 2011, which incorporated data from 42 published studies, demonstrated insufficient specificity (81%, 95% CI 76–85%) to rule-in the diagnosis of appendicitis.[8] Moreover, the modified Alvarado score has insufficient sensitivity to rule out appendicitis when a cut-off value of 4 is used.[24] In fact, unstructured physician judgment had a higher sensitivity (93%, 95% CI 82–98%) when compared with the Alvarado score (72%, 95% CI 58–84%) in that study.

Our results reinforce and expand upon previously reported findings. None of the clinical scoring systems had sufficient positive predictive value to guide surgical intervention. Had one followed through with appendectomy simply based on these scores, the negative laparotomy rate would have ranged from 61% in the case of the RIPASA score to as low as 46% for the modified Alvarado score. Conversely, had one used a negative result as definitive evidence that the patient did not have appendicitis, 11–25% of patients would have been falsely negative. Physician-estimated likelihood of appendicitis was just as accurate (Table 2). As suggested by a recent meta-analysis, we also evaluated the test characteristics of these clinical scores when used in combination with physician-estimated likelihoods, but did not find substantial improvements (Table 4).

We acknowledge several limitations with this study. First, it is a single-center study, which may limit generalizability. However, other studies have yielded similar results.[24] Secondly, we used a convenience sample of patients, amounting to 26% of the eligible population. The chief reason for the relatively low recruitment was the fact that we relied on physicians to recruit patients for most of the study period. Clinical demands and other competing interests for these physicians' time precluded a more uniform approach to enrollment. We were, however, able to collect some basic data by performing frequent audits of the medical record to see which patients would have been eligible for enrollment based on the minimum criteria of age and CT order. These groups were found to have some statistically significant differences including the incidence of appendicitis (18% vs 33%) and age (33 years vs 39 years), though the gender distribution was not different. These differences were likely due to physicians thinking of enrollment in patients with more "classic" presentations of appendicitis, which more commonly occurs in younger patients, increasing the prevalence of disease while decreasing the average age for the study cohort. We would argue that the age difference is likely not clinically relevant and the incidence of appendicitis in our study population is similar to that of previous reports.[6] Finally, physician-determined likelihood of appendicitis was asked after the physician ascertained all components necessary to calculate the clinical scores (though the method of calculation was not provided to them). This may have encouraged physicians to incorporate data helpful in estimating pre-test probability that they otherwise may not have consciously or explicitly considered. This may have biased our results, likely in favor of physician-estimated probability. It is also possible that individual physician practice includes the calculation of one of these scores prior to ordering CT, which may cause physician-determined estimates to be more congruent with scoring systems. However, as mentioned in the methods section, this is not typical practice at our center, particularly outside of the pediatric patient population.

In summary, our results do not support using either the high or low thresholds of clinical scoring systems for the diagnostic evaluation of patients with possible appendicitis when a physician has already determined that medical imaging is clinically warranted. Physician-determined probability estimates were as accurate as these systems, yet none of these approaches was sufficiently accurate to direct the management of these patients and would have led to both negative appendectomies and inappropriate discharge of patients with appendicitis. Additionally, using a combination of physician-determined pre-test probabilities and clinical scores did not enhance the accuracy enough to obviate the need for imaging at either the high or low thresholds. Based on these findings, the clinical practice at our center has not changed. In particular, when deemed necessary by a clinician, CT scans continue to be routinely ordered. Future efforts to limit patient exposure to ionizing radiation should be aimed at using medical imaging tests with limited radiation exposure (low-dose CT) or no exposure at all (ultrasound and MRI).

Acknowledgments

The project described was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427 and KL2TR000428. Additional support was provided by the National Institute on Aging, grant K23AG038352, and the National Institute on Drug Abuse, grant K23DA032306. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Support was also provided by the University of Wisconsin – Madison, Department of Radiology.

Appendix 1: Test characteristics for each of the scoring systems at each cut-off level. Results are reported as point estimates with 95% confidence intervals

	Alvarado Score				Modified Alvarado Score				RIPASA Score			
Score	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
1	1 (1–1)	0.01 (0–0.02)	0.34 (0.29–0.39)	1 (1–1)	1 (1–1)	0.01 (0–0.02)	0.34 (0.29–0.39)	1 (1–1)				
2	1 (1–1)	0.02 (0–0.04)	0.34 (0.28–0.4)	1 (1–1)	1 (1–1)	0.02 (0–0.04)	0.34 (0.28–0.4)	1 (1–1)				
3	0.99 (0.97–1)	0.07 (0.03–0.11)	0.35 (0.29–0.41)	0.93 (0.8–1)	0.99 (0.97–1)	0.07 (0.03–0.11)	0.35 (0.29–0.41)	0.93 (0.8–1)	1 (1–1)	0 (0–0)	0.34 (0.28–0.4)	0 (0–0)
4	0.94 (0.89–0.99)	0.23 (0.17–0.29)	0.38 (0.32–0.44)	0.88 (0.79–0.97)	0.92 (0.86–0.98)	0.24 (0.18–0.3)	0.38 (0.32–0.44)	0.85 (0.75–0.95)	1 (1–1)	0.01 (0–0.02)	0.34 (0.28–0.4)	1 (1–1)
5	0.85 (0.78–0.92)	0.45 (0.38–0.52)	0.44 (0.37–0.51)	0.86 (0.79–0.93)	0.83 (0.76–0.91)	0.47 (0.4–0.54)	0.44 (0.37–0.51)	0.85 (0.78–0.92)	0.99 (0.97–1)	0.04 (0.01–0.07)	0.34 (0.28–0.4)	0.89 (0.68–1.1)
6	0.73 (0.64–0.82)	0.6 (0.53–0.67)	0.48 (0.4–0.56)	0.81 (0.75–0.87)	0.69 (0.6–0.78)	0.63 (0.56–0.7)	0.49 (0.41–0.57)	0.8 (0.74–0.86)	0.97 (0.93–1)	0.13 (0.08–0.18)	0.36 (0.3–0.42)	0.89 (0.77–1.01)
7	0.59 (0.49–0.69)	0.73 (0.67–0.79)	0.53 (0.44–0.62)	0.78 (0.72–0.84)	0.46 (0.36–0.56)	0.81 (0.75–0.87)	0.54 (0.43–0.65)	0.75 (0.69–0.81)	0.84 (0.77–0.92)	0.3 (0.23–0.37)	0.38 (0.31–0.45)	0.79 (0.69–0.89)

	Alvarado Score				**Modified Alvarado Score**				**RIPASA Score**			
Score	**Sensitivity**	**Specificity**	**Positive Predictive Value**	**Negative Predictive Value**	**Sensitivity**	**Specificity**	**Positive Predictive Value**	**Negative Predictive Value**	**Sensitivity**	**Specificity**	**Positive Predictive Value**	**Negative Predictive Value**
8	0.42 (0.32–0.52)	0.85 (0.8–0.9)	0.58 (0.46–0.7)	0.74 (0.68–0.8)	0.19 (0.11–0.27)	0.93 (0.89–0.97)	0.58 (0.41–0.75)	0.7 (0.64–0.76)	0.73 (0.64–0.82)	0.49 (0.42–0.56)	0.42 (0.34–0.5)	0.78 (0.7–0.86)
9	0.18 (0.1–0.26)	0.94 (0.91–0.97)	0.61 (0.43–0.79)	0.69 (0.63–0.75)	0.01 (0–0.03)	0.99 (0.98–1)	0.33 (0–0.86)	0.67 (0.62–0.72)	0.7 (0.61–0.79)	0.63 (0.56–0.7)	0.49 (0.4–0.58)	0.81 (0.75–0.87)
10	0.01 (0–0.03)	0.99 (0.98–1)	0.33 (0–0.86)	0.67 (0.62–0.72)	0 (0–0)	1 (1–1)	0 (0–0)	0.67 (0.62–0.72)	0.46 (0.36–0.56)	0.74 (0.68–0.8)	0.47 (0.37–0.57)	0.73 (0.67–0.79)
11									0.39 (0.29–0.49)	0.84 (0.79–0.89)	0.55 (0.43–0.67)	0.73 (0.67–0.79)
12									0.23 (0.15–0.32)	0.93 (0.89–0.97)	0.64 (0.48–0.8)	0.7 (0.64–0.76)
13									0.09 (0.03–0.15)	0.97 (0.95–0.99)	0.62 (0.36–0.88)	0.68 (0.62–0.74)
14									0.04 (0–0.08)	0.99 (0.98–1)	0.8 (0.45–1)	0.67 (0.61–0.73)
15									0 (0–0)	1 (1–1)	0 (0–0)	0.66 (0.6–0.72)
					Physician-Estimated Likelihood							
				Score	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value				
				<40%	1 (1–1)	0 (0–0)	0.33 (0.28–0.38)	0 (0–0)				
				40–60%	0.89 (0.83–0.95)	0.34 (0.27–0.41)	0.4 (0.33–0.47)	0.86 (0.78–0.94)				
				60–80%	0.69 (0.6–0.78)	0.7 (0.64–0.76)	0.54 (0.45–0.63)	0.82 (0.76–0.88)				
				>80%	0.23 (0.15–0.32)	0.95 (0.92–0.98)	0.71 (0.55–0.87)	0.71 (0.65–0.77)				

Results are reported as point estimates with 95% confidence intervals.

References

1. Mason RJ. Surgery for appendicitis: is it necessary? Surg Infect. 2008; 9(4):481–488. DOI: 10.1089/sur.2007.079
2. Birnbaum BA, Wilson SR. Appendicitis at the millennium. Radiology. 2000; 215(2):337–348. DOI: 10.1148/radiology.215.2.r00ma24337 [PubMed: 10796905]
3. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med. 1986; 15(5):557–564. [PubMed: 3963537]
4. 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):418–419. [PubMed: 7702329]
5. N N, Mohammed A, Shanbhag V, Ashfaque K, S A P. A Comparative Study of RIPASA Score and ALVARADO Score in the Diagnosis of Acute Appendicitis. J Clin Diagn Res JCDR. 2014; 8(11):NC03–NC05. DOI: 10.7860/JCDR/2014/9055.5170
6. Schneider C, Kharbanda A, Bachur R. Evaluating Appendicitis Scoring Systems Using a Prospective Pediatric Cohort. Ann Emerg Med. 2007; 49(6):778–784.e1. DOI: 10.1016/j.annemergmed.2006.12.016 [PubMed: 17383771]
7. 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(3): 165–169. DOI: 10.1097/PEC.0b013e31820d6460 [PubMed: 21346681]
8. Ohle R, O'Reilly F, O'Brien KK, Fahey T, Dimitrov BD. The Alvarado score for predicting acute appendicitis: a systematic review. BMC Med. 2011; 9:139.doi: 10.1186/1741-7015-9-139 [PubMed: 22204638]
9. Ebell MH, Shinholser J. What Are the Most Clinically Useful Cutoffs for the Alvarado and Pediatric Appendicitis Scores? A Systematic Review. Ann Emerg Med. Apr.2014 doi: 10.1016/j.annemergmed.2014.02.025

10. Raja AS, Wright C, Sodickson AD, et al. Negative appendectomy rate in the era of CT: an 18-year perspective. Radiology. 2010; 256(2):460–465. DOI: 10.1148/radiol.10091570 [PubMed: 20529988]

11. Coursey CA, Nelson RC, Patel MB, et al. Making the diagnosis of acute appendicitis: do more preoperative CT scans mean fewer negative appendectomies? A 10-year study. Radiology. 2010; 254(2):460–468. DOI: 10.1148/radiol.09082298 [PubMed: 20093517]

12. Raman SS, Osuagwu FC, Kadell B, Cryer H, Sayre J, Lu DSK. Effect of CT on false positive diagnosis of appendicitis and perforation. N Engl J Med. 2008; 358(9):972–973. DOI: 10.1056/NEJMc0707000 [PubMed: 18305278]

13. Pickhardt PJ, Lawrence EM, Pooler BD, Bruce RJ. Diagnostic performance of multidetector computed tomography for suspected acute appendicitis. Ann Intern Med. 2011; 154(12):789–796. W - 291. DOI: 10.7326/0003-4819-154-12-201106210-00006 [PubMed: 21690593]

14. Pooler BD, Lawrence EM, Pickhardt PJ. Alternative diagnoses to suspected appendicitis at CT. Radiology. 2012; 265(3):733–742. DOI: 10.1148/radiol.12120614 [PubMed: 23023965]

15. Otero HJ, Ondategui-Parra S, Erturk SM, Ochoa RE, Gonzalez-Beicos A, Ros PR. Imaging utilization in the management of appendicitis and its impact on hospital charges. Emerg Radiol. 2008; 15(1):23–28. DOI: 10.1007/s10140-007-0678-x [PubMed: 17972120]

16. Pines JM. Trends in the rates of radiography use and important diagnoses in emergency department patients with abdominal pain. Med Care. 2009; 47(7):782–786. DOI: 10.1097/MLR.0b013e31819748e9 [PubMed: 19536032]

17. Brenner DJ, Doll R, Goodhead DT, et al. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. Proc Natl Acad Sci U S A. 2003; 100(24):13761–13766. DOI: 10.1073/pnas.2235592100 [PubMed: 14610281]

18. Tan WJ, Acharyya S, Goh YC, et al. Prospective comparison of the Alvarado score and CT scan in the evaluation of suspected appendicitis: a proposed algorithm to guide CT use. J Am Coll Surg. 2015; 220(2):218–224. DOI: 10.1016/j.jamcollsurg.2014.10.010 [PubMed: 25488354]

19. 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. Arch Surg Chic Ill 1960. 2011; 146(1):64–67. DOI: 10.1001/archsurg.2010.297

20. 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): 489–493. DOI: 10.1016/j.ajem.2006.08.020 [PubMed: 17543650]

21. Probst MA, Kanzaria HK, Schriger DL. A conceptual model of emergency physician decision making for head computed tomography in mild head injury. Am J Emerg Med. 2014; 32(6):645–650. DOI: 10.1016/j.ajem.2014.01.003 [PubMed: 24560384]

22. Rohacek M, Buatsi J, Szucs-Farkas Z, et al. Ordering CT pulmonary angiography to exclude pulmonary embolism: defense versus evidence in the emergency room. Intensive Care Med. 2012; 38(8):1345–1351. DOI: 10.1007/s00134-012-2595-z [PubMed: 22584801]

23. Kline JA, Jones AE, Shapiro NI, et al. Multicenter, randomized trial of quantitative pretest probability to reduce unnecessary medical radiation exposure in emergency department patients with chest pain and dyspnea. Circ Cardiovasc Imaging. 2014; 7(1):66–73. DOI: 10.1161/CIRCIMAGING.113.001080 [PubMed: 24275953]

24. Meltzer AC, Baumann BM, Chen EH, Shofer FS, Mills AM. Poor sensitivity of a modified Alvarado score in adults with suspected appendicitis. Ann Emerg Med. 2013; 62(2):126–131. DOI: 10.1016/j.annemergmed.2013.01.021 [PubMed: 23623557]

25. Jo YH, Kim K, Rhee JE, et al. The accuracy of emergency medicine and surgical residents in the diagnosis of acute appendicitis. Am J Emerg Med. 2010; 28(7):766–770. DOI: 10.1016/j.ajem.2009.03.017 [PubMed: 20837252]

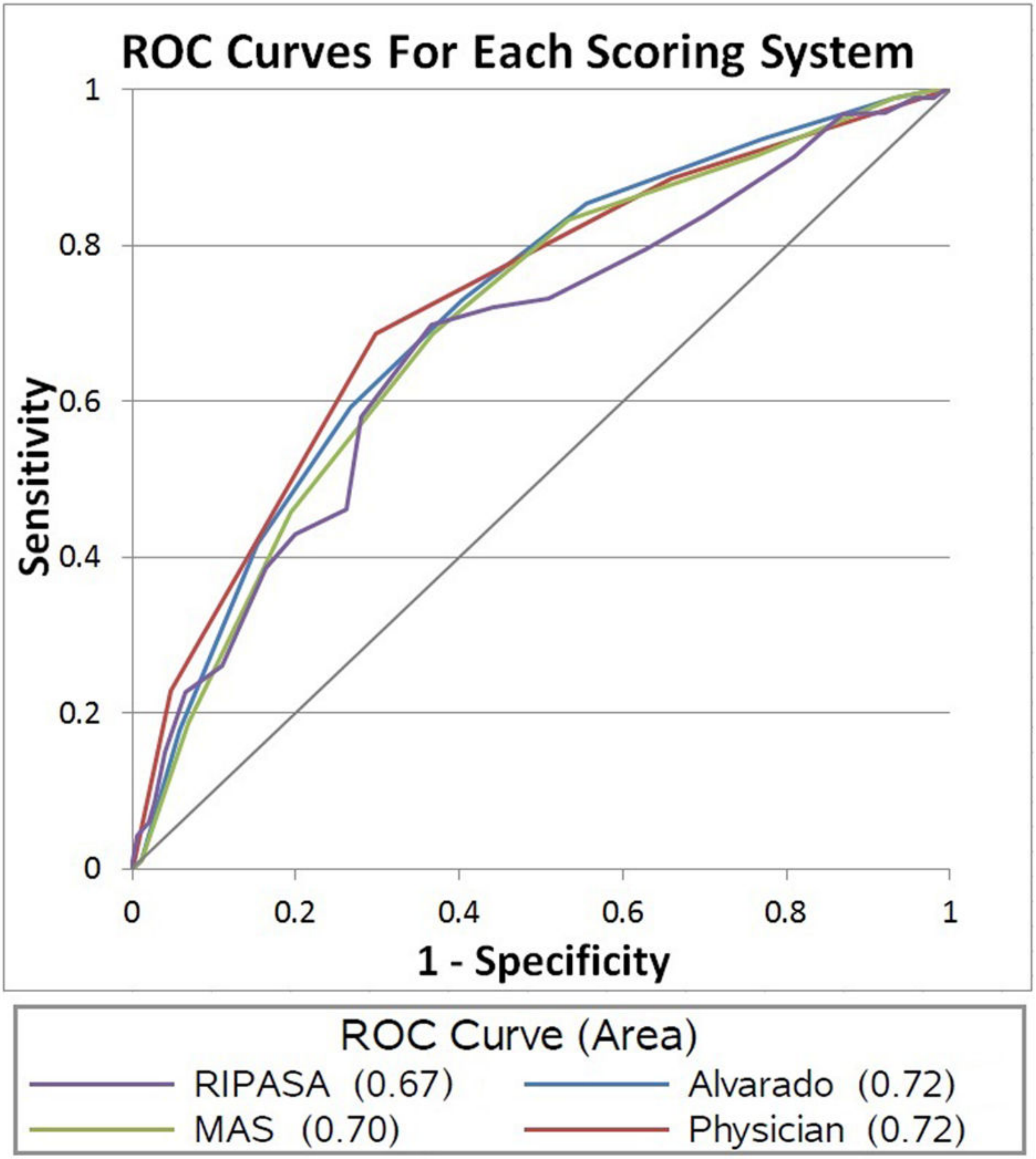

Figure. Receiver operator characteristic (ROC) curve for each clinical scoring system and physician impression

Area under the curve (AUC) is also reported. MAS = Modified Alvarado Score; Physician = Physician-determined likelihood of appendicitis.

Table 1

Demographic information about patients included versus those eligible, but not included.

	Included (N=287)	Eligible, Not Included (N = 805)	p-value
Female (%)	172 (60%)	473 (58.7%)	0.72
Less than 18 years old (%)	23 (8%)	76 (9.4%)	0.47
Over 65 years old (%)	10 (3.5%)	76 (9.4%)	0.001
Prevalence of appendicitis	94 (33%)	146 (18.2%)	<0.0001
Mean age in years (range, SD)	33 (12–88, 15.2)	39 (12–95, 18.2)	<0.0001
Racial identification:			
White	82.5%	87.8%	0.02
Black or African American	8.8%	5.9%	0.09
Asian	4.4%	3.5%	0.49
American Indian or Alaska Native	1.3%	0.4%	0.1
Native Hawaiian or Other Pacific Islander	0.6%	0%	0.03
Declined to Answer	1.2%	0.7%	0.42
Not listed	1.2%	1.7%	0.56

Table 2

Test characteristics and receiver operator characteristic (ROC) curve results for each clinical scoring system

Results are presented as point estimates with 95% confidence intervals in parentheses. Each scoring system’s ROC curve’s area under the curve (AUC) was compared by the Chi-square test based on the asymptotic chi-square distribution of the Wald statistic. Cutoff values are listed with their corresponding test characteristics.

Clinical Score	Area under the curve (AUC)	P-value compared with Physician	Cutoff Value	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Positive Likelihood Ratio	Negative Likelihood Ratio
RIPASA	0.67 (0.6–0.74)	0.11	7.5	0.78 (0.68–0.86)	0.36 (0.29–0.44)	0.39 (0.32–0.47)	0.76 (0.66–0.84)	1.3 (1.1–1.5)	0.5 (0.4–0.8)
			6.5	0.88 (0.82–0.95)	0.22 (0.16–0.28)	0.36 (0.29–0.42)	0.8 (0.69–0.91)	1.1 (1.0–1.2)	0.4 (0.2–0.9)
			5	0.99 (0.94–1)	0.04 (0.02–0.09)	0.35 (0.29–0.41)	0.89 (0.82–1)	1 (1.0–1.1)	0.2 (0.0–2.2)
Alvarado	0.72 (0.66–0.78)	0.93	7	0.61 (0.51–0.71)	0.74 (0.67–0.8)	0.53 (0.43–0.62)	0.79 (0.73–0.85)	2.2 (1.7–3.0)	0.6 (0.4–0.7)
			4	0.94 (0.89–0.99)	0.23 (0.17–0.29)	0.37 (0.31–0.43)	0.88 (0.79–0.97)	1.2 (1.1–1.3)	0.3 (0.1–0.6)
Modified Alvarado	0.7 (0.64–0.77)	0.54	7	0.47 (0.37–0.57)	0.81 (0.75–0.86)	0.54 (0.43–0.65)	0.76 (0.70–0.82)	2.4 (1.6–3.4)	0.7 (0.6–0.8)
			4	0.97 (0.93–1)	0.05 (0.02–0.08)	0.33 (0.27–0.39)	0.75 (0.46–1)	1.2 (1.1–1.3)	0.4 (0.2–0.7)
Physician Estimate	0.72 (0.66–0.78)	N/A	60%	0.88 (0.82–0.95)	0.34 (0.27–0.4)	0.39 (0.33–0.46)	0.86 (0.77–0.94)	1.3 (1.2–1.5)	0.3 (0.2–0.6)

Results are presented as point estimates with 95% confidence intervals in parentheses. Each scoring system’s ROC curve’s area under the curve (AUC) was compared by the Chi-square test based on the asymptotic chi-square distribution of the Wald statistic. Cutoff values are listed with their corresponding test characteristics.

Table 3

Test characteristics of each scoring system, stratified by age group and gender

Results are reported as point estimates with 95% confidence intervals.

Score	Cohort	<18 years old	≥18 years old	Males	Females
Alvarado Score	Sensitivity	0.8 (0.55–1)	0.57 (0.47–0.68)	0.63 (0.5–0.76)	0.55 (0.4–0.71)
	Specificity	0.69 (0.44–0.94)	0.74 (0.68–0.8)	0.72 (0.61–0.83)	0.74 (0.66–0.82)
	Positive Predictive Value	0.67 (0.4–0.94)	0.51 (0.41–0.61)	0.67 (0.54–0.8)	0.4 (0.27–0.53)
	Negative Predictive Value	0.82 (0.59–1)	0.78 (0.72–0.84)	0.69 (0.58–0.8)	0.83 (0.76–0.9)
	Positive Likelihood Ratio	2.6 (1.09–6.22)	2.16 (1.59–2.93)	2.26 (1.44–3.55)	2.09 (1.4–3.11)
	Negative Likelihood Ratio	0.29 (0.08–1.06)	0.58 (0.45–0.75)	0.51 (0.35–0.75)	0.61 (0.43–0.86)
Modified Alvarado Score	Sensitivity	0.7 (0.42–1)	0.43 (0.33–0.54)	0.48 (0.35–0.62)	0.43 (0.28–0.59)
	Specificity	0.77 (0.54–1)	0.81 (0.75–0.87)	0.77 (0.66–0.88)	0.82 (0.75–0.89)
	Positive Predictive Value	0.7 (0.42–0.98)	0.52 (0.4–0.64)	0.65 (0.5–0.8)	0.44 (0.29–0.59)
	Negative Predictive Value	0.77 (0.54–1)	0.75 (0.69–0.81)	0.63 (0.52–0.74)	0.82 (0.75–0.89)
	Positive Likelihood Ratio	3.03 (1.04–8.85)	2.25 (1.53–3.32)	2.1 (1.23–3.59)	2.42 (1.45–4.03)
	Negative Likelihood Ratio	0.39 (0.14–1.05)	0.7 (0.57–0.85)	0.67 (0.5–0.9)	0.69 (0.52–0.91)
RIPASA Score	Sensitivity	0.78 (0.51–1)	0.8 (0.71–0.89)	0.81 (0.7–0.92)	0.78 (0.65–0.91)
	Specificity	0.31 (0.06–0.56)	0.38 (0.31–0.45)	0.29 (0.17–0.41)	0.41 (0.32–0.5)
	Positive Predictive Value	0.44 (0.2–0.68)	0.39 (0.32–0.46)	0.51 (0.4–0.62)	0.3 (0.21–0.39)
	Negative Predictive Value	0.67 (0.29–1)	0.79 (0.7–0.88)	0.62 (0.43–0.81)	0.85 (0.76–0.94)
	Positive Likelihood Ratio	1.12 (0.68–1.85)	1.28 (1.09–1.5)	1.13 (0.91–1.4)	1.32 (1.06–1.64)

Score	Cohort	<18 years old	≥18 years old	Males	Females
	Negative Likelihood Ratio	0.72 (0.17–3.13)	0.54 (0.34–0.86)	0.67 (0.33–1.34)	0.54 (0.29–1)
Physician-Estimated Likelihood>60%	Sensitivity	0.6 (0.3–0.96)	0.7 (0.6–0.8)	0.65 (0.52–0.78)	0.74 (0.61–0.88)
	Specificity	0.46 (0.19–0.73)	0.72 (0.65–0.79)	0.74 (0.63–0.85)	0.68 (0.6–0.76)
	Positive Predictive Value	0.46 (0.19–0.73)	0.55 (0.46–0.64)	0.69 (0.56–0.82)	0.43 (0.32–0.54)
	Negative Predictive Value	0.6 (0.3–0.9)	0.83 (0.77–0.89)	0.7 (0.59–0.81)	0.89 (0.83–0.95)
	Positive Likelihood Ratio	1.11 (0.54–2.27)	2.48 (1.89–3.26)	2.47 (1.55–3.93)	2.34 (1.71–3.19)
	Negative Likelihood Ratio	0.87 (0.33–2.27)	0.42 (0.3–0.59)	0.48 (0.32–0.71)	0.38 (0.23–0.64)

Results are reported as point estimates with 95% confidence intervals.

Table 4

Test characteristics for each clinical scoring system when combined with physician-determined pre-test probability of appendicitis

Each row represents a combination of physician estimated probability of appendicitis with either the high (“rule-in”) or low (“rule-out”) cut-off threshold for each of the scoring systems. Results are reported as point estimates with 95% confidence intervals in parentheses.

Clinical Score	Cutoff Value	Physician-Estimated Likelihood of Appendicitis	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Positive Likelihood Ratio	Negative Likelihood Ratio
Alvarado	7	≥60%	0.51 (0.41–0.61)	0.87 (0.82–0.91)	0.65 (0.54–0.76)	0.78 (0.73–0.84)	3.67 (2.44–5.53)	0.58 (0.47–0.71)
	4	<60%	0.63 (0.53–0.73)	0.71 (0.64–0.78)	0.53 (0.44–0.63)	0.79 (0.72–0.85)	2.36 (1.8–3.09)	0.48 (0.36–0.64)
RIPASA	7.5	≥60%	0.6 (0.49–0.7)	0.72 (0.65–0.79)	0.53 (0.43–0.63)	0.78 (0.7–0.84)	2.33 (1.75–3.1)	0.51 (0.39–0.67)
	6.5	≥60%	0.63 (0.53–0.73)	0.71 (0.64–0.77)	0.53 (0.43–0.63)	0.79 (0.71–0.85)	2.31 (1.76–3.04)	0.48 (0.36–0.64)
	5	<60%	0.65 (0.55–0.75)	0.71 (0.64–0.77)	0.54 (0.44–0.63)	0.8 (0.72–0.85)	2.38 (1.82–3.12)	0.45 (0.33–0.61)
Modified Alvarado	7	≥60%	0.39 (0.29–0.49)	0.9 (0.86–0.94)	0.66 (0.53–0.79)	0.75 (0.7–0.81)	3.87 (2.36–6.35)	0.68 (0.58–0.8)
	4	<60%	0.63 (0.53–0.73)	0.72 (0.65–0.78)	0.54 (0.44–0.63)	0.79 (0.72–0.85)	2.41 (1.83–3.17)	0.47 (0.35–0.63)

Each row represents a combination of physician estimated probability of appendicitis with either the high (“rule-in”) or low (“rule-out”) cut-off threshold for each of the scoring systems. Results are reported as point estimates with 95% confidence intervals in parentheses.