Diagnostic Performance of Emergency Physician Gestalt for Predicting Acute Appendicitis in Patients Age 5 to 20 Years

Laura E. Simon[1,2], Mamata V. Kene, MD, MPH[3], E. Margaret Warton, MPH[1], Adina S. Rauchwerger, MPH[1], David R. Vinson, MD[1,4], Mary E. Reed, DrPH[1], Uli K. Chettipally, MD, MPH[5], Dustin G. Mark, MD[1,6], Dana R. Sax, MD, MPH[6], D. Ian McLachlan, MD, MPH[5], Dale M. Cotton, MD[7], James S. Lin, MD[8], Gabriela Vazquez-Benitez, PhD[9], Anupam B. Kharbanda, MD, MSc[10], Elyse O. Kharbanda, MD, MPH[9], and Dustin W. Ballard, MD, MBE[1,11]

ABSTRACT

Objectives: Pediatric appendicitis remains a challenging diagnosis in the emergency department (ED). Available risk prediction algorithms may contribute to excessive ED imaging studies. Incorporation of physician gestalt assessment could help refine predictive tools and improve diagnostic imaging decisions.

Methods: This study was a subanalysis of a parent study that prospectively enrolled patients ages 5 to 20.9 years with a chief complaint of abdominal pain presenting to 11 community EDs within an integrated delivery system between October 1, 2016, and September 30, 2018. Prior to diagnostic imaging, attending emergency physicians enrolled patients with ≤5 days of right-sided or diffuse abdominal pain using a Web-based application embedded in the electronic health record. Predicted risk (gestalt) of acute appendicitis was prospectively entered using a sliding scale from 1% to 100%. As a planned secondary analysis, we assessed the performance of gestalt via c-statistics of receiver operating characteristic (ROC) curves; tested associations between gestalt performance and patient, physician, and facility characteristics; and examined clinical characteristics affecting gestalt estimates.

Results: Of 3,426 patients, 334 (9.8%) had confirmed appendicitis. Physician gestalt had excellent ROC curve characteristics (c-statistic = 0.83, 95% confidence interval = 0.81 to 0.85), performing particularly well in the low-risk strata (appendicitis rate = 1.1% in gestalt 1%–10% range, negative predictive value of 98.9% for appendicitis diagnosis). Physicians with ≥5 years since medical school graduation demonstrated improved gestalt performance over those with less experience (p = 0.007). All clinical characteristics tested, except pain <24 hours, were significantly associated with physician gestalt value (p < 0.05).

From the [1]Division of Research, Kaiser Permanente, Oakland, CA; the [2]University of California San Diego School of Medicine, La Jolla, CA; [3]The Permanente Medical Group, Kaiser Permanente San Leandro Medical Center, San Leandro, CA; [4]The Permanente Medical Group, Kaiser Permanente Roseville Medical Center, Roseville, CA; [5]The Permanente Medical Group, Kaiser Permanente San Francisco Medical Center, San Francisco, CA; [6]The Permanente Medical Group, Kaiser Permanente Oakland Medical Center, Oakland, CA; [7]The Permanente Medical Group, Kaiser Permanente South Sacramento Medical Center, Sacramento, CA; [8]The Permanente Medical Group, Kaiser Permanente Santa Clara Medical Center, Santa Clara, CA; the [9]HealthPartners Institute, Bloomington, MN; the [10]Children's Hospitals and Clinics of Minnesota, Minneapolis, MN; and [11]The Permanente Medical Group, Kaiser Permanente San Rafael Medical Center, San Rafael, CA.
Received November 7, 2019; revision received January 31, 2020; accepted February 1, 2020.
Presented at Pediatric Academic Societies Annual Meeting, San Francisco, CA, May 2017.
Supported by the National Institutes of Health (R01 HD079463 [Kharbanda]).
The authors have no potential conflicts to disclose.
Author contributions: LES, DRV, and DWB conceptualized and designed the study; LES, EOK, ABK, and DWB performed manual chart review; LES drafted the initial manuscript and reviewed and revised the manuscript; EMW performed the programming; UKC, ASR, DRV, and DWB designed the data collection instrument; MER, EOK, ABK, DRV, DWB, and ASR obtained research funding; DWB oversaw the study as a whole; LES drafted the manuscript and all authors contributed substantively to its critical revision; all authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.
Supervising Editor: Alice M. Mitchell, MD, MS.
Address for correspondence and reprints: Dustin W. Ballard, MD, MBE; e-mail: dustin.ballard@kp.org.
ACADEMIC EMERGENCY MEDICINE 2020;00:1–11.

doi: 10.1111/acem.13931
ISSN 1553-2712

Conclusion: Physician gestalt for acute appendicitis diagnosis performed well, especially in low-risk patients and when employed by experienced physicians.

Pediatric abdominal pain with concern for acute appendicitis is a common clinical scenario in the emergency department (ED). Acute appendicitis symptoms overlap with other conditions, making the assessment challenging.[1] Clinical prediction risk scores, such as the Pediatric Appendicitis Score (PAS), can aid in diagnosis. However, some scores assign a large proportion of patients to intermediate-risk categories, leading to the potential overutilization of computed tomography (CT) and ultrasound (US) imaging.[2–4]

Physician gestalt can be defined as a physician's implicit probability estimation based on a synthesis of provider experience and clinical perception in the absence of definitive diagnostic testing.[5] Assessments of physician gestalt across various medical conditions, such as pulmonary embolism and acute coronary syndrome, demonstrate variable accuracy; for some conditions, such as pulmonary embolism, studies suggest that gestalt can perform similarly to clinical prediction rules.[2,4–10] However, physician gestalt of diagnostic probability is rarely incorporated into risk-stratification tools. Additionally, although it has been shown to perform well in many scenarios, physicians do not always behave consistently with their reported gestalt.[11] This may be due to concern for adverse consequences of a missed diagnosis and the limited number of validation assessments of physician gestalt performance.[11,12]

To our knowledge, only a handful of studies have described the diagnostic performance of physician gestalt for acute appendicitis, and only one has assessed emergency physician gestalt exclusively in a pediatric population.[10,13–15] This four-center Australian study in academic EDs (two tertiary pediatric centers and two mixed) reported reasonable diagnostic accuracy for emergency physicians (70%–82%) that did not vary with experience.[13] Our investigation had a similar objective but in a U.S. community ED population with a secondary goal of providing data that could inform clinicians when gestalt is a reliable diagnostic tool and when to utilize other clinical decision support (CDS) tools, imaging, or consultants.[13]

In this secondary analysis to a larger prospective cohort study, we sought to 1) characterize the diagnostic performance of general emergency physician gestalt for acute appendicitis in patients age 5 to 20.9 years presenting to a community ED with acute abdominal pain; 2) characterize the association between patient-, physician-, and facility-level characteristics and the receiver operating characteristic (ROC) curve characteristics of physician gestalt; and 3) examine clinical characteristics associated with gestalt assessments. We hypothesized that emergency physician gestalt would have a good c-statistic for predicting acute appendicitis and that more experienced physicians would show superior ROC curve characteristics.

METHODS

Study Design and Setting

Kaiser Permanente Northern California (KPNC) is a large, integrated health care delivery system that provides care to approximately four million members across 21 medical facilities with multiple clinics and ancillary services.[16] KPNC members represent approximately 33% of the insured population in areas served and are comparable to the surrounding and statewide population with respect to age, sex, and race/ethnicity.[17] KPNC utilizes a comprehensive integrated electronic health record (EHR; Epic, Verona, WI), fully implemented in 2009.[18]

This study was conducted as a secondary analysis of a larger prospective study evaluating a CDS system for pediatric abdominal pain evaluation in 11 KPNC EDs (NCT02633735). This larger investigation consisted of a pre–post cluster-randomized trial of providing CDS with the pediatric Appendicitis Risk Calculator (pARC) score to providers. Detailed implementation methods of the larger study are reported elsewhere.[19]

At study EDs, care was provided by board-certified or board-eligible emergency physicians. Table S1 in Data Supplement S1 (available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1111/acem.13931/full) shows facility-specific characteristics. All facilities had access to CT and US during regular business hours; however, after-hours US availability varied across facilities. Four of the study facilities had pediatric inpatient units.

Participant Selection

Treating emergency physicians enrolled eligible patients through a Web-based application embedded

in the EHR. Patients were eligible if they were 5 to 20 years old with ≤5 days of right-sided or diffuse abdominal pain. These inclusion criteria were based on the original derivation/validation cohorts of the pARC.[3] The age range, with an upper limit of 20 years, was chosen based on the inclusion criteria of the parent study. Exclusion criteria included abdominal trauma, known appendicitis or history of appendectomy, current pregnancy, or other uncommon chronic or confounding conditions described previously.[3,19] To ensure that gestalt assessment was not influenced by imaging results, patients were excluded if enrollment occurred after ordering advanced abdominal imaging (US or CT). Only the first patient encounter between October 1, 2016, and September 30, 2018, was included in this analysis and enrollments made by providers listed as residents, students, or physician assistants were removed from the cohort post hoc.

To facilitate enrollment, promotional posters were placed in EDs, emergency physicians were sent automated text-message alerts when assigned a potentially eligible patient, and physicians received a small incentive ($5 gift card) for each completed enrollment.[20] For the last 15 months of the study period, six of the 11 facilities also received CDS based on the pARC with care pathway recommendations (following gestalt entry) as part of the larger cluster-randomized trial. Other risk-stratification tools such as the Alvarado and PAS were not provided to, or routinely used by, our clinicians.

This study was approved by the KPNC Institutional Review Board with a waiver of informed consent. Patient safety was monitored by an independent data safety monitoring board.

Data Collection

Clinical variables of interest were identified based on previously reported associations with appendicitis and incorporation in validated risk scores.[3,21,22] Data were collected from the EHR using automated data collection techniques and from physician-entered enrollment responses. Clinical characteristics entered by the emergency physician at the time of ED visit were based on predetermined definitions adapted from Kharbanda et al.[3] (Table S2) and required for the pARC. Physicians prospectively entered gestalt on a continuous sliding scale of 1% to 100% after reporting the variables for the pARC but prior to ordering abdominal imaging (Figure S1). Gestalt could be entered before or after a white blood cell (WBC) count was determined. Gestalt estimates were not permitted post hoc.

Laboratory and abdominal imaging results were extracted from EHR data. Emergency physician data included age, sex, years since medical school graduation, and years as a KPNC physician. Facility characteristics included the presence of a pediatric inpatient unit and teaching hospital designation.

Outcomes

Our primary outcome was physician gestalt performance for the diagnosis of acute appendicitis. Patients were considered to have acute appendicitis if the diagnosis was made at the index ED visit or within 7 days. Appendicitis verification was performed via manual EHR review of operative and pathology reports with outcome definitions based on prior work by the study team.[3,23,24] If the patient had a diagnosis of appendicitis in the EHR but no operative or pathology reports were available, the patient record was manually reviewed by a trained study abstractor. Patients transferred out of the KPNC system with an ED diagnosis of appendicitis (n = 6) were assumed to have appendicitis based on review of their encounter notes. As a subset of appendicitis cases, missed appendicitis was determined as a safety outcome and defined as appendicitis within 7 days after the initial ED enrollment and not part of the initial encounter or immediate transfer. All outcomes were reviewed by two trained study investigators with adjudication by a third investigator as needed. All cases of missed appendicitis were reviewed by four study investigators.

Secondary outcomes were analyzed to further assess the safety of physician gestalt assessment and included the rate of negative appendectomy and perforation. Negative appendectomy was defined as an appendectomy without a confirmed diagnosis of appendicitis based on operative or pathology notes. Perforation was defined as perforated appendicitis confirmed by operative and pathology notes.[19]

Patients Not Enrolled

We assessed for potentially missed eligible patients via EHR database query and calculated the estimated appendicitis rate using principal diagnosis and appendectomy procedural codes in the missed eligible and excluded patient populations.[25] Additionally, an audit was conducted at the start of the study to assess the characteristics of missed eligible patients.

Data Analysis

We generated initial predicted probabilities of appendicitis for each patient with a logistic model regressed on provider gestalt. We then ran logistic regression models of the outcome on the predicted probabilities to generate area under the curve (AUC) estimates and standard errors for each comparison group separately and compared the difference in AUC estimates using a chi-square distribution. A calibration plot was graphed and a Hosmer-Lemeshow test was used to determine goodness of fit. We compared differences in the c-statistics for physician gestalt by facility characteristics and by physician experience measures including age (≤40 vs. >40), years since medical school graduation (<5 vs. ≥5), and years with the medical group (<5 vs. ≥5). Age of 40 years was chosen based on median emergency physician age and experience cutoffs were based on a prior study in the same care setting.[26] To analyze differences in these independent groups within the cohort, we compared differences in area under the ROC curves using chi-square tests with gestalt treated as a continuous variable.[27] In addition to comparisons by facility and physician characteristics, we compared distributions of clinical characteristics across physician gestalt categories with chi-square tests for categorical variables and ANOVA for continuous variables. Gestalt categories of 1% to 10%, 11% to 49%, 50% to 89%, and 90% to 100% were chosen for descriptive purposes a priori because of their potential for clinical relevance. Test characteristics were calculated for the gestalt 1% to 10% category as a diagnostic predictor of appendicitis. A power analysis was conducted based on preliminary data and demonstrated that differences in c-statistics of 0.06 could be detected with 93% power with a sample size of 2,250 patients.[26] All analyses were conducted in SAS version 9.4.

Sensitivity Analysis

As a planned sensitivity analysis, we assessed the c-statistic for gestalt after excluding cases where the WBC count was resulted prior to gestalt entry, determined using time stamps in the EHR.

RESULTS

We enrolled 3,426 patients (Figure 1) over the 24-month period; 436 physicians (mean age of 40.6 years, 60.6% male) completed enrollments. Physician gestalt estimates ranged from 1% to 97% (median = 18%, interquartile range = 5% to 43%). Of the eligible patients, 1,493 (43.6%) were in the physician gestalt category of 1% to 10%, 1,121 (32.7%) were 11% to 49%, 744 (21.7%) were 50% to 89%, and 68 (2.0%) were 90% to 100%. A total of 1,938 (56.6%) patients had a WBC count determined in the ED, 385 (11.2%) determined before gestalt entry, and 1,774 (51.8%) patients received US and/or CT imaging. Of those with low gestalt (1%–10%), 341 (22.8%) had imaging done in the ED (CT 1.7%, US 20.2%, both 0.9%). Sixty-six percent of patients in the 11% to 49% gestalt category received imaging (CT 7.3%, US 52.3%, both 6.3%).

Among eligible patients, 334 (9.8%) had confirmed acute appendicitis. Gestalt was found to be an excellent predictor of acute pediatric appendicitis with a c-statistic of 0.83 (95% CI = 0.81 to 0.85). Physician gestalt categorized 43.6% of patients in the low-gestalt category of 1% to 10% with an appendicitis rate of 1.1% and perforation rate of 0.3% (1%–10%—negative predictive value = 98.9% [95% CI = 98.3% to 99.3%]; >10%—sensitivity = 95.2% [95% CI = 92.3%–97.2%] and specificity = 47.8% [95% CI = 46.0%–49.6%] for diagnosis of appendicitis). However, gestalt demonstrated poor calibration due to overestimation of risk at the higher end of the spectrum (Hosmer-Lemeshow $p < 0.001$; Figure S2): appendicitis incidences were 7.6% in gestalt 11% to 49% range, 26.9% in gestalt 50% to 89% range, and 48.5% in gestalt 90% to 100% range. Distribution of physician gestalt by appendicitis outcome is shown in Figure 2. There was no evidence of temporal trends in gestalt ROC performance (quarterly comparisons) across the study time period.

Physician-level characteristics are presented in Table 1. Analysis of physician gestalt performance showed variation associated with years of physician experience. Physicians with ≥5 years since medical school graduation had improved c-statistics compared to those with <5 years since medical school graduation (c-statistic = 0.84 vs. 0.74, $p = 0.007$; Table 2). Other physician-level characteristics were not significantly associated with gestalt performance: years with the medical group ($p = 0.06$), sex ($p = 0.10$), and age ($p = 0.11$; Table 2). Facility pediatric inpatient unit availability ($p = 1.00$) and teaching hospital designation ($p = 0.49$) were not significantly associated with physician gestalt performance.

All clinical variables tested, except for duration of pain <24 hours, were significantly associated with physician gestalt assessment ($p < 0.05$; Table 3).

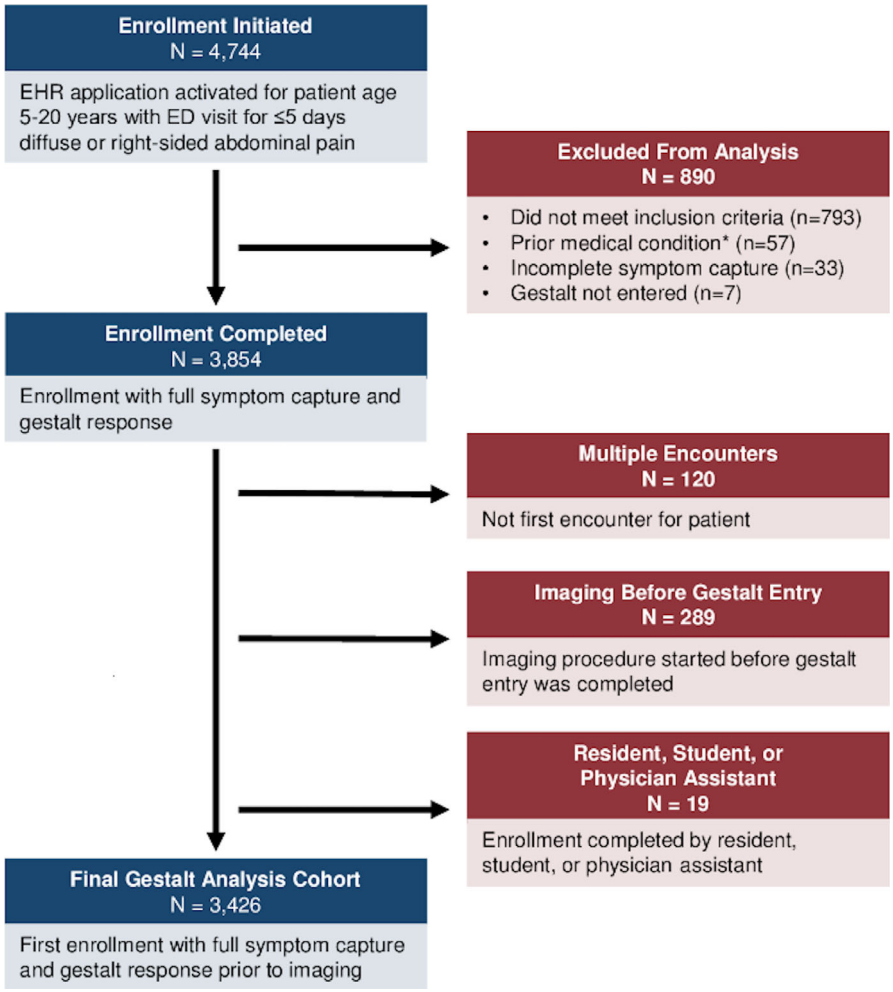

Figure 1 Cohort assembly for physician gestalt analysis of pediatric appendicitis for patients presenting to the ED with abdominal pain. EHR = electronic health record. *Includes acute or chronic pancreatitis; prior intraabdominal surgery; volvulus; intestinal atresia/stenosis; inflammatory bowel disease; ulcerative enterocolitis; Hirschsprung's disease; sickle cell disease; cancer; lupus; Henoch Schonlein purpura; juvenile rheumatoid arthritis; cystic fibrosis; human immunodeficiency virus; mental retardation; chromosomal anomaly; bone marrow, heart, kidney, or liver transplant; kidney failure/dialysis.

There were notable increases in prevalence between the low- and high-gestalt strata for anorexia (increased by 55.1%), guarding (57.8%), migration of pain to right lower quadrant (RLQ; 72.2%), pain with coughing/hopping/walking (66.3%), and maximal tenderness in the RLQ (88.8%; Table 3). The highest percentage of ED imaging was for gestalt category 50% to 89% ($p < 0.001$; Table 3). Sensitivity analysis indicated insignificant variation in gestalt performance between those with no WBC count determined before gestalt entry ($n = 3,043$) and the overall cohort (c-statistic = 0.84 vs. 0.83, 95% CI = 0.82 to 0.87 vs. 0.81 to 0.85).

Safety and secondary outcomes are presented in Table 4. Of the 334 patients with appendicitis, 56 (16.8%) had a perforation. The negative appendectomy rate was 6.2% (22/356) and the missed appendicitis rate was 0.4% (15/3,426). Chart review analysis of low-gestalt (1%–10%) appendicitis cases ($n = 16$) revealed that 13 (81.0%) of these cases were early presentations of appendicitis (pain <24 hours). Chart review determined characteristics of low-gestalt appendicitis (1.1%), negative appendectomy (15.8%), and missed appendicitis (0.3%) patients are presented in Table 5. The three patients in the 1% to 10% gestalt category with perforated missed appendicitis all had pain <24 hours, no migration of pain, no pain with walking, no RLQ tenderness, and no guarding at the time of gestalt entry. These three patients returned to the ED between 7 and 72 hours following the index ED visit.

The appendicitis rate of nonenrolled patients was 1.1% (252/22,902) and audits assessing patient characteristics confirmed only a limited number of nonenrolled patients were truly eligible for the study.[25] In a separate analysis by Cotton et al.[25] examining a subset

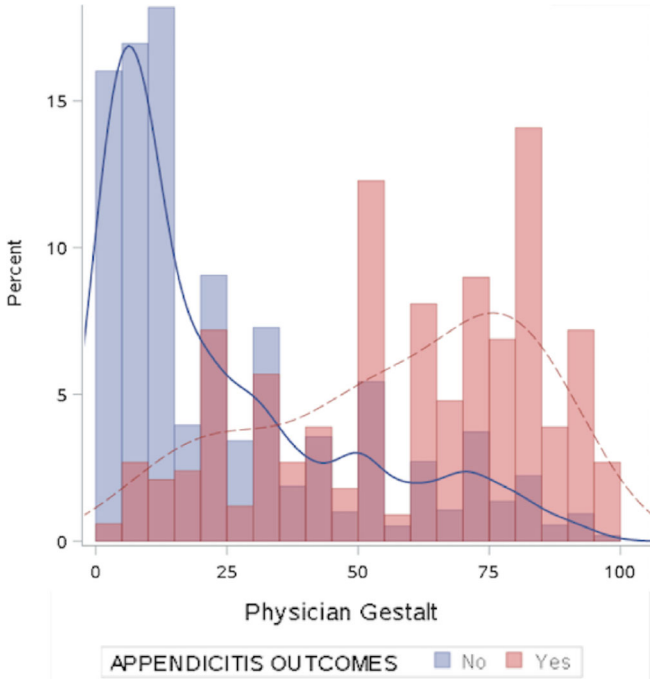

Figure 2 Distribution of physician gestalt by appendicitis outcome.

Table 1
Characteristics of ED Physicians Who Enrolled Patients in the Gestalt Analysis Cohort (*N* = 436)

Provider characteristics	
Age (years), mean (±SD)	40.6 (±7.5)
Categorical	
<31	17 (3.9)
31–40	218 (50.0)
41–50	157 (36.0)
51–60	38 (8.7)
>60	6 (1.4)
Sex	
Female	172 (39.4)
Male	264 (60.6)
Years since medical school graduation, mean (±SD)	12.3 (±7.3)
Categorical	
0–4 years	61 (14.0)
≥5 years	375 (86.0)
Years with the medical group, mean (±SD)	7.1 (±6.7)
Categorical	
0–4 years	189 (43.3)
≥5 years	247 (56.7)

Data are reported as *n* (%) unless otherwise specified.

of this population, enrolled and nonenrolled cohorts did not differ significantly by age, sex, or race.

DISCUSSION

In this prospective study, we describe the diagnostic performance of emergency physician gestalt for the diagnosis of acute pediatric appendicitis and the association of physician gestalt with patient, physician, and facility characteristics.

Table 2
Provider and Facility Area Under the Receiver Operating Characteristics Curve Comparisons for Gestalt Performance

Provider characteristics	AUC (95% CI)	p-value
Overall	0.83 (0.81–0.85)	
Age (years)		
≤40	0.82 (0.79–0.85)	0.11
>40	0.85 (0.82–0.88)	
Sex		
Female	0.81 (0.77–0.84)	0.10
Male	0.85 (0.82–0.87)	
Years since medical school graduation		
0–4	0.74 (0.66–0.81)	0.007
≥5	0.84 (0.82–0.86)	
Years with the medical group		
0–4	0.81 (0.77–0.84)	0.06
≥5	0.85 (0.82–0.87)	
Pediatric inpatient unit available		
Yes	0.83 (0.80–0.86)	1.00
No	0.83 (0.81–0.86)	
Teaching hospital designation		
Major	0.81 (0.77–0.85)	0.49
Not major	0.83 (0.80–0.86)	

AUC = area under the curve.

Emergency physician gestalt in our community setting was found to have excellent ROC curve characteristics (c-statistic = 0.83), although with poorer discrimination at the higher end of the spectrum. Figure 2 demonstrates the especially good performance in the low-gestalt strata. This performance is notably better than that reported in a prior study of patients age 11 years and older (not restricted to pediatrics) who underwent CTs in the ED for possible appendicitis and used a dichotomous gestalt cutoff of 60%.[10] The variation in performance between our study and theirs is multifactorial. Most prominently, our study focused on pediatric patients and treated gestalt as both a categorical and continuous variable. A recent study by Lee et al.[13] found comparable physician gestalt performance to ours (c-statistic = 0.84), although this study was conducted at four EDs (two pediatric only) in Australia, where training pathways and clinical practices (e.g., CT is rarely used in pediatric abdominal pain evaluation) are significantly different from those in the U.S. community ED setting.

Emergency physician gestalt had good discriminatory ability in assigning patients to the low-risk (1%–

Table 3
Patient Characteristics, ED Laboratory Values, ED Imaging, and Appendicitis Diagnosis by Physician Gestalt Category (N = 3,426)

		Physician Gestalt Category			
	n (%)*	1%–10% (n = 1,493)	11%–49% (n = 1,121)	50%–89% (n = 744)	90%–100% (n = 68)
Clinical characteristics (% yes)					
Age (years), mean (±SD)	11.0 (±4.2)†	10.5 (±4.3)	11.0 (±4.1)	11.8 (±4.1)	12.0 (±4.2)
Male	1,590 (46.4)*	656 (43.9)	514 (45.9)	371 (49.9)	49 (72.1)
Temperature > 38°C	270 (7.9)*	95 (6.4)	100 (8.9)	62 (8.3)	13 (19.1)
Nausea/vomiting	2,271 (66.3)*	940 (63.0)	738 (65.8)	542 (72.9)	51 (75.0)
Anorexia	1,695 (49.5)*	539 (36.1)	600 (53.5)	494 (66.4)	62 (91.2)
Pain <24 hours	1,974 (57.6)	860 (57.6)	645 (57.5)	431 (57.9)	38 (55.9)
Guarding	609 (17.8)*	59 (4.0)	204 (18.2)	304 (40.9)	42 (61.8)
Pain migrating to RLQ	804 (23.5)*	85 (5.7)	259 (23.1)	407 (54.7)	53 (77.9)
Pain with coughing/hopping/walking	1,255 (36.6)*	241 (16.1)	446 (39.8)	512 (68.8)	56 (82.4)
Maximal tenderness to RLQ	1,214 (35.4)*	144 (9.7)	399 (35.6)	604 (81.2)	67 (98.5)
Diagnostic test utilization (% yes)					
Labs					
WBC count determined	1,938 (56.6)*	517 (34.6)	748 (66.7)	617 (82.9)	56 (82.4)
WBC > 10×10^9/L‡	923 (47.6)*	206 (39.9)	330 (44.1)	347 (56.2)	40 (71.4)
PMN count done	1,836 (53.6)*	484 (32.4)	715 (63.8)	585 (78.6)	52 (76.5)
PMN > 7.5×10^9/L§	806 (43.9)*	175 (36.2)	286 (40.0)	311 (53.2)	34 (65.4)
Imaging					
Any ED abdominal imaging	1,774 (51.8)*	341 (22.8)	739 (65.9)	642 (86.3)	52 (76.5)
Ultrasound only¶	1,338 (75.4)*	301 (88.3)	586 (79.3)	418 (65.1)	33 (63.5)
CT only¶	219 (12.3)	26 (7.6)	82 (11.1)	105 (16.4)	6 (11.5)
Both¶	217 (12.2)	14 (4.1)	71 (9.6)	119 (18.5)	13 (25.0)
Appendicitis diagnosis (% yes)					
Confirmed appendicitis	334 (9.8)	16 (1.1)	85 (7.6)	200 (26.9)	33 (48.5)

PMN = polymorphonuclear leukocytes; RLQ = right lower quadrant; WBC = white blood cell.
*Statistically significant variation between gestalt categories ($p < 0.001$).
†$p < 0.05$ for all categoric comparisons except 11% to 49% vs. 90% to 100% and 50% to 89% vs. 90% to 100%.
‡Percentage of those with a WBC done per gestalt category.
§Percentage of those with a PMN done per gestalt category.
¶Percentage of those with any imaging performed in the ED per gestalt category.

Table 4
Secondary Outcome Events by Physician Gestalt Category

		Physician Gestalt Category				
	n (%) (N = 3,426)	1%–10% (n = 1,493)	11%–49% (n = 1,121)	50%–89% (n = 744)	90%–100% (n = 68)	p-value
Negative appendectomy*	22/356 (6.2)	3/19 (15.8)	6/91 (6.6)	12/212 (5.7)	1/34 (2.9)	<0.001
Perforation†	56/334 (16.8)	4/16 (25.0)	14/85 (16.5)	35/200 (17.5)	3/33 (9.1)	<0.001
Missed appendicitis‡	15/3426 (0.4)	5/1493 (0.3)	5/1121 (0.5)	5/744 (0.7)	0/68 (0)	0.61

*Negative appendectomy related to index ED visit, percentage of those with an appendectomy in each gestalt cohort.
†Perforation within 7 days of index ED visit, percentage of those with appendicitis in each gestalt cohort.
‡Chart reviewed confirmed appendicitis within 7 days not as part of index visit or immediate transfer.

10%) category. The low appendicitis rate in the low-gestalt category (1.1%) provides confidence in gestalt performance at the low end of the spectrum. Even in cases where initial gestalt was 1% to 10% and the patient had a final diagnosis of appendicitis, including those with perforations, chart review of the ED notes often revealed a progression of disease symptoms throughout the ED visit. However, emergency physicians often acted conservatively, even when their gestalt was low—as evidenced by the high imaging rate

Table 5
Clinical Characteristics of Patients in Gestalt Category 1% to 10% With Confirmed Appendicitis or Confirmed Secondary Outcomes Based on Chart Review

Outcome	*n*	Pain <24 Hours	WBC Count Obtained in ED	Imaging During Index ED Visit			Perforation
				CT Only	US Only	CT and US	
Appendicitis	16	13	13	1	5	1*	4
Missed appendicitis†	5‡	3	3	0	0	0	3
Negative appendectomy	3	0	3	0	3§	0	—

US = ultrasound; WBC = white blood cell.
*Patient also had intestinal malrotation.
†These patients are a subset of the appendicitis cases.
‡One case had chronic abdominal pain but was not excluded so as not to introduce bias due to selected chart review.
§All had equivocal/nondiagnostic imaging.

(22.8%) in the low-gestalt cohort. Reducing imaging for those deemed to be at low risk of appendicitis has the potential to decrease ED length of stay and resource utilization and, in the case of CT, mitigate a child's exposure to radiation.[28,29] Of note, our integrated health care system, with its good follow-up capability, is conducive to this care model. In select care settings with higher prevalence of appendicitis or other surgical diagnoses, for example, tertiary pediatric EDs, an US to magnetic resonance imaging (MRI) algorithm may be appropriate. However, during our study period, abdominal MRI was not readily, rapidly, and consistently available at our community EDs for the pediatric abdominal pain diagnostic algorithm.[30,31] Notably, the gestalt category 50% to 89% had the highest imaging rate (86.3%), demonstrating a high level of concern regarding an appendicitis diagnosis in this patient strata. The somewhat lower imaging rates in the gestalt 90% to 100% category (76.5%) suggest that in this highest estimated risk decile, physicians may have been somewhat more confident in their diagnosis and the low negative appendectomy rate (2.9%) supports this contention. We were underpowered to robustly evaluate gestalt in this highest risk decile, but our results suggest that it may perform well as an adjunct to existing decision aids for this patient population.

Risk overestimation, especially in the intermediate gestalt categories, likely contributes to the overutilization of imaging. Overestimation may be due to concern for the ramifications of a missed diagnosis, both legal and adverse patient outcomes, and the relatively low-risk tolerance often prevalent in emergency physicians.[11,12,32] Risk-minimizing behavior by emergency physicians may also contribute to the overutilization of advanced imaging due to the perceived risk of missing a high-consequence diagnosis.[33]

We did not design this study to compare emergency physician gestalt performance to the pARC and PAS, which would not be a fair comparison because not all physicians who entered a gestalt ordered a WBC count in the ED, and we could not verify if those with a WBC count viewed the result prior to entering gestalt. However, recent work from our study team has reported on the performance of pARC and PAS in the same setting with distinct inclusion criteria (requiring the presence of a determined ED WBC count). The reported c-statistics range from 0.85 to 0.89 for pARC and 0.77 to 0.80 for PAS.[25] While, comparatively, gestalt performed slightly better than the PAS and slightly worse than the pARC, we remind the reader that gestalt overestimated risk in the intermediate ranges (in which imaging rates were high) and as such is likely most useful in identifying low-risk (1%–10%) patients for whom no further ED workup is necessary. As such, the incorporation of gestalt for low-risk patients into CDS tools may facilitate provider buy-in and integration into provider workflow, thus increasing uptake in clinical practice.[34,35] However, for cases falling in higher gestalt categories further evaluation may be necessary, including surgical consultation and/or imaging. CDS tools may help correct for the overestimation of risk at the higher end of the spectrum and provide reassurance to the provider when deciding if imaging is necessary.

Assessment of emergency physician characteristics and gestalt performance showed no significant variation by physician age, sex, or years with the medical group. Gestalt performance improved for physicians with ≥5 years since medical school graduation in all risk strata (Figure S3). This finding of enhanced gestalt performance with physician experience aligns with other studies on the performance of gestalt for pulmonary embolism diagnosis.[36,37] This finding

supports targeting the use of CDS tools toward more junior clinicians, who have also been reported to be more accepting of prediction rules than more experienced providers.[38] Our evaluation also demonstrated that physician gestalt performance was not associated with specific facility variables.

Our results also provide insight into how physicians formulate their gestalt. For example, the presence of RLQ maximal tenderness was dramatically higher in the 90% to 100% gestalt category compared with the 1% to 10% gestalt category ($p < 0.001$), while pain <24 hours was not significantly associated with increased gestalt ($p = 0.99$). Performance of physician gestalt is known to vary in a condition-specific manner, and it is possible that pediatric appendicitis is associated with better performance due to the presence of trademark physical examination findings such as RLQ tenderness.[5,9,32,39] Interestingly, 57.6% of patients in the lowest gestalt subgroup had pain <24 hours, accounting for 81% of appendicitis cases in the low-gestalt cohort. Our finding that pain <24 hours has poor correlation with gestalt demonstrates the difficulty of appendicitis diagnosis in patients with a brief duration of pain, potentially due to a lower likelihood of pain concentration in the RLQ within a short pain duration period.

LIMITATIONS

Several study limitations deserve mention. First, this analysis was undertaken as a component of a larger study on pediatric abdominal pain. The presence of this parent study may have increased physician awareness around the diagnostic evaluation and management of appendicitis, which may have, over time, impacted gestalt estimates. However, the publication of the pARC validation study was in April 2018, near the end of our study period, and at no time during the study was the pARC calculator available on publicly available Web-based platforms (i.e., MDCalc, New York, NY).

Enrollment for this study was initiated by the emergency physician and consequently did not capture all providers at the 11 KPNC EDs and only a sample of the total eligible patient population is represented. Study enrollment for the parent study and this subanalysis was performed on an opt-in basis by the treating physicians to capture an appropriate patient population at risk for appendicitis and meeting all eligibility criteria as defined above. Audits of missed eligible patients demonstrated that less than a quarter of potentially eligible patients were actually eligible for the larger study, and the low rate of appendicitis in this population suggests the we captured a representative risk pool. It is also unclear how the inability to compare physicians who enrolled patients in our study versus those who did not, as well as our specialized practice setting, affect study generalizability. Additionally, due to the necessary data collection design, physicians were asked about the presence of the patient's clinical variables immediately prior to entering their gestalt. Theoretically, this may have increased the association between clinical variables and gestalt; however, this effect is likely mitigated since the assessed clinical variables are standard components of acute appendicitis evaluation in the ED. Since we could not control for physician gestalt being entered before or after attaining relevant clinical data, we did not consider a "gestalt-only" model and, instead, the availability of these clinical data and determination of gestalt were treated as a single step. Also, physicians could enter their gestalt before or after ordering a WBC count, but only 11% of enrollments had WBC counts determined at the time of gestalt entry. We were also unable to discern if imaging was requested by a consultant, such as a surgeon. Finally, there was the potential for providers to calculate the PAS or other risk scores on their own prior to completing the gestalt form; however, these scores require a WBC count and we are unaware of their regular use by KPNC emergency physicians.

CONCLUSION

Emergency physician gestalt for possible pediatric appendicitis presenting to the ED had excellent receiver operating characteristic curve characteristics. Emergency physicians with less experience showed decreased c-statistics. The very low rate of appendicitis in the low-gestalt risk category (1%–10%) provides support for providers' decisions to forgo imaging in these patients. In higher-risk gestalt categories, the overestimation of risk suggests a possible benefit of utilizing prediction algorithms to mitigate imaging studies of limited value.

REFERENCES

1. Rentea RM, Peter SD. Pediatric appendicitis. Surg Clin North Am 2017;97:93–112.

2. Kollár D, McCartan D, Bourke M, Cross K, Dowdall J. Predicting acute appendicitis? A comparison of the Alvarado score, the appendicitis inflammatory response score and clinical assessment. World J Surg 2015;39:104–9.
3. Kharbanda AB, Vazquez-Benitez G, Ballard DW, et al. Development and validation of a novel pediatric Appendicitis Risk Calculator (pARC). Pediatrics 2018;141: e20172699.
4. Schriger DL, Elder JW, Cooper RJ. Structured clinical decision aids are seldom compared with subjective physician judgment, and are seldom superior. Ann Emerg Med 2017;70:338–44.
5. Hendriksen JM, Lucassen WA, Erkens PM, et al. Ruling out pulmonary embolism in primary care: comparison of the diagnostic performance of "gestalt" and the Wells rule. Ann Fam Med 2016;14:227–34.
6. Kline JA, Kahler ZP, Beam DM. Outpatient treatment of low-risk venous thromboembolism with monotherapy oral anticoagulation: patient quality of life outcomes and clinician acceptance. Patient Prefer Adherence 2016;10:561–9.
7. Stein J, Louie J, Flanders S, et al. Performance characteristics of clinical diagnosis, a clinical decision rule, and a rapid influenza test in the detection of influenza infection in a community sample of adults. Ann Emerg Med 2005;46:412–9.
8. Glas AS, Pijnenburg BA, Lijmer JG, et al. Comparison of diagnostic decision rules and structured data collection in assessment of acute ankle injury. Can Med Assoc J 2002;166:727–33.
9. Penaloza A, Verschuren F, Meyer G, et al. Comparison of the unstructured clinician gestalt, the Wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism. Ann Emerg Med. 2013;62:117–24.
10. Golden SK, Harringa JB, Pickhardt PJ, et al. Prospective evaluation of the ability of clinical scoring systems and physician-determined likelihood of appendicitis to obviate the need for CT. Emerg Med J 2016;33:458–64.
11. Quinn JV, Stiell IG, McDermott DA, Kohn MA, Wells GA. The San Francisco Syncope Rule vs physician judgment and decision making. Am J Emerg Med 2005;23:782–6.
12. Palchak MJ, Holmes JF, Kuppermann N. Clinician judgment versus a decision rule for identifying children at risk of traumatic brain injury on computed tomography after blunt head trauma. Pediatr Emerg Care 2009;25:61–5.
13. Lee WH, O'Brien S, Skarin D, et al. Accuracy of clinician gestalt in diagnosing appendicitis in children presenting to the emergency department. Emerg Med Australas 2019;31:612–8.
14. Bal A, Anil M, Narturk M, et al. Importance of clinical decision making by experienced pediatric surgeons when children are suspected of having acute appendicitis: the reality in a high-volume pediatric emergency department. Pediatr Emerg Care 2017;33:e38–42.
15. Leung YK, Chan CP, Graham CA, Rainer TH. Acute appendicitis in adults: diagnostic accuracy of emergency doctors in a university hospital in Hong Kong. Emerg Med Australas 2017;29:48–55.
16. Vinson DR, Lugovskaya N, Warton EM, et al. Ibutilide effectiveness and safety in the cardioversion of atrial fibrillation and flutter in the community emergency department. Ann Emerg Med 2018;71:96–108.e102.
17. Gordon N, Lin TJ. The Kaiser Permanente Northern California adult member health survey. Perm J 2016;20:34.
18. Bornstein S. An integrated EHR at Northern California Kaiser Permanente: pitfalls, challenges, and benefits experienced in transitioning. Appl Clin Inform 2012;3:318–25.
19. Ekstrom HL, Kharbanda EO, Ballard DW, et al. Development of a clinical decision support system for pediatric abdominal pain in emergency department settings across two health systems within the HCSRN. eGEMS 2019;7:15.
20. Simon LE, Rauchwerger AS, Chettipally UK, et al. Real-time text message alerts to emergency physicians identifying potential study candidates increases clinical trial enrollment. J Patient Cent Res Rev 2018;5(Suppl 1):57.
21. Samuel M. Pediatric Appendicitis Score. J Pediatr Surg 2002;37:877–81.
22. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 1986;15:557–64.
23. Vinson DR, Ballard DW, Huang J, et al. Outpatient management of emergency department patients with acute pulmonary embolism: variation, patient characteristics, and outcomes. Ann Emerg Med 2018;72:62–72.e63.
24. Vinson DR, Ballard DW, Mark DG, et al. Risk stratifying emergency department patients with acute pulmonary embolism: does the simplified Pulmonary Embolism Severity Index perform as well as the original? Thromb Res 2016;148:1–8.
25. Cotton DM, Vinson DR, Vazquez-Benitez G, et al. Validation of the pediatric Appendicitis Risk Calculator (pARC) in a community emergency department setting. Ann Emerg Med 2019;74:471–80.
26. Kene M, Ballard D, Liu M, et al. Performance of clinical gestalt in predicting pediatric appendicitis: does experience matter? Ann Emerg Med 2017;70:S98.
27. Gönen M. Analyzing Receiver Operating Characteristic Curves with SAS. Cary, NC: SAS Institute, 2007.
28. Kanzaria HK, Probst MA, Ponce NA, Hsia RY. The association between advanced diagnostic imaging and ED length of stay. Am J Emerg Med 2014;32:1253–8.
29. Bookman K, West D, Ginde A, et al. Embedded clinical decision support in electronic health record decreases use of high-cost imaging in the emergency department: Emb ED study. Acad Emerg Med 2017;24:839–45.

30. Herliczek TW, Swenson DW, Mayo-Smith WW. Utility of MRI after inconclusive ultrasound in pediatric patients with suspected appendicitis: retrospective review of 60 consecutive patients. AJR Am J Roentgenol 2013;200:969–73.
31. Aspelund G, Fingeret A, Gross E, et al. Ultrasonography/MRI versus CT for diagnosing appendicitis. Pediatrics 2014;133:586–93.
32. Kline JA, Stubblefield WB. Clinician gestalt estimate of pretest probability for acute coronary syndrome and pulmonary embolism in patients with chest pain and dyspnea. Ann Emerg Med 2014;63:275–80.
33. Kanzaria HK, Hoffman JR, Probst MA, Caloyeras JP, Berry SH, Brook RH. Emergency physician perceptions of medically unnecessary advanced diagnostic imaging. Acad Emerg Med 2015;22:390–8.
34. Arnold LK, Alomran H, Anantharaman V, et al. Knowledge translation in international emergency medical care. Acad Emerg Med 2007;14:1047–51.
35. Prevedello LM, Raja AS, Ip IK, Sodickson A, Khorasani R. Does clinical decision support reduce unwarranted variation in yield of CT pulmonary angiogram? Am J Med 2013;126:975–81.
36. Kabrhel C, Camargo CA, Goldhaber SZ. Clinical gestalt and the diagnosis of pulmonary embolism: does experience matter? Chest 2005;127:1627–30.
37. Rosen MP, Sands DZ, Morris J, Drake W, Davis RB. Does a physician's ability to accurately assess the likelihood of pulmonary embolism increase with training? Acad Med 2000;75:1199–205.
38. Ballard DW, Rauchwerger AS, Reed ME, et al. Emergency physicians' knowledge and attitudes of clinical decision support in the electronic health record: a survey-based study. Acad Emerg Med 2013;20:352–60.
39. das Virgens CM, Lemos L Jr, Noya-Rabelo M, et al. Accuracy of gestalt perception of acute chest pain in predicting coronary artery disease. World J Cardiol 2017;9:241.

Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1111/acem.13931/full

Data Supplement S1. Supplemental material.