Interrater Reliability of Clinical Findings in Children With Possible Appendicitis

AUTHORS: Anupam B. Kharbanda, MD, MSc,^{a,b} Michelle D. Stevenson, MD, MS,^c Charles G. Macias, MD, MPH,^d Kelly Sinclair, MD,^e Nanette C. Dudley, MD,^f Jonathan Bennett, MD,^g Lalit Bajaj, MD, MPH,^h Manoj K. Mittal, MD,ⁱ Craig Huang, MD,^j Richard G. Bachur, MD,^k and Peter S. Dayan, MD, MSc,^a for the Pediatric Emergency Medicine Collaborative Research Committee of the American Academy of Pediatrics

^aDepartment of Pediatrics, Columbia University College of Physicians and Surgeons, New York, New York; ^bDepartment of Pediatrics, University of Minnesota, Minneapolis, Minnesota; ^cDepartment of Pediatrics, University of Louisville, Louisville, Kentucky; ^dDepartment of Pediatrics, Baylor College of Medicine, Houston, Texas; ^eDepartment of Pediatrics, University of Missouri, Kansas City, Missouri; ^fDepartment of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah; ^gDepartment of Pediatrics, Alfred I. duPont Hospital for Children, Wilmington, Delaware; ^hDepartment of Pediatrics, University of Colorado School of Medicine, Denver, Colorado; ⁱDepartment of Pediatrics, The Children's Hospital of Philadelphia and Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; ^jDepartment of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas; and ^kDivision of Emergency Medicine, Children's Hospital Boston, Harvard Medical School, Boston, Massachusetts

KEY WORDS

appendicitis, interrater reliability, clinical prediction rules

ABBREVIATIONS

CI—confidence interval

CT—computed tomography

PED—pediatric emergency department, PEM, pediatric emergency medicine

PEM-CRC—Pediatric Emergency Medicine Collaborative Research Committee

RLQ—right lower quadrant

All authors made a substantial contribution to the concept and design of the study, including acquisition of data and data analysis; all authors contributed to drafting of the manuscript, made revisions, and approved the final version that has been submitted for review.

(Continued on last page)

WHAT'S KNOWN ON THIS SUBJECT: Few studies have examined the reliability of clinical findings in pediatric appendicitis. Clinical prediction rules are most useful if the included variables are reliable across practice settings and practitioners.

WHAT THIS STUDY ADDS: Among children who present with possible appendicitis, the interrater reliability varied considerably for patient history and physical examination variables. Those variables with the highest degree of reliability may be best suited for inclusion in appendicitis clinical prediction rules.

abstract

OBJECTIVE: Our objective was to determine the interrater reliability of clinical history and physical examination findings in children undergoing evaluation for possible appendicitis in a large, multicenter cohort.

METHODS: We conducted a prospective, multicenter, cross-sectional study of children aged 3–18 years with possible appendicitis. Two clinicians independently evaluated patients and completed structured case report forms within 60 minutes of each other and without knowing the results of diagnostic imaging. We calculated raw agreement and assessed reliability by using the unweighted Cohen κ statistic with 2-sided 95% confidence intervals.

RESULTS: A total of 811 patients had 2 assessments completed, and 599 (74%) had 2 assessments completed within 60 minutes. Seventy-five percent of paired assessments were completed by pediatric emergency physicians. Raw agreement ranged from 64.9% to 92.3% for history variables and 4 of 6 variables had moderate interrater reliability ($\kappa > .4$). The highest κ values were noted for duration of pain ($\kappa = .56$ [95% confidence intervals .51–.61]) and history of emesis (.84 [.80–.89]). For physical examination variables, raw agreement ranged from 60.9% to 98.7%, with 4 of 8 variables exhibiting moderate reliability. Among physical examination variables, the highest κ values were noted for abdominal pain with walking, jumping, or coughing (.54 [.45–.63]) and presence of any abdominal tenderness on examination (.49 [.19–.80]).

CONCLUSIONS: Interrater reliability of patient history and physical examination variables was generally fair to moderate. Those variables with higher interrater reliability are more appropriate for inclusion in clinical prediction rules in children with possible appendicitis. *Pediatrics* 2012;129:695–700

Appendicitis is the most common surgical emergency in children, with more than 75 000 appendectomies performed each year in the United States alone.^{1,2} Appendicitis remains difficult to diagnose and the source of considerable health care expense.^{2,3} Clinicians frequently use abdominal computed tomography (CT) to evaluate children for possible appendicitis despite concerns related to the risks of radiation exposure.⁴ Substantial controversy exists among experts regarding the optimal diagnostic approach to children with acute abdominal pain.^{3,4}

Clinical prediction rules hold promise in the evaluation of children with possible appendicitis to reduce reliance on imaging.^{5,6} An important factor to consider in the development and implementation of clinical prediction rules is the reliability of potentially important variables.⁷ Previous studies to assess the reliability of clinical findings for a variety of pediatric complaints and illnesses have demonstrated substantial variability across assessors.^{8–11} If patient history and physical examination variables cannot be reliably elicited by clinicians with varying levels of training, the prediction rule will neither be generalizable across practice settings nor gain widespread acceptance.

Although previous studies have examined the reliability of patient history and physical examination findings in patients with abdominal pain, these studies have been limited by single-center data collection, small sample sizes, or samples not specific to those with possible appendicitis.^{8,10,11} Therefore, in this study we aimed to determine the interrater reliability of clinical history and physical examination findings in children undergoing evaluation for possible appendicitis in a large, multicenter cohort.

METHODS

Study Design and Setting

We performed a prospective, multicenter, cross-sectional study as part of a larger study to validate and refine a low-risk clinical decision rule for pediatric appendicitis. The study was conducted from March 2009 to April 2010 at 10 pediatric emergency departments (PEDs) that are members of the Pediatric Emergency Medicine Collaborative Research Committee (PEM-CRC). The study was approved by the institutional review board at each participating site with waiver of written informed consent at 7 sites, while written informed consent and assent were required from 3 sites. We obtained verbal consent and assent from children aged ≥ 7 years as required by the local institutional review board.

Study Participants

We enrolled children aged 3 to 17 years, inclusive, who presented to the PED with acute abdominal pain of < 96 hours' duration and underwent evaluation for possible appendicitis. "Possible appendicitis" was defined as the treating physician obtaining blood tests, radiologic studies (CT and/or ultrasound), or a surgical consult for the purpose of diagnosing/excluding appendicitis. Radiologic studies or surgical consults were obtained at the discretion of the treating physician. We excluded patients with any of the following conditions: pregnancy, previous abdominal surgery (eg, gastrostomy tube, abdominal hernia repair, previous appendectomy), chronic abdominal illness or pain (eg, inflammatory bowel disease, chronic pancreatitis, chronic or recurrent appendicitis), sickle cell anemia, cystic fibrosis, or a medical condition affecting the provider's ability to obtain an accurate history (eg, substantial language or developmental delay). We also excluded patients who had radiologic studies (CT or ultrasound) of the abdomen performed

before PED arrival or a history of abdominal trauma within 7 days of the PED evaluation. The current study was composed of a convenience sample of patients from the main cohort for whom 2 clinicians were able to perform independent evaluations.

Study Protocol

Each site principal investigator conducted group and one-on-one instructional sessions with clinicians in each respective PED before initiation of the study to provide training on the completion of the case report forms. Our goal in these sessions was to provide an understanding of the case report forms without overtraining on the clinical examination variables. The first clinician assessor (faculty physician, fellow, resident, nurse practitioner, or physician assistant) completed a standardized history and physical examination on a structured case report form. A second clinician assessor (who could have been any of the clinician types just noted) completed a similar structured case report form, independent of the first examination. Clinicians were instructed to complete the second evaluation within 60 minutes of the first assessment and to record the date and time of their examination on each case report form. Both clinician assessors affirmed on the case form that their evaluation was completed without knowledge of CT or ultrasound results.

We categorized clinicians based on their current position (attending physician, fellow, resident, nurse practitioner, or physician assistant) and their specialty training (pediatrics, pediatric emergency medicine [PEM], emergency medicine, family medicine, or other). We categorized clinicians as PEM specialists if they were board certified in the subspecialty or had engaged in a PEM fellowship. All residents were trainees in pediatrics, emergency medicine, or family medicine.

Measurements

The case report forms assessed information regarding the patient's presenting history and specific physical examination findings. The clinical findings selected were based on an extensive review of literature and through detailed discussions with faculty physicians in PEM, pediatrics, and pediatric surgery. For specific findings, we anticipated that developmental immaturity would make assessment by the physician difficult for young children. Therefore, we included "don't know," "preverbal," or "unsure" as choices on the case report form for appropriate variables.

Data Analysis and Sample Size

We collected and analyzed most clinical variables as dichotomous parameters. Duration of abdominal pain was collected and analyzed as ordinal data (collected as <12, 12–23, 24–35, 36–47, 48–72, and >72 hours). We also collected the presence of abdominal tenderness on examination in ordinal categories (none, mild, moderate, or severe pain) and conducted analyses of abdominal tenderness both by using the original ordinal categories and as a dichotomous variable (presence or absence of abdominal tenderness). For each variable, we conducted the κ analysis in 3 ways. First, we excluded from the κ analysis any paired observations for which data were missing or at least 1 assessor was unable to determine the presence or absence of the clinical finding. Second, we categorized "unsure" or "don't know" responses as if the patient had the respective sign or symptom. For both of these approaches, we excluded patients for whom the time interval between the first and second assessments was either longer than 60 minutes or could not be determined (eg, time of evaluation was not recorded). Finally, we conducted a separate analysis in which we removed preschool children (aged 3–5 years) from the κ

calculations to understand the effect that young age may have on interrater reliability.

We calculated raw agreement and assessed reliability (chance-corrected agreement) by using the unweighted Cohen κ statistic with 2-sided 95% confidence intervals (CIs).¹² We categorized the interobserver reliability, based on κ point estimates, as slight ($\kappa = .0$ –.20), fair ($\kappa = .21$ –.40), moderate ($\kappa = .41$ –.60), substantial ($\kappa = .61$ –.80), and almost perfect ($\kappa = .81$ –1.0).^{12,13} We did not perform formal sample size calculations based on the measurement of interobserver reliability as this was not the primary aim of the larger validation study. We expected that specific physical examination findings would rarely be normal (eg, presence of right lower quadrant [RLQ] pain) in our patient sample and therefore anticipated that some variables would have high measures of raw agreement but not necessarily high κ statistic values.

RESULTS

Study Population and Patient Characteristics

We present data from 8 sites; interrater observations were not performed at 1 site, and another site was excluded from the data analysis due to a low overall capture rate (<40%). In total, 811 subjects had paired independent observations by 2 clinicians, representing 31% of the 2625 patients who were enrolled in the parent prediction rule study. Of those with a second

assessment, 599 (74%) patients had the evaluation performed within 60 minutes of the first evaluation, 82 (10%) within 1 to 2 hours, and 50 (6%) at >2 hours; 80 (10%) patients had no time recorded for the second assessment. The number of second assessments completed at the 8 sites ranged from 7.5% to 94.5% of enrolled subjects, while the rate of appendicitis at enrolling sites ranged from 27% to 45.8%. Patients for whom a second assessment was conducted were similar in terms of age, gender, race, performance of imaging, and the proportion with appendicitis compared with the overall cohort (Table 1).

Further analysis was completed only for those 599 patients in whom the second assessment was performed within 60 minutes of the first. Of these 599 assessments, the skill level was known for both assessors in 576 (96%) cases. For the first evaluation, 475 (79%) were performed by a PEM attending or fellow physician, 73 (12%) by a resident, and 28 (5%) by a nurse practitioner. For the second evaluation, 551 (92%) were completed by a PEM attending or fellow physician. Paired PEM physicians performed 450 (75%) of the evaluations.

The κ statistics are detailed in Tables 2 (patient history variables) and 3 (physical examination variables). Overall raw agreement ranged from 64.9% (duration of abdominal pain) to 92.3% (history of emesis). In general, raw agreement was high, with 7 of the 14 variables having >75% agreement. For patient history variables, agreement beyond chance was at least moderate (>.40)

TABLE 1 Patient Characteristics and Comparison With All Patients Enrolled in Prediction Rule Population

	Interrater Reliability Sample (599)	Prediction Rule Study Population (2625)
Median age, y (IQR)	10.8 (8.0–13.5)	10.9 (7.9–14.1)
Gender, % male	53	50
Race, % white	75	64
Diagnostic imaging (CT, ultrasonography) performed, %	80	81
Appendicitis, %	45	39

IQR, interquartile range.

TABLE 2 Interrater Reliability of Patient History Variables

Criterion	No. of Patients Evaluated	Characteristic Present per First Assessor, <i>n</i> (%)	Patients for Whom at Least One Assessor Marked "Don't Know," <i>n</i>	Missing Data, <i>n</i>	Overall Agreement, %	κ (95% CI) ^a	κ (95% CI) "Don't know" Responses Recoded ^b
Duration of pain ^c	599	N/A	0	0	64.9	.56 (.51–.61)	.56 (.51–.61)
Focal RLQ pain	576	482 (80.8)	22	1	84.9	.48 (.39–.58)	.45 (.35–.54)
Migration of pain to RLQ	522	287 (47.9)	76	1	68.4	.37 (.29–.45)	.29 (.21–.36)
Anorexia	568	389 (64.9)	30	1	67.6	.32 (.24–.40)	.30 (.22–.37)
Nausea	562	321 (53.6)	35	2	72.0	.44 (.37–.52)	.43 (.36–.50)
Emesis	596	343 (57.3)	3	0	92.3	.84 (.80–.89)	.84 (.80–.89)

^a "Don't know" responses excluded from analysis.

^b "Don't know" responses recoded as the presence of particular finding.

^c Duration of abdominal pain collected as <12, 12–23, 24–35, 36–47, 48–72, and >72 h.

TABLE 3 Interrater Reliability for Physical Examination Findings

Criterion	No. of Patients Evaluated	Characteristic Present per First Assessor, <i>n/N</i> (%)	Patient for Whom at Least 1 Assessor Marked "Unsure", <i>n</i>	Missing Data, <i>n</i>	Overall Agreement, %	κ (95% CI) ^b	κ (95% CI) "Unsure" Responses Recoded ^c
Abdominal tenderness ^a (none, mild, moderate, severe)	598	N/A	1	0	60.9	.33 (.26–.40)	.33 (.26–.40)
Abdominal tenderness ^a (presence/absence)	598	593 (99.0)	1	0	98.7	.49 (.19–.80)	.49 (.19–.80)
Right-sided abdominal tenderness	583	562 (93.8)	15	1	91.1	.30 (.16–.44)	.31 (.17–.44)
Maximal abdominal tenderness in RLQ	554 ^d	447 (80.7)	29	0	80.0	.45 (.36–.54)	.42 (.33–.51)
Rebound	548	224 (37.4)	51	0	68.4	.32 (.24–.40)	.30 (.22–.38)
Guarding	584	384 (64.1)	14	1	67.5	.30 (.23–.38)	.30 (.22–.37)
Abdominal pain with walking, jumping, or coughing	494	418 (69.8)	103	2	83.6	.54 (.45–.63)	.44 (.35–.53)
Right-sided abdominal pain with walking, jumping, or coughing	416 ^e	311 (74.8)	76	2	77.6	.52 (.43–.60)	.45 (.37–.53)

^a Abdominal tenderness variable analyzed as ordinal data (none, mild, moderate, or severe tenderness) as originally collected and as a dichotomous variable (presence or absence of tenderness).

^b "Unsure" responses excluded from analysis.

^c "Unsure" responses recoded as the presence of particular finding.

^d Excludes patients in whom clinician marked "unsure" for right-sided abdominal pain.

^e Excludes those patients in whom clinician marked "unsure" for abdominal pain with walking, jumping, or coughing.

for 4 of 6 variables. Only 1 patient history variable, history of emesis, had substantial or greater reliability. Migration of pain was the patient history variable with the largest number of responses of "don't know."

For physical examination variables, reliability was at least moderate ($>.40$) for 4 of 8 variables. The highest interrater reliability was seen for presence of abdominal pain with walking, jumping, or coughing ($\kappa = .54$) and for right-sided abdominal pain during walking, jumping, or coughing ($\kappa = .52$). However, these 2 variables also had the most responses for which at least 1 assessor indicated "unsure." In secondary analyses, recoding the "don't know" or "unsure" responses as if the patient

had the respective sign or symptom resulted in slightly lower κ values. This effect was most pronounced for physical examination variables.

The effect of age on the κ statistics is demonstrated in Table 4. When we limited the analysis to children aged ≥ 6 years, the interrater reliability improved, although marginally. In this subgroup analysis, the most substantial increases in κ values were observed for the history of migration of pain to the RLQ and for the presence of any abdominal tenderness on physical examination.

DISCUSSION

In this large, multicenter study, we found that the interrater reliability of

patient history and physical examination findings among children being evaluated for possible appendicitis was generally fair to moderate. Interrater reliability between clinician assessors was highest for the duration of abdominal pain, history of emesis, presence of any abdominal tenderness, and pain with walking, jumping, or coughing. Those variables with at least moderate reliability may be better suited for inclusion in clinical prediction rules and guidelines to assess for appendicitis so as to improve generalizability.

One earlier study examined the interrater reliability of historical findings in children specifically being evaluated for possible appendicitis. In this single-center study, 350 children with acute

TABLE 4 Interrater Reliability of Patient History Variables in Patients 6 to 18 Years Old

Criterion	School-Aged Children (≥ 6 y) $N = 529$		
	"Don't know" or "Unsure," n	Overall Agreement, %	κ (95% CI)
Duration of pain, h	0	66.0	.57 (.52–.62)
Focal RLQ pain	14	84.8	.46 (.36–.56)
Migration of pain to RLQ	40	69.0	.38 (.30–.46)
Anorexia	25	67.4	.32 (.24–.40)
Nausea	18	71.5	.43 (.35–.51)
Emesis	2	91.8	.84 (.79–.88)
Abdominal tenderness ^a (presence/absence)	1	99.1	.61 (.30–.92)
Abdominal tenderness ^a (none, mild, moderate, severe)	1	62.1	.35 (.27–.42)
Right-sided abdominal tenderness	13	92.6	.37 (.21–.53)
Maximal abdominal tenderness in RLQ	19	81.3	.47 (.38–.57)
Rebound	32	68.6	.34 (.25–.42)
Guarding	10	66.7	.30 (.21–.38)
Abdominal pain with walking, jumping, or coughing	72	83.7	.53 (.44–.63)
Right-sided abdominal pain with walking, jumping, or coughing	66	77.6	.51 (.42–.60)

^a Abdominal tenderness variable analyzed as ordinal data (none, mild, moderate, or severe tenderness) as originally collected and as a dichotomous variable (presence or absence of tenderness).

abdominal pain were evaluated by PEM faculty and fellow physicians and surgical residents. The investigators found moderate interrater reliability for duration of pain ($\kappa = .53$, 95% CI .43–.63), presence of nausea ($\kappa = .54$, 95% CI .45–.64), emesis ($\kappa = .82$, 95% CI .76–.89), history of migration of pain ($\kappa = .43$, 95% CI .33–.53), and history of focal RLQ pain ($\kappa = .52$, 95% CI .42–.62).⁸ Our findings are similar, although we noted slightly lower reliability for a history of nausea and for a history of migration of pain.

Two other relevant studies have examined the interrater reliability of physical exam parameters in patients with acute abdominal pain, although these studies were not specific to appendicitis.^{10,11} In a study of adult patients with abdominal pain, the investigators reported κ values ranging from .27 to .82 for physical examination findings. The presence of abdominal tenderness had a κ of .42 (95% CI .23–.61), consistent with our results. In a separate study of children, investigators described generally low interrater reliability between the PEM attending and residents or the PEM attending and surgeons for physical examination findings, with κ values ranging

from .15 to .54. The relatively lower reliability found in that study may have been in part due to skewed responses to some of the variables (eg, most physicians found the sign or symptom present). In circumstances in which there is a high level of agreement due to chance, a statistical paradox can occur, such that the absolute agreement is high and the calculated κ value is low.^{14,15} Additionally, the somewhat higher κ values in our study may have resulted from both assessments being performed most frequently by PEM physicians and within 1 hour of each other.

Our findings and those of previous studies highlight the variability in eliciting historical and physical examination findings in children. Age and maturity of the patient being examined likely impact the assessment of some of the clinical factors. Though not studied here, the variability in physical examination findings may also reflect the fluid nature of the physical examination and the need to perform multiple examinations to gain a better perspective of a patient's clinical status. As researchers look to develop clinical prediction rules and clinical pathways to evaluate

children with acute abdominal pain, they will need to account for this variability. It is possible that composite variables, such as a combination variable "nausea and emesis," may be more reliable and have increased usefulness over individual clinical variables.⁹

Regardless of the variability noted, our results reveal that some physical examination parameters are more reliable than others. For instance, the ability to determine whether a patient had any abdominal tenderness or maximal tenderness in the RLQ had a moderate degree of reliability. This is in comparison with rebound tenderness or presence of guarding, both of which had only fair interrater reliability. In addition, the physical examination variables had differing rates of clinicians marking "unsure." For example, clinicians noted being unsure if the patient had pain with walking, jumping, or coughing in >15% of cases. We assessed this variable as it was included in the Samuel's score; however, it is possible that clinicians were not accustomed to eliciting this sign and felt less confident in their examination accuracy.⁶ Although our findings support that this variable should be considered for inclusion in future appendicitis prediction rules, it is possible that a patient's age and his or her degree of abdominal pain along with physician comfort assessing this finding may limit its reproducibility and clinical usefulness.

Our study had the following limitations. Although many different physicians performed the clinical assessments, most were conducted by PEM physicians. Therefore, our results may not be generalizable to other clinical settings such as community hospitals or acute care facilities that rely on nonpediatric providers. Additionally, we allowed the 2 clinician assessments to occur up to 60 minutes apart, even this amount of time may decrease reliability due to fluctuations in the patients' symptoms and

signs. Furthermore, we did not reliably capture on our case report forms the timing of pain medication administered between clinician assessor examinations. This information would have been useful, although excluding those patients who received pain medication between examinations would likely have led to higher interrater reliability. It should be noted that the CIs for our κ statistics were wide; thus, the point estimates reported may overestimate the true level of agreement. Finally, there is debate as to the appropriateness of using the κ statistic when there is a large skew in

the distribution of responses (eg, most responses were yes or no), as it was for several variables in our study.¹⁴ The benefit of the κ statistic is that it provides an index that can be compared across settings.

CONCLUSIONS

The interrater reliability of patient history and physical examination findings of children with possible appendicitis was variable and generally fair to moderate. This modest reliability should be understood and accounted for when

developing prediction rules and clinical pathways that rely on specific patient findings to guide clinical management. Duration of pain, history of emesis, presence of any abdominal tenderness, and pain with walking, jumping, or coughing were the variables with the highest degree of interrater reliability in children with possible appendicitis.

ACKNOWLEDGMENTS

We thank all of the clinicians who enrolled patients into this study and the research coordinators who greatly facilitated study completion.

REFERENCES

1. Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. *Am J Epidemiol*. 1990;132(5):910–925
2. Guthery SL, Hutchings C, Dean JM, Hoff C. National estimates of hospital utilization by children with gastrointestinal disorders: analysis of the 1997 kids' inpatient database. *J Pediatr*. 2004;144(5):589–594
3. Bundy DG, Byerley JS, Liles EA, Perrin EM, Katznelson J, Rice HE. Does this child have appendicitis? *JAMA*. 2007;298(4):438–451
4. Brenner D, Elliston C, Hall E, Berdon W. Estimated risks of radiation-induced fatal cancer from pediatric CT. *AJR Am J Roentgenol*. 2001;176(2):289–296
5. Kharbanda AB, Taylor GA, Fishman SJ, Bachur RG. A clinical decision rule to identify children at low risk for appendicitis. *Pediatrics*. 2005;116(3):709–716
6. Samuel M. Pediatric appendicitis score. *J Pediatr Surg*. 2002;37(6):877–881
7. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med*. 1999;33(4):437–447
8. Kharbanda AB, Fishman SJ, Bachur RG. Comparison of pediatric emergency physicians' and surgeons' evaluation and diagnosis of appendicitis. *Acad Emerg Med*. 2008;15(2):119–125
9. Gorelick MH, Atabaki SM, Hoyle J, et al. Interobserver agreement in assessment of clinical variables in children with blunt head trauma. *Acad Emerg Med*. 2008;15(9):812–818
10. Pines J, Uscher Pines L, Hall A, Hunter J, Srinivasan R, Ghaemmaghami C. The interrater variation of ED abdominal examination findings in patients with acute abdominal pain. *Am J Emerg Med*. 2005;23(4):483–487
11. Yen K, Karpas A, Pinkerton HJ, Gorelick MH. Interexaminer reliability in physical examination of pediatric patients with abdominal pain. *Arch Pediatr Adolesc Med*. 2005;159(4):373–376
12. Cicchetti D, Bronen R, Spencer S, et al. Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers. *J Nerv Ment Dis*. 2006;194(8):557–564
13. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*. 1977;33(2):363–374
14. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543–549
15. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol*. 1990;43(6):551–558

(Continued from first page)

www.pediatrics.org/cgi/doi/10.1542/peds.2011-2037

doi:10.1542/peds.2011-2037

Accepted for publication Nov 29, 2011

This work was presented in part at the annual meeting of Pediatric Academic Societies; April 30–May 3, 2011; Denver, CO.

Address correspondence to Anupam Kharbanda, MD, MSc, Division of Pediatric Emergency Medicine, University of Minnesota, Amplatz Children's Hospital, East Building, Room M6532450 Riverside Ave, Minneapolis, MN 55454. Email: ak493@columbia.edu or abk@umn.edu

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275).

Copyright © 2012 by the American Academy of Pediatrics

FINANCIAL DISCLOSURE: The authors have indicated they have no financial relationships relevant to this article to disclose.

FUNDING: Supported by grant UL1 RR024156 from the National Center for Research Resources (NICRR), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research. Dr Kharbanda received salary support from the Empire Clinical Scholars Program (New York State Department of Health). The PEM-CRC data center is supported in part by the Center for Clinical Effectiveness at Baylor College of Medicine/Texas Children's Hospital. Funded by the National Institutes of Health (NIH).