Interexaminer Reliability in Physical Examination of Pediatric Patients With Abdominal Pain

Kenneth Yen, MD; Anna Karpas, MD; Heidi J. Pinkerton, MD; Marc H. Gorelick, MD, MSCE

Objective: To test the interexaminer reliability of abdominal examinations performed by pediatric emergency medicine physicians and surgeons in an emergency department.

Methods: A prospective cross-sectional study in which 3 different types of physicians (pediatric emergency department residents, pediatric emergency department attending physicians, and pediatric surgeons in training) independently examined a convenience sample of children (aged 3-19 years) with initial complaint of abdominal pain. The interexaminer reliability of 6 components of the abdominal examination (the presence or absence of abdominal distension, abdominal tenderness to percussion, abdominal tenderness to palpation, abdominal guarding, rebound tenderness, and bowel sounds) and the clinical diagnosis of peritonitis was tested.

Results: Sixty-eight patients were examined by pediatric emergency department residents and pediatric emergency department attending physicians. All 3 physician types examined 46 of these 68 patients. When comparing residents and attending physicians, the components of the abdominal examination showed less than moderate chance-adjusted agreement ($\kappa$ range, −0.04 to 0.38). When comparing attending physicians and surgeons, the presence of rebound tenderness showed moderate agreement ($\kappa=0.54$). The rest of the components demonstrated less than moderate chance-adjusted agreement ($\kappa$ range, −0.04 to 0.34).

Conclusions: The components of the abdominal examination are poorly reliable between physician types. Only the “rebound tenderness” component of the abdominal examination shows moderate agreement between the pediatric emergency department attending physicians and the surgeon. No component of the abdominal examination appears to be consistently reliable. Interexaminer agreement must be considered when developing management strategies for acute abdomen. Interventions to improve reliability should be developed.

Arch Pediatr Adolesc Med. 2005;159:373-376
The primary objective of this study was to test the interexaminer reliability of abdominal examinations performed by pediatric emergency medicine physicians and surgeons in an emergency department in the evaluation of children with abdominal pain.

METHODS

This was a cross-sectional study conducted at a university-affiliated, urban pediatric emergency department with an annual emergency department census of 43,000. During the study period (May 1, 2002, to April 1, 2003), 3- to 19-year-old patients seen in the pediatric emergency department with a chief complaint of abdominal pain of less than 72 hours’ duration were eligible for enrollment and prospectively identified. Patients with prior abdominal surgery and/or developmental delay were excluded. Study subjects were a convenience sample of eligible patients who were seen when a research assistant was present to enroll the patients and when a pediatric surgeon was available to examine the patient during the study time frame. Formal surgical consultation was not required. Informed consent and assent were obtained prior to study commencement. The study was approved by the hospital’s institutional review board.

Three different physician types examined the patient: residents rotating in the emergency department (including pediatric, emergency medicine, or family practice residents), pediatric emergency medicine physicians (attending physician or fellow), and pediatric surgeons in training (senior surgical resident or fellow).

The examiners were allowed to ask the patient’s medical history. The 3 examiners independently performed the physical examination (including the abdominal examination). Each examination was performed with only 1 examiner in the room with the patient at a time. Each of the examiners was blinded to the findings of the other examiners. The interval between examinations was less than 30 minutes. The order of examiners was resident, pediatric emergency medicine physician, and then pediatric surgeon. Each physician completed an abdominal examination form eliciting information about the presence of the clinical findings shown in the Table. If any part of the examination was not possible to complete, the physician had the option of marking “unable to assess.” During the session, the patient was not told the results of the examination.

The prevalence of each of the clinical findings for each of the physician types was calculated. For each clinical finding, 2 × 2 tables comparing the rating (present or absent) by each of 2 observers were constructed. Two pairwise comparisons were evaluated: pediatric emergency medicine physician vs resident and pediatric emergency medicine physician vs pediatric surgeon. For each 2 × 2 table, the total overall agreement and a k statistic (chance-adjusted agreement) were calculated. By convention, a k value of 0.6 or greater is considered to indicate good to excellent agreement, while a k of less than 0.4 is considered poor agreement. For the examination components of the abdominal examination, we considered moderate agreement (k range, 0.4–0.6) as a reasonable goal. Bootstrapping was used to calculate 95% confidence intervals. All statistical computations were performed with Intercooled Stata 7 for Windows (STATA Corp, College Station, Tex).

RESULTS

A total of 68 patients had examinations performed by both a pediatric emergency medicine resident and the pediatric emergency medicine attending physician. Forty-six patients had examinations performed by both the pediatric emergency medicine attending physician and a pediatric surgeon. The pediatric surgeon did not complete the abdominal examination within the half-hour window for 22 of the patients. The mean age of subjects was 10 years, and 68% of the subjects were female. Six patients were admitted for surgical reasons (observation for possible appendicitis or actual appendicitis), 5 patients were admitted for medical reasons (eg, rehydration), and the rest were discharged home.

The prevalence of positive examination finding for each of the components is presented in the Table. Overall agreement between pediatric emergency medicine attending physicians and pediatric emergency department residents was calculated and presented in the Table. For each of the examination components, there is relatively high raw agreement (ranging from 69.1% for abdominal tenderness with palpation to 95.5% for clinical diagnosis of peritonitis).

A similar pattern of agreement exists between pediatric emergency medicine attending physicians and pediatric surgeons. Overall agreement between these groups ranged from 65.2% for abdominal tenderness with palpation to 95.5% for clinical diagnosis of peritonitis.

However, chance-adjusted agreement was generally fair to poor, as presented in the Table. The k statistic for the agreement between pediatric emergency medicine physicians and residents on examination components

Table. Percentage Overall Agreement and k Statistic for Each Clinical Examination Finding and Diagnosis of Peritonitis

<table>
<thead>
<tr>
<th>Clinical Finding</th>
<th>PEM Attending Physician and Resident</th>
<th>PEM Attending Physician and Surgeon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal distension</td>
<td>87.5</td>
<td>84.4</td>
</tr>
<tr>
<td>Tenderness to percussion</td>
<td>69.7</td>
<td>71.1</td>
</tr>
<tr>
<td>Tenderness to palpation</td>
<td>69.1</td>
<td>65.2</td>
</tr>
<tr>
<td>Abdominal guarding</td>
<td>73.1</td>
<td>91.3</td>
</tr>
<tr>
<td>Rebound tenderness</td>
<td>86.9</td>
<td>93.5</td>
</tr>
<tr>
<td>Absent bowel sounds</td>
<td>91.0</td>
<td>88.6</td>
</tr>
<tr>
<td>Clinical diagnosis of peritonitis</td>
<td>95.5</td>
<td>95.5</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; PEM, pediatric emergency medicine.
ranged from −0.04 for absence of bowel sounds to 0.38 for the diagnosis of peritonitis. Similar values for $\kappa$ were found between pediatric emergency medicine physicians and pediatric surgeons. The only examination component with $\kappa$ higher than 0.4 was rebound tenderness, with a $\kappa$ of 0.54, indicating intermediate agreement.

We found poor interexaminer reliability among a broad range of examiners in the clinical examination of acute abdominal pain. Only the presence of rebound pain showed moderate agreement between the pediatric surgeons in training and the pediatric emergency medicine attending physician.

These differences may be the result of differences in training and levels of experience. Since the study did not define what is considered a positive finding, each of the physician types may have different definitions of what a true, positive finding is.

A possible explanation for the difference between pediatric emergency medicine physicians and surgeons is that these groups may, in effect, calibrate their examinations differently because of different intrinsic thresholds. The presence of positive clinical examination components tended to be higher with the pediatric emergency medicine attending physician compared with the pediatric surgeon. This may be related to the fact that positive clinical examination results would mean different things for different physician types. For the pediatric emergency medicine physician, a positive examination finding would mean a pediatric surgery consultation; false-positive results bear relatively little cost compared with a false-negative result, which could lead to a missed diagnosis. For a pediatric surgeon, in contrast, a positive examination result may mean surgery. Surgeons would be expected to want to minimize false-positive results. Thus, pediatric emergency medicine physicians would interpret findings in a way to maximize sensitivity, while the surgeons may aim for greater specificity.

There are other limitations to this study. Since most of the patients had normal test results for many of the clinical examination components, considerable agreement may occur when the abnormal examination findings are rare. Under such circumstances, the $\kappa$ statistic may appear artificially low.9,10

Furthermore, our study did not look at the severity or final outcome of the patient. For example, we did not look at who went to surgery or whether patients who were discharged from the hospital returned with an acute abdomen. Agreement may improve if the abdominal pain is more severe. Thomas et al11 demonstrated fairly good agreement for the need for opioid analgesia between patients and physicians. Also, physicians may agree more frequently in patients with more serious diagnoses, in whom findings may be less equivocal. Relatively few patients in this study were thought to have peritoneal signs by any of the examiners.

The implications of this are that interexaminer agreement must be considered when developing management strategies for acute abdominal pain. Recent increases in the development and use of clinical guidelines and practice pathways are a prime example of how determination of interexaminer reliability of physical examination components may be important. Clinical
guidelines that consider the abdominal physical examination in their decision trees must be aware of the poor reliability of the examination components between physician types.

More research on the reliability of abdominal and other physical examination components needs to be performed to understand the limitations of the clinical examination. Interventions that improve abdominal examination reliability should be developed and investigated. As stated earlier, our study did not define what was a normal or abnormal examination finding. Providing definitions would probably be a first and essential step to reduce variability, after which, standardizing the method of examination should also be investigated and considered.

Accepted for Publication: December 7, 2004.

Correspondence: Kenneth Yen, MD, Department of Pediatrics, Medical College of Wisconsin, 9000 W Wisconsin Ave, Milwaukee, WI 53224 (kyen@mail.mcw.edu).

Acknowledgment: We thank Jo Bergholte, MS, and her staff of research assistants for assistance with patient enrollment, data collection, data entry, and review of the manuscript.

REFERENCES