Inter-rater agreement in the evaluation of abdominal radiographs for necrotizing enterocolitis

Sherif El-Kady, Dara Petel, Robert Baird *

Department of Pediatric General and Thoracic Surgery, The Montreal Children's Hospital, McGill University, Montreal Quebec, Canada
McGill University, Montreal Quebec, Canada
Department of Radiology, The Montreal Children's Hospital, McGill University, Montreal Quebec, Canada

A R T I C L E   I N F O
Article history:
Received 6 February 2014
Accepted 13 February 2014

Key words:
Necrotizing enterocolitis
Abdominal radiograph
Inter-rater agreement

ABSTRACT
Purpose: Abdominal radiographs are frequently employed in the surveillance of patients with necrotizing enterocolitis (NEC), with typical findings well described. Clinicians interpret and act upon these films at different intervals, however, and inter-rater agreement has not been evaluated to date.

Methods: Thirty abdominal radiographs of premature infants were distributed to attending radiologists (4), pediatric surgeons (4), and trainees (4), who evaluated for findings of NEC from a list of eight potential choices (1 = normal, 8 = perforation). Fleiss's Kappa (FK) was used to evaluate concordance between multiple raters with 0–0.2 = slight association and 0.8–1 = almost perfect agreement.

Results: Practicing surgeons had an FK of 0.77 overall (95% CI: 0.67–0.87), but demonstrated poor agreement when evaluating decubitus films (FK: 0.39, 95% CI: 0.12–0.65). Radiologists had excellent inter-rater agreement (FK: 0.81, 95% CI: 0.74–0.88), but had only modest agreement with surgeons (FK: 0.59, 95% CI: 0.56–0.63) and poor agreement for decubitus films (FK: 0.15, 95% CI: 0.47–0.26). Surgical and radiology trainees had fair agreement with their respective attendings (0.60, 95% CI: 0.55–0.65 and 0.64, 95% CI: 0.60–0.69, respectively).

Conclusions: While inter-rater agreement was good–excellent among attending staff, it was only moderate between radiologists and surgeons and between trainees and their attendings. This highlights the importance of inter-disciplinary and hierarchical communication to optimize clinical decision-making. Decubitus films may be of limited value in evaluating patients with NEC.

© 2014 Elsevier Inc. All rights reserved.

Necrotizing enterocolitis (NEC) is a devastating surgical disease of prematurity that causes significant morbidity and mortality [1,2]. NEC has a variable course, and can quickly progress from abdominal distension to hemodynamic instability, sepsis, intestinal perforation and death. Vigilant monitoring of disease progression is crucial due to the extreme velocity of clinical change and overall fragility of this patient population. Ultimately, surgical intervention is required in up to 50% of patients diagnosed with NEC, with mortality rates reaching 40% in some series [1,3].

Diagnosis, staging, choice of therapy and monitoring of disease progression are generally supported through radiological abdominal imaging. Typical findings of NEC on radiograph include an abnormal gas pattern, pneumatosis intestinalis (PI) and gas in the portal vein, among others [4]. Radiologists, surgeons and trainees as well as members of the NICU team routinely evaluate these films, although the timing of interpretation typically differs, with point-of-care clinicians frequently called upon to make significant treatment decisions before the reading of a radiologist. Inter-rater agreement between clinicians has yet to be evaluated for the interpretation of abdominal radiographs for NEC; we hypothesize that the inter-rater agreement is suboptimal. Given that radiographs for NEC have been shown to increase the potential long-term malignancy risk [5], gross discordance amongst clinicians could signal a need to reduce superfluous radiographs while standardizing the interpretation of imaging performed in a more judicious manner.

1. Methods

With institutional review board approval (11-636-PED), this study was performed by recruiting attending pediatric surgeons and radiologists, as well as trainees in each of these departments, to voluntarily evaluate digital abdominal radiographs for findings of NEC. Each participant evaluated a total of thirty radiographs from neonates in the neonatal intensive care unit census at the Montreal Children's Hospital between May 2007 and May 2013. These radiographs included 15 abdominal anteroposterior (AP) views and 15 lateral shoot-through films or lateral decubitus views. All infants were <29 days old at the time of imaging and had NEC (any Bell stage) as their discharge diagnosis. Imaging was captured using typical portable x-ray equipment, and stored on a Picture Archiving and

http://dx.doi.org/10.1016/j.jpedsurg.2014.02.058
0022-3468/© 2014 Elsevier Inc. All rights reserved.
Communication System (PACS). The radiographs were selected, converted into digital files and then prepared by the senior author, with the infant’s identity obscured by cropping patient identification information from each image.

The radiographs were interpreted in the participants’ offices in the hospital, under similar illuminating conditions between 9 A.M. and 5 P.M. according to the participants’ convenience. No time limit was set, although most viewing sessions lasted less than 30 min. All observers used Microsoft Office Picture Manager (Redmond, WA, USA) software to adjust contrast, focus, brightness and zoom of the images. All radiographs were examined during one session by each observer, and the sequence of reading the radiographs was similar for all the observers, i.e., from case 1 to case 30. The interpretations were done independently, and observers did not have access to other observers’ interpretations. All observers were unaware of the clinical course and outcome of patients. Each observer was asked to evaluate for the absence, suspicion, or presence of the following radiologic signs: (A) Normal, (B) Bowel dilatation (C) Bowel wall thickening, (D) Asymmetric gas pattern, (E) Cystic pneumatosis intestinalis (F) Linear pneumatosis intestinalis and (G) Portal venous air and (H) Free intra-peritoneal air. Findings were not deemed mutually exclusive (except for “A” and others), and multiple findings were allowed. No attempt was made to enforce a uniform interpretation criterion for individual radiologic signs of NEC.

Fleiss’s Kappa (FK) was used to evaluate concordance between multiple raters within and between each group [6]. This technique evaluates the reliability of agreement amongst ≥ 1 rater, with 0–0.2 = slight association and 0.8–1 = almost perfect agreement [7]. When respondents provided multiple interpretations, concordance was deemed to exist when one interpretation matched in order to maximize the potential for concordant interpretations. Key radiologic findings were also individually assessed, including the interpretation of the presence or absence of free air or PI (a summary of cystic and linear PI). Differences in these were evaluated using the Student t test, with p < 0.05 considered significant.

2. Results

Four pediatric surgeons and four radiologists participated in this study, along with two trainees from each discipline for a total of 360 independent interpretations performed. Surgeons had an FK of 0.77 overall (95% CI: 0.67–0.87), but demonstrated poor agreement when evaluating decubitus films (FK: 0.39, 95% CI: 0.13–0.65) (Table 1). Radiologists had excellent inter-rater agreement (FK: 0.81, 95% CI: 0.74–0.88) (Table 2). On the other hand, radiologists had only modest agreement with surgeons (FK: 0.59, 95% CI: 0.56–0.63) and poor agreement with surgeons for decubitus films (FK: 0.15, 95% CI: 0.47–0.26), see Table 3. Fair agreement was seen with radiologists and surgeons for AP and lateral shoot-through films (FK: 0.65, 95% CI: 0.59–0.71 and FK: 0.59, 95% CI: 0.51–0.67, respectively). Surgical and radiology trainees had fair agreement with their respective attending faculty (0.60, 95% CI: 0.55–0.65 and 0.64, 95% CI: 0.60–0.69, respectively).

There were two discordant readings for the presence of free air on lateral films, three on AP views. In all cases, the discordance suggested a trainee failed to identify free air while all attending physicians did.

Table 1
Inter-rater agreement among Pediatric Surgeons (n = 4).

<table>
<thead>
<tr>
<th>Study type</th>
<th>FK (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP radiographs</td>
<td>0.726 (0.581 to 0.87)</td>
</tr>
<tr>
<td>Lateral</td>
<td>0.801 (0.652–0.959)</td>
</tr>
<tr>
<td>Decubitus</td>
<td>0.389 (0.128–0.651)</td>
</tr>
<tr>
<td>All</td>
<td>0.771 (0.673–0.869)</td>
</tr>
</tbody>
</table>

AP = Anteroposterior, FK = Fleiss’s Kappa.

Table 2
Inter-rater Agreement among Radiologists (n = 4).

<table>
<thead>
<tr>
<th>Study type</th>
<th>FK (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP radiographs</td>
<td>0.828 (0.717 to 0.938)</td>
</tr>
<tr>
<td>Lateral</td>
<td>0.832 (0.671–0.994)</td>
</tr>
<tr>
<td>Decubitus</td>
<td>0.51 (0.251–0.769)</td>
</tr>
<tr>
<td>All</td>
<td>0.81 (0.738–0.881)</td>
</tr>
</tbody>
</table>

AP = Anteroposterior, FK = Fleiss’s Kappa.

The interpretation of PI was highly variable, with attending radiologists identifying its presence more often than attending surgeons (11.2 ± 4.3 vs 8.3 ± 5.2, p < 0.01). No difference was detected when characterizing PI as either linear or cystic.

3. Discussion

NEC is a relatively common disease of prematurity characterized by inflammation and necrosis of the gastro-intestinal tract that is associated with significant morbidity and mortality [8]. Perforation of the intestines occurs in up to 31% of patients; greatly increasing the associated mortality rate, which despite advances in neonatal and surgical care continues to hover between 20% and 40%. NEC is a very unpredictable disease that can worsen precipitously [1,2]. Although three stages of NEC have been described – suspected disease, definite disease, and advanced disease [9] – these do not necessarily connote a predictable step-wise progression.

Surveillance of NEC is typically performed using clinical indicators and radiological imaging. Practice patterns vary in the frequency of imaging, but typically multiple radiographs are acquired per day for patients with active disease [2]. These films are evaluated by neonatologists, pediatric surgeons and pediatric radiologists. Clinical care decision-making is based greatly on the interpretation of these films, particularly the presence or absence of pneumoperitoneum. In this study, inter-rater agreement was very good amongst attending staff, and moderate between radiologists and surgeons and between trainees and their attendings. Whereas AP and lateral shoot-through films demonstrated fair inter-rater agreement, decubitus films demonstrated poor inter-rater agreement and may therefore have limited value in evaluating patients with NEC.

The evaluation of inter-observer agreement for plain radiographs has been performed by pediatric radiologists in the context of other disease processes. A study evaluating plain x-rays and intussusception was able to demonstrate that long-held beliefs about the optimal patient positioning was incorrect; practice patterns changed accordingly [10]. Other examples include evaluations of inter and intra-rater evaluations of pediatric chest radiographs for signs of lower respiratory infections and neonatal imaging for atelectasis [11–13]. In these instances, caution was advised in confidently converting radiologic imaging into diagnostic criteria. Furthermore, differences in interpretations between radiologists and clinicians led to increased emphasis on communication between departments when making decisions of clinical consequence. Based on the variable interpretation of abdominal radiographs in our study, combined with the profound magnitude of clinical decisions in neonates, a similar improvement in cross-discipline communication appears warranted.
There are few reports about the reliability of neonatal abdominal imaging in necrotizing enterocolitis. The review by Buonomo et al. illustrates all of the classic findings associated with NEC with examples, although no descriptive data are provided to ascertain incidence and reliability of radiographic findings [14]. Frey et al. noted that of nineteen confirmed perforations over a 5.5 year period, only 63% had demonstrable findings on abdominal plain films [12]. The authors make the point that the timing of radiographs likely resulted in a significant false negative rate, although interpretative error cannot be excluded as an explanatory factor [15]. The spectra of missing radiological pneumoperitoneum cannot be understated, as clinical deterioration typically ensues shortly thereafter [16]. Our study demonstrated that trainees might still miss the diagnosis of free air; vigilance must therefore be maintained in ensuring multiple experienced clinicians evaluate the radiographs of high-risk neonates.

Recently, the Dukes Abdominal Assessment Scale (DAAS) has been described as an effort to address the significant discordance in the interpretation of abdominal radiographs in neonates with suspected or confirmed NEC [17]. It is a 10-point rating scale that enumerates all of the typical findings observed in NEC, and was evaluated using PACS, as in our study. The mean Intra- and Inter-observer agreement among radiologists (Cohen’s Kappa value, equivalent to the Fleiss’s Kappa value for single raters) was 0.79 and 0.66, respectively. These findings are similar to the results reported in this study, and suggest that further refinement of the scale is required to improve reproducibility. These findings were significantly superior to those reported in the era before digital imagery, which ranged from 0.31 to 0.56 [18-20]. In addition, the DAAS has been shown to correlate with clinical outcome, with higher scores predicting worse outcomes [21].

Given the suboptimal performance of radiographs demonstrated in this study, practice pattern alterations appear warranted. Assessment of radiographs could be improved by implementing greater cross-discipline communication, as well as by having the imaging of high-risk neonates reviewed by multiple clinicians in order to minimize errors and to standardize interpretation [22]. Increased emphasis on alternative surveillance strategies for neonates with NEC should also continue to be explored. Abdominal ultrasonography has been shown to be equivalent and possibly superior to abdominal radiography at detecting intramural air, gas in the portal vein, and thinning of the bowel wall [23]. Abdominal ultrasonography has been found useful for guiding treatment for NEC, and to have equal or superior sensitivity to abdominal radiography with regards to the detection of intra-peritoneal air [24]. Finally, non-radiologic adjuncts to traditional clinical surveillance may ultimately improve the timeliness and precision of treatment. These include serial utilization of a clinical severity score [25], dynamic heart rate monitoring [26], or evaluation of plasma, urinary or fecal markers [27,28].

This study had several limitations. There were only a small number of participating clinicians; their assessments of the provided images might not have been representative of clinicians at large. In addition, it was not possible to directly predict the impact of inter-rater agreement on the clinical course of these patients as the reported findings were not correlated with the evolution of the disease. It may also have been useful to assess inter-rater agreement in the use of other imaging techniques, such as abdominal ultrasonography, in order to provide a more complete picture of the monitoring of NEC as it is carried out in our facility.

Radiography remains a key adjunct in the surveillance for progression of NEC. This study demonstrates variable inter-rater agreement when evaluating neonatal abdominal radiographs, suggesting a need for better inter-disciplinary and hierarchical communication. Additional efforts to improve concordance and adopt objective radiologic criteria appear warranted, as does increasing the utilization of alternative surveillance strategies.

References