

Question Sheet

Prospective Validation of the Pediatric Appendicitis Score in a Canadian Pediatric Emergency Department

Bhatt M, Joseph L, Ducharme FM et al. *Acad Emerg Med* 2009;16(7):591-596

1. Provide a no more than 200 word summary of this paper in the box provided. Only the first 200 words will be considered – short bullet points are acceptable. Maximum of 7 marks available.

Answer

- *Prospective diagnostic observational study (must mention prospective) (1 Mark)*
- *Study population – children 4-18 years old presenting to the ED with suspected acute appendicitis (1 Mark)*
- *Test of interest – Pediatric appendicitis score (1 mark)*
- *Reference standard – diagnosis of appendicitis based on histology (1/2 mark) AND telephone follow-up at 30 days for children who did not undergo surgery (1/2 mark)*
- *Results*
 - Single pediatric appendicitis score cut point of 6:*
 - Sensitivity 92.8% (95% CI 85.1 to 96.6)*
 - Specificity 69.3% (95% CI 61.9 to 75.9)*
 - (1 mark for all three including the cut off – CI not mandatory)*
 - AND*
 - Using two cut points:*
 - PAS less than/equal to 4 – Sens. 97.6%, NPV 97.7% (1/2 mark for either)*
 - PAS greater than/equal to 8 – Spec 95.1%, PPV 85.2% (1/2 mark for either)*
- *Author's conclusion – Single cut point didn't perform sufficiently to allow discharge or transfer to theatre decisions (1/2 mark) need to say single cut off for this half mark*
 - Using two cut points usefully identifies a group that can be safely discharged (PAS 4 or less), a group that need surgery (PAS 8 or more) and a group in the middle that require further evaluation. (1/2 mark – must have two of the three scenarios)*

If mention needs validation using multiple cutoffs gave one half as well

The conclusion should be the authors' rather than the candidate's.

2. The primary objective of this study was to determine the diagnostic properties of the Pediatric Appendicitis Score cut-point of 6 for diagnosing appendicitis.

List four strengths of the study design in this paper. (4 marks)

Answer

- *Prospective design (1 mark)*
- *Appropriate study group – those suspected of the target disorder (acute appendicitis) i.e. appropriate inclusion criteria (1 mark) inclusion criteria correct even if population skewed as tertiary centre*
- *Appropriate exclusion criteria (1 mark)*
- *Score components recorded blinded to the score (1 mark)*
- *Interobserver agreement assessed (1 mark)*
- *Appropriate reference standard (1 mark) and applied to all (1 mark)*
- *Intended to follow-up (1 mark)*
- *Standardized data collection form*
- *Appropriate training in use of score*
- *Scoring done before ordering tests or referring*
- *ED assessment not Surgical*
- *Intended follow up of all non operated*
- *Did not interfere with normal practice*
- *Test under scrutiny was independent of the gold standard and applied to all as was the gold standard (hybrid)*
- *One person entered all data – independent of the scoring*

Score other answers deemed appropriate by examiners 1 mark each to a total of 4 marks

3. The paper does not mention whether those ascertaining the outcome diagnosis ('appendicitis' or 'no appendicitis') were blinded to the Pediatric Appendicitis Score.
- (a) Explain why a lack of such blinding may introduce possible bias into the results. (2 marks).

Answer

- (a) *Lack of blinding to the test of interest result may bias those ascertaining the reference diagnosis if the means of making the reference diagnosis is in any way subjective e.g. interpreting a scan. In that case, knowledge of the test of interest result might make the assessor more or less likely to make a particular 'call' on the reference standard (1 mark).*

Reference standards that are objective (e.g. automated blood result) are unlikely to give rise to bias due to lack of such blinding (1 mark)

4. (a) The results section of the paper reports that a Pediatric Appendicitis Score cut-point of 6 or more had a sensitivity of 92.8% and a specificity of 69.3% for the diagnosis of appendicitis.

Comment on the utility of this cut point in ruling out appendicitis.
(2 marks)

- (b) With reference to the discussion section of the paper, what is the probability that a child with a Pediatric Appendicitis Score of 8 or more does not have appendicitis? (2 marks)

Answer

- (a) *The best measure to assess the ability of a test to rule out in a population is sensitivity. A sensitivity of 92.8% is not good enough to rule out appendicitis as it will miss 7 out of 100 with the diagnosis. As the consequences of missing the diagnosis are significant, this miss rate is not acceptable.*

(1 mark for identifying sensitivity as the appropriate measure and 1 mark for saying that rule out ability not good enough)

- (b) *The positive predictive value is the relevant measure to look at as it describes the probability of disease in those with a positive result.*

The PPV for a cut point of 8 is 85.2%. The probability of having appendicitis using this cut point is 85.2%; the probability of not having appendicitis is 14.8%.

(1 mark for identifying PPV as the relevant measure and 1 mark for 14.8%)

5. Figure 2 in the paper presents a Receiver operating characteristic (ROC) curve.
- (a) List 2 ways by which ROC curves add to the understanding of diagnostic tests. (2 marks)

Answer

- (a) *The greater the area under the curve, the better the overall performance of the test (1 mark)*

A straight line with a slope of 1 passing through 0 denotes a test that has no diagnostic value (1 mark)

The point of the graph closest to the upper left hand corner denotes the cut point with the highest sensitivity and specificity. (1 mark)

Score other answers deemed acceptable by examiners 1 mark each to a total maximum of 2 marks.

6. Table 2 of the paper reports that 45% of those with appendicitis and 37% with no appendicitis had imaging investigations. The difference (95% CI) is 12% (-1 to 24).

(a) Is this a statistically significant difference? (1 mark)

(b) Explain your answer. (1 mark)

Answer

(a) *No (1 mark)*

(b) *The confidence interval crosses 0 suggesting that it is possible the true difference is 0. (1 mark – must say interval includes 0)*

7. The following is a quote from the results section of the paper:

'Interobserver scores were obtained in 37 (14.6%) of the 246 patients. The kappa coefficient was 0.65 (95% CI = 0.48 to 0.81) ...' (The kappa coefficient is used to express level of agreement between observers)

Comment on the level of agreement between observers in terms of the point estimate (0.65) and the 95% confidence interval (0.48 to 0.81). (2 marks)

Answer

The authors decided a priori that a kappa value greater than 0.6 was a reliable estimate of agreement. On that basis, the point estimate 0.65 suggests an acceptable level of agreement. (1 mark)

The 95% confidence interval runs from 0.48 to 0.81. This suggests that the true level of agreement could be as good as 0.81 (well above the authors' threshold of 0.6), or it could be as low as 0.48 (well below the authors' threshold). In view of the latter we are uncertain that there is actually a satisfactory level of agreement. (1 mark for understanding along these lines)

8. Give four reasons why you would not adopt this test in your Emergency Department.

Answers

- *The study used a convenience sample. It is possible that a non-representative sample was enrolled e.g. those with equivocal findings. (1 mark)*
- *The cut-point of 6 (the primary objective of the study) is not sensitive enough to rule out – patients with appendicitis could be sent home (1 mark)*
- *The cut-point of 6 (the primary objective of the study) is not specific enough to rule in and guide the decision to operate – patients without appendicitis could be operated on (1 mark)*
- *While the point estimate suggested satisfactory interobserver agreement (0.65), the 95% confidence interval fell well below the threshold of 0.6. (1 mark)*
- *We are unsure if those ascertaining the outcome diagnosis were blinded to the score (1 mark) or blinded to the workup – have to specify what they should be blinded to*
- *Single centre so not generalisable*
- *Senior clinicians (more senior than UK FY2) so not generalisable*
- *Failed to validate previous paper – and not conclusive enough*

Not powered at all and sample size too small to trust- confidence intervals too wide

Score quite subjective – ie intraobserver variability .

IR based on small sample – selective therefore ?Skewed