COMMENTARY

Appendiceal ultrasound: the importance of conveying probability of disease

Andrew T. Trout[1] · David B. Larson[2]

Received: 12 June 2015 / Accepted: 2 July 2015 / Published online: 18 August 2015

We were pleased to see the article "Pediatric Appendiceal Ultrasound: Accuracy, Determinacy and Clinical Outcomes" by Dr. Binkovitz and colleagues [1] in which the authors describe a six-category interpretive scheme for ultrasound of the appendix. The scheme that they describe is very similar to a five-category scheme that we recently published [2].

For too long, the sonographic evaluation of appendicitis has been treated as a binary interpretive scheme, which is an oversimplification that both diminishes the value of the test and predisposes to false interpretation (both positive and negative). Any radiologist who interprets appendiceal sonograms regularly (and who conscientiously follows up on his or her cases) knows that there is a subset of cases that cannot be definitively categorized as positive or negative. This is likely a reflection of the evolving character of the disease, the spectrum of disease severity, differences in host response, and technical factors that influence visualization of the appendix and surrounding soft tissues. Equivocal cases of suspected appendicitis are not unique to radiology but also manifest clinically. Two of the most well-known clinical stratification schemes, the Pediatric Appendicitis Score and the Alvarado score, include an intermediate-probability category in addition to low- and high-probability categories [3, 4].

Despite the known existence of radiologically and clinically equivocal cases, there has been an expectation that the result of an ultrasound exam should be definitive in all cases [5]. Such a desire is understandable because our clinical colleagues want a definitive result to guide patient management. The desire for a definitive positive or negative result also reflects the most common statistical approach embraced by the medical field over the last few decades, an approach that uses binary prediction and outcomes classifiers. Under this model, determinations of diagnostic accuracy depend on classification of the outcome of a test as a true positive, false positive, true negative or false negative. Such models, however, dilute the available diagnostic information by forcing all results into positive or negative categories, thereby taking what is a probability distribution and converting it to, in effect, a less confident "most likely appendicitis" and "most likely not appendicitis" [6].

Binary interpretive schemes ignore the value of knowing the confidence of the results of a test, which is especially important when multiple tests and other sources of information are available to drive decision-making, and when multiple management strategies are available depending on the likelihood of the condition. This is very much the case in acute appendicitis where history, physical exam, and laboratory and imaging data all contribute to the diagnosis and possible management strategies include further imaging, nonintervention, antibiotic therapy, immediate surgery and delayed surgery. Data from the current study by Binkovitz et al. [1] and from our prior study [2] demonstrate the value of information that is available in equivocal ultrasound results. In both studies, the likelihood of appendicitis in a sonographically equivocal case (12.9–24.3% in Binkovitz et al. [1], 25.8–39.3% in Larson et al. [2]) is distinctly different from the likelihood in a clearly positive (86.5% in Binkovitz et al. [1], 92.6% in Larson et al. [2]) or negative case (2% in Binkovitz et al. [1], 2.6% in Larson et al. [2]). An equivocal result conveying an intermediate probability of appendicitis adds

✉ Andrew T. Trout
andrew.trout@cchmc.org

[1] Department of Radiology,
Cincinnati Children's Hospital Medical Center,
3333 Burnet Ave., MLC 5031, Cincinnati, OH 45229-3026, USA

[2] Department of Radiology, Stanford University, Stanford, CA, USA

diagnostic value when synthesized with other clinical and historical data that provide a pre-test probability of disease [7].

The interpretive scheme described by Dr. Binkovitz and colleagues [1] is a step toward establishing definitive interpretive categories for appendiceal ultrasound that convey probabilities of disease. Their scheme had similar overall diagnostic accuracy to the scheme we previously described (96% versus 96.8%) despite originating in distinct patient populations [2]. The similarity in reported diagnostic accuracy between our study [2] and the study by Binkovitz et al. [1] reflects a similar means of calculating diagnostic performance in which equivocal cases are excluded from performance statistic calculations.

However, in addition to the diagnostic performance calculation that produced the similar result to our study, Dr. Binkovitz and colleagues [1] also included another calculation of test performance, described as the "intention-to-diagnose" model. This approach is based on an article by Schuetz et al. [8] and allows the performance of a test with non-binary outcomes to be captured with a single number: accuracy. This is accomplished by counting indeterminate results (indeterminate negative [IN] or indeterminate positive [IP]) as incorrect results in the calculation of accuracy using the following equation: (TP+TN)/(TP+FP+IP+TN+FN+IN). Schuetz et al. [8] argue that excluding equivocal/indeterminate results from the calculation of performance parameters overestimates the test's performance and that the intention-to-diagnose model provides "a more realistic picture of the clinical potential of diagnostic tests." We argue, however, that including equivocal categories in the calculation of performance parameters of a test is even more erroneous than not including the equivocal categories because this represents the worst case performance of the test and obscures the performance of the test when it commits to positive or negative. One test may be extremely accurate when it commits to positive or negative whereas another test may be less accurate when it commits. Ignoring equivocal cases altogether in the calculation of performance statistics, on the other hand, is also misleading, because the test may be very good when it commits, but it may frequently avoid committing, which is less helpful to clinicians.

We argue that the calculation of performance parameters of sensitivity, specificity and accuracy should exclude equivocal cases but that the frequency of equivocal cases should also be reported along with those performance statistics. Binkovitz et al. [1] themselves support this approach by the way they frame their results in their conclusion, stating that "pediatric appendiceal US can definitively rule in or rule out acute appendicitis in approximately 70% of patients with an accuracy of 96%." Furthermore, in the source that the authors cite, Fedko et al. [9] themselves recommend against using the intent-to-diagnose method in calculating performance parameters of diagnostic tests.

Finally, it should be borne in mind that the test under consideration, ultrasonography of the appendix, also necessitates a substantial element of human judgment. Such performance is not immutable; it can be studied and improved through feedback and deliberate practice. Reporting of equivocal cases separately better frames the problem to encourage practices to review those cases to identify ways to decrease the number of equivocal cases while maintaining a high level of accuracy. For example, in our study [2] a definitive result was provided in 88% of cases compared to 70% of those of Binkovitz et al. [1], while the overall accuracy in our study was higher than that reported by Binkovitz. This suggests an opportunity for improvement, which is exactly what performance measures are designed to reveal.

We applaud the work of Dr. Binkovitz and colleagues in embracing the use of equivalent categories in reporting the results of ultrasonography of the appendix. Nevertheless, we respectfully assert that the so-called intent-to-diagnose approach to calculating performance statistics of diagnostic ultrasound is misleading and should be abandoned.

Conflicts of interest None

References

1. Binkovitz LA, Unsdorfer KM, Thapa P et al (2015) Pediatric appendiceal ultrasound: accuracy, determinacy and clinical outcomes. Pediatr Radiol. doi:10.1007/s00247-015-3432-7
2. Larson DB, Trout AT, Fierke SR et al (2015) Improvement in diagnostic accuracy of ultrasound of the pediatric appendix through the use of equivocal interpretive categories. AJR Am J Roentgenol 204: 849–856
3. Alvarado A (1986) A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 15:557–564
4. Samuel M (2002) Pediatric appendicitis score. J Pediatr Surg 37: 877–881
5. Ang A, Chong NK, Daneman A (2001) Pediatric appendicitis in 'real-time': the value of sonography in diagnosis and treatment. Pediatr Emerg Care 17:334–340
6. Trout AT, Towbin AJ, Fierke SR et al (2015) Appendiceal diameter as a predictor of appendicitis in children: improved diagnosis with three diagnostic categories derived from a logistic predictive model. Eur Radiol. doi:10.1007/s00330-015-3639-x
7. Gilbert R, Logan S, Moyer VA et al (2001) Assessing diagnostic and screening tests: Part 1. Concepts. West J Med 174:405–409
8. Schuetz GM, Schlattmann P, Dewey M (2012) Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. BMJ 345:e6717
9. Fedko M, Bellamkonda VR, Bellolio MF et al (2014) Ultrasound evaluation of appendicitis: importance of the 3x2 table for outcome reporting. Am J Emerg Med 32:346–348