Optimizing the Interpretation of CT for Appendicitis: Modeling Health Utilities for Clinical Practice

C. Craig Blackmore, MD, MPH, Teruhiko Terasawa, MD

Purpose: Error in radiology can be reduced by standardizing the interpretation of imaging studies to the optimum sensitivity and specificity. In this report, the authors demonstrate how the optimal interpretation of appendiceal computed tomography (CT) can be determined and how it varies in different clinical scenarios.

Methods: Utility analysis and receiver operating characteristic (ROC) curve modeling were used to determine the trade-off between false-positive and false-negative test results to determine the optimal operating point on the ROC curve for the interpretation of appendicitis CT. Modeling was based on a previous meta-analysis for the accuracy of CT and on literature estimates of the utilities of various health states. The posttest probability of appendicitis was derived using Bayes’s theorem.

Results: At a low prevalence of disease (screening), appendicitis CT should be interpreted at high specificity (97.7%), even at the expense of lower sensitivity (75%). Conversely, at a high probability of disease, high sensitivity (97.4%) is preferred (specificity 77.8%). When the clinical diagnosis of appendicitis is equivocal, CT interpretation should emphasize both sensitivity and specificity (sensitivity 92.3%, specificity 91.5%).

Conclusions: Radiologists can potentially decrease medical error and improve patient health by varying the interpretation of appendiceal CT on the basis of the clinical probability of appendicitis. This report is an example of how utility analysis can be used to guide radiologists in the interpretation of imaging studies and provide guidance on appropriate targets for the standardization of interpretation.

Key Words: ROC curves, medical error, appendicitis, CT scan, health utilities

J Am Coll Radiol 2006;3:115-121. Copyright © 2006 American College of Radiology

INTRODUCTION

The interpretation of diagnostic imaging studies remains a subjective art as well as a science. Agreement between imagers is seldom perfect, and differences in interpretation may lead to medical errors. One important cause of interpretive disagreement is the trade-off between sensitivity and specificity. Some imagers will choose to interpret studies with high sensitivity, identifying as abnormal any subtle findings that may identify the presence of disease. In contrast, others may choose to use a high specificity threshold, labeling as abnormal only those cases in which the findings are well defined. The trade-off between these 2 is familiar to all radiologists. However, there has been very little empirical work in defining which is optimal, high sensitivity, or high specificity.

In the past 2 decades, there has been a dramatic increase in the availability of imaging for patients with abdominal pain and in particular to exclude the presence of appendicitis [1]. Computed tomography (CT) in such patients (appendiceal CT) has been shown in multiple studies to be highly accurate for diagnosing appendicitis [2]. However, despite its high accuracy, data on the impact of CT on patient management and outcome are contradictory [3-5]. The trade-off between sensitivity and specificity contributes to variability in interpretation and may decrease the effectiveness of CT in improving outcomes.

In this analysis, we make explicit the trade-off between the utility of high-sensitivity interpretation and that of high-specificity interpretation. We demonstrate the appropriate sensitivity and specificity thresholds for the interpretation of CT scans in patients with possible appendicitis from various clinical scenarios. We also derive
posttest probabilities of appendicitis on the basis of the receiver operating characteristic (ROC) curve threshold and pretest probabilities from the different scenarios. Our intent is both to provide information to guide the interpretation and utilization of CT studies and to demonstrate how this approach may be used to potentially improve interpretation and prevent medical errors.

METHODS

This report is an extension of our previously published meta-analysis of CT and ultrasound for the detection of appendicitis in adults. In that meta-analysis, we identified 12 prospective studies of the accuracy of CT that met our methodologic inclusion criteria, involving 1,172 subjects, of whom 533 (45%) had appendicitis. The summary sensitivity and specificity of CT were 0.94 and 0.95, respectively. After exploring for heterogeneity and curve symmetry, we combined the results of these 12 studies using the fixed-effects diagnostic odds ratio model [2]. From the meta-analysis, the summary ROC curve for appendiceal CT is defined by the relationship

\[ \text{sensitivity} = \frac{1}{1 + \frac{1}{128.9071 \times [(1 - \text{specificity})/\text{specificity}]]} \]

We make several assumptions in this analysis. The first is that for the determination of the optimal ROC operating point, patient management will be based on the CT report. We assume therefore that positive CT results will lead to laparotomy and negative CT results will not lead to laparotomy, except after some substantial delay.

Modeling the ROC Threshold

The optimal ROC operating point is dependent on 3 factors—the prevalence of disease, the net disutility of false-negative diagnoses, and the net disutility of false-positive diagnoses—and is defined by the following expression:

\[ m = \frac{(1 - p)}{p} \times \frac{(U[FP]/U[FN])}{(1 - (U[FP]/U[FN]))} \]

where \( m \) is the slope of the ROC curve corresponding to the optimal sensitivity and specificity, \( p \) is the prevalence of disease, and \( U[FP] \) and \( U[FN] \) are the net disutility of false-positive and false-negative diagnoses compared to true-negative and true-positive diagnoses, respectively [6,7]. In this example, the net disutility is negative for both, reflecting the fact that adverse outcomes result from false diagnoses.

Implicit in this equation is the balance between the utility, or value to the patient, of the health states that arise from false-positive and false-negative test results and the frequency with which false-positive and false-negative diagnoses occur. If the adverse consequences of the false-positive test results outweigh the consequences of the false-negative test results, then interpretation at high specificity will be favored at the expense of sensitivity and vice versa.

Prevalence of Disease

For this analysis, we evaluate the spectrum of possible uses of imaging, including screening for a rare disease; circumstances of clinical uncertainty, when the disease is possible but not definite; and the situation of near clinical certainty, when a test might be performed only to rule out disease before some definitive therapy. In the model, we consider all possible pretest probabilities of disease but focus on these 3 clinical scenarios.

In clinical scenario 1, we consider a screening situation in which imaging is applied to all subjects with abdominal pain who present to emergency departments. Because most abdominal pain is due to some benign condition, the prevalence of appendicitis in this scenario is low. We assume the prevalence of appendicitis in this group to be 5%.

Clinical scenario 2 is the situation in which the diagnosis of appendicitis is clinically equivocal. These are subjects in whom surgery is not definitely indicated but is being considered. This scenario corresponds to the published CT literature, when CT is performed for suspected but not definite appendicitis. Therefore, to estimate the prevalence of appendicitis in this group, we use the prevalence of appendicitis in the studies included in our recent meta-analysis of this issue, 45% [2].

Clinical scenario 3 is when appendicitis is clinically certain, and subjects are to undergo surgery for the diagnosis of acute appendicitis. The prevalence of appendicitis in this group comes from published studies of the rate of therapeutic or necessary laparotomy for appendicitis. Because approximately 15% of subjects who undergo laparotomy for appendicitis have normal appendixes [3, 8-10], we estimate the prevalence of appendicitis in this group to be 85%.

Net Disutility of False-Positive and False-Negative Test Results

The net disutility of false-negative and false-positive test results represents the loss in value of the health states resulting from either of these test results. To determine the net disutility of false-positive and false-negative test results, we rely on the literature for mortality from laparotomy and from appendicitis. For simplicity, we limit the disutility of false diagnoses to mortality only. Data for morbidity are harder to quantify. Furthermore, because the optimal interpretation threshold is based on the ratio of the disutility of false-positive and false-negative diagnoses, the inclusion of morbidity would not change the overall results if in the same ratio as mortality.

There are several large administrative database series with mortality data for subjects who undergo laparotomy
for perforated and nonperforated appendicitis. The most recent of these is the report by Blomqvist et al [10]. This study used an administrative database in Sweden that included over 117,000 appendectomies for the years 1987 to 1996 and determined the mortality from appendectomy in subjects with nonperforated appendicitis (0.076%) and perforated appendicitis (0.51%). Other studies have reported a range of mortality rates for nonperforated appendicitis (0.05% to 0.24%) and perforated appendicitis (0.36% to 2.7%) [11-13]. Age distribution may explain some of the differences between these estimates, because mortality is much higher in older subjects.

Given sufficient time, nonperforated appendicitis will progress to perforation in a proportion of the cases. If perforation does not occur, then we assume that there is no excess mortality associated with a delay in the diagnosis of nonperforated appendicitis. However, in the cases that perforate, the mortality will be greater. To estimate the mortality rate from false-negative CT results, we consider the difference in mortality between perforated and nonperforated appendicitis, as well as the expected rate of progression to perforation after missed diagnoses for subjects at the age of 23 years, the median age in the report from Blomqvist et al [10]. We arbitrarily assign a delay in diagnosis of 24 hours to false-negative CT results. On the basis of these estimates and the regression model of Koepsell et al [14], 8.8% of nonperforating appendicitis will progress to perforation after false-negative CT results, resulting in a fatality rate of 0.028% attributable to progression to perforation after false-negative diagnosis (Table 1). In addition, there is some mortality associated with delay in diagnosis after perforation. For the base case, we assume that half of the mortality in the perforation group is attributable to this delay.

Table 1. Values defining the net disutility of false-negative and false-positive test results with ranges explored in sensitivity analyses

<table>
<thead>
<tr>
<th>Variable</th>
<th>Base (%)</th>
<th>Low Mortality From False-Negative Test Result (%)</th>
<th>High Mortality From False-Negative Test Result (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality from nonperforated appendicitis</td>
<td>0.076a</td>
<td>0.076</td>
<td>0.24b</td>
</tr>
<tr>
<td>Mortality from perforated appendicitis</td>
<td>0.51a</td>
<td>0.51</td>
<td>1.66b</td>
</tr>
<tr>
<td>Mortality from normal appendectomy</td>
<td>0.076c</td>
<td>0.076</td>
<td>0.14b</td>
</tr>
<tr>
<td>Rate of perforation from 24-hour delay</td>
<td>8.8d</td>
<td>8.8</td>
<td>30.0e</td>
</tr>
<tr>
<td>Mortality increase from developing perforation</td>
<td>0.028f</td>
<td>0.028</td>
<td>0.316g</td>
</tr>
<tr>
<td>Mortality increase from delay in subjects after perforation</td>
<td>0.0655h</td>
<td>0.0328i</td>
<td>0.214j</td>
</tr>
<tr>
<td>Percentage perforated at presentation</td>
<td>25.8k</td>
<td>25.8</td>
<td>25.8</td>
</tr>
<tr>
<td>Net disutility of false-positive diagnosis</td>
<td>-0.076</td>
<td>-0.076</td>
<td>-0.076</td>
</tr>
<tr>
<td>Net disutility of false-negative diagnosis</td>
<td>-0.094</td>
<td>-0.061</td>
<td>-0.53</td>
</tr>
<tr>
<td>Ratio of utility of false-positive to false-negative</td>
<td>0.81</td>
<td>1.2</td>
<td>0.14</td>
</tr>
</tbody>
</table>

aFrom Blomqvist et al [10].
bFrom Velanaovich and Satava [11].
cFrom Blomqvist et al’s [10] estimate of mortality from nonperforated appendicitis.
dFrom Koepsell et al [14] for age 23.
eFrom Koepsell et al [14] for age 50.
hDerived from Blomqvist et al’s [10] mortality estimates as equal to (mortality rate from perforated appendicitis * 0.5).
iDerived from average of Blomqvist et al’s [10] and Velanovich and Satava’s [11] mortality estimates as equal to (mortality rate from perforated appendicitis * 0.25).
jDerived from Velanovich and Satava’s [11] mortality estimates as equal to (mortality rate from perforated appendicitis * 0.5).
kFrom Flum et al [3].
erature summary [11] suggested that the mortality rate for normal appendectomy is lower than for nonperforated appendicitis. Regardless, no reliable estimates for the mortality attributable to the normal appendectomy are available. For this analysis, we assume that normal appendectomy will carry the same mortality rate as non-perforated appendicitis. A range of values for normal appendectomy are considered in the sensitivity analysis.

There are assumptions intrinsic to our analysis, so we performed sensitivity analyses to estimate the impact of these assumptions on the results (Table 1). For each of the 3 clinical scenarios, sensitivity analysis was performed on the ratio of net disutility of false-positive and false-negative diagnoses, including mortality increase from missed perforation and mortality decrease from spontaneous resolution. The extreme of the sensitivity analysis for severe consequences of false-negative diagnoses (and mild consequences of false-positive diagnoses) was derived by using higher estimates for progression to perforation (on the basis of a median subject age of 50 years) and the higher estimates of mortality from appendicitis from Velanovich and Satava [11]. The extreme of the sensitivity analysis for mild consequences of false-negative diagnoses (and severe consequences of false-positive diagnoses) was derived from the lower attribution of mortality from perforation to diagnostic delay. Because the equation for the optimal ROC operating point involves the ratio of the net disutility of false-positive to false-negative diagnoses, the sensitivity analysis for high mortality for false-negative test results applies to the opposite situation of lower mortality for false-positive diagnoses. Overall, the sensitivity analysis covered an almost 5-fold range in the ratio of the utilities of false-positive and false-negative studies.

**Posttest Probability**

Reverend Thomas Bayes, an 18th-century Lutheran minister, is credited with developing the idea that the interpretation of test results is not a deterministic process but rather is probabilistic, depending on the likelihood of disease before testing. According to Bayes’s theorem, the probability of disease after a test result is dependent on the probability of disease before the test result, as well as the sensitivity and specificity of the test [6]. The probability of disease after a positive test result can be defined from Bayes’s theorem as

\[ p_{\text{posttest}} = \frac{p_{\text{pretest}} \times \text{sensitivity}}{(p_{\text{pretest}} \times \text{sensitivity}) + (1 - p_{\text{pretest}}) \times (1 - \text{specificity})} \]

where \( p_{\text{posttest}} \) is the probability of disease after a positive test result, and \( p_{\text{pretest}} \) is the probability of disease before testing (adapted from Sox et al [6]).

The pretest probability of disease by definition varies in the 3 clinical scenarios. Also, because the optimal interpretation threshold changes with the pretest probability, the sensitivity and specificity used for Bayes’s theorem will change. We calculate the probability of appendicitis in subjects with positive and negative CT results for the range of possible prevalence of disease, using the prevalence specific optimal sensitivity and specificity.

**RESULTS**

The optimal operating point on the ROC curve varies with the prevalence of disease, as expected (Figure 1). In clinical scenario 1, the screening situation, interpretation at high specificity is preferred. At a prevalence of disease of 5%, the optimal interpretation of appendiceal CT would be at a sensitivity of 75% to achieve a specificity of 97.7%. In clinical scenario 2, an indeterminate probability of appendicitis, the optimal operating point occurred at a sensitivity of 92.3% and a specificity of 91.5%. Finally, in clinical scenario 3, a high clinical probability of disease, the optimal operating point was at 97.4% sensitivity and 77.8% specificity (Table 2).

On the basis of these results, radiologists should decrease the positivity threshold for diagnosing appendicitis on CT scans as the pretest probability increases. Thus, knowledge of clinical suspicion can aid radiologists in optimizing interpretation.

The ranges explored in the sensitivity analysis produced changes in the optimal operating point, reflecting the uncertainty in some of the variables (Table 3). For the screening and high clinical suspicion scenarios, the changes from the range encompassed in the sensitivity analysis was minimal. However, in the equivocal appendicitis group, there was some variation in the optimal operating point, with sensitivity emphasized under high mortality from false-negative test results and specificity...
more important under low mortality of false-negative test results.

Posttest probabilities of appendicitis after positive and negative test results are shown in Figure 2.

**DISCUSSION**

In this paper, we present an example of how health utility assessment can be used to guide the optimum interpretation of an imaging test. We use CT for possible appendicitis as a clinical demonstration of differing test interpretation optimization and posttest probabilities of disease depending on the pretest probability of disease. When appendicitis is unlikely, CT should be interpreted with high specificity, whereas when appendicitis is highly likely, CT should be interpreted with high sensitivity. Information about the specific clinical scenario can enable a radiologist to interpret an imaging study in a manner that optimizes patient utility.

Radiologists have the ability to alter the sensitivity and specificity of their interpretations. For objective positivity criteria, different thresholds can be chosen to consider test results positive. For example, with appendicitis, one important factor in CT interpretation is the size of the appendix. Using a lower size threshold (eg, 6 mm) to consider an appendix abnormal will result in higher sensitivity for appendicitis, at the expense of lower specificity. Using a higher size measurement to consider the appendix abnormal (eg, 8 mm) would result in lower sensitivity but higher specificity. For subjective criteria such as periappendiceal fat stranding, the process is the same but less explicit. Individual radiologists can alter how much increase in the density of the fat is necessary to be considered abnormal “stranding.” A recent paper by Daly et al [17] discussed the positive predictive values of various individual CT signs of appendicitis and may help guide radiologists in altering the sensitivity and specificity of their interpretations. Further research is necessary, however, to define the exact positivity criteria that are optimal in subjects at different pretest probabilities of appendicitis.

In 1999, the Institute of Medicine [18] released the report *To Err Is Human*, in which it was estimated that over 44,000 people in the United States die each year as a direct consequence of medical error. The standardization of practice, with the elimination of variability, has been embraced as a means of decreasing error [19]. However, to be effective, standardization must occur to the appropriate point. Error may actually increase if standardization occurs to a less optimal state. In this report, we present an approach to determining the appropriate target for this standardization. This study uses simplifying assumptions but does represent an approach to improving imaging interpretation on the basis of the best available evidence.

In general, when a disease is rare and there are substantial costs to a false-positive diagnosis, interpreting a test with high specificity will maximize patient benefit. In contrast, when a disease is common and there are substantial costs for a false-negative diagnosis, interpretation at high sensitivity will maximize utility. Our analysis reinforces this concept with information specific to appendiceal CT. However, the optimal operating point on the ROC curve may be different for other diseases. For example, in cancer screening, the health cost of false-negative diagnosis may be much greater than that for false-positive diag-

**Table 2.** Optimal interpretation of computed tomography for possible appendicitis in different clinical scenarios

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Prevalence of Appendicitis</th>
<th>Optimal Sensitivity</th>
<th>Optimal Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening</td>
<td>0.05</td>
<td>75.0</td>
<td>97.7</td>
</tr>
<tr>
<td>Equivocal appendicitis</td>
<td>0.45</td>
<td>92.3</td>
<td>91.5</td>
</tr>
<tr>
<td>Clinical appendicitis</td>
<td>0.85</td>
<td>97.4</td>
<td>77.8</td>
</tr>
</tbody>
</table>

**Table 3.** Sensitivity analysis on consequences of false-negative test results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Low Mortality From False-Negative Test Result</th>
<th>High Mortality From False-Negative Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Optimal Sensitivity</td>
<td>Optimal Specificity</td>
</tr>
<tr>
<td>Screening</td>
<td>73.4</td>
<td>97.9</td>
</tr>
<tr>
<td>Equivocal appendicitis</td>
<td>90.7</td>
<td>93.0</td>
</tr>
<tr>
<td>Clinical appendicitis</td>
<td>96.7</td>
<td>81.5</td>
</tr>
</tbody>
</table>
nosis, leading toward optimization with a more sensitive interpretation.

One aspect of the care for appendicitis that may differ from other diseases is the relatively high health cost of a false-positive diagnosis. Patients with false-positive diagnoses of appendicitis will undergo laparotomy. However, laparotomy with even a normal appendix is not a benign procedure, and particularly in vulnerable populations, including children and the elderly, it has a significant mortality risk [10,15]. A false-negative CT result, resulting in a delayed diagnosis of nonperforated appendicitis, carries an increase in mortality if perforation occurs [9-11]. However, although data on the rate of perforation are limited, it is likely that perforation after initial presentation is the uncommon exception rather than the general rule [9,14]. In addition, the spontaneous resolution of nonperforating appendicitis has not been well quantified but is increasingly accepted as important [20,21].

Our prior meta-analysis indicated that radiologists interpret CT scans with approximately equal sensitivity and specificity. The current analysis indicates that this is an appropriate threshold at the intermediate probability of disease at which CT is commonly used today. If CT is to be used in populations at higher or lower probabilities of disease, then different imaging thresholds will be appropriate. However, our current analysis does not investigate the circumstances under which it is appropriate to image. We can provide information only on the optimal interpretation.

This report represents a modeling analysis and as such has limitations. The analysis is based on assumptions for determining the shape of the ROC curve and on estimates for the net disutility of false-positive and false-negative diagnoses. Accordingly, there is some uncertainty inherent in the analysis. To model this, we included broad ranges for the variables in the sensitivity analysis. An additional aspect of this study is that we look at the utility of the various health states from the perspective of patients and society. From the hospital and physician standpoint, the avoidance of litigation may be a major concern affecting interpretation.

Furthermore, the analysis is based on a meta-analysis of the medical literature regarding test characteristics of appendiceal CT. As such, it is dependent on the quality of the underlying studies. Unfortunately, the appendiceal CT literature suffers from several flaws. Predominant among these is the use of different reference standards in determining the final diagnosis. Subjects with appendicitis were almost always diagnosed using pathology as a reference standard, whereas subjects without appendicitis were generally diagnosed by clinical follow-up. This differential verification bias likely led to some overestimation of both sensitivity and specificity [2]. Finally, until there are validated clinical prediction rules enabling the definition of subject pretest probability of appendicitis, it is difficult to apply this analysis to clinical practice.

Our prior meta-analysis indicated that radiologists interpret CT scans with approximately equal sensitivity and specificity. The current analysis indicates that this is an appropriate threshold at the intermediate probability of disease at which CT is commonly used today. If CT is to be used in populations at higher or lower probabilities of disease, then different imaging thresholds will be appropriate. However, our current analysis does not investigate the circumstances under which it is appropriate to image. We can provide information only on the optimal interpretation.

This report represents a modeling analysis and as such has limitations. The analysis is based on assumptions for determining the shape of the ROC curve and on estimates for the net disutility of false-positive and false-negative diagnoses. Accordingly, there is some uncertainty inherent in the analysis. To model this, we included broad ranges for the variables in the sensitivity analysis. An additional aspect of this study is that we look at the utility of the various health states from the perspective of patients and society. From the hospital and physician standpoint, the avoidance of litigation may be a major concern affecting interpretation.

Furthermore, the analysis is based on a meta-analysis of the medical literature regarding test characteristics of appendiceal CT. As such, it is dependent on the quality of the underlying studies. Unfortunately, the appendiceal CT literature suffers from several flaws. Predominant among these is the use of different reference standards in determining the final diagnosis. Subjects with appendicitis were almost always diagnosed using pathology as a reference standard, whereas subjects without appendicitis were generally diagnosed by clinical follow-up. This differential verification bias likely led to some overestimation of both sensitivity and specificity [2]. Finally, until there are validated clinical prediction rules enabling the definition of subject pretest probability of appendicitis, it is difficult to apply this analysis to clinical practice.

We do not consider morbidity in this report, only mortality. Morbid events affecting health utility can be expected to occur with higher frequency than mortality. However, our analysis is based on the ratio of mortality between false-positive and false-negative diagnoses. Morbid events occurring with the same ratio would not affect the results, regardless of frequency. An investigation of optimal interpretation for other conditions and tests might require a greater emphasis on morbidity.

Fryback and Thornbury [22] introduced a tiered efficacy model for understanding the usefulness of an imaging study, whereby the value of an imaging study increases as efficacy at each of the successive levels is demonstrated. Under their approach, the basic level of efficacy of imaging is the ability to produce an image-technical efficacy. This is followed by diagnostic accuracy, the mainstay of the radiology research literature. However, Fryback and Thornbury’s [22] hierarchy goes further, suggesting that to be useful, diagnostic tests must also affect diagnostic certainty, medical decision making, and even patient and societal outcomes. In this paper, we extend the diagnostic accuracy information from a meta-analysis of appendiceal CT. Using modeling and Bayes’s theorem, we estimate the potential effect of CT on diagnostic certainty and medical decision making. Concurrently, we provide guidance for interpretation of CT through optimization of the trade-off between sensitivity and specificity. Although analyses of this type are by nature dependent on assumptions, we believe that this approach can be used to improve patient care through a
decrease in practice variation and the optimization of both imaging utilization and interpretation.

In conclusion, modeling health utilities can provide radiologists with evidence to guide the interpretation of diagnostic tests. For CT in suspected appendicitis, interpretation at high specificity is preferred when the disease is considered unlikely, and interpretation at high sensitivity will maximize patient health in situations in which the disease is considered highly likely.

REFERENCES