Comparative Effectiveness Review
Number 157

Effective Health Care Program

Diagnosis of Right Lower Quadrant Pain and Suspected Acute Appendicitis

Executive Summary

Background

Abdominal pain is a common presenting symptom for patients seeking care at emergency departments, with approximately 3.4 million expected cases per year in the United States.[1] Appendicitis is a frequent cause of abdominal pain and occurs in approximately 8 to 10 percent of the population over a lifetime.[2,3] Appendicitis has its highest incidence between the ages of 10 and 30 years. The ratio of incidence in men and women is 3:2 through the mid-20s and then equalizes after age 30. Appendicitis is the most common abdominal surgical emergency, with over 250,000 appendectomies performed annually in the United States. The risk of acute appendicitis in pregnant women is not much lower than that of the general population, making appendicitis the most common nonobstetric emergency during pregnancy.[4-7] Untreated appendicitis can lead to perforation of the appendix, which typically occurs within 24 to 48 hours of the onset of symptoms.[8] Perforation of the appendix can cause intra-abdominal infection, sepsis, intraperitoneal abscesses, and rarely death.[4] In order to avoid the sequelae of perforated appendicitis, a low percentage of "negative" appendectomies (i.e., removing a normal noninflamed appendix in patients mistakenly diagnosed with appendicitis) is generally accepted from a surgical standpoint.

Effective Health Care Program

The Effective Health Care Program was initiated in 2005 to provide valid evidence about the comparative effectiveness of different medical interventions. The object is to help consumers, health care providers, and others in making informed choices among treatment alternatives. Through its Comparative Effectiveness Reviews, the program supports systematic appraisals of existing scientific evidence regarding treatments for high-priority health conditions. It also promotes and generates new scientific evidence by identifying gaps in existing scientific evidence and supporting new research. The program puts special emphasis on translating findings into a variety of useful formats for different stakeholders, including consumers.

The full report and this summary are available at **www.effectivehealthcare.ahrq.gov/reports/final.cfm**.

Clinical symptoms and signs suggestive of appendicitis include a history of central abdominal pain migrating to the right lower quadrant (RLQ), anorexia, fever, and nausea/vomiting. On examination, RLQ tenderness, along with "classical" signs of

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA
AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov
Effective Health Care

peritoneal irritation (e.g., rebound tenderness, guarding, rigidity, referred pain), may be present. Other signs (e.g., the psoas or obturator signs) may help the clinician localize the inflamed appendix.[9-11] However, many patients have a less typical presentation, necessitating the use of laboratory or imaging tests to establish a diagnosis. Laboratory evaluations potentially useful for the diagnosis of appendicitis include the white blood cell and granulocyte counts, the proportion of polymorphonuclear blood cells, and serum C-reactive protein.[10-12] Imaging tests, such as ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI), are used extensively for the diagnosis of appendicitis.[13-19] Imaging tests can be used alone or in combination. For example, US is sometimes used as a triage test to separate patients in whom sonography alone is adequate to establish a diagnosis from those who require further imaging.[20] Different factors may affect the performance of alternative tests and their impact on clinical outcomes. For example, US examination is considered to be highly operator dependent[21] and is technically challenging in obese patients or women in late pregnancy. CT scanning can be performed with or without the use of contrast agents, and contrast can be administered orally, rectally, intravenously, or via combinations of these routes.[20]

Clinical symptoms and signs, along with the results of laboratory or imaging tests, can be combined into multivariable diagnostic scores (sometimes referred to as "clinical prediction rules") that synthesize the findings of different investigations to determine the most likely diagnosis.[22] In adults, the most commonly used diagnostic score for appendicitis is the Alvarado score,[23] which is based on eight items: pain migration, anorexia, nausea, RLQ tenderness, rebound pain, elevated temperature, leukocytosis, and shift of white blood cell count to the left.[24] Although the Alvarado score is also used in pediatric populations,[25,26] the Pediatric Appendicitis Score has been specifically developed and validated for use in children.[27]

Diagnostic laparoscopy is also used for the evaluation of patients with RLQ pain and suspected acute appendicitis, primarily when a diagnosis cannot be established via other means. Although diagnostic laparoscopy is generally considered safe, studies have reported variable rates of morbidity and mortality from the procedure.[28]

In general these diagnostic tests are widely available in the United States. Clinical symptoms and signs can be evaluated relatively easily and inexpensively. Evidence from the National Hospital Ambulatory Medical Care Survey suggested that CT and complete blood counts are obtained in the majority of patients presenting to the emergency department with abdominal pain. The survey also showed that over time (between 1992 and 2006) the use of CT for both adults and children has increased. Over the same period, the use of the complete blood count increased in adults but decreased in children.[29,30] Various sources suggest that the use of US and MRI is increasing in populations in which exposure to ionizing radiation is of particular concern (e.g., children and pregnant women).[31-37]

As with all diagnostic tests, the modalities used in the diagnostic investigation of patients with RLQ pain affect clinical outcomes indirectly through their impact on clinicians' diagnostic thinking and decisionmaking.[38] More accurate and timely diagnosis of appendicitis can minimize the time to the indicated intervention (e.g., surgery), thus reducing the time patients are in pain and improving clinical outcomes (e.g., reducing the rate of perforated appendicitis and its attendant complications).[39] Conversely, time-consuming or unnecessary diagnostic workup (an important outcome, but hard to operationalize) may delay the indicated treatment and increase the risk of complications or result in false-positive results and more negative appendectomies. Furthermore, diagnostic testing can impact resource use for the management of patients with acute abdominal pain. For example, examination with CT may reduce length of stay by avoiding prolonged observation in cases in which a diagnosis cannot be established clinically or by eliminating the need for additional diagnostic testing.[18] In some cases, CT can also facilitate direct therapeutic intervention. For example, in patients with perforated appendicitis complicated by an abscess, the radiologist can not only detect but also treat the abscess by percutaneous drainage, thus avoiding the need for immediate operative intervention.

The diagnostic workup of acute appendicitis is complex because patients with acute abdominal pain of different etiologies can present with similar symptoms. Diagnosis is particularly challenging in children, women of reproductive age, pregnant women, and frail or elderly patients.[20,40,41] In young children (especially toddlers and preschool-age children), acute appendicitis is often diagnosed after perforation has occurred.[42-44] Children have a thinner appendiceal wall and less developed omentum, and thus may not readily wall off a perforation. In addition, many common childhood illnesses have symptoms similar to those of early acute appendicitis. Young children may also have difficulty communicating about their discomfort or describing their symptoms.[11] In addition, the use of modalities that involve ionizing radiation (e.g., CT) entails greater risks for children than for older patients.[20] A large proportion of women of reproductive age with appendicitis are misdiagnosed.[41,45] Establishing a diagnosis in this patient group can be particularly challenging because symptoms of acute appendicitis can mimic those of common gynecologic diseases (e.g., pelvic inflammatory disease, ectopic pregnancy). In pregnant women the diagnosis of suspected acute appendicitis can also be challenging because some symptoms of appendicitis (nausea and vomiting) are common in normal pregnancies and because enlargement of the uterus can alter the location of the appendix, which often moves higher and to the back.[46] Anatomic changes induced by pregnancy make the clinical examination of pregnant patients with abdominal pain more challenging and result in technical difficulties when using US.[37,47,48] Tests involving ionizing radiation (e.g., CT) are also generally avoided during pregnancy to prevent exposure of the fetus to radiation. Finally, obtaining a white blood cell count may not be helpful in the diagnosis of acute appendicitis because leukocytosis is common during pregnancy. The elderly typically present with appendicitis in a more advanced stage because they may delay seeking care, and definitive diagnosis is sometimes delayed further because competing etiologies for abdominal pain (e.g., malignancy or diverticulitis) are considered more likely.[49] Therefore, the performance of diagnostic tests may be modified by patient age, and elderly and frail individuals with appendicitis have a higher complication rate and a higher risk of mortality than younger and less frail patients.

Rationale for Evidence Review

Accurate testing of patients presenting with symptoms consistent with acute appendicitis to identify those who need treatment can improve clinical outcomes and reduce resource use. There is a lack of specific guidance for selecting diagnostic modalities, particularly in patient subgroups in whom the diagnosis is known to be particularly challenging (e.g., children, women of reproductive age, pregnant women, and the elderly). Existing systematic reviews typically assess a single diagnostic modality, focus almost exclusively on test performance outcomes rather than patient-relevant outcomes, and do not address factors that may modify test performance. No review to date has comprehensively examined all tests of interest or focused on comparisons between alternative strategies.

Key Questions

With input from clinical experts, we developed the following Key Questions to clarify the focus of the proposed systematic review.

Key Question 1: What is the performance of alternative diagnostic tests, alone or in combination, for patients with RLQ pain and suspected acute appendicitis?

a. What are the performance and comparative performance of alternative diagnostic tests in the following patient populations: children, adults, nonpregnant women of reproductive age, pregnant women, the elderly (age ≥65 years)?

b. What factors modify the test performance and comparative test performance of available diagnostic tests in these populations?

Key Question 2: What is the comparative effectiveness of alternative diagnostic tests, alone or in combination, for patients with RLQ pain and suspected acute appendicitis?

a. For the populations listed under Key Question 1a, what is the effect of alternative testing strategies on diagnostic thinking, therapeutic decisionmaking, clinical outcomes, and resource utilization?

b. What factors modify the comparative effectiveness of testing for patients with RLQ pain and suspected acute appendicitis?

Key Question 3: What are the harms of diagnostic tests per se, and what are the treatment-related harms of test-directed treatment for tests used to diagnose RLQ pain and suspected acute appendicitis?

Methods

We performed a systematic review of the published literature using established methods as outlined in the Agency for Healthcare Research and Quality (AHRQ) “Methods Guide for Effectiveness and Comparative Effectiveness Reviews” (Methods Guide).[50] We followed the reporting requirements of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[51] All key methodological decisions were made a priori. The protocol was developed with input from external clinical and methodological experts in consultation with the AHRQ Task Order Officer (TOO) and was posted online to solicit additional comments. The review’s PROSPERO registration number is CRD42013006480.

AHRQ TOO and External Stakeholder Input

A panel of Key Informants, including patients and other stakeholders, gave input on the Key Questions to be examined. These Key Questions were posted on AHRQ’s Effective Health Care Web site for public comment and revised in response to comments. A Technical Expert Panel, including representatives of professional societies and experts in the diagnosis and treatment of RLQ abdominal pain and appendicitis, provided input to help further refine the Key Questions and protocol, identify important issues, and define the parameters for the review of evidence. The AHRQ TOO was responsible for overseeing all aspects of this project. Discussions among the Evidence-based Practice Center, TOO, and Technical Expert Panel occurred during a series of teleconferences and via email.

Analytic Framework

We used an analytic framework (Figure A) that maps the Key Questions within the context of populations, interventions, comparators, and outcomes of interest.

Inclusion and Exclusion Criteria

Populations and Conditions of Interest

The population of interest for all Key Questions was patients with acute RLQ abdominal pain (≤7 days duration) for whom appendicitis was considered in the differential diagnosis. Separate analyses were performed for children (age <18 years), adults (age ≥18 years), women of reproductive age, pregnant women, and the elderly. We initially planned to separately examine the subgroup of very young children (<2 years and 2–5 years of age); however, information for these subgroups was poorly reported and we were unable to perform these subgroup analyses.

Interventions

For all Key Questions, the interventions of interest were diagnostic tests (alone or in combination) for diagnosing appendicitis, including clinical symptoms, clinical signs, laboratory tests, multivariable diagnostic scores, imaging tests, nuclear imaging studies, and diagnostic laparoscopy.

Comparators (Index and Reference Standard Tests)

For all Key Questions, the comparators were alternative tests or test combinations (listed previously) or clinical observation.

Outcomes

For Key Question 1, the outcome of interest was test performance, using pathology or clinical followup as the reference standard. For Key Question 2, we examined the impact of testing on diagnostic thinking, on therapeutic decisionmaking, and on patient-centered and resource use outcomes (negative appendectomy rate, bowel perforation, fistula formation, infectious complications, delay in diagnosis, length of hospital stay, fetal/maternal outcomes, and mortality). For Key Question 3, we considered adverse effects, including direct harms of testing and harms of test-directed treatment. When outcome definitions were not provided by the included studies, we adopted the terms used by the studies at face value.

Timing

Studies were considered regardless of duration of followup.

Setting

All health care settings were considered.

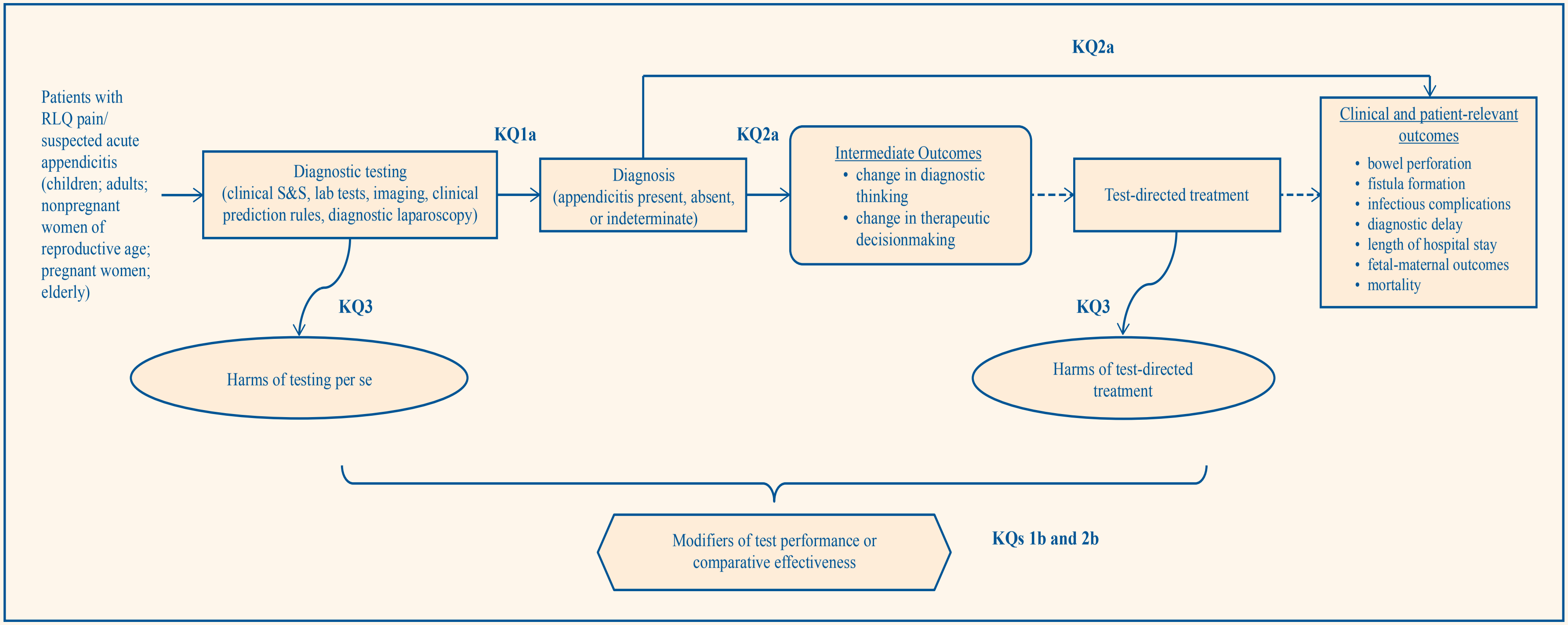
Study Design and Additional Criteria

For studies assessing test performance, we used previously completed systematic reviews to identify relevant studies and obtain specific data items. We updated these reviews to include more recent studies identified through literature searches. For index tests for which no relevant systematic review of test performance meeting our selection criteria could be identified, we performed a de novo systematic review. We accepted both randomized and nonrandomized comparative studies but analyzed them separately. We included only English-language studies because our preliminary searches indicated that non–English-language studies represented a small portion of the evidence base for any given test modality and were unlikely to change conclusions.

Literature Search and Abstract Screening

Appendix A in the full report describes our literature search strategies. Searches were conducted in PubMed®, Embase®, the Cochrane Central Register of Controlled Trials, and the Cumulative Index to Nursing and Allied Health Literature (CINAHL®) databases to identify primary research studies meeting our criteria (last search on August 6, 2014, for PubMed; August 12, 2014, for all other databases). We also used the PubMed search results to identify systematic reviews of the tests of interest (last search, July 31, 2013; search for systematic reviews not updated). All reviewers screened a common set of 200 abstracts, and discrepancies were discussed in order to standardize screening practices and ensure understanding of screening criteria. The remaining citations were split into nonoverlapping sets, each screened by two reviewers independently. Discrepancies were resolved by consensus involving a third investigator. We asked the Technical Expert Panel to provide citations of potentially relevant articles and identified additional studies through the

Figure A. Analytic framework


KQ2a
Patients with RLQ pain/ suspected acute appendicitis (children; adults; nonpregnant women of reproductive age; pregnant women; elderly)
KQ1a
Diagnostic testing (clinical S&S, lab tests, imaging, clinical prediction rules, diagnostic laparoscopy)
Diagnosis (appendicitis present, absent, or indeterminate)
KQ2a
Intermediate Outcomes
• change in diagnostic thinking
• change in therapeutic decisionmaking
Test-directed treatment
Clinical and patient-relevant outcomes
• bowel perforation
• fistula formation
• infectious complications
• diagnostic delay
• length of hospital stay
• fetal-maternal outcomes
• mortality
KQ3
Harms of testing per se
KQ3
Harms of test-directed treatment
Modifiers of test performance or comparative effectiveness
KQs 1b and 2b

KQ = Key Question; RLQ = right lower quadrant; S&S = symptoms and signs perusal of reference lists of eligible studies, clinical practice guidelines, relevant reviews, and conference proceedings. The Technical Expert Panel reviewed the final list of included studies to ensure that no key publications had been missed.

Study Selection and Data Abstraction

Potentially eligible citations were reviewed in full text for eligibility. A single reviewer examined each article; a second reviewer independently examined a subset of 350 articles. Disagreements were resolved by consensus involving a third reviewer. We included only English-language studies during full-text review because our preliminary searches indicated that non–English-language studies had small sample sizes and represented a small portion of the evidence base for any given test modality, so their exclusion is unlikely to have affected our conclusions. We excluded studies published exclusively in abstract form because they are typically not peer reviewed, they report only partial results, and their findings may change substantially when fully published. A detailed description of quality control measures is available in the protocol (www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1827). The lists of included and excluded studies (organized by reason for exclusion) are in Appendix B of the full report.

Previously published reviews were used as sources of eligible studies of test performance and as sources of data for objective data elements from these studies (bibliographic information, characteristics of included populations, and counts of individuals stratified by diagnostic test result and disease status). We verified all data from studies included in previously published systematic reviews against the full text of the corresponding publications. Because of the large number of studies, a single reviewer extracted data from each eligible noncomparative study of test performance; for nonrandomized comparative studies (NRCSs) and randomized controlled trials (RCTs), one reviewer extracted and a second reviewer verified the data. For RCTs, when possible, data were extracted according to the intention-to-treat principle. We verified the data extraction and risk-of-bias assessment in a random sample of 368 noncomparative test performance studies (1,487 separate estimates of test performance). Overall, agreement was excellent on items capturing information about the index and reference standard tests and numerical information on test performance. Agreement was less good for some risk-of-bias items; information on these items was reextracted for all included studies following a series of standardization exercises.

Assessment of the Risk of Bias of Individual Studies

We assessed the risk of bias for each study using the assessment methods detailed by the AHRQ Methods Guide.[50] We used items from the updated QUADAS (Quality Assessment of Diagnostic Accuracy Studies) 2 instrument to assess the risk of bias of the diagnostic test studies included in the review.[52-55] For studies of other designs, we used appropriate items to assess risk of bias: for NRCSs, we used items from the Newcastle-Ottawa scale;[56] for RCTs, we used items from the Cochrane Risk of Bias tool.[57] We rated each study as having low, intermediate, or high risk of bias on the basis of adherence to accepted methodological principles.

Evidence Synthesis

We summarized the included studies qualitatively and present important features of the study populations, designs, interventions, outcomes, and results in summary tables in the full report and its appendixes. All studies evaluating the test performance of the same single index test in a similar patient population were synthesized jointly, regardless of their source (our own literature searches or previously published reviews). Analyses were performed separately for the following patient populations: children, women of reproductive age, pregnant women, and the elderly. For each comparison of interest, we judged whether the eligible studies were sufficiently similar for meta-analysis on the basis of clinical heterogeneity of patient populations and testing strategies, as well as methodological heterogeneity of study designs and outcomes reported.

When five or more sufficiently similar studies evaluated the test performance of the same test in the same population, we used a bivariate-bivariate normal meta-analysis model to obtain summary sensitivity and specificity estimates.[58,59] We used the model estimates to calculate summary positive and negative likelihood ratios (LRs)[60] and to construct summary receiver operating characteristic (ROC) curves.[61,62] Meta-analyses were conducted using Bayesian methods with flat (minimally informative) priors.[63] We assessed heterogeneity by inspecting plots of study estimates in the ROC space and by examining the posterior distribution of the between-study heterogeneity parameters (for logit-sensitivity and logit-specificity). We explored heterogeneity using subgroup and meta-regression analyses. There were not enough studies comparing the same test strategies to allow meta-analysis for clinical outcomes and resource use.

In cases in which only a subset of the available studies could be quantitatively combined, we synthesized findings across all studies qualitatively by taking into account the magnitude and direction of effects and estimates of performance.

Grading the Strength of Evidence and Assessing Applicability

We followed the Methods Guide[50] to evaluate the strength of the body of evidence for each Key Question with respect to the following domains: risk of bias, consistency, directness, precision, and reporting bias.[50,64] Briefly, we assessed risk of bias (low, medium, or high) on the basis of the study design and the methodological quality of the studies. We rated the consistency of the data on the basis of the direction, magnitude, and statistical significance of all studies and made a determination. We assessed directness of the evidence on the basis of the use of surrogate outcomes or the need for indirect comparisons. We assessed the precision of the evidence on the basis of the degree of certainty surrounding each effect estimate. The potential for reporting bias was evaluated with respect to publication bias, selective outcome reporting bias, and selective analysis reporting bias. For all types of reporting bias, we made qualitative dispositions rather than performing formal statistical tests to evaluate differences in the effect sizes between more precise (larger) and less precise (smaller) studies. Instead of relying on statistical tests, we evaluated the reported results across studies qualitatively on the basis of completeness of reporting, number of enrolled patients, and numbers of observed events.[63,64] Judgment on the potential for selective outcome reporting bias was based on reporting patterns for each outcome of interest across studies. Finally, we rated the overall strength of the body of evidence using four levels: high, moderate, low, and insufficient.[47]

We followed the Methods Guide[50] to evaluate the applicability of included studies to patient populations of interest. We considered important population subgroups separately and evaluated the duration of symptoms before enrollment, outcomes reported, and setting of care.

Results

We reviewed the full text of 5,187 publications, of which 969 were considered eligible for inclusion in the review. Figure B presents the literature flow; our search strategies are presented in Appendix A; the lists of included and excluded studies (organized by reason for exclusion) are provided in Appendix B of the full report.

Key Question 1: What is the performance of alternative diagnostic tests, alone or in combination, for patients with RLQ pain and suspected acute appendicitis?

In total, 903 studies published between 1956 and 2014 met the inclusion criteria for Key Question 1. In this Executive Summary we present information on the tests that we thought were most clinically relevant on the basis of our reading of the literature, discussions with local clinical experts, and discussions with Key Informants and the Technical Expert Panel. The full report and appendixes present complete data on all tests we examined. Throughout, results for each test are presented separately for adults, children, women of reproductive age, pregnant women, and mixed populations (typically including male and female patients of all ages).

Studies of Test Performance

In general, studies of test performance were deemed to be at moderate to high risk of bias. Estimates of test performance often appeared to be affected by characteristics of study design that may be related to risk of bias, particularly partial and incomplete verification. In most cases, factors indicative of high risk of bias were associated with higher values of estimated test performance. These findings suggest that study conduct may have affected estimates of test performance in our meta-analyses. However, the assessment of the impact of risk of bias had to rely on information that was often poorly reported in the primary studies. Because each risk-of-bias item was examined individually and because different items may be correlated with each other and with other study characteristics that may affect test performance, we do not believe that definitive conclusions about specific items can be reached at this time.

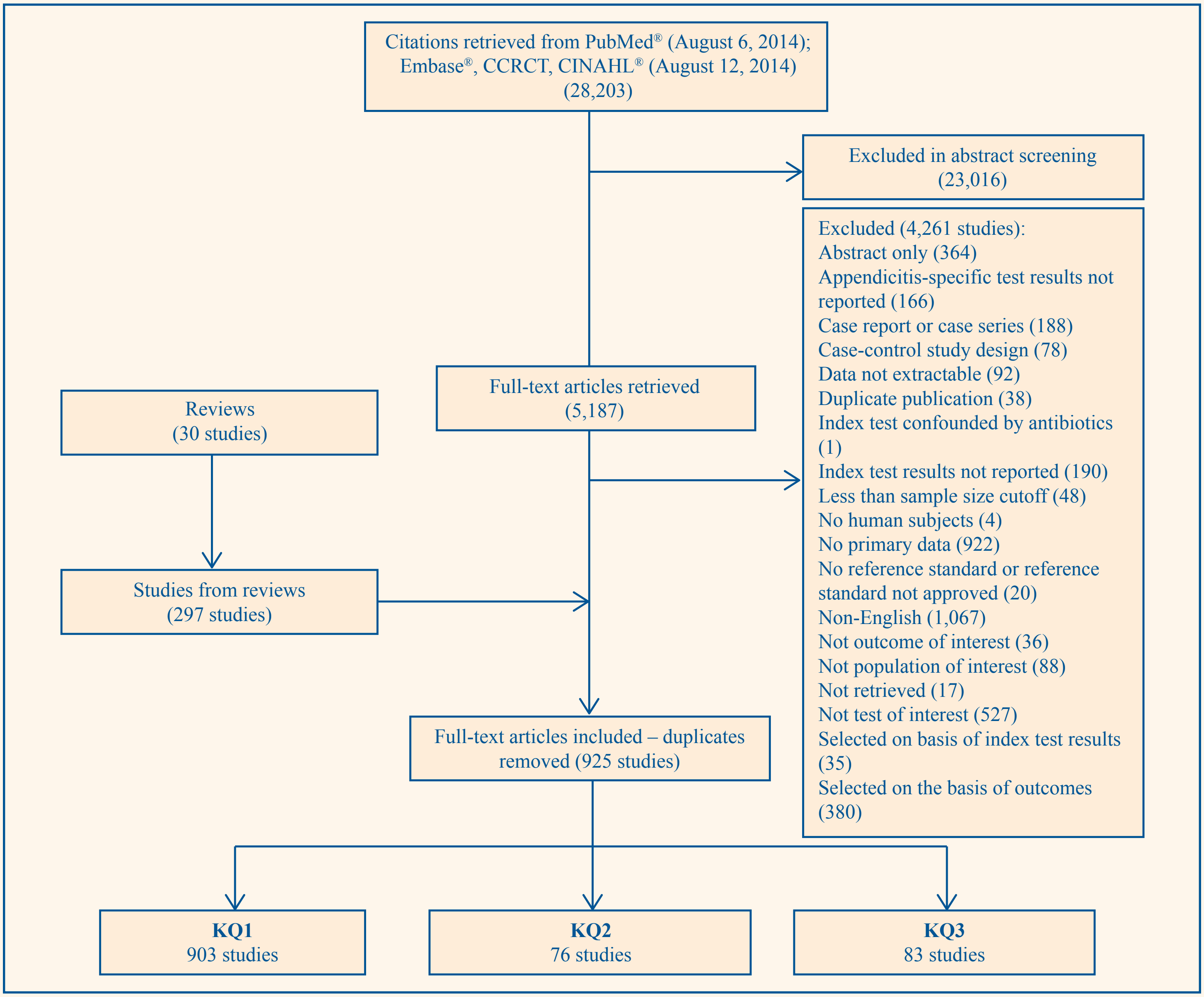
Test Performance of Clinical Symptoms and Signs (in Isolation)

Table A presents key test performance results for selected clinical symptoms and signs. Symptoms and signs had limited test performance when used in isolation. There was substantial heterogeneity in sensitivity and specificity for most clinical symptoms and signs.

Test Performance of Laboratory Tests (in Isolation)

Table B presents key test performance results for selected laboratory tests. The performance of individual laboratory tests was also rather limited, but it was better than that

Figure A. Flow chart of included studies


Citations retrieved from PubMed® (August 6, 2014); Embase®, CCRCT, CINAHL® (August 12, 2014) (28,203)
Excluded in abstract screening (23,016)
Excluded (4,261 studies):
Abstract only (364)
Appendicitis-specific test results not reported (166)
Case report or case series (188)
Case-control study design (78)
Data not extractable (92)
Duplicate publication (38)
Index test confounded by antibiotics (1)
Index test results not reported (190)
Less than sample size cutoff (48)
No human subjects (4)
No primary data (922)
No reference standard or reference standard not approved (20)
Non-English (1,067)
Not outcome of interest (36)
Not population of interest (88)
Not retrieved (17)
Not test of interest (527)
Selected on basis of index test results (35)
Selected on the basis of outcomes (380)
Full-text articles retrieved (5,187)
Reviews (30 studies)
Studies from reviews (297 studies)
Full-text articles included – duplicates removed (925 studies)
KQ1 903 studies
KQ2 76 studies
KQ3 83 studies

CCRCT = Cochrane Central Register of Controlled Trials; KQ = Key Question

of clinical symptoms and signs. There was substantial heterogeneity in sensitivity and specificity for most laboratory tests. Nevertheless, in most cases, summary ROC lines appeared to fit the data relatively well.

Test Performance of Multivariable Diagnostic Scores

Information on one or more multivariable diagnostic scores was reported in 127 studies. The authors usually proposed two types of cutpoints for these scores: a low value, below which patients might be safely discharged or observed (we refer to this cutpoint as the "low-risk cutoff"), and a high value, above which patients should be referred for treatment without additional investigation (we refer to this cutoff as the "high-risk cutoff"). The low- and high-risk cutoff values can be used to define three patient groups at different risk for appendicitis: low, intermediate, and high risk. If the diagnostic score has adequate classification performance and good calibration, the preferred test-and-treat strategy for each group will be different. When studies reported results at multiple cutpoints, we performed analyses at low-risk and high-risk cutpoints suggested by the original score developers or recommended in studies conducted after the ones examined. For scores developed specifically for binary classification, we used a single cutpoint. Test performance

Table A. Summary estimates of test performance of clinical symptoms and signs for the diagnosis of acute appendicitis

Sympotom or Sign	Population	N Studies (N Affected/ N Unaffected)	Sensitivity (95% CrI or Range*)	Specificity (95% CrI or Range*)
Fever	Adults	15 (2,082/1,796)	0.46 (0.29 to 0.64)	0.63 (0.47 to 0.77)
	Children	22 (3,952/3,845)	0.51 (0.41 to 0.61)	0.72 (0.66 to 0.77)
	Children <5 years	2 (196/77)	0.88 (0.83 to 0.93)	0.34 (0.29 to 0.39)
	Women of reproductive age	2 (37/36)	0.36 (0.20 to 0.53)	0.94 (0.89 to 1.00)
	Pregnant women	10 (309/166)	0.33 (0.14 to 0.59)	0.65 (0.37 to 0.86)
	Mixed	33 (8,766/5,386)	0.50 (0.39 to 0.61)	0.72 (0.62 to 0.80)
Guarding	Adults	5 (771/1,158)	0.67 (0.36 to 0.89)	0.69 (0.43 to 0.87)
	Children	8 (870/1,554)	0.64 (0.49 to 0.77)	0.69 (0.54 to 0.81)
	Women of reproductive age	1 (17/27)	0.76	0.85
	Pregnant women	4 (144/103)	0.63 (0.14 to 0.76)	0.55 (0.43 to 0.74)
	Mixed	18 (3,151/4,231)	0.63 (0.47 to 0.78)	0.69 (0.53 to 0.81)
Pain Migration	Adults	11 (1,831/864)	0.56 (0.45 to 0.67)	0.65 (0.50 to 0.78)
	Children	15 (2,049/3,535)	0.57 (0.39 to 0.73)	0.74 (0.66 to 0.81)
	Women of reproductive age	1 (17/27)	0.53	0.67
	Pregnant women	1 (42/14)	0.57	0.86
	Mixed	23 (4,475/6,156)	0.61 (0.49 to 0.71)	0.67 (0.56 to 0.76)
Tenderness	Children	2 (206/474)	0.63 (0.26 to 1.00)	0.57 (0.46 to 0.68)
	Children <5 years	1 (155/28)	0.98	0.25
	Women of reproductive age	1 (17/27)	1.00	0.04
	Mixed	10 (1,450/1,510)	0.99 (0.95 to 1.00)	0.30 (0.08 to 0.67)
Rebound Tenderness	Adults	11 (1,423/1,540)	0.67 (0.50 to 0.81)	0.70 (0.51 to 0.83)
	Children	11 (1,013/1,895)	0.60 (0.43 to 0.77)	0.73 (0.57 to 0.84)
	Children <5 years	1 (155/28)	0.85	0.86
	Women of reproductive age	1 (26/79)	0.42	0.65
	Pregnant women	5 (160/111)	0.71 (0.36 to 0.92)	0.58 (0.21 to 0.88)
	Mixed	30 (5,859/6,738)	0.74 (0.65 to 0.82)	0.60 (0.48 to 0.72)

*We report sensitivity and specificity values as medians and report central 95% credible intervals (95% CrI) when ≥5 studies were available. We report medians and minimum-to-maximum values when <5 studies were available. When a single study was available, we report the estimate from that study.

results for commonly used scores are presented in Tables C and D.

The majority of multivariable diagnostic scores were developed prior to the widespread use of diagnostic imaging with CT and US. More recently developed scores were designed with the intention of identifying a low-risk group in which imaging can be omitted. Furthermore, multivariable models were often developed and evaluated in the same patient sample. It is likely that the lack of separation between the training and testing datasets led to optimistic estimates of test performance. Lack of external validation also limited our ability to assess the generalizability of many diagnostic scores.

Test Performance of Imaging Tests

Table E presents key test performance results for selected imaging tests. Positive and negative LRs were generally higher for CT and MRI than for US, but all three tests had LRs that are clinically relevant (>5 and <0.2 for positive and negative LRs, respectively). US had substantially higher rates of nondiagnostic exams. The median percentage of nondiagnostic scans for CT was lower than 6% for all populations examined; the median proportion was substantially higher for US (ranging from 0% in women of reproductive age to 77.3% in pregnant women). However, the reporting of information on nondiagnostic scans was inconsistent across studies, raising concerns

Table B. Summary estimates of test performance of laboratory values for the diagnosis of acute appendicitis

Laboratory Value	Population	N Studies (N Affected/ N Unaffected)	Sensitivity (95% CrI or Range*)	Specificity (95% CrI or Range*)
CRP	Adults	15 (1,541/983)	0.84 (0.73 to 0.92)	0.67 (0.50 to 0.81)
	Children	22 (2,226/1,635)	0.73 (0.66 to 0.80)	0.72 (0.61 to 0.81)
	Elderly	2 (213/72)	0.91 (0.91 to 0.92)	0.21 (0.17 to 0.25)
	Women of reproductive age	3 (169/133)	0.79 (0.44 to 0.97)	0.70 (0.33 to 0.93)
	Pregnant women	1 (31/8)	0.68	0.50
	Mixed	52 (8,742/5,903)	0.79 (0.74 to 0.83)	0.65 (0.57 to 0.72)
WBC	Adults	26 (4,070/2,452)	0.81 (0.74 to 0.87)	0.54 (0.42 to 0.64)
	Children	41 (6,595/4,473)	0.80 (0.73 to 0.85)	0.65 (0.56 to 0.73)
	Elderly	3 (287/82)	0.71 (0.69 to 0.77)	0.50 (0.38 to 0.70)
	Women of reproductive age	2 (49/18)	0.64 (0.60 to 0.69)	0.67 (0.67 to 0.67)
	Pregnant women	6 (197/82)	0.63 (0.21 to 0.92)	0.75 (0.38 to 0.95)
	Mixed	84 (19,074/10,883)	0.78 (0.75 to 0.82)	0.62 (0.58 to 0.66)
WBC + CRP	Adults	2 (194/68)	0.93 (0.86 to 1.00)	0.62 (0.37 to 0.86)
	Children	5 (566/132)	0.81 (0.42 to 0.96)	0.73 (0.54 to 0.85)
	Elderly	1 (77/8)	0.96	0.13
	Women of reproductive age	1 (29/9)	0.93	0.44
	Mixed	15 (4,145/1,734)	0.72 (0.42 to 0.91)	0.73 (0.54 to 0.88)

CRP = C-reactive protein; WBC = white blood cell count

*We report sensitivity and specificity values as medians and report central 95% credible intervals (95% CrI) when ≥5 studies were available. We report medians and minimum-to-maximum values when <5 studies were available. When a single study was available, we report the estimate from that study.

Table C. Summary estimates of test performance of Alvarado diagnostic score test (low-risk cutoff) for the diagnosis of acute appendicitis

Test	Population	N Studies (N Affected/ N Unaffected)	Sensitivity (95% CrI or Range*)	Specificity (95% CrI or Range*)
Alvarado	Adults	3 (407/264)	0.91 (0.89 to 0.93)	0.31 (0.24 to 0.78)
	Children	6 (674/898)	0.99 (0.92 to 1.00)	0.48 (0.24 to 0.74)
	Mixed	20 (3,986/4,073)	0.96 (0.92 to 0.98)	0.46 (0.34 to 0.58)
	Women of reproductive age	2 (89/50)	0.99 (0.98 to 1.00)	0.24 (0.22 to 0.25)
	Children <5 years	1 (17/10)	1.00	0.20

*We report sensitivity and specificity values as medians and report central 95% credible intervals (95% CrI) when ≥5 studies were available. We report medians and minimum-to-maximum values when <5 studies were available. When a single study was available, we report the estimate from that study.

about reporting bias. The full report presents the results of sensitivity analyses for the test performance of imaging tests under different assumptions about nondiagnostic scans. Heterogeneity in sensitivity and specificity was moderate or high for most tests with adequate data for assessment, yet in most cases summary ROC lines appeared to fit the data relatively well. CT had high sensitivity (summary estimates ranging from 0.95 to 1) and specificity (0.91 to 0.99) in all populations of interest for this report. MRI had high sensitivity (0.91 to 1) but appeared to have variable specificity (0.86 to 1), mainly because of the smaller number of available studies, and the findings are most applicable to pregnant women. In adult populations, US had lower sensitivity (0.83) and specificity (0.89) than CT and MRI, and produced more nondiagnostic scans. In

Table D. Summary estimates of test performance of diagnostic score tests (high-risk cutoff) for the diagnosis of acute appendicitis

Test	Population	N Studies (N Affected/ N Unaffected)	Sensitivity (95% CrI or Range*)	Specificity (95% CrI or Range*)
Alvarado	Adults	16 (2,354/1,212)	0.75 (0.59 to 0.87)	0.75 (0.57 to 0.87)
	Children	9 (855/1,163)	0.85 (0.75 to 0.93)	0.84 (0.61 to 0.96)
	Mixed	30 (4,475/4,337)	0.77 (0.69 to 0.84)	0.79 (0.74 to 0.84)
	Women of reproductive age	5 (202/177)	0.70 (0.35 to 0.92)	0.91 (0.65 to 0.99)
	Children <5 years	1 (17/10)	0.76	0.60
Alvarado Modified	Adults	4 (254/126)	0.68 (0.54 to 0.89)	0.60 (0.14 to 0.89)
	Children	5 (109/110)	0.89 (0.71 to 0.98)	0.80 (0.37 to 0.97)
	Elderly	1 (7/10)	0.86	0.80
	Mixed	6 (412 /139)	0.82 (0.63 to 0.93)	0.62 (0.24 to 0.89)
	Women of reproductive age	4 (186/69)	0.60 (0.17 to 0.91)	0.50 (0.17 to 1.00)
PAS	Children	1 [108/18)	0.95	0.11

PAS = Pediatric Appendicitis Score

*We report sensitivity and specificity values as medians and report central 95% credible intervals when ≥5 studies were available. We report medians and minimum-to-maximum values when <5 studies were available. When a single study was available, we report the estimate from that study.

Table E. Summary estimates of test performance of diagnostic imaging for acute appendicitis

Test	Population	N Studies (N Affected/ N Unaffected)	Sensitivity (95% CrI or Range*)	Specificity (95% CrI or Range*)
CT	Adults	72 (7,833/14,469)	0.96 (0.95 to 0.97)	0.96 (0.93 to 0.97)
	Children	34 (3,581/3,122)	0.96 (0.94 to 0.98)	0.92 (0.85 to 0.96)
	Elderly	4 (144/582)	1.00 (0.94 to 1.00)	1.00 (0.43 to 1.00)
	Women of reproductive age	11 (596/652)	0.99 (0.96 to 1.00)	0.91 (0.75 to 0.97)
	Pregnant women	5 (26/84)	0.99 (0.96 to 1.00)	0.91 (0.75 to 0.97)
	Mixed	93 (9,341/10,357)	0.96 (0.95 to 0.97)	0.94 (0.91 to 0.95)
MRI	Adults	7 (512/467)	0.95 (0.88 to 0.98)	0.92 (0.87 to 0.95)
	Children	7 (359/665)	0.97 (0.87 to 1.00)	0.96 (0.84 to 0.99)
	Women of reproductive age	1 (50/88)	1.00	0.86
	Pregnant women	11 (76/570)	0.98 (0.92 to 1.00)	0.98 (0.96 to 1.00)
	Mixed	5 (243/141)	0.94 (0.83 to 0.99)	1.00 (0.97 to 1.00)
US	Adults	38 (3,560/3,656)	0.85 (0.79 to 0.90)	0.90 (0.83 to 0.95)
	Children	85 (8,539/15,167)	0.89 (0.86 to 0.92)	0.91 (0.89 to 0.94)
	Women of reproductive age	11 (516/539)	0.72 (0.51 to 0.88)	0.92 (0.75 to 0.98)
	Pregnant women	13 (188/198)	0.72 (0.45 to 0.92)	0.95 (0.84 to 0.99)
	Mixed	125 (11,902/14,314)	0.86 (0.83 to 0.89)	0.90 (0.87 to 0.92)

CT = computed tomography; MRI = magnetic resonance imaging; US = ultrasound

*We report sensitivity and specificity values as medians and report central 95% credible intervals when ≥5 studies were available. We report medians and minimum-to-maximum values when <5 studies were available. When a single study was available, we report the estimate from that study.

children, the specificity of US was similar to that of CT (0.92 vs. 0.91), but CT had greater sensitivity (0.89 vs. 0.96); these results were based on a large number of studies (72 for US and 32 for CT). In the same patient population, MRI had a specificity of 0.99 and sensitivity of 1, but data were derived from only three studies and are therefore less reliable than those for other imaging tests. Among pregnant women, CT (5 studies), MRI (10 studies), and US (10 studies) had similar specificity (0.98, 0.98, and 0.95, respectively), but CT and MRI had higher sensitivity than US (0.95, 0.98, and 0.73, respectively).

Test Performance of Diagnostic Laparoscopy

Fifty-five studies published between 1974 and 2014 reported information on the test performance of diagnostic laparoscopy. The reporting of methods and outcomes in these studies was less complete than that of studies of other tests. When possible to discern such information from the reported data, patients undergoing diagnostic laparoscopy often presented atypically and had already been examined with a number of other diagnostic modalities. In addition, studies of laparoscopy did not fully report information on the final diagnosis of patients for whom the procedure did not reveal an inflamed appendix. Studies often did not report operational definitions for the absence of any pathology and had heterogeneous management policies for such cases. These features of the studies can influence the estimates of test performance; for this reason, we did not perform any quantitative synthesis for the test performance of diagnostic laparoscopy. It is important to note that patients included in studies of diagnostic laparoscopy are likely to be different from patients included in studies of noninvasive tests, even if the selection criteria are not clearly presented. They may, for example, have more severe symptoms or have atypical findings on other tests. Thus, indirect comparisons of diagnostic laparoscopy with noninvasive tests are not meaningful.

Sensitivity and Specificity

For the 54 studies for which they could be calculated, the median sensitivity and specificity were 100 and 89 percent, respectively. However, there was a wide range, with sensitivity ranging from 37 to 100 percent (25th percentile, 95%; 75th percentile, 100%) and specificity ranging from 0 to 100 percent (25th percentile, 73%; 75th percentile, 100%). This variability likely reflects the heterogeneous populations evaluated in these studies. In the 16 studies that reported on women of reproductive age, the median sensitivity was 100 percent (25th percentile, 100%; 75th percentile, 100%), and the median specificity was 89 percent (25th percentile, 79%; 75th percentile, 100%).

Tests Positive for Other Pathology

Forty-one studies reported some information on other pathology diagnosed at laparoscopy. The median proportion of patients identified with nonappendiceal pathology was 22 percent (25th percentile, 11.5%; 75th percentile, 34%). Only six small studies reported that other pathology was found when appendicitis was also present. The median was 5 percent (25th percentile, 2%; 75th percentile, 13%). In studies of women of reproductive age, the median proportion of patients identified with nonappendiceal pathology was 23 percent (25th percentile, 18%; 75th percentile, 26%); no nonappendiceal pathologies were found in patients who had appendicitis.

Other: Information on other test performance outcomes of diagnostic laparoscopy—for example, the proportion of cases in which the appendix could not be visualized and the proportion of cases in which no cause of pain was identified (i.e., nonproductive abdominal explorations)—is presented in the full report.

Modifiers of Test Performance

The vast majority of studies did not report adequate data to assess factors that may affect test performance; for this reason we relied on comparisons across studies via meta-regression analyses to identify such factors. Overall, no distinct pattern emerged to establish a particular factor as a modifier of test performance. For all clinically relevant factors examined, credible intervals were wide, indicating substantial uncertainty regarding the relative performance of tests over levels of the modifiers. Details on the impact of patient- and test-related characteristics on the test performance of various tests in specific subpopulations are presented in the full report.

Comparative Assessments of Test Performance

Our assessment of comparative test performance relied on randomized and nonrandomized direct (i.e., within-study) comparisons of tests. Overall, on the basis of items from the Cochrane risk–of-bias tool, RCTs were deemed to be at moderate risk of bias. NRCSs were at high risk of bias because they either did not make any attempt to address differences among groups receiving different test strategies or failed to consider at least some important factors (e.g., age, sex, or duration and severity of symptoms).

Randomized Comparisons of Alternative Tests

Although 36 RCTs reported information on comparative test performance, each possible comparison was examined by only one or two small trials, and these trials did not report information on the same outcomes. Therefore, it was not possible to draw strong conclusions about the comparative performance of different tests.

Nonrandomized Comparisons of Alternative Tests

Nonrandomized Comparisons of Diagnostic Scores. Eight studies reported direct comparisons among alternative diagnostic scores for appendicitis. Three studies included only children, one included women of reproductive age, and four included mixed populations. Across all eight studies, differences in test performance between scores were small; this was particularly true for the comparison of the Alvarado score and Pediatric Appendicitis Scores applied to children with suspected acute appendicitis. The one exception was a study that compared a multivariable diagnostic score based on clinical symptoms and signs versus a score combining the same clinical variables with the addition of US: incorporation of imaging information improved test performance substantially with respect to both sensitivity and specificity.

Nonrandomized Comparisons of CT and US. Fifty-three studies reported results in cohorts using both CT and US as index tests, potentially permitting direct nonrandomized comparisons of these modalities. Ten studies investigated CT as a replacement for US, 13 investigated US as a triage test for CT, and 30 studies were unclear about the actual role of testing that was being evaluated (often using convenience samples of patients selected using criteria that were poorly reported). Nine of the studies had a paired design and 44 had a parallel-group design. In general, CT had better test performance than US when used as a replacement test or when the role of testing being evaluated was unclear. In the triage context, CT had high test performance (diagnostic odds ratios higher than 10 and often higher than 100) in patient populations selected on the basis of US results (typically, patients with nondiagnostic US findings or negative US findings in the presence of symptoms suggestive of appendicitis).

Nonrandomized Comparisons of MRI and US. Eight studies reported results in cohorts using both MRI and US as index tests. Four studies investigated MRI as a replacement for US, one investigated US as a triage test for MRI, and three studies were unclear about the actual role of testing that was being evaluated (tending to use convenience samples of patients selected for a specific test using criteria that were poorly reported). Four of the studies had a paired design and four had a parallel-group design. MRI, when used as a replacement test for US, had greater test performance; however, the available studies are few and, when combined, produce rather imprecise results.

Key Question 2: What is the comparative effectiveness of alternative diagnostic tests, alone or in combination, for patients with RLQ pain and suspected acute appendicitis?

Of 925 included studies, 54 reported information on comparative effectiveness outcomes related to diagnostic tests (36 RCTs and 18 NRCSs). Many of the included RCTs were small and may have produced unstable estimates of event rates and treatment effects. Furthermore, selection criteria differed substantially among trials, rendering cross-study comparisons uninformative.

Key Question 3: What are the harms of diagnostic tests per se, and what are the treatment-related harms of test-directed treatment for tests used to diagnose RLQ pain and suspected acute appendicitis?

Of 925 included studies, only 83 mentioned harms related to diagnostic tests: 17 RCTs, 13 NRCSs, and 53 diagnostic cohort studies. Eight studies (3 RCTs and 5 diagnostic cohort studies) reported an absence of adverse events for all tests except diagnostic laparoscopy. The fact that so few studies reported harms raises concerns about selective outcome reporting.

Contrast-Related Adverse Events

Eight studies (3 RCTs and 5 diagnostic cohort studies) reported on adverse events related to contrast administration. Of these, three reported that the contrast was well tolerated. The others reported a combination of nonfatal adverse events.

Exposure to Ionizing Radiation

No studies reported direct evidence on the effect of ionizing radiation on patient-relevant outcomes. Twelve studies (3 RCTs, 4 NRCSs, and 5 diagnostic cohort studies) reported radiation doses for CT, and three of these discussed strategies to reduce CT-related radiation exposure in a population, but they did not link this information with clinical outcomes.

Maternal/Fetal Adverse Events

Six studies (3 studies of US, 3 of MRI, and 2 of multiple clinical and lab tests, some studies evaluating more than 1 test) reported information on maternal outcomes. One study of MRI reported that 17 patients without appendicitis progressed to uneventful labor and delivery. A second study of MRI reported that not using oral contrast sped up the imaging process. The remaining four studies reported that there was no maternal mortality.

Seven studies (5 studies of US, 2 of MRI, 1 of CT, and 1 of clinical symptoms and signs, some studies evaluating more than 1 test) reported information on fetal outcomes. One study of US reported that 18 of 22 patients had a normal-term delivery; there were two spontaneous abortions (in patients with no clinical or sonographic evidence of acute appendicitis) and two elective abortions. The second study examined US and clinical and laboratory tests, and found that all 20 women delivered healthy infants. The third study gave fetal outcomes for only 2 of the 45 participants who underwent US for the diagnosis of appendicitis. One was a spontaneous abortion in a woman with surgically confirmed acute appendicitis without perforation, and the other was a premature delivery in a patient with no evidence of appendicitis at followup through delivery. The fourth study reported a total of nine adverse fetal outcomes (5/31 who had MRI and 4/44 in the US or clinical group); none were in the perioperative period. The fifth study reported outcomes for US, MRI, and CT. In the US group, one patient, who had an open appendectomy in the first trimester, developed severe preeclampsia and had a premature delivery at 33 weeks. (This patient also had a diagnostic CT.) There was one fetal death after a negative open appendectomy, but neither the fetal death nor the early delivery was related directly to the appendectomy, and one patient with perforated appendicitis had abruptio placentae and vaginal hemorrhage. Only one patient had MRI, and she delivered a healthy baby at term. Of 13 patients who had a diagnostic CT, 9 delivered healthy infants; 1, who had an open appendectomy in the first trimester, developed severe preeclampsia and had a premature delivery at 33 weeks (previously mentioned); and 3 were lost to followup. The sixth study reported fetal outcomes for 55 of 80 patients who had CT. Fifty-one had a live infant at or near term, one had a premature delivery of a live 30-week infant 3 days after CT-diagnosed gastric cancer, two had spontaneous vaginal delivery of a nonviable fetus (1 at 18 weeks with sepsis after normal CT and normal laparotomy, and 1 at 22 weeks with chorioamnionitis, 5 days after normal CT). There was one fetal death at 26 weeks (4 weeks after a CT examination with normal findings). The seventh study reported that in a group evaluated using symptoms and signs, there were seven therapeutic abortions and two perioperative spontaneous abortions (first trimester), and four women without appendicitis had severe perinatal morbidity or mortality.

Surgical Complications in Studies of Diagnostic Laparoscopy

Thirty-four studies of diagnostic laparoscopy mentioned surgery-related harms. Eight RCTs (469 patients) and 8 NRCSs (4,084 patients) described complications related to laparoscopy compared with open appendectomy; 25 diagnostic cohort studies (5,553 patients) reported on complications of diagnostic laparoscopy. In general, the rates of specific complications were low (generally less than 10% and in most cases less than 2%). Few studies attributed specific adverse events to diagnostic laparoscopy (as opposed to additional surgical intervention). Nine studies, including five RCTs, reported that there were no complications related to the diagnostic laparoscopic procedure.

Discussion

Key Findings and Strength-of-Evidence Assessment

The literature on the test performance of various clinical symptoms and signs, laboratory and imaging tests, and diagnostic scores is vast but consists almost exclusively of studies assessing the test performance of individual tests. Information on test performance of multiple tests applied jointly and conditional test performance (i.e., test performance among patients already examined with other tests) was limited. The few studies that provided information on more than one index test were typically not designed with the goal of providing comparative information, and cross-study comparisons cannot provide reliable evidence on relative performance. Studies meeting our selection criteria provided limited information on the test performance or comparative effectiveness of diagnostic pathways (i.e., well-defined sequences of diagnostic and treatment steps). We assessed the strength of evidence for key outcomes selected on the basis of our reading of the literature and discussions with Key Informants and Technical Experts. Our assessment integrates subjective judgments on risk of bias, consistency of findings, directness of the available information, and precision of estimates.

Test Performance

Clinical symptoms and signs used in isolation, including classical signs of peritoneal irritation, fever, and various assessments of abdominal pain, appeared to have limited test performance for all the populations of interest to this report. Among laboratory tests, white blood cell count, C-reactive protein, and tests derived from combinations of measurements on the complete blood count and differential had test performance that was generally higher than that of clinical symptoms and signs (especially with respect to sensitivity using a low-risk threshold) but still rather limited (e.g., in terms of summary LRs). These observations were relatively stable across the patient populations examined. Because studies did not allow an examination of the performance of multiple tests applied jointly and because conditional test performance was not reported uniformly across studies, the clinical implications of the relatively limited test performance of many nonimaging tests is not clear. Furthermore, symptoms and signs are variable in a patient (over the course of disease) and among patients, and it is hard to assess their clinical usefulness based on test performance. Importantly, the clinical examination forms the basis of the investigation of acute abdominal pain and suspected acute appendicitis and, even if poorly reported, all studies of imaging tests use some form of clinical examination (e.g., for patient selection). Multivariable diagnostic scores appeared to have test performance that was superior to the individual clinical signs, symptoms, or laboratory tests they included but still rather limited (e.g., in terms of summary LRs). Of note, the majority of studies assessed scores that had been developed before the widespread availability of CT and US imaging, suggesting that their results may be less applicable to current clinical practice.

Among imaging tests, CT and MRI had high sensitivity and specificity, resulting in clinically relevant summary LRs. CT has been investigated in a large number of diagnostic cohort studies, leading to precise estimates of test performance in all populations of interest for this report. In contrast, MRI has been investigated in a relatively small number of studies, mainly focused on pregnant women; therefore, the results may not be applicable to other populations. US has been investigated in a large number of studies and results were somewhat heterogeneous, suggesting that the average estimate of test performance may not apply to all populations for which US is considered. Possible explanations for this heterogeneity are the operator dependence of the test performance of US and the fact that studies were conducted in different settings. Despite the heterogeneity, the data suggest that US had lower overall test performance than CT and MRI, and resulted in a substantially greater proportion of nondiagnostic examinations. Diagnostic laparoscopy appeared to have good test performance, but studies were poorly reported and differed in their policies regarding removal of the appendix when no pathology was macroscopically visible, which may bias test performance results. Furthermore, patients included in studies of diagnostic laparoscopy are likely to be very different from patients included in studies of noninvasive tests. Therefore, our results for the test performance of laparoscopy should not be compared with the other diagnostic tests reviewed in this report. Table F summarizes our findings regarding the strength of evidence for the diagnostic performance of selected tests. When interpreting these results, readers should remember that test performance is not directly related to clinical outcomes, and high sensitivity and specificity do not necessarily imply better patient-relevant outcomes.

Comparisons among tests with respect to test performance relied on a small number of RCTs with moderate risk of bias, a relatively small number of direct comparisons among index tests in diagnostic cohort studies that were not designed to obtain comparative information, and indirect comparisons across single index test studies enrolling diverse populations in heterogeneous clinical settings. There was moderate-strength evidence that CT has superior overall test performance compared with US and produces fewer nondiagnostic results. Similarly, MRI appeared to have better test performance than US, but the strength of evidence was deemed low. The strength of evidence on comparisons among other imaging tests and among multivariable diagnostic scores was deemed insufficient. The evidence regarding the effect of patient- and test-related characteristics on test performance was also deemed insufficient. There were indications that aspects of study design characteristics affect test performance, but the effects are often unpredictable in direction and do not have direct clinical relevance.

Patient-Relevant Outcomes

We based our assessment of the comparative effectiveness of alternative tests on randomized studies (with the exception of outcomes among pregnant women), because indirect (across studies) comparisons of outcomes other than test performance are susceptible to bias resulting from differences among the populations included. We found a few RCTs with moderate risk of bias that provided information on the comparative effectiveness of alternative testing strategies. These studies assessed various comparisons across different modalities (or different versions of the same modality) and therefore did not provide definitive evidence for any of the possible pairwise contrasts they evaluated.

Adverse Events of Testing

Information on harms was often incomplete and poorly reported. Only a minority of the included studies provided

Table F. Assessment of the strength of evidence for test performance of individual tests			
Test or Score	**Strength of Evidence**	**Test Sensitivity in Key Subgroups — Subgroup (N Studies): Sensitivity (95% CrI)**	**Test Specificity in Key Subgroups — Subgroup (N Studies): Specificity (95% CrI)**
WBC count	Moderate	Adults (26): 0.81 (0.74 to 0.87) Children (41): 0.80 (0.73 to 0.85) Elderly (3): 0.71 (0.69 to 0.77) Women of reproductive age (2): 0.64 (0.60 to 0.69) Pregnant women (6): 0.63 (0.21 to 0.92)	Adults (26): 0.54 (0.42 to 0.64) Children (41): 0.65 (0.56 to 0.73) Elderly (3): 0.50 (0.38 to 0.70) Women of reproductive age (2): 0.67 (0.67 to 0.67) Pregnant women (6): 0.75 (0.38 to 0.95)
CRP	Low	Adults (15): 0.84 (0.73 to 0.92) Children (22): 0.73 (0.66 to 0.80) Elderly (2): 0.91 (0.91 to 0.92) Women of reproductive age (3): 0.79 (0.44 to 0.97) Pregnant women (1): 0.68	Adults (15): 0.67 (0.50 to 0.81) Children (22): 0.72 (0.61 to 0.81) Elderly (2): 0.91 (0.91 to 0.92) Women of reproductive age (3): 0.79 (0.44 to 0.97) Pregnant women (1): 0.68
Measures based on the CBC and differential	Low	Please see the Results section for the test performance of various test combinations	—
Alvarado score (low-risk cutoff)	Moderate	Adults (3): 0.91 (0.89 to 0.93) Children (6): 0.99 (0.92 to 1.00) Women of reproductive age (2): 0.99 (0.98 to 1.00)	Adults (3): 0.31 (0.24 to 0.78) Children (6): 0.48 (0.24 to 0.74) Women of reproductive age (2): 0.24 (0.22 to 0.25)
Alvarado score (high-risk cutoff)	Moderate	Adults (16): 0.80 (0.60 to 0.93) Children (9): 0.83 (0.73 to 0.91) Women of reproductive age (5): 0.70 (0.35 to 0.92)	Adults (16): 0.71 (0.50 to 0.85) Children (9): 0.81 (0.63 to 0.92) Women of reproductive age (5): 0.91 (0.65 to 0.99)
PAS	Low	Children (5): 0.03 (0.00 to 0.13)	Children (5): 1.00 (0.99 to 1.00)
CT	Moderate–high	Adults (72): 0.96 (0.95 to 0.97) Children (34): 0.96 (0.94 to 0.98) Elderly (4): 1.00 (0.94 to 1.00) Women of reproductive age (11): 0.99 (0.96 to 1.00) Pregnant women (5): 0.99 (0.96 to 1.00)	Adults (72): 0.96 (0.93 to 0.97) Children (34): 0.92 (0.85 to 0.96) Elderly (4): 1.00 (0.94 to 1.00) Women of reproductive age (11): 0.91 (0.75 to 0.97) Pregnant women (5): 0.91 (0.75 to 0.97)
MRI	Low	Adults (7): 0.95 (0.88 to 0.98) Children (7): 0.97 (0.87 to 1.00) Women of reproductive age (1): 1.00 Pregnant women (11): 0.98 (0.92 to 1.00)	Adults (7): 0.92 (0.87 to 0.95) Children (7): 0.96 (0.84 to 0.99) Women of reproductive age (1): 0.86 Pregnant women (11): 0.98 (0.96 to 1.00)
US	Moderate	Adults (38): 0.85 (0.79 to 0.90) Children (85): 0.89 (0.86 to 0.92) Women of reproductive age (11): 0.72 (0.51 to 0.88) Pregnant women (13): 0.72 (0.45 to 0.92)	Adults (38): 0.90 (0.83 to 0.95) Children (85): 0.91 (0.89 to 0.94) Women of reproductive age (11): 0.92 (0.75 to 0.98) Pregnant women (13): 0.95 (0.84 to 0.99)
Laparoscopy	Moderate	Please see the Results section in the main report for a full description of results related to diagnostic laparoscopy	—

CBC = complete blood count; CrI = credible interval; CRP = C-reactive protein; CT = computed tomography; MRI = magnetic resonance imaging; PAS = Pediatric Appendicitis Score; US = ultrasound; WBC = white blood cell count information on test-related harms, raising concerns about selective outcome and analysis reporting. The majority of the studies providing information on adverse events did not report the definitions or ascertainment methods they used. Importantly, no information was available from studies meeting our selection criteria regarding the effects of ionizing radiation. This is particularly important, as there is substantial variation in the levels of radiation delivered with newer multiphase CT scans performed for evaluation of appendicitis. Information was particularly limited on fetal and maternal outcomes of various diagnostic modalities applied during pregnancy for the investigation of acute appendicitis. Overall, we rated the strength of evidence on the harms of tests for acute appendicitis to be insufficient, primarily because of concerns about outcome reporting bias and the sparseness of available evidence.

Limitations of the Evidence Base

The evidence base regarding the diagnosis of acute appendicitis is limited in the following ways:

- Studies reporting information on test performance outcomes were at moderate to high risk of bias. Differential verification (the use of different reference-standard tests depending on the results of the index test) and partial verification (the failure to apply the reference standard to all of the included patients) were common, particularly in studies that were not surgical series (generally, studies with a lower prevalence of appendicitis). Studies with complete and nondifferential verification tended to be surgical cohorts reporting exclusively on patients undergoing appendectomy and so are not representative of all patients presenting with acute RLQ pain. In addition, poor reporting of information on study design hampered our risk-of-bias assessment.
- Studies provided limited information to assess the impact of various factors related to patients, technical implementation, operators, or systems on the performance of the tests of interest.
- Information on the comparative effectiveness of alternative testing strategies (e.g., sequential use of tests as part of a diagnostic algorithm) with respect to test performance, patient-relevant outcomes, and resource use was limited. Direct (within study) comparisons of test performance and the impact of testing strategies on clinical outcomes were scarce. Studies have not compared diagnostic algorithms (e.g., combinations of tests applied in sequence, such that the results of earlier tests determine the choice of subsequent tests). When two or more index tests were evaluated in the same study, the role of testing that was being examined (add-on, replacement, triage) was often unclear.
- In studies of diagnostic scores, multivariable models were often developed and evaluated in the same patient sample. The lack of separation between the training and testing datasets (or any attempt at internal validation of the model) generally leads to optimistic (too high) estimates of test performance. The lack of external validation (replication) also limited our ability to assess the generalizability of many diagnostic scores.
- Few RCTs compared alternative test strategies with respect to patient-relevant outcomes. The few trials reporting patient-relevant outcomes were fragmented across heterogeneous comparisons of alternative testing strategies. The trials often used suboptimal methods for randomized sequence generation, allocation concealment, and blinding, or they provided information that was too limited to assess these aspects of study design. Many had sample sizes that were too small to reliably detect small or moderate differences between the strategies being compared.
- In contrast to the RCTs, NRCSs of alternative testing strategies attained large sample sizes but often reported unadjusted analyses (or analyses adjusted for only a small number of potential confounders) that do not allow strong conclusions about the comparative effectiveness of alternative test strategies to be drawn.

Strengths and Limitations of This Review

Previous reviews on this topic have focused on special patient populations, have almost exclusively focused on test performance outcomes, have not assessed harms systematically, or have focused on a very limited spectrum of study designs. Our work provides a comprehensive up-to-date summary of the evidence on the diagnosis of RLQ pain and suspected acute appendicitis. For many of the examined tests and patient populations, this review is the first to be conducted. For some important modalities that have been investigated to some extent in previous meta-analyses (e.g., CT, MRI, US, and multivariable diagnostic scores), our work includes a much larger number of studies (and a greater total number of patients) than previous reviews. This allows us to provide accurate estimates of test performance in different patient populations that can be used to inform clinical decisions (especially if used as inputs in decision and simulation models) and to identify evidence gaps to inform the planning of future research.

Nonetheless, several limitations, which to a large extent reflect the limitations of the underlying evidence base, must be considered when interpreting our results.

- The evidence base has a number of limitations, detailed in the preceding section: quality was often poor, patient-relevant outcomes and harms were incompletely and inconsistently reported, and information on study- or population-level characteristics that could modify test performance and patient-relevant outcomes was also incomplete.
- We assumed that pathological diagnosis and clinical followup have negligible error (i.e., that they represent a "gold" standard). It is unlikely that this assumption is exactly true. Consequently, it is likely that estimates of test performance are biased, and the direction of this bias is hard to predict, particularly at the meta-analysis level. However, we believe that the error rate of these reference standards is low enough that its influence on our estimates is relatively small.
- Finally, we did not address contextual factors (e.g., availability of equipment, trained readers) that are important determinants of the adoption of specific diagnostic strategies in particular settings.

Applicability of Review Findings

In general, the existing evidence on alternative diagnostic tests for the diagnosis of acute RLQ pain and suspected acute appendicitis appears to be applicable to clinical practice in the United States. The included studies enrolled patients representative of the age and sex distribution of patients seeking care for RLQ abdominal pain in the United States, and evidence on test performance was available for all commonly used modalities. Information on adults and children was often separately reported, allowing the assessment of test performance in these patient subgroups. However, information was more limited for patients at the extremes of age (i.e., children younger than 5 years or the elderly), pregnant women, and women of reproductive age; in some cases, decisions for these will have to rely on extrapolation of results from population subgroups with more available information, and thus applicability assessments are not possible. Approximately one-third of the studies in this review were conducted in the United States, and the vast majority were carried out either in the United States or in industrialized European or Asian countries. Care settings varied across studies, including academic and nonacademic centers, and patient populations included those sampled at emergency departments, in surgical cohorts, or from mixed populations.

Assessing the applicability of studies on clinical symptoms and signs was challenging: the pathophysiologic rationale for many of these tests is well established, but many of the relevant studies were conducted before the widespread availability of imaging modalities, and thus their findings may reflect test performance in a population with more advanced disease or populations selected for a high probability of appendicitis (e.g., surgical cohorts). Studies of laboratory and imaging tests evaluated "stable" technologies (e.g., white blood cell count) or were conducted in recent years; for example, many studies of C-reactive protein, CT, and US were conducted from 2005 onward. In meta-regression analyses comparing test performance in the last decade against earlier years, there was no evidence that the performance of laboratory or imaging tests has changed significantly over time; however, the indirect nature of metaregression comparisons and the low precision of metaregression estimates limit the strength of these results. In contrast, the applicability of the evidence on most multivariable diagnostic scores may be somewhat limited because most were developed before the era of widespread availability of imaging. The lack of external validation for most diagnostic scores also limits the applicability of these results. The findings of studies on diagnostic laparoscopy may also be less applicable because many of the studies were conducted before the widespread availability of diagnostic imaging.

Future Research Needs

Studies of Diagnostic Test Performance

- Cohort studies of test performance would provide useful information, particularly for diagnostic tests that have not been studied adequately (e.g., MRI in all relevant patient populations) and to compare the performance of tests for which comparative information is limited (e.g., direct comparisons of CT vs. US; comparisons between CT with contrast administered via alternative routes).
- Such diagnostic cohort studies (and comparative studies in particular) are also needed to evaluate the test performance of combinations of tests and testing strategies by estimating conditional test performance and by developing and validating multivariable diagnostic tools internally and in independent datasets. For example, they could examine the use of US as a triage test for CT or MRI, or the use of multivariable diagnostic scores to select patients who can be monitored without immediate imaging or treatment (e.g., low-risk patients who can be managed with

wait-and-see strategies), those who need imaging, and those who need the initiation of treatment without imaging. They can also provide information to determine how patient- and test-related factors affect performance (i.e., to examine whether test performance depends on easily identifiable patient characteristics).

- Research is needed on the natural history of acute appendicitis, specifically on whether (and how often) cases of appendicitis resolve on their own and the rate of recurrence among such cases. Studies of natural history (e.g., among patients deemed to be appropriate candidates for medical management or wait-and-see strategies) are necessary for evaluating the impact of tests in decision and simulation modeling studies (discussed later) and also to inform the design of studies of alternative test-and-treatment strategies, including studies of the sequencing of multiple tests and the timing of examinations. Of note, the test performance of diagnostic tests may vary during different timepoints in the development of acute appendicitis; for instance, laboratory tests may be highly sensitive for cases associated with more severe inflammation.
- Paired test study designs, in which all index tests are applied to all enrolled patients (so that each patient has results from every test of interest), are generally more efficient than parallel-arm designs and should be considered when planning future studies.[65]
- Cohort studies assessing the performance of tests that have been evaluated extensively (e.g., CT and US) are most needed for specific patient populations (e.g., pregnant women, young children, and the elderly); for other tests (e.g., MRI) further research is needed in all patient populations. Comparative studies are needed for all tests and all populations. Ideally, future studies of test performance will be large (powered to achieve adequate precision), prospectively designed, multicenter investigations enrolling patients representative of those seen in clinical practice. Studies should prespecify the criteria for a positive test, use standardized diagnostic criteria for the diagnosis of appendicitis, use followup for an adequate period of time (1–2 weeks) for patients who do not undergo surgery, and have as complete followup as possible. Studies that evaluate two or more index tests should provide a detailed description of the role of testing they are evaluating (triage, add-on, replacement) and report data in enough detail to allow statistical analyses appropriate for that evaluation.[66]
- Multivariable diagnostic scores provide an appealing way to combine information from multiple clinical symptoms and signs, laboratory tests, and possibly US. Multivariable scores may be particularly useful in identifying patients who are at low risk for appendicitis and who may be candidates for wait-and-see strategies or less aggressive imaging strategies. Cohort studies for the development and validation of such scores should use state-of-the science methods for model development and internal and external validation.
- Future research needs to be better reported and studies should adhere to established reporting guidelines (e.g., STAndards for the Reporting of Diagnostic accuracy studies; www.stard-statement.org/).

Studies of Patient-Relevant Outcomes and Resource Use

- Cohort studies of diagnostic test strategies can also be used to study the impact of tests on patient-relevant and resource use outcomes. For tests with well-understood performance characteristics, such studies may use randomized designs. In many cases, however, randomized comparisons of alternative test strategies are unlikely to be fruitful because existing studies indicate that many of the competing tests have sensitivities and specificities that are fairly similar and close to 1.[67] Under these conditions, RCTs comparing alternative test strategies would need to enroll very large numbers of participants to allow reliable comparisons. If randomized studies are deemed necessary, consideration should be given to paired randomized designs because they are more efficient than parallel-arm trials.
- Large-scale observational prospective studies could be used to evaluate the effectiveness of alternative test strategies with respect to short- and long-term patient-relevant outcomes and to explore factors that may modify the effect of tests on these outcomes. Such studies would need to collect detailed information on baseline factors that may be associated with the choice of test strategy and the outcomes of interest in order to attempt to address confounding bias. Comparisons across methods should be performed only among patients who would be candidates for assessment with all methods being compared.
- Decision and simulation modeling can be used to determine whether randomized or nonrandomized cohort studies assessing patient-relevant outcomes and resource use are necessary and to guide their design.

Models can also be used to synthesize evidence on test performance, impact of tests on clinical decisions, treatment effectiveness, resource use (and, when relevant, economic costs), and patient preferences to guide clinical decisionmaking. We think that the results of the current review provide a solid basis for conducting such modeling studies.

Studies of Test-Related Adverse Events

- Future studies should report complete information on test-related adverse events, using prespecified criteria and careful ascertainment methods.
- Mathematical modeling studies can be used to combine data on the effective radiation dose received during alternative CT-based approaches with external information on long-term radiation effects.[67]

Conclusions

The literature on the test performance of clinical symptoms and signs, laboratory and imaging tests, and multivariable diagnostic scores for the diagnosis of acute appendicitis is large, but it consists almost exclusively of studies at moderate risk of bias, primarily because of differential and incomplete verification. The few studies that assess multiple tests are typically not designed with the goal of providing comparative information. Thus, the available evidence supports fairly strong conclusions about the performance of individual tests, but it is largely insufficient to support conclusions about comparative effectiveness, especially with respect to clinical outcomes. Clinical symptoms and signs and laboratory tests have relatively limited test performance when used in isolation. Their combination in multivariable scores is promising, but the best studied scores were developed before the widespread use of imaging modalitie, and more recently developed scores have not yet been studied adequately. All three major imaging modalities have adequate test performance. Evidence on CT is mature for most patient populations of interest. In contrast, MRI has been investigated in fewer studies, many of which focus on its use for pregnant women. US produces nondiagnostic scans more often than CT or MRI, and when a diagnosis is possible, its performance appears to be somewhat worse than CT and MRI. Beyond test performance, information on patient-relevant outcomes and resource use is very limited. Information on test-related harms (e.g., adverse events due to radiation) is provided by only a minority of studies and is poorly reported. More research, much of which could be accomplished through nonrandomized studies, is needed to establish the performance in understudied patient populations (very young children, women of reproductive age, the elderly) and modalities (e.g., MRI, multivariable scores); compare competing tests; identify factors that affect performance; and evaluate the impact of testing strategies on patient-relevant outcomes, resource use, and harms. Perhaps most importantly, given the large volume of accumulated evidence on the performance of various tests, decision and simulation modeling (e.g., decision analysis, simulation modeling of the impact of radiation on long-term outcomes) should be used to guide decisionmaking and to inform the design of future studies.

References

1. Paulson EK, Kalady MF, Pappas TN. Clinical practice. Suspected appendicitis. N Engl J Med. 2003 Jan 16;348(3):236-42. PMID: 12529465.
2. Addiss DG, Shaffer N, Fowler BS, et al. The epidemiology of appendicitis and appendectomy in the United States. Am J Epidemiol. 1990 Nov;132(5):910-25. PMID: 2239906.
3. Ferri FF. Ferri's Clinical Advisor: Instant Diagnosis and Treatment, 2002. W.B. Saunders Company; 2003.
4. Parks NA, Schroeppel TJ. Update on imaging for acute appendicitis. Surg Clin North Am. 2011 Feb;91(1):141-54. PMID: 21184905.
5. Wolfe J, Henneman P. Acute appendicitis. In: Marx J, Hockberger R, Walls R, eds. Rosen's Emergency Medicine. Philadelphia: Mosby/Elsevier; 2010:1193-9.
6. Andersson RE, Lambe M. Incidence of appendicitis during pregnancy. Int J Epidemiol. 2001 Dec;30(6):1281-5. PMID: 11821329.
7. Zingone F, Sultan AA, Humes DJ, et al. Risk of acute appendicitis in and around pregnancy: a population-based cohort study from England. Ann Surg. 2015 Feb;261(2):332-7. PMID: 24950289.
8. Bickell NA, Aufses AH Jr, Rojas M, et al. How time affects the risk of rupture in appendicitis. J Am Coll Surg. 2006 Mar;202(3):401-6. PMID: 16500243.
9. Humes DJ, Simpson J. Clinical presentation of acute appendicitis: clinical signs - laboratory findings - clinical scores, Alvarado score and derivative scores. In: Keyzer C, Gevenois PA, eds. Imaging of Acute Appendicitis in Adults and Children. Springer-Verlag; 2011.
10. Andersson RE. Meta-analysis of the clinical and laboratory diagnosis of appendicitis. Br J Surg. 2004 Jan;91(1):28-37. PMID: 14716790.
11. Bundy DG, Byerley JS, Liles EA, et al. Does this child have appendicitis? JAMA. 2007 Jul 25;298(4):438-51. PMID: 17652298.
12. Hallan S, Asberg A. The accuracy of C-reactive protein in diagnosing acute appendicitis--a meta-analysis. Scand J Clin Lab Invest. 1997 Aug;57(5):373-80. PMID: 9279962.

13. Al-Khayal KA, Al-Omran MA. Computed tomography and ultrasonography in the diagnosis of equivocal acute appendicitis. A meta-analysis. Saudi Med J. 2007 Feb;28(2):173-80. PMID: 17268692.

14. Barger RL Jr, Nandalur KR. Diagnostic performance of magnetic resonance imaging in the detection of appendicitis in adults: a meta-analysis. Acad Radiol. 2010 Oct;17(10):1211-6. PMID: 20634107.

15. Blumenfeld YJ, Wong AE, Jafari A, et al. MR imaging in cases of antenatal suspected appendicitis--a meta-analysis. J Matern Fetal Neonatal Med. 2011 Mar;24(3):485-8. PMID: 20695758.

16. Doria AS, Moineddin R, Kellenberger CJ, et al. US or CT for diagnosis of appendicitis in children and adults? A meta-analysis. Radiology. 2006 Oct;241(1):83-94. PMID: 16928974.

17. Krajewski S, Brown J, Phang PT, et al. Impact of computed tomography of the abdomen on clinical outcomes in patients with acute right lower quadrant pain: a meta-analysis. Can J Surg. 2011 Feb;54(1):43-53. PMID: 21251432.

18. Terasawa T, Blackmore CC, Bent S, et al. Systematic review: computed tomography and ultrasonography to detect acute appendicitis in adults and adolescents. Ann Intern Med. 2004 Oct 5;141(7):537-46. PMID: 15466771.

19. Greenhalgh R, Punwani S, Taylor SA. Is MRI routinely indicated in pregnant patients with suspected appendicitis after equivocal ultrasound examination? Abdom Imaging. 2008 Jan-Feb;33(1):21-5. PMID: 17874265.

20. Howell JM, Eddy OL, Lukens TW, et al. Clinical policy: critical issues in the evaluation and management of emergency department patients with suspected appendicitis. Ann Emerg Med. 2010 Jan;55(1):71-116. PMID: 20116016.

21. Carroll PJ, Gibson D, El-Faedy O, et al. Surgeon-performed ultrasound at the bedside for the detection of appendicitis and gallstones: systematic review and meta-analysis. Am J Surg. 2013 Jan;205(1):102-8. PMID: 22748292.

22. Liu JL, Wyatt JC, Deeks JJ, et al. Systematic reviews of clinical decision tools for acute abdominal pain. Health Technol Assess. 2006 Nov;10(47):1-167, iii-iv. PMID: 17083855.

23. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med. 1986 May;15(5):557-64. PMID: 3963537.

24. Ohle R, O'Reilly F, O'Brien KK, et al. The Alvarado score for predicting acute appendicitis: a systematic review. BMC Med. 2011;9:139. PMID: 22204638.

25. Kulik DM, Uleryk EM, Maguire JL. Does this child have appendicitis? A systematic review of clinical prediction rules for children with acute abdominal pain. J Clin Epidemiol. 2013 Jan;66(1):95-104. PMID: 23177898.

26. Escriba A, Gamell AM, Fernandez Y, et al. Prospective validation of two systems of classification for the diagnosis of acute appendicitis. Pediatr Emerg Care. 2011 Mar;27(3):165-9. PMID: 21346681.

27. Samuel M. Pediatric appendicitis score. J Pediatr Surg. 2002 Jun;37(6):877-81. PMID: 12037754.

28. Korndorffer JR Jr, Fellinger E, Reed W. SAGES guideline for laparoscopic appendectomy. Surg Endosc. 2010 Apr;24(4):757-61. PMID: 19787402.

29. Fahimi J, Herring A, Harries A, et al. Computed tomography use among children presenting to emergency departments with abdominal pain. Pediatrics. 2012 Nov;130(5):e1069-75. PMID: 23045569.

30. Tsze DS, Asnis LM, Merchant RC, et al. Increasing computed tomography use for patients with appendicitis and discrepancies in pain management between adults and children: an analysis of the NHAMCS. Ann Emerg Med. 2012 May;59(5):395-403. PMID: 21802777.

31. Drake FT, Florence MG, Johnson MG, et al. Progress in the diagnosis of appendicitis: a report from Washington State's Surgical Care and Outcomes Assessment Program. Ann Surg. 2012 Oct;256(4):586-94. PMID: 22964731.

32. Bachur RG, Hennelly K, Callahan MJ, et al. Advanced radiologic imaging for pediatric appendicitis, 2005-2009: trends and outcomes. J Pediatr. 2012 Jun;160(6):1034-8. PMID: 22192815.

33. Mittal MK, Dayan PS, Macias CG, et al. Performance of ultrasound in the diagnosis of appendicitis in children in a multicenter cohort. Acad Emerg Med. 2013 Jul;20(7):697-702. PMID: 23859583.

34. Hryhorczuk AL, Mannix RC, Taylor GA. Pediatric abdominal pain: use of imaging in the emergency department in the United States from 1999 to 2007. Radiology. 2012 Jun;263(3):778-85. PMID: 22535565.

35. Dingemann J, Ure B. Imaging and the use of scores for the diagnosis of appendicitis in children. Eur J Pediatr Surg. 2012 Jun;22(3):195-200. PMID: 22767172.

36. Jaffe TA, Miller CM, Merkle EM. Practice patterns in imaging of the pregnant patient with abdominal pain: a survey of academic centers. AJR Am J Roentgenol. 2007 Nov;189(5):1128-34. PMID: 17954650.

37. Wallace GW, Davis MA, Semelka RC, et al. Imaging the pregnant patient with abdominal pain. Abdom Imag. 2012 Oct;37(5):849-60. PMID: 22160283.

38. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991 Apr-Jun;11(2):88-94. PMID: 1907710.

39. SCOAP Collaborative, Cuschieri J, Florence M, et al. Negative appendectomy and imaging accuracy in the Washington State Surgical Care and Outcomes Assessment Program. Ann Surg. 2008 Oct;248(4):557-63. PMID: 18936568.

40. Clinical policy: critical issues for the initial evaluation and management of patients presenting with a chief complaint of nontraumatic acute abdominal pain. Ann Emerg Med. 2000 Oct;36(4):406-15. PMID: 11020699.

41. Flum DR, Morris A, Koepsell T, et al. Has misdiagnosis of appendicitis decreased over time? A population-based analysis. JAMA. 2001 Oct 10;286(14):1748-53. PMID: 11594900.

42. Ponsky TA, Huang ZJ, Kittle K, et al. Hospital- and patient-level characteristics and the risk of appendiceal rupture and negative appendectomy in children. JAMA. 2004 Oct 27;292(16):1977-82. PMID: 15507583.

43. Nance ML, Adamson WT, Hedrick HL. Appendicitis in the young child: a continuing diagnostic challenge. Pediatr Emerg Care. 2000 Jun;16(3):160-2. PMID: 10888451.

44. Newman K, Ponsky T, Kittle K, et al. Appendicitis 2000: variability in practice, outcomes, and resource utilization at thirty pediatric hospitals. J Pediatr Surg. 2003 Mar;38(3):372-9; discussion -9. PMID: 12632352.

45. Rothrock SG, Green SM, Dobson M, et al. Misdiagnosis of appendicitis in nonpregnant women of childbearing age. J Emerg Med. 1995 Jan-Feb;13(1):1-8. PMID: 7782616.

46. Brown JJ, Wilson C, Coleman S, et al. Appendicitis in pregnancy: an ongoing diagnostic dilemma. Colorectal Dis. 2009 Feb;11(2):116-22. PMID: 18513191.

47. Katz DS, Klein MA, Ganson G, et al. Imaging of abdominal pain in pregnancy. Radiol Clin North Am. 2012 Jan;50(1):149-71. PMID: 22099493.

48. Long SS, Long C, Lai H, et al. Imaging strategies for right lower quadrant pain in pregnancy. AJR Am J Roentgenol. 2011 Jan;196(1):4-12. PMID: 21178041.

49. Kraemer M, Franke C, Ohmann C, et al. Acute appendicitis in late adulthood: incidence, presentation, and outcome. Results of a prospective multicenter acute abdominal pain study and a review of the literature. Langenbecks Arch Surv. 2000 Nov;385(7):470-81. PMID: 11131250.

50. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Chapters available at www.effectivehealthcare.ahrq.gov.

51. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. Int J Surg. 2010;8(5):336-41. PMID: 20171303.

52. Whiting P, Westwood M, Beynon R, et al. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. J Clin Epidemiol. 2011 Jun;64(6):602-7. PMID: 21075596.

53. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9. PMID: 16519814.

54. Whiting P, Rutjes AW, Dinnes J, et al. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004 Jun;8(25):iii, 1-234. PMID: 15193208.

55. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 Nov 10;3:25. PMID: 14606960.

56. Wells GA, Shea B, O'Connell J, et al. The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-Analysis. www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed July 23, 2014.

57. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928. PMID: 22008217.

58. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006 Dec;59(12):1331-2; author reply 2-3. PMID: 17098577.

59. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005 Oct;58(10):982-90. PMID: 16168343.

60. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med. 2008 Feb 28;27(5):687-97. PMID: 17611957.

61. Arends LR, Hamza TH, van Houwelingen JC, et al. Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008 Sep-Oct;28(5):621-38. PMID: 18591542.

62. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001 Oct 15;20(19):2865-84. PMID: 11568945.

63. Dahabreh IJ, Trikalinos TA, Lau J, et al. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. Methods Research Report. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) AHRQ Publication No. 12(13)-EHC136-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

64. Singh S, Chang SM, Matchar DB, et al. Grading a body of evidence on diagnostic tests. In: Methods Guide for Medical Test Reviews. Rockville. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality. June 2012. www.effectiveghealthcare.ahrq.gov/ reports/final.cfm.

65. Lu B, Gatsonis C. Efficiency of study designs in diagnostic randomized clinical trials. Stat Med. 2013 Apr 30;32(9):1451-66. PMID: 23071073.

66. Hayen A, Macaskill P, Irwig L, et al. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol. 2010 Aug;63(8):883-91. PMID: 20079607.

67. Hall EJ, Brenner DJ. Cancer risks from diagnostic radiology. Br J Radiol. 2008 May;81(965):362-78. PMID: 18440940.

Full Report

This executive summary is part of the following document: Dahabreh IJ, Adam GP, Halladay CW, Steele DW, Daiello LA, Weiland LS, Zgodic A, Smith BT, Herliczek TW, Shah N, Trikalinos TA. Diagnosis of Right Lower Quadrant Pain and Suspected Acute Appendicitis. Comparative Effectiveness Review No. 157. (Prepared by the Brown Evidence-based Practice Center under Contract No. 290-2012-00012-I.) AHRQ Publication No. 15(16)-EHC025-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2015. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

AHRQ Pub. No. 15(16)-EHC025-1-EF
December 2015