Comparative Effectiveness Review Disposition of Comments Report

Research Review Title: *Diagnosis of Right Lower Quadrant Pain and Suspected Acute Appendicitis*

Draft review available for public comment from November 4, 2014 to December 2, 2014.

Research Review Citation: 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.

Comments to Research Review

The Effective Health Care (EHC) Program encourages the public to participate in the development of its research projects. Each research review is posted to the EHC Program Web site in draft form for public comment for a 4-week period. Comments can be submitted via the EHC Program Web site, mail or E-mail. At the conclusion of the public comment period, authors use the commentators' submissions and comments to revise the draft research review.

Comments on draft reviews and the authors' responses to the comments are posted for public viewing on the EHC Program Web site approximately 3 months after the final research review is published. Comments are not edited for spelling, grammar, or other content errors. Each comment is listed with the name and affiliation of the commentator, if this information is provided. Commentators are not required to provide their names or affiliations in order to submit suggestions or comments.

The tables below include the responses by the authors of the review to each comment that was submitted for this draft review. The responses to comments in this disposition report are those of the authors, who are responsible for its contents, and do not necessarily represent the views of the Agency for Healthcare Research and Quality.

Source: https://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2158
Published Online: December 14, 2015

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	General Comments	Is the report clinically meaningful? Yes - very much	Thank you. No action required.
Peer Reviewer #1	General Comments	The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers" I find the report very technical and probably difficult to read and understand for most of the audience stated above. This is probably more directed to researchers	We have simplified the exposition in the executive summary and have provided additional data summaries that may be useful for clinicians. It is true that EPC reports have a diverse audience and several components of the reports may be of particular interest only to researchers (e.g., Recommendations for Future Research).
Peer Reviewer #1	General Comments	Are the key questions appropriate and explicitly stated? yes	Thank you. No action required.
Peer Reviewer #1	General Comments	The diagnosis of appendicitis remains an enigma. It is a large group of patients and the diagnosis consumes large resources. The present review is very welcome. The task is ambitious and enormous and I can only congratulate the authors for this great effort. I am impressed by the approach and methodology which is to be commended. I thank for the opportunity to review this work. I hope my input may have some impact on the final version as I understand that this is only a preliminary draft.	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #2	General Comments	This is an important topic and the authors should be commended for addressing the state of the literature on the diagnostic approach to right lower quadrant pain. As noted, this is a difficult area to summarize despite the abundance of literature. Although the report shows great effort at reviewing and weighing the science, I would hope that the authors consider providing more clinical perspective to their conclusions; for example, current application of multivariate clinical scores are for risk stratification to guide management rather than diagnosis. Although these may be valuable for incorporation into diagnostic algorithms, a review of the interrater reliability of the components should be noted. Similarly, the authors note the varied populations included in studies assessing diagnositc imaging, but the report should acknowledge that advanced imaging was never intended in every patient and near impossible to study because of the time-dependent (signs often change over hours of the evaluation), and highly variable presentation of individual patients.	We have expanded our exposition on the clinical implications of using multivariate scores for risk stratification and as components of management strategies. Unfortunately, a review of the inter-rater reliability of score components (or composite scores) was deemed out of the scope of the current report. A review of this sort would be a major undertaking, but we agree with the reviewer that its implications are potentially clinically important. We have suggested such a review as a potential item for future research. We also agree with the reviewer's points about imaging tests and the large variability of clinical signs and symptoms over the course of disease. We have incorporated some of the suggested ideas in the Discussion section of the revised manuscript.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #2	General Comments (continued from row above)	Finally, the value of diagnostic imaging cannot be considered in isolation as if appendicitis is the only diagnostic consideration; it is important to judge clinical outcomes in the context of all important potential etiologies of right lower quadrant pain. Although the most important appendicitis-related outcomes are negative appendectomy rate and perforation rate, future research will need to weigh timeliness of diagnosis, hospitalization and ED revisit rates, timely and efficient identification of other serious diagnoses, quality of life outcomes, and cost effective care. I realize that the evaluation of suspected appendicitis is evolving, but any report informing future investigation needs to consider the potential medical treatment of appendicitis since the evaluation of suspected appendicitis may create a different dependence on imaging and other new diagnostics.	We agree with these important points and have incorporated them in the Discussion section of the report, particularly with respect to future research needs.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	General Comments	The key questions are appropriate and explicitly stated. The rationale for evidence review (ES-11) is clear. However, the target audience is not well defined. It is assumed to be clinicians who evaluate patients for right lower quadrant pain and acute appendicitis. However, the report, as it exists is highly unlikely to be useful to the overwhelming majority of such clinicians. It contains a wealth of information, but is difficult to read and process. I do not believe the average clinician will find this useful. If this is to be directed at payers and policy-makers, it may be more relevant.	Thank you for your comments regarding the key questions and rationale of the review. No action is required in response to these comments. EPC reports aim to serve the needs of broad and diverse audiences, including patients, practicing clinicians, researchers, payers, policy makers, test manufacturers, etc. We have streamlined the Executive Summary to the extent it was possible.
Peer Reviewer #4	General Comments	This study was extremely well written, is comprehensive, and although broad in the scope of appendicitis it has sufficient depth to provide meaningful conclusions. I found only minor unclarities throughout the text which require minor revision. Congratulations on the nice work of the authors!	Thank you. We have attempted to address all issues identified by the reviewer.
Peer Reviewer #5	General Comments	Thank you for the opportunity to review this manuscript on the accuracy of tests to diagnose patients with acute appendicitis. In general the summary report is lengthy and challenging to read. I made some specific recommendations to condense the text and move some data into tables to allow the reader to make comparisons more easily.	Thank you for these suggestions. We have attempted to streamline the text and simplify the tables. Please see below for our responses to your specific suggestions.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #1	General Comments	Report is clinically relevant and all the appropriate target populations are explicitly defined and key questions are appropriate and explicitly stated.	Thank you. No action required.
TEP Reviewer #2	General Comments	This is a very detailed report that contains a significant amount of useful information. However, the problem with the report is that including so many articles and information they overwhelm the readers. I am concerned that many will simply ignore the report given its very detailed information without any clear consensus on any part of the evaluation of patients with abdominal pain.	We appreciate the reviewer's concern that the size of the report may be intimidating to some readers. To address this issue, we have attempted to better highlight the report's clinical messages in the Executive Summary. Please also note that we are preparing several manuscripts that will present our findings in a more easily digestible format.
TEP Reviewer #2	General Comments	The Key questions listed are appropriate. I have focused my review on KQ1 as this is my area of expertise. The entire document was reviewed, however, the items related to KQ1 require the greatest degree of attention.	Thank you. Please see below for our response to the specific comments.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #3	General Comments	It is disappointing that after such exhaustive review that the final conclusion is basically that more/better studies are needed. The clinically relevant findings (that multivariable scores plus selective imaging improve results fewer negative appendices and perforations- at least in the few studies that look at that and those studies are not strong ones methodologically) are not easily seen. I do not fault the authors for either of these outcomes since the literature is what they are assessing as much as the actual problem of RLQ pain	We appreciate the reviewer's concern that the complexity of the clinical condition and the size of the report may be intimidating for some readers. To address this issue, we have attempted to better highlight the report's clinical messages in the Executive Summary. Please also note that we are preparing several manuscripts that will report the report findings in a more digestible format.
TEP Reviewer #4	General Comments	The Agency for Healthcare Research and Quality has performed a systematic review on the comparative effectiveness of various diagnostic modalities used in the diagnosis of right lower quadrant pain and suspected acute appendicitis. There are many options available to clinicians and there is a great need for evidence to inform us about the impact of testing strategies.	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #4	General Comments	Overall, the manuscript is logical, concise, and straightforward. There are minimal (if any) errors in grammar, style, and punctuation. The methodology is robust and the search strategy comprehensive. This was a well-performed systematic review which, unfortunately, could only document that there is insufficient existing evidence to support comparative effectiveness conclusions and can only point out the obvious fact that more research is needed.	Thank you. We agree that the available evidence does not support strong comparative conclusions. We have expanded the Future Research Needs section of the report in response to this comment.
TEP Reviewer #4	General Comments	As a clinician who deals with this presentation commonly, I approached this study with great eagerness, but was disappointed with the lack of conclusive evidence. This is not the fault of the authors, obviously, but rather due to the limitations of the evidence base.	Thank you. As noted by the reviewer, our ability to draw strong conclusions is limited by the amount and "quality" of the available research.
TEP Reviewer #5	General Comments	ES 10 lines 43-44. I would add that the ionizing radiation concern may delay the ordering of a CT which could then lead to an increase in the rate of perforation.	We have added this information to the Background section of the report.

Commentator & Affiliation	Section	Comment	Response
Public Reviewer #1 Advisory Council for General Surgery	General Comments	The review focuses on diagnostic considerations. We concur with the step wise use of imaging studies when indicated. If imaging studies are necessary then the treating physician should be responsible for determining the best test (s) if needed at all. We would emphasize that physical examination is the cornerstone of diagnosis and should direct imaging. In addition, we believe that the best diagnostic algorithm may vary among hospitals and that the utilization of diagnostic modalities is best individualized according to institutional factors and, most importantly, physician experience.	We agree that clinical examination is the cornerstone of diagnosis and should inform the use of imaging tests. We have explicitly mentioned this in the revised report. We also agree that the "best diagnostic algorithm" may vary among hospitals, physicians, and patients. We have emphasized the importance of such "heterogeneity" in the Discussion section of the review, but we also note the difficulties in disentangling the attendant complexity.
Public Reviewer #1 Advisory Council for General Surgery	General Comments	It is difficult to broadly apply the findings of this review to clinical practice at this time. The review does not address the time pressures for diagnosis and the methods of diagnosis in a widely applicable manner. For example, CT scans are often widely and immediately available for this diagnosis and many surgeons are very comfortable with determining the need for appendectomy on their own interpretation, not needing a delay in reading by a radiologist. This may not be true for MRI which is much more time consuming and may more often require a radiologist to interpret rather than a treating physician.	The challenges in determining the applicability of review findings in practice are due to the breadth of the questions addressed by our review and the current status of the literature. With respect to the former, we have provided some guidance for interpreting our findings in the Applicability subsection of the Discussion. With respect to the latter, we hope that our recommendations for future research will contribute to the generation of more clinically relevant evidence. We agree that the timing of tests is important in evaluating the evidence and in making clinical decisions. We have noted some of the issues pointed out by the reviewer in the Discussion section of the report.

Commentator & Affiliation	Section	Comment	Response
Public Reviewer #1 Advisory Council for General Surgery	General Comments (continued from row above)	The members of the advisory council rarely use MRI and only consider it as a diagnostic alternative in the pregnant patient. Additionally, MRI often requires additional time to perform and additional personnel/interpreters which are not always available. Ultrasound is technique that is highly operator-dependent, a fact lightly referenced in this review. Although we did not review all of the ultrasound (US) references, we suspect that the reports primarily reviewed US performed by experienced radiologists. Surgeon or emergency medicine physicians perform US but it is a generally unregulated/ unsupervised/un-reviewed technique. In the hands of an experienced operator this may be a very useful technique, but the availability of such personnel 24/7 is uncertain and not applicable to many hospitals across the USA.	We agree with these modality-specific points and have incorporated them in the Background and Discussion sections of the report. Of note, though we appreciate the importance of contextual factors (e.g., availability of equipment and trained readers), our review was not designed to address them.
Public Reviewer #1 Advisory Council for General Surgery	General Comments	We have a concern about statements regarding diagnostic laparoscopy as part of the diagnostic tests. It would be uncommon for a patient with RLQ pain who at laparoscopy is found to have a normal appendix in the absence of any other pathology not to have a concurrent appendectomy. There was no outcome evidence presented to support avoidance of appendectomy in this setting.	We have revised our wording regarding diagnostic laparoscopy for clarity; we believe that what we state is compatible with the reviewers' point.

Commentator & Affiliation	Section	Comment	Response
Public Reviewer #1 Advisory Council for General Surgery	General Comments	The review focuses on sensitivity and specificity for test evaluation. What the clinician wants to know is the positive or negative predictive value of a test. In other words, if the test is positive, how likely is it that the patient has the disease? If the test is negative, how likely is it that the patient does not have the disease? PPV and NPV are dependent on the prevalence of the disease in a given population and more useful. In other words, if the patient is more likely to have the disease based on clinical presentation (periumbilical pain to RLQ, focal RLQ tenderness, etc) a negative test is less accurate/NPV lower. If the patient is less likely to have the disease (nonfocal pain, pain onset after emesis, etc) a positive test is less accurate/PPV lower. Sensitivity and specificity remain the same. The studies reviewed did not contain patient level data to sort this out.	Because PPV and NPV are prevalence-dependent, we refrain from synthesizing these measures across studies. Instead, for the main tests we reviewed, we have generated curves that show the dependence of PPV and NPV over prevalence, for the estimated summary values of sensitivity and specificity. This information is presented in the Executive Summary and the main text of the report.
Peer Reviewer #4	Abstract	Study appraisal and synthesis methods: I would recommend to include in the Abstract the methods that the authors used in this review for assessment of preferred reporting items for systematic reviews and meta-analyses and for evaluating risk of bias of primary studies.	We have added the requested information in abbreviated form in the Abstract.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Abstract	Results: Please clarify. ….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 US). Is this sentence correct? Would not it be 72 for US and 32 for CT?	Thank you for spotting this typographical error. We have corrected the sentence and updated the numbers of studies based on the results of our updated searches.
Peer Reviewer #5	Abstract	Abstract: I think it would be useful for the authors to report the sensitivity and specificity of the clinical symptoms, signs, and laboratory tests when used in isolation and when used in combination, instead of simply stating their low performance characteristics.	We provide complete test-specific information in the main text of the report. Unfortunately, it is often not possible to meaningfully assess the test performance of combinations of diagnostic tests because such information is either not reported or is reported inconsistently (e.g., different combinations of tests are assessed in each study or conditional test performance is not reported). That said, we agree that clinical signs and symptoms are almost always used in combination (with other signs and symptoms and with other tests). We think that it would be very cumbersome for primary studies to report all possible conditional test performance measures, and this probably explains the poor and inconsistent reporting. However, we think that studies could fit multivariable diagnostic models and report their results (e.g., using standard logistic regression methods). We have suggested this as a possibility in the Future Research Needs section of the report.
Peer Reviewer #5	Abstract	Conclusions: This section does not state one of your major findings, which is that the individual tests, other than advanced imaging, have low sensitivity and specificity to accurately diagnose acute appendicitis.	We have added this information to the “Conclusions” paragraph of the Abstract.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #4	Abstract	No major changes are necessary. The results section is very wordy and quite lengthy. If possible, please try to reduce the length.	We have tried to streamline the Results section of the Executive Summary by referring readers to the main text of the report for some of the more detailed results.
Peer Reviewer #1	Introduction	I miss an explanation for your focus on screening and early stage cancer only. I miss some information about the natural history of appendicitis which has important implications to the diagnosis and management of appendicitis. See my attached document.	We agree that a consideration of the natural history of appendicitis is important for both the interpretation of our results and the planning of future research. We have emphasized this point in the Discussion section of the revised report.
Peer Reviewer #1	Introduction	The risk of appendicitis is lower in pregnant women. See PMID: 11821329 and PMID: 24950289	Thank you for these citations. We have revised our wording to indicate the conflicting data on the relative incidence of appendicitis.
Peer Reviewer #2	Introduction	Clearly written	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	Introduction	ES-10 (14-22). The authors make a somewhat sweeping statement regarding the relationship of diagnostic testing to outcomes. In fact, this is unsupported by evidence and better left out. There is no evidence that an emergent appendectomy (immediately after diagnosis) is associated with better outcomes than urgent appendectomy. Multiple papers in the literature in children and adults have supported the premise that appendiceal rupture and subsequent outcomes are defined by the patient's presentation and access to care and NOT events following the patient's presentation, diagnostic or otherwise.	We agree with the reviewer that the impact of tests on clinical outcomes is indirect, that the existing evidence is generally insufficient to draw conclusions on clinical outcomes, and that the results of studies on test performance should be extrapolated to clinical outcomes with extreme caution. We have clearly stated this in the Introduction and Discussion sections of the report. That said, we also believe that (to a large extent) primary research on diagnostic test performance and systematic reviews of such research are motivated by the belief that a connection between test performance and clinical outcomes is possible.
Peer Reviewer #3	Introduction	The authors also assume that appendectomy is the definitive intervention upon confirmation of diagnosis. Although this remains the most common approach to this disease, a significant body of emerging evidence in children and adults is strongly raising the possibility that acute non-perforated appendicitis may respond to antibiotic treatment without operation. The end point should therefore be states as initiation of treatment for appendicitis and not surgery.	Thank you for this important point. As stated in the report, we included studies that used non-pathology based reference standards and explored whether the use of such reference standards influenced estimates of test performance. We have used the suggested wording regarding the initiation of treatment.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Introduction	Scope of the Review: Page ES-13: Populations and Conditions of Interest: What is the rationale for considering as very young children those <2 years and 2.5 years of age? In Tables A (for fever, page ES-18 and for tenderness, page ES-19) and D (page ES-21) the authors use a cutoff of <5 years (instead of 2 or 2.5 years) to summarize estimates of test performance of clinical symptoms and signs for the diagnosis of acute appendicitis, and estimates of test performance of diagnostic score tests. Again, in page ES-36 Applicability of Review Findings, 1st paragraph) the authors comment on a cutoff of 5 years of age.	We had *a priori* decided to examine two strata within the group of children aged less than 5 years of age on the basis of input from clinical experts and other stakeholders. However, our ability to assess test performance and clinical outcomes subgroups was limited by the lack of reporting of relevant information in the included studies. We have clarified this in the revised report.
Peer Reviewer #4	Introduction	Key Questions: Page 4: Key Question 1: For clarify purposes please categorize the a priori determined age range for children and adults (similar to what is available for the elderly).	We had *a priori* decided to examine two strata within the group of children aged less than 5 years of age on the basis of input from clinical experts and other stakeholders. However, our ability to explore such subgroups was limited by the lack of reporting of relevant information in the included studies. We have clarified this in the revised report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Introduction	Introduction: I would recommend making this introduction section more concise. What is considered to be an acceptable low rate of 'negative' appendectomies? I don't think there is an acceptable rate, other than the range that is often reported in the literature. I would simply call it a false negative appendectomy rate, instead of using all the words to explain it, to make it more concise. Also, please list what is the reported range in the literature instead of stating that there is an "acceptable" rate.	We have streamlined the Introduction section in the Executive Summary. We have also made changes to our wording regarding the negative appendectomy rate, as suggested. We agree that defining an "acceptable" rate of negative appendectomy is challenging. However, we also think that the "rate reported in the literature" is not well defined (e.g., which treatment centers should contribute to the estimation of the rate? Over which years?). For this reason we have simply stated that – in general – non-zero negative appendectomy rates are tolerated (e.g., in order to avoid delayed intervention in cases where surgery is needed).
Peer Reviewer #5	Introduction	In your description of the clinical symptoms and signs suggestive of appendicitis, I would describe these as "classic" symptoms and signs or "classic" teaching about the symptoms and signs of appendicitis. I don't think it's necessary to elaborate on all the signs of peritoneal irritation, but rather important to mention that a large percentage of patients present atypically, so clinicians rely on the use of testing for diagnosis. Ultrasound is also used as the initial test particularly in patients whom clinicians want to avoid the radiation risk of CT, such as pregnant patients and children.	We have used the suggested wording regarding signs and symptoms and have stated that many patients present atypically. We have already mentioned that US is often used to avoid exposing children and pregnant women to ionizing radiation.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Introduction	Where does clinical observation come into your analysis? You mention it as a comparator, but do not mention how it is used to diagnose appendicitis.	We have described in detail our operational definitions for what constistutes an acceptable reference standard. We do not conceptualize deferred treatment strategies as diagnostic tests (that is, we only consider diagnostic procedures that result in a determination regarding the presence of disease – i.e., positive or negative result). However, we accepted studies using observation as the reference standard (for cases deemed negative by the index test). Please note that we did not evaluate clinical observation as a treatment strategy because the assessment of the comparative effectiveness of interventions was not considered within the scope of the report.
Peer Reviewer #5	Introduction	Your final five paragraphs describe the diagnostic difficulty in various populations, which I think can be consolidated into fewer paragraphs and highlight the importance of using advanced imaging because of this.	We have streamlined the Executive Summary section of the report on special populations.
Peer Reviewer #5	Introduction	Key Questions: I don't quite understand Q1—is this supposed to be "available" diagnostic tests and not "alternative" diagnostic tests? If not, what are alternative diagnostic tests used to diagnose appendicitis?	We did not use the term "alternative" to mean "unconventional"; instead we were referring to the use of the word to mean "express a choice" [Merriam-Webster dictionary]. We consider this to be standard usage and have retained the original wording.
TEP Reviewer #1	Introduction	no additional comments	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #2	Introduction	Well written. Major points of comment are:	Thank you. Please see below for our responses to specific comments.
TEP Reviewer #2	Introduction	mean time to perforation is more accurately 36-48 hours. Line 18, page ES-9	We have used the suggested information.
TEP Reviewer #2	Introduction	In general, the use of high risk rules for appendicitis is not entirely useful as the decision to operate with our without imaging is based on the surgeon. This is a difficult decision point to impact. On the other hand, the decision to defer imaging or choose a path of observation is something that non-surgical clinicians can consider. It is for this reason that I believe that the authors should include the utility of low PAS score as well as the "Low Risk Appendicitis Rule in this document. This is especially true for Children. Lines 51-53, ES-9.	We agree with the point regarding the practical usefulness of risk scores and have adopted some of the reviewer's thinking in the Discussion section of the report. The studies suggested by the reviewer are included in the report. We have not made changes to the Results section because we cannot devote disproportionate space to studies selected using somewhat arbitrary criteria. However, we have alluded to the importance of these studies in the Discussion section of the report.
TEP Reviewer #2	Introduction	I was surprised by the lack of focus on the pediatric population in this document. Appendicitis is largely a disease of children, and that does not come across in the text.	We respectfully disagree with the assertion that the report lacks focus on the pediatric population. We have stratified all analyses by age group and have addressed challenges and drawn conclusion conclusions specific to this age group. Given the (very) large number of included studies and the overall length of the document, it is easy to feel that a particular topic of primary interest has not received adequate attention. We hope to address this to some extent by seeking the publication of a separate peer-reviewed paper focusing exclusively on our findings for the pediatric population.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #3	Introduction	Very fair overall statement of the problem and nature of the work	Thank you. No action required.
TEP Reviewer #4	Introduction	Regarding "negative" appendectomies, the concept of allowing (and even desiring) a certain rate of negative appendectomies is held over from an era before advanced imaging like CT and MRI. Removing a normal appendix is less acceptable these days. "Up to a third of women of reproductive age with appendicitis are misdiagnosed" (page ES-10, line 45). This statement is supported by a reference that is 20 years old. Is this still true in 2014?	We have provided a more recent reference for the high rate of misdiagnosis among women of reproductive age.
TEP Reviewer #4	Introduction	Rationale for Evidence Review: no major changes necessary	Thank you. No action required.
TEP Reviewer #4	Introduction	Key Questions: No major changes necessary	Thank you. No action required.
TEP Reviewer #5	Introduction	Excellent	Thank you. No action required.
Public Reviewer #1 Advisory Council for General Surgery	Introduction	We wondered if there is a reference to the comment, "Untreated appendicitis can lead to perforation of the appendix, which typically occurs within 24 to 36 hours of the onset of symptoms." It was interesting that out of over 3000 references this statement had no citation. This may be a generally held opinion but, with increasing experience, it may be less true that previously thought.	Thank you for this comment. We have provided a reference for this statement.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods	Are the inclusion and exclusion criteria justifiable? The inclusion criteria is OK, but maybe there should be also some exclusion critera? I assume that case-control studies are excluded. I would propose that studies based on operated patients should also be excluded.	We have excluded studies that sampled participants on the basis of the presence of disease (these are often referred to as “case-control” studies, but we prefer to reserve this term for studies of etiology, not diagnosis). We have not excluded studies of operated patients, instead we have conducted and reported detailed subgroup analyses by this variable.
Peer Reviewer #1	Methods	Are the search strategies explicitly stated and logical? Yes	Thank you. No action required.
Peer Reviewer #1	Methods	Are the definitions or diagnostic criteria for the outcome measures appropriate? Not always. See my attached document.	Rather than exclude studies on the basis of their reference standards (which is equivalent to discounting all information from these studies), we have retained them in the dataset and conducted subgroup analyses.
Peer Reviewer #1	Methods	Are the statistical methods used appropriate? I guess so. I trust the AHRQ on this point.	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods	My main critique regard the selection of studies. A large part of the literature on the diagnosis of appendicitis are made retro- or prospectively on populations that were appendectomized. Such studies have a high prevalence of appendicitis. In table 1 and table 6 (and on other places) it is clear that the median number of appendicitis is larger than the non-appendicitis patients in almost all the subgroups. I assume that in some studies there may be <10% non-appendicitis patients. This does not represent the kind of population on which the test is thought to be applied on – namely patients with abdominal pain and suspected appendicitis. A typical prevalence of appendicitis in such a population is about 30%. So we have a problem of spectrum bias in the "surgical cohort". We do not know what mechanisms that worked in the selection for the decision to operate, but it will certainly lead to selection bias. Findings that were used for the selection will have lower discriminating capacity in such studies. This applies to clinical variables, which were probably used for the selection of patients for surgery. The discriminating capacity of those variables will be underestimated in thses studies.	Rather than exclude studies on the basis of disease prevalence (which is equivalent to discounting all information from these studies), we have retained them in the dataset and conducted appropriate subgroup analyses. We note that spectrum bias arises only if sensitivity and specificity vary over some characteristics of the patients or the severity of the disease. The reviewer seems to imply that (1) such variability is present and (2) that appendicitis prevalence is a good proxy for between-study variation in the bias-inducing characteristic. We assessed the potential association between appendicitis prevalence and summary sensitivity and specificity in appropriate subgroup analyses (reported in the main report for every test and every patient population). We have provided additional guidance on how the results of such analyses should be interpreted in the Discussion section of the revised report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods (continued from row above)	This is also evident in some of the separately reported results of the metaregression analyses, like rebound tenderness, CRP, neutrophil%, WBC and CT. It is also consistent with the finding that CrIs for a risk of bias item did not include the null value in studies of white blood cell count that had complete verification of index test results had lower specificity than studies with incomplete verification. I would say that the results in the group with complete verification (surgical cohort) is biased. We also do not know the number of false positive diagnoses in the non-operated patients. The specificity diagnostic capacity will therefore be overestimated. One argument in support of keeping the surgical cohorts is the problem of verification. However, what is important is if the patient had appendicitis needing surgical treatment, which is clarified by a follow up of non-operated patients. I comment on that further down I think all those studies should be excluded. This is a firm recommendation. At least the studies based on patients with suspicion of appendicitis should be reported separately. I am aware that this will reduce the available evidence drastically, but better to have valid than biased evidence.	We have discussed these *potential* sources of bias extensively in the Results and Discussion sections of the report. We prefer to include studies and perform appropriate subgroup analyses (reported in detail in the main text of the report) rather than exclude studies *a priori*.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods	One crucial point in all the studies is the gold-standard for the diagnosis. You write that "we assumed that pathological diagnosis and clinical followup have negligible measurement error. It is unlikely that this assumption is exactly true (e.g., pathologic examination may have some diagnostic error, and clinical followup provides less than perfectly accurate information). However, we believe that the error rate of this reference standard is low enough that its influence on our estimates is unlikely to be substantial."	We have discussed these *potential* sources of bias extensively in the Results and Discussion sections of the report. We prefer to include studies and perform appropriate subgroup analyses (reported in detail in the main text of the report) rather than exclude studies *a priori*. We have softened our wording regarding measurement error in the reference standard test.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods (continued from row above)	I do not agree. This is a big problem. The appendicitis diagnosis is routine for pathologists. They know that the surgeon will be glad if he gets a confirmation on the diagnosis, and the pathologist may always find some neutrophils in the appendix, showing "inflammation" Pathologists use very different criteria for the histopathological diagnosis and there is no generally accepted standard. Carr (Annals of Diagnostic Pathology, Vol4, No 1 (February), 2000: pp 46-56) writes "it seems surprising and somewhat ironic that the diagnosis of acute appendicitis, one of the most common made by anatomic pathologists, remains poorly defined, subject to misconceptions, and prone to variable terminology." I think that this review should give a section to this problem. I suspect that studies reporting very low rates of negative appendectomy practice fuzzy criteria for the diagnosis. This may be especially true if there may be economic incentives for avoiding negative appendectomies. One quality item could be if the study reported the histopathologic criteria used for the diagnosis. According to Carr there should be transmural invasion of neotrophils.	We have discussed these *potential* sources of bias extensively in the Results and Discussion sections of the report. We prefer to include studies and perform appropriate subgroup analyses (reported in detail in the main text of the report) rather than exclude studies *a priori*. We have softened our wording regarding measurement error in the reference standard test. In addition, please note that we strongly prefer to *not* use aspects of study reporting (e.g., "if the study reported the histopathologic criteria used for the diagnosis") as markers of study validity (i.e., as items indicative of risk of bias).

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods	In my opinion, the methodological quality items you use are unclear, incomplete and need to be adapted to the specific situation of diagnosing appendicitis. - Could the selection of patients have introduced bias? The surgical cohort clearly breaks this item. Hospitals only receiving referrals is another. - Could the conduct or interpretation of the index test have introduced bias? The answer here is almost always Yes. I know about one blinded study of CT and US (Poortman 2003) which showed rather low sens and spec and no difference between CT and US. I think some clinical scoring systems have been validated without calculating the score-sum. This is one kind of linding. Have you identified more blinded studies? I think this is an important quality Item that you could comment on and use in meta-regression. - Is the reference standard likely to correctly classify the target condition? You have answered about 50% Yes and 50% no to this item. How do you interpret this question in this situation? This is very unclear to me.	Thank you for these points. We have provided the operational definitions of all risk of bias items in the report's appendix. We believe that our criteria are reasonable and consistent with current methodological guidance for systematic reviews of diagnostic tests.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods (continued from row above)	Were the reference standard results interpreted without knowledge of the results of the index test? This is a tricky question. The pathologists often regard appendicitis as a routine diagnosis and they know that the surgeon will be happier if he reports appendicitis than if he negates it. So there is a risk of false positive reference result in the operated cases. This does not apply for the non-operated. - Did all patients receive a reference standard? Did all patients receive the same reference standard? This question needs to be adapted to the diagnosis of appendicitis. The reference standard for operated cases is the histopathology examination. For non-operated cases it is a follow up within at least 2 weeks. This means that they have two different reference standard but I do not think this is a quality problem if they both have had them. What reference standard did you crave for in the nonoperated patients? - Could the reference standard, its conduct, or its interpretation have introduced bias? Certainly. The quality of the histopathology reports are very variable. How many of the studies declared the criteria for the histopathologic diagnosis?	Thank you for these points. We have provided the operational definitions of all risk of bias items in the report's Appendix. We believe that our criteria are reasonable and consistent with current methodological guidance for systematic reviews of diagnostic tests.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods (continued from row above)	- Could the patient flow have introduced bias? Hospitals only receiving referral patients will already have a selection. Some studies have compared the preliminary clinical diagnosis before imaging with the final diagnosis after imaging. I do not know if you have included such studies among the clinical indicators? Ng 2002 is one such study. In his follow up Ng 2010 he notes that "it can be appreciated that the passage of time (24 h – i.e. standard practice) can bring with it some increase in diagnostic confidence, since acute abdominal symptoms and signs either evolve or regress, helping to confirm or deny one ' s initial diagnosis."	Thank you for these points. We have provided the operational definitions of all risk of bias items in the report's Appendix. We believe that our criteria are reasonable and consistent with current methodological guidance for systematic reviews of diagnostic tests. We have also discussed the importance of considering the natural history of appendicitis when interpreting our results and when planning future research.
Peer Reviewer #1	Methods	Andersson 2000 PMID:11071167 is based on the same patients as Andersson 1999 PMID: 9880421. The difference is that Andersson 2000 analyse the predictors of a negative outcome among all patients with suspicion of appendicitis (including those having appendicitis and the nonoperated non-appendicitis patients). It does not report any new information. It should be excluded	Both cited studies are included in the report, and this is consistent with our selection criteria. We have included in the report all studies that met our selection criteria and used them in descriptive analyses. However, in inferential analyses (e.g., meta-analyses), we excluded studies reporting results on the same test applied to the same or partially overlapping patient populations from inferential analyses. Thus, there is no double counting related to the cited studies.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Methods	I do not understand the meaning of Community and Ambulatory setting? Are these referrals? I guess it should have an impact on the prevalence of appendicitis which should be a more meaningful measure.	We no longer use this variable in descriptive or inferential analyses.
Peer Reviewer #1	Methods	What is meant by Clinical presentation consistent with appendicitis? How do you define that?	To some extent we had to adopt the definitions provided in the primary studies. This is stated in the methods section of the report.
Peer Reviewer #2	Methods	One of the key features of interpreting the performance of diagnostic imaging is understanding how equivocal or non-diagnostic studies were handled in the reporting of test performance. Because this is a major source of bias in prior studies, please address this in the current report.	We have provided details about the handling of indeterminate test results in the Methods section of the report. We now also report the results of an extensive sensitivity analysis in an Appendix.
Peer Reviewer #3	Methods	I am concerned about exculsion of non-Enlgish publications ES-14 (54). The authors state that these were " few and had small sample sizes", yet report 951 studies that were excluded due to language! This is of particular concern since many publications on US in particular have been published in European and Asian languages where there is much more expertise with US than in the US. The other inclusion and exclusion criteria appear reasonable.	At this point we cannot expand the scope of the report to include non-English language publications. We have mentioned this as a potential limitation.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	Methods	Some of the outcome measures are simply non-sensical and should be removed. Why did the authors choose fistula formation as an outcome. This is exceedingly rare, and cannot be correlated with diagnosis by any stretch of the imagination. Stump appendicitis (a complication almost exclusive to lap appe) is another inappropriate outcome measure. Maternal/fetal outcomes are also inappropriate outcome measures in my opinion	Outcomes were chosen *a priori* in consultation with clinical experts and other stakeholders. We believe that the outcomes we considered represent all major clinical events of interest to patients with acute abdominal pain and suspected acute appendicitis. We also believe that diagnostic tests are used to inform management in order to reduce the incidence of such events (and that this indirect link does not represent a "stretch of the imagination"). We have made no changes to the outcomes of the review.
Peer Reviewer #3	Methods	In studies looking at clinical findings without composite scoring (e.g. Alvarado), the authors look at each symptom or clinical finding separately. This is not clinically relevant, as each of theses studies simply looks at the value of the clinical evaluation without imaging, rather than considering any one symptom or sign.	It is unfortunately not possible to meaningfully assess the test performance of combinations of diagnostic tests because such information is either not reported or is reported inconsistently at the study level (e.g., different combinations of tests are assessed in each study and conditional test performance is never reported). We provide information on this issue in the Discussion section of the report.
Peer Reviewer #3	Methods	In my opinion, diagnostic laparoscopy should be completely removed as a diagnostic test. In contemporary practice, in the presence of multiple imaging modalities, laparoscopy is not considered a diagnostic maneuver for possible appendicitis. Rather, it is added to the procedure of lap appe if the appendix is found to be normal. I think combining a surgical procedure with radiologic tests and clinical evaluations in the review creates significant confusion.	The inclusion of studies on diagnostic laparoscopy was decided a priori, in consultation with clinical experts and other stakeholders. We generally prefer to include studies and carefully assess what conclusions can be drawn from them, instead of ignoring them entirely. We address the problems of studies of diagnostic laparoscopy (some of which the reviewer points out) in the Discussion section of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	Methods	In multiple areas, the authors comment that further quality control exercises will be completed that will review inclusion criteria, data acquisition, etc. I am curious therefore whether some results and conclusions may not be changed after this draft report.	We have undertaken extensive quality control activities, and we report all of them in the revised version of the report.
Peer Reviewer #4	Methods	Figure A. Analytic Framework: Page ES-13: Please clarify if in this figure the box that states "Diagnosis (appendicitis present or not)" which lies between KQ1a and KQ2a items relates to histology (appendicitis present or not) OR to diagnostic testing (then you should consider appendicitis present, negative or equivocal).	We have clarified the semantics of the analytic framework graph.
Peer Reviewer #4	Methods	Literature Search and Abstract Screening; Page ES-14: Please clarify if you have used MEDLINE (instead of PubMed or both) through an OVID-power search process. Who (number of reviewers, background of reviewers, participation of an experienced librarian? level – number of years – of experience of reviewers) performed the literature search? Page ES-16: Please clarify if in cases where only a subset of the available studies could be quantitatively combined, if you have used meta-narrative reviews (Wong et al. BMC Medicine 2013: 11(20): 1-15)	We have provided the requested information. We did not use meta-narrative or realist review methods because we do not think they are applicable to the topic.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Methods	Scope of the Review: Populations: While using RLQ abdominal pain is one way to select for patients with possible appendicitis, it is fairly narrow and does not account for patients who present atypically (particularly the very young, pregnant, and very old patients) or people who present early and may have only periumbilical or diffuse abdominal pain.	The population of interest included both patients with RLQ pain and patients otherwise suspected of having acute appendicitis. We have clarified this in the revised report.
TEP Reviewer #1	Methods	methodological section is thorough and clearly defined; inclusion/exclusion criteria clearly stated and justified; diagnositic criteria for outcome measures are appropriate; statistical methods are appropriate.	Thank you. No action required.
TEP Reviewer #2	Methods	The search criteria listed are easy to follow. However, I believe the authors have made a mistake by including so many articles. I would prefer that the authors not include any articles prior to 2000. The quality of these older studies as well as the use of diagnostic imaging has considerably changed the management of appendicitis over the past 10-15 years. For KQ1, I would prefer that the authors focus on the recent literature. The authors already take this approach for KQ2 and KQ3 as they focus on MRI, CT and US and the majority of this literature is recent.	We generally prefer to include all relevant studies and then perform appropriate subgroup analyses. We think that year of publication cut-offs are not appropriate because technology diffusion happens at different rates in different healthcare settings and because it is very unlikely that the same cut-off year is appropriate for all technologies and test types. Please note that we have performed subgroup analyses by year of publication for all tests and all populations considered.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #2	Methods	Furthermore, I would encourage the authors to include all relevant publications from 2000 on. For example, the decision to limit studies from the discussion for multivariate scores (to 5 studies) was confusing to me. This approach eliminated the Low Risk appendicitis score, but this particular study was the largest conducted over the past 15 years	We had reviewed and included the study on the "Low Risk appendicitis score" in the report – it is omitted from one of the tables in the main report because it is the only study evaluating this particular score. The complete data are provided in the Appendix of the report.
TEP Reviewer #2	Methods	Lastly, for KQ1 I would like to have seen a section on risk stratification using clinical pathways.	Any evaluation of diagnostic strategies that met our selection criteria has been considered in KQ1. Many studies evaluating clinical pathways did not meet our selection criteria because they were not comparative (i.e., did not provide information on comparative performance), did not allow estimation of the test performance of the entire strategy, and did not provide information on harms. This is because clinical pathways are best conceptualized not as diagnostic tests, but as complete management strategies (that involve testing as a component -- test-and-treat strategies). We have made recommendations about the assessment of clinical pathways in the Future Research Needs section.
TEP Reviewer #3	Methods	I have no argument with their methods	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #4	Methods	Methods: no major changes necessary. Although it is too late now, I question whether it is really relevant to include studies published since 1956. The practice of medicine is vastly different now and technology has progressed so much in the past half century. CT scans are much more sensitive now than they were 10 years ago. Why would I base my clinical decisions on an old study on an outdated technology? Additionally, the scientific quality and methodology of most studies prior to the 1990s is poor and likely to be at high risk of bias. Why choose 1956 as a cut-off? Did you notice any differences based on publication year?	Arguably the test performance of signs and symptoms has not changed much over time (and we have assessed formally whether test performance differs by year of study publication for all tests we considered). Please note that 1956 was not a cut-off year chosen by us, it is simply a reflection of the indexing of the various databases we used and the publication patterns in the field.
TEP Reviewer #5	Methods	Excellent and thorough. Strategies were well stated and logical. Definitions were appropriate.	Thank you. No action required.
Peer Reviewer #1	Results	Is the amount of detail presented in the results section appropriate? Sometimes too much on not relevant detail, and sometimes I miss other more relevant outcomes - see attached file	We have streamlined the Executive Summary. For pre-specified outcomes we report all information available in the studies (please consider that many studies do not report information on all outcomes of interest).

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	Are the characteristics of the studies clearly described? No. I do not understand some of the characteristics: Community, Ambulatory setting? Clinical presentation consistent with appendicitis? Some of the risk of bias elements are unclear: "Quality Item 4 = Is the reference standard likely to correctly classify the target condition?"	We have provided operational definitions for all risk of bias items. We have omitted some information on healthcare setting that was poorly reported across studies.
Peer Reviewer #1	Results	Are the key messages explicit and applicable? I do not find any clear or concise key message. And definitely not any applicable	We respectfully disagree with this comment. We provide a comprehensive review of single test performance and extensive comparative analyses of test performance and the impact of tests on clinical outcomes. Some of the limitations of our results are a reflection of limitations in the literature. We provide detailed suggestions for future research, which we hope can move the field forward.
Peer Reviewer #1	Results	Are figures, tables and appendices adequate and descriptive? Yes	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	Did the investigators overlook any studies that ought to have been included or conversely did they include studies that ought to have been excluded? I think Morino PMID: 17122613 should be included. The study is on patients with non-specific pain, but I think appendicitis is definitely an alternative diagnosis in this group of patients. Andersson 2000 PMID:11071167 should be excluded. see my attached document	PMID 17122613 has been included in the report. PMID 11071167 was included in descriptive analyses, consistent with our selection criteria (i.e., both 11071167 and 9880421 are described in the report). We have included in the report all studies that met our selection criteria and used them in descriptive analyses. However, we excluded studies reporting results regarding the same test applied to the same or partially overlapping patient populations from inferential analyses (e.g., meta-analyses). Thus, there is no double counting with respect to the cited study in analyses for which we rated the strength of evidence.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	Test performance of diagnostic laparoscopy is a special case. It is an expensive and invasive procedure with morbidity that can be life threatening. After the introduction of diagnostic laparoscopy the negative appendectomy rate has lost its meaning. In the era of open appendectomy it was a measure of diagnostic error, but now when a non-inflamed appendix is left in situ it does not reflect the quality of the preoperative diagnosis. Each diagnostic laparoscopy is an abdominal exploration, and if no pathology is found it was a preoperative diagnostic error. I would propose that you use the term "non-productive abdominal exploration" instead. A high rate of non-productive abdominal exploration is a sign of an overuse of this modality.	We agree with the reviewer's point and have treated diagnostic laparoscopy as a "special case". We have used the term "non-productive abdominal exploration," as suggested.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	Further I miss one study in the comparison of early diagnostic laparoscopy with conventional treatment - Morino 2006. This is excluded as not relevant. However, I think it should be included together with Gaitan and Decadt. You only report time to diagnosis, which is obviously shorter in the early diagnostic group. This outcome is not meaningful or has very little importance. You compare patients with 100% exploration with others having selective exploration. Most would conclude that laparoscopy was associated with a higher rate of correct diagnosis, but a close examination will reveal that this is due to a 4-5 time higher frequency of appendicitis in the laparoscopic arm. This is a strong evidence that resolving appendicitis is common, and that the detection and treatment of these resolving cases will give completely misleading results in terms of proportion of perforation and negative appendectomy.	We have reported many outcomes regarding the use of diagnostic laparoscopy, both in the Executive Summary and the main report text. Morino 2006 (PMID 17122613) has been included in the report (please also see our reply to comments from the same reviewed regarding this study). We agree with the reviewer's comment regarding the importance of considering the natural history of diagnostic laparoscopy and have elaborated on this point in the Discussion section of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	Randomized trials comparing CT versus routine management. For some reason you report mortality for 4, frequency of perforation for 3 and negative appendectomy, the most important outcome, for 3 of the 7 studies. I think you could underline that there was no difference in negative appendectomy. I think you miss to report that the proportion of operation for appendicitis was higher in four of the studies, once again a possible effect of resolving appendicitis (Ng, Lee, Lehtimäki, Sala). A large part of that passage describe the outcome in non-randomised clinical studies. These studies compare the outcome in patients that had a CT scan with those that had not. All such studies have a very high risk of selection bias and I think should not be included in this systematic review aiming at defining results based on high quality studies.	For pre-specified outcomes we report all information available in the studies (please consider that many studies do not report information on all outcomes of interest). We believe that the reviewer is thinking about "confounding" when referring to "selection bias" (these two sources of bias are distinct). We have considered the possibility of both confounding and selection bias and have rated the strength of evidence after considering the limitations of the studies. We prefer not to exclude studies because of presumed risk of bias; instead we include all relevant studies and modify our conclusions based on the assessment of the risk of bias of the studies.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	US versus routine management or clinical assessment. Also here you give rather detailed figures on the proportion of perforation and the time to operation. I think other results can be more important. Douglas mention that there was more appendicitis cases in the US arm (probably because spontaneous resolution) and that "the diagnosis of acute appendicitis aided by graded compression ultrasonography has not been shown to produce better outcomes than clinical diagnosis alone."	We have reported information on the proportion of patients diagnosed with appendicitis in comparative studies.
Peer Reviewer #1	Results	From my points above I would argue that a more intense diagnostic workup may give an increase in number of operations for appendicitis that will resolve with conventional management. This can be regarded as an adverse event.	We agree that unnecessary diagnostic investigations are an important outcome but studies do not report the proportion of patients receiving such diagnostic investigations.
Peer Reviewer #1	Results	Table 5 has a wrong heading - performance of imaging tests for the diagnosis of acute appendicitis. This error is also in the text section page 68.	Thank you for pointing this out. We have corrected the Table headings.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	I do not understand why you define the low risk cutoff point in clinical score as the "not low risk cutoff"? Do I understand correctly that Table 15 is the diagnostic characteristics at the low and table 14 at the high cutoff points? I think it should be noted that the combination of sensitivity at the low cut off (table 15) and specificity at high cut off (table 14) is in many instances at par with ultrasound, at least for the sensitivity.	We have clarified our wording for cut-off values in studies of multivariate diagnostic tools. We have also discussed how multivariable diagnostic tools can be used to separate patients with suspected appendicitis into groups at high, low, or intermediate risk of appendicitis.
Peer Reviewer #1	Results	There must be something wrong with the calculations of LR- for the AIR- and Bengezi scores in table 15, and in table 14 for the AIR-, Alvarado in Women of reproductive age, Alvarado modified in Adults, Alvarado modified in Women of reproductive age and Bengezi. All these have the same results in LR+ and LR- column. I wonder if there may be more errors. I find one study for the Alvarado score for Adults in table 15 (low cut off) but 12 in table 14 (high cutoff).	Thank you for this comment. We have verified every table entry by comparing it against the automatically generated meta-analysis output.
Peer Reviewer #1	Results	For diagnostic imaging there is a problem how to treat non-diagnostic scans. Some reports exclude them from the analysis. Others include them among the negative scans. How did you treat those scans?	We have reported the proportion of non-diagnostic scans when available. We have also performed extensive sensitivity analyses using data from studies that reported information on the proportion of scans that were non-diagnostic. This information is provided in the Appendix of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	One apparent difference between imaging and clinical diagnoses is the low proportion of perforations in the imaging studies. This suggest that imaging studies are made on a different population than the clinical studies. An alternative is that more cases with non-perforated appendicitis was detected and operated on, leading to a lower proportion of perforations. I think this should be commented on. How can this have influenced the result?	We have mentioned this point in the Discussion section of the report. We believe that this is a primary reason for examining comparative studies, both for test performance and for impact on clinical outcomes.
Peer Reviewer #1	Results	I think that in some of the imaging studies the final diagnosis was done after re-examining the CT-study, ie the next day by a more senior radiologist? Did you take that in to consideration?	We have provided details about our methods for handling studies of multiple test raters in the Methods section of the main report. We have also performed sensitivity analyses for studies of multiple raters; these results are now provided in the Appendix.
Peer Reviewer #1	Results	In the meta-regression of imaging tests you found some factors that had an impact, but you did not make any conclusion. Can these differences be related to the surgical cohort?	We have discussed our meta-regression results more extensively.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	For the Test Performance of Diagnostic Laparoscopy the diagnostic yield is often reported and a high yield is regarded as positive. I often miss information on the number of these diagnoses that lead to a change in treatment which I think is more important. If some diagnosis is made that does not need treatment or resolves spontaneously the patient had no benefit of the exploration. Non-productive exploration is one term that is used for such situations. Diagnostic laparoscopy versus immediate appendectomy. One advantage of laparoscopy is diagnostic. You have reported on the rate of negative appendectomy, which is obviously lower with laparoscopy. A more important result would be if other pathology was identified, and if the laparoscopy lead to any other change in treatment.	We report information on whether any pathology was identified by diagnostic laparoscopy. Information on treatment changes is typically not available in the form that the reviewer suggests (in contrast, information on yield is available and has been reported in the Results section). For this reason, we have not attempted to extract additional information on this outcome. We have used the term “non-productive exploration”, as suggested.
Peer Reviewer #2	Results	Well organized based on key questions. Tables and figures easy to read and follow.	Thank you. No action required.
Peer Reviewer #3	Results	The results are presented mostly in very detailed tables that are quite extensive, so extensive as to render them less than useful to the average clinician. In the text, specific studies are summarized, without enough effort to describe the pooled data	We have made our best effort to report our findings concisely and at the same time avoid selective reporting. We believe that the current version admits to various readings, ranging from a focused review of key findings (as summarized in the Executive Summary) to an extremely detailed presentation of test-, population-, and cut-off-specific results (as presented in the Appendix).

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Results	Modifiers of Test Performance: Pages ES-23 – 26: Did the authors consider analysing the data according to the generation of CT and ultrasound scanners of primary studies? Changes in technology over time may be a modifier of test performance.	Unfortunately, information on scanner types was often poorly reported. We have attempted to capture the evolution of the technology by examining subgroups by year of study publication.
Peer Reviewer #4	Results	Test Performance of Multivariable Diagnostic Scores: Page 40 It was difficult to me to understand the difference between the terms “not low risk cutoff” (if it is not low risk would not it be high or intermediate risk?) and “high risk cutoff”. Please consider re-visiting the terminology for clarity purposes.	We have revised the wording used to describe score cut-offs.
Peer Reviewer #4	Results	Test Performance of Diagnostic Laparoscopy: Page 57 Please emphasize the reference standard test used which I presume it was histology.	We have provided the requested information.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Results	Page 58. I have concerns about the terminology used: Appendix not visualized (=partial or complete inability to visualize the appendix) – 3rd paragraph of page 58 – I would consider this category as "indeterminate". No cause found – 7th paragraph of page 58 – 33 studies gave information on indeterminate (false positive) findings, in which the appendix appeared normal and no other pathology was found. Typo: 4th paragraph of page 58: "In the studies of women…).	We have reworded these sentences using more uniform terminology.
Peer Reviewer #4	Results	Table 27: Page 69 It would be of interest to provide information about whether the CT scans were focused or full (in terms of area of scanning and exposure to radiation) for some categories (e.g. IV and oral contrast-enhanced CT vs nonenhanced CT – focused or full?; mandatory CT vs selective CT – focused or full?; etc). This information is available only for some categories of this table.	We have reported this information whenever available for comparative studies.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Results	Included Studies with Information on Adverse Events: Page 77 Adverse events most likely relate to the use of contrast agents (CT and potentially MRI if gadolinium is used), radiation (CT), surgical procedure (laparoscopy), intolerance of the imaging test (CT, MRI) and discomfort (ultrasound, CT). Adverse events' risks related to unknown bioeffects or noise from MRI are minimal, but can result in maternal/fetal adverse events. It would be helpful if the authors could mention at the beginning of each sub-section the number of the articles that mentioned harms related to diagnostic tests within the articles for that imaging modality. For example, out of ….. studies of MRI six (…%) reported information on maternal outcomes. This would facilitate the flow of reading, would avoid the need for the reader to have to go back and forth to the tables, and would provide more comprehensive reporting of the available data on adverse events per imaging modality.	We have added the requested information.
Peer Reviewer #5	Results	In Table A, please clarify the two "mixed" rows under "pain migration." Why are there separate "mixed" populations?	Thank you for pointing out this issue. We have corrected the table (one of the entries was redundant).

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Results	Test Performance: The paragraph describing Table E is too challenging to read, especially because the table that follows contains the same information. I would recommend taking out some of the numbers, such as the number of studies that the data was based on.	We have opted to retain Table E in its original form for consistency with other parts of the Discussion section.
Peer Reviewer #5	Results	Diagnostic laparoscopy: I would recommend the authors report this in a table just like the other tests, with the same breakdown for children, pregnant women, etc.	We have opted to retain the original presentation for this section. We reported the information in the text because studies often used mixed populations and because no inferential statistical analyses were performed for diagnostic laparoscopy.
Peer Reviewer #5	Results	Comparative assessments of test performance: In the comparisons of US and CT or US and MRI, could you make any conclusions about test performances in kids vs. adults vs. pregnant patients or elderly? Were these populations all mixed or only adults?	We have reported the additional information requested when available in the primary studies.
Peer Reviewer #5	Results	Key Questions 2: Could the authors summarize the papers in the CT vs. clinical assessment in a table instead of writing it all out in paragraphs? It would enable the readers to more clearly compare the studies and their results.	We have used our best judgment to decide how to organize study-specific information. For randomized studies, we have opted to provide additional descriptive information in the body of the text (not only in tables) because that information cannot be forced into tabular format.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Results	Key Question 3: This should be highlighted in the conclusions of this paper—that the studies of each of the diagnostic imaging studies do not report adverse events or ionizing radiation exposure. I would recommend condensing this into a table summarizing the studies that discuss maternal/fetal adverse events, again to allow the reader to make some comparisons. Otherwise, this long paragraph is difficult to read and draw conclusions from.	We have emphasized this point in the Conclusions section of the report. We have streamlined the text on fetal maternal outcomes, but have not used a tabular format because we do not think this information can be easily forced into such a format.
TEP Reviewer #1	Results	Data are clearly presented in tabular format. In Tables A-E under the heading "N studies", add "affected/unaffected" for the column heading.	We have added this descriptor, as suggested.
TEP Reviewer #2	Results	The amount of information presented is overwhelming and I believe will serve to render this document useless to many readers.	We have streamlined the Executive Summary quite aggressively. We also plan to pursue the publication of separate manuscripts based on the report for specific populations; we hope that this will make our main findings more accessible.
TEP Reviewer #2	Results	I would like the authors to make some more conclusions. The manner in which the information is presented is, in my opinion, too non-objective. For example, the Alvarado score is essentially useless in children but the authors present the data in a form that some readers may consider this to an objective scoring system to utilize in the pediatric population.	We have tried to outline the clinical and future research implications of our work for practice and future research. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #2	Results	Page 40, test performance of multivariate diagnostic scores - authors should include a table to describe low risk rules. The decision to solely focus on the high risk rules, in my opinion, is a mistake.	We have reported this information, both in the Executive Summary and the main text of the report. Perhaps the reviewer failed to notice the second table on low risk rules.
TEP Reviewer #2	Results	Tables 1-12 - difficult to follow. Lots of extraneous information.	The tables presented in the Executive Summary have been streamlined; we have opted to keep the original level of detail in the main text (and we report complete data in the Appendix).
TEP Reviewer #2	Results	The decision to limit Table #13 to studies with five or more studies seems arbitrary. I am confused as to why the authors have done this.	The tables would become very unwieldy if we presented all information in the main text tables. The complete information is summarized in the Appendix.
TEP Reviewer #3	Results	Overall feel this was appropriate. The complexity of the literature and variables are such that extensive tables are useful for confirming the textual statements.	Thank you. No action required.
TEP Reviewer #4	Results	Figure B – the first box states “Citations retrieved from MEDLINE”. The search strategies describe searching PubMed, EMBASE, Cochrane, and CINAHL. Pardon my ignorance, but do all these databases fall under the global MEDLINE heading?- where did the Reviews (30 studies) come from?	These databases are separate (but their content highly overlapping). As stated in the Draft Report, we only presented evidence from MEDLINE in the original version. We have now included evidence from all databases. Systematic reviews were identified in MEDLINE, but each review had used a different set of databases to identify studies. We have provided complete information in the search flow chart of the Final Report.
TEP Reviewer #5	Results	The amount of detail was appropriate. Characteristics of the studies were clearly described. Key messages were explicit and applicable. Figures, tables and appendices were adequate.	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
Public Reviewer #2 Stefan Sauerland	Results	All analyses on diagnostic laparoscopy in the present report examine negative appendectomy rates but only the number of patients undergoing appendectomy are taken as denominator when calculating the rate. For example the negative appendectomy rate for the study by Bruwer et al. on page 74 is stated to be 2 of 11 but the laparoscopic group included 18 patients so 2 of 18 would be the correct rate. Using only a subgroup of patients as denominator violates the intention to treat principle and neglects the fact that laparoscopy has spared some patients from unnecessary appendectomy by obviating appendectomy completely. In the analysis of diagnostic laparoscopy versus no laparoscopy page 74 only two RCTs are included Decadt et al. and Gaitn et al.. It remains unclear why the RCT by Morino et al. was excluded Ref.No. 2650 as it examined exactly the same medical question.	We have applied the intention to treat principle where relevant (e.g., in randomized studies) and where applicable (e.g., when the number of events and the number of patients randomized refer to the same group, regardless of actual treatment received). The (common) practice of using a numerator of observed events among those complying with treatment (sometimes this is all that is available from the studies) and a denominator of total patients assigned (while excluding events in non-compliant patients) is not valid. Treatment effect estimators (e.g., risk differences, odds ratios, etc.) that rely on these proportions do not estimate a meaningful causal effect. The study by Morino et al. has been included in the Final Report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Are the implications of the major findings clearly stated? Not that I can see.	We have attempted to describe our findings as clearly as possible. We consider the following to be our key findings: • The literature on the diagnosis of acute appendicitis is large, but consists almost exclusively of studies assessing the performance of individual tests. • The evidence can support fairly strong conclusions about the performance of individual tests. Imaging tests have adequate test performance, while clinical signs and symptoms and laboratory tests used in isolation have lower discriminatory capacity. • The evidence is largely insufficient to support comparative effectiveness conclusions because studies assessing more than two test strategies on the same population are few and have evaluated different test comparisons. • More research is needed to evaluate the comparative performance and effectiveness of individual tests, test combinations and integrated diagnostic algorithms, to identify potential modifiers, and to evaluate the impact of testing strategies on patient-relevant outcomes, resource utilization, and harms. • Decision and simulation models using information from this review could inform the design of future studies and guide decisionmaking. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Are the limitations of the review/studies described adequately? Yes, but they could be more concise and using a language which is easier to understand. How do I interpret that studies consist "almost exclusively at moderate risk of bias (primarily due to differential and incomplete verification)"? I can understand it but not everyone. And What is the implication of such a statement? I guess the researchers should make an analysis on the subset of studies with high quality!.	We have streamlined the exposition, as suggested. Definitions of risk of bias items (including our operational definitions of differential and incomplete verification) are presented in the Appendix. Detailed analyses that have been limited to studies of low risk of bias for various items are described in the main report.
Peer Reviewer #1	Discussion/ Conclusion	In the discussion, did the investigators omit any important literature? I think the investigators can make a better job in analysing the diagnostic process in patients with suspicion of appendicitis and put the results into this framework. What algorithm is most optimal? What role does the clinical diagnosis, score, imaging, laparoscopy has in that algoritm? Selective vs routine imaging?	We discuss the need for future research on diagnostic evaluation algorithms (largely adopting the reviewer's perspective). We do not make clinical recommendations. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Is the future research section clear and easily translated into new research? The language is once again very technical and can be clearer. What is paired test vs parallel arm design? There is only a general notion that future research should adhere to established reporting guidelines. I think the authors could be more precise and give clear and concise information on what criteria is needed - prospective, patients with abdominal pain of <7 days duration, not only operated patients, criteria for histopathological diagnosis, follow up of non-operated patients, etc.	There is a balance between making specific future research needs recommendations (which are by necessity technical) and maintaining a simple level of exposition. We have tried to simplify a bit more and have addressed some of the specific research needs outlined by the reviewer.
Peer Reviewer #1	Discussion/ Conclusion	Another crucial point is the natural history of untreated appendicitis. You mention that "the impact of differential and partial verification on our results depends on the natural history of appendicitis; specifically on whether – and how often – cases of appendicitis can resolve on their own and the rate of recurrence among such cases". We know from indirect evidence that spontaneous resolution is common. Many patients with abdominal pain can have appendicitis that can resolve untreated. If detected they will probably be operated. We therefore need to define what instances of appendicitis that need treatment. This is a research need for the future. Other consequences are:	We agree that the natural history of appendicitis is important in interpreting existing evidence and in planning future studies. We have addressed this issue in the Discussion section of the report. We do not make recommendations about studies comparing alternative treatments because the comparative effectiveness of treatment strategies was out of the scope of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	A. Randomised trials between early imaging or laparoscopy with conventional management tend to have more diagnosed cases of non-perforated appendicitis in the imaging or laparoscopic arm. This is a sort of work-up bias. For the studies of laparoscopy this difference is very large (I will come back to that). So imaging leads to more operations for non-perforated appendicitis and as a consequence the proportion of perforations and negative appendectomy gets smaller as the denominator gets larger. As a consequence the proportion of perforations is a questionable outcome. At least I would propose that you also report on the proportion of operations (or treatment) for appendicitis when comparing intense with less intense diagnostic workup. You will find that it is larger in imaging and laparoscopy compared with conventional management.	We do not think that the issue identified by the reviewer is a source of bias in the statistical/ epidemiological sense (at least not in well-designed randomized trials). We agree however that the results may be interpreted in a “biased” way by some investigators – we do not believe that our interpretation has been affected by such “bias”.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	B. Another consequence is that the observed increasing proportion of perforation with increasing duration of symptoms can be a consequence of selection of the more advanced cases due to the resolution of non-perforated appendicitis. You write “Untreated appendicitis can lead to perforation of the appendix, which typically occurs within 24 to 36 hours of the onset of symptoms” . There are two alternative hypothesis to the observed increasing proportion of perforation with duration of symptoms: 1) appendicitis is a progressing disease and eventually all appendicitis cases may perforate; 2) there are two kinds of appendicitis, one progressing to perforation which usually occurs before the arrival to hospital, and another that will resolve without treatment within 2-3 days . See PMID: 17180556. These hypothesis have very important implications on the interpretation of studies in appendicitis diagnosis. I think these hypothesis could be included in the background section.	As stated above, we agree that the natural history of appendicitis is important in interpreting existing evidence and in planning future studies. We have addressed this issue in the Discussion section of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	C. You also write that “More accurate and timely diagnosis of appendicitis can minimize the time to the indicated intervention (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)”. I agree that a timely diagnosis and treatment is important for many reasons, but no one has shown that the rate of perforation can be reduced. On the contrary. There is quite an extensive litterature that has not shown an increase in the proportion of perforations or complications associated with doctor’s delay. The most recent is PMID: 24509193. This shows that some hours of observation is safe. I think this may also be included in the background section.	We agree with the general point that the link between diagnostic performance and clinical outcomes is currently not empirically established. We believe that our current wording is consistent with this view and have highlighted this issue in the Background and Discussion sections of the report.
Peer Reviewer #1	Discussion/ Conclusion	D. This also has an implication for the value of repeat examination after in-hospital observation. You have not included this in you review, but there are some studies that show an improved diagnostic capacity for clinical variables when repeated after some hours of observation. Repeat examination after observation is safe and gives an improved clinical diagnosis. I think the value of repeat examination should have a section in this review.	As stated above, we agree that the natural history of appendicitis is an important factor in interpreting existing evidence and in planning future studies. This is the key idea behind repeat testing with deferred treatment, an issue that is now discussed more explicitly in the Discussion section of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	One of your conclusions is “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 compared to imaging.” I think you need to differentiate how diagnostic tests are used when you evaluate their diagnostic performance. The test with the highest discriminating capacity cannot be applied to all patients for whom appendicitis is considered as a differential diagnosis. I rather think we need to use many different diagnostic modalities, including clinical diagnosis and selective imaging.	We agree with this point. We have discussed the need for selective use of imaging tests, and the need to judiciously combine imaging tests with clinical and laboratory tests.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion (continued from above)	There must be some structure based on the probability of appendicitis. Terasawa has pointed out that CT will not perform well in populations with low or high prevalence of appendicitis. You mention that US can be used for triage, but also this is a waste of resources and illogical as US has low sensitivity. A screening test for triage should have high sensitivity to rule out appendicitis. You always need a clinical diagnosis first. So I think you play down the important findings that for instance combination of WBC+CRP has very high sensitivity. And this is also the most important role for clinical scoring systems. When you write that scores does not perform as well as imaging, it is not completely true. Many scores has a very low LR- in some groups. The scores can thus rule out appendicitis with high accuracy. I think you should comment on that in the text.	Thank you for these important points. We have discussed these issues in the report, but we have not fully adopted the suggested binary approach (sensitivity is “high” or “low”); instead we have paid more attention on quantitative results regarding the relative performance of various modalities.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Another important issue related to the natural history of appendicitis and the structured management is that I would encourage you to differentiate the diagnosis of advanced appendicitis, which need immediate surgery. Inflammatory variables perform much better in those cases. So if you can rule out advanced appendicitis (gangrenous or perforated) there is no immediate need for an early diagnosis or treatment and repeat examination after observation may even allow spontaneous resolution. There are some studies that have reported results for advanced appendicitis separately. It would give a better picture of the importance of the clinical diagnosis if these results were also analysed here.	We have addressed this issue in the Discussion section of the report. At this point, we do not think it is meaningful to focus on a selected set of studies that report on the suggested subgroup analysis. In addition, we would need to revisit more than 1000 full text papers to make sure we identify all relevant information for this post-hoc analysis.
Peer Reviewer #1	Discussion/ Conclusion	Under Limitations of the Evidence Bias you mention the problems with unadjusted analyses in the NRCS. I think you should mention the unclear selection mechanisms that makes the comparisons very unreliable.	It is not clear what the reviewer means by “selection”. The term is often used to denote “confounding” (which is the term we prefer), a problem we have discussed at length in the report. Selection bias, in the structural sense, has also been addressed.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Under Applicability of Review Findings you discuss the lack of studies in some populations. I think that this is even more evident if you would exclude “the surgical cohort” studies. There is a very high need for well conducted prospective studies in the elderly, pregnant women, and small children with abdominal pain where appendicitis is a possible cause.	We agree with this point and have stated it explicitly in the Discussion section of the report.
Peer Reviewer #1	Discussion/ Conclusion	Under the same heading you also write “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).” I agree completely about the surgical cohort but wonder about the time before the widespread use of imaging? I have claimed that the widespread use of imaging has led to a higher detection rate of milder appendicitis cases that does not need surgery. This may explain the low proportion of perforated appendicitis in the imaging studies. But other things has also happened since the 1980s.	We agree that this is a pertinent issue and have mentioned it in the Discussion. We also agree with the reviewer that the currently available evidence cannot address this question.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Under Conclusions p38 and p145. I am not sure what you mean by "The literature on clinical symptoms and signs, laboratory and imaging tests, and multivariable diagnostic scores is very large, but consists almost exclusively of studies at moderate risk of bias (primarily due to differential and incomplete verification) assessing the test performance of individual tests." I suspect that you mean that only populations operated for suspicion of appendicitis have complete verification? I would say on the contrary – the problem with the available studies is not incomplete verification (if you mean non-operated) but that the majority are retrospective, only operated patients are included, the gold standard is fuzzy.	We think that incomplete verification (a problem that is impossible to elimintate for a condition such as appendicitis) and the relatively poor design of available studies are potential sources of bias. We have clarified this in the Discussion section of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Discussion section Key Findings and Strength of Evidence Assessment As a clinician I am somewhat disappointed here. I need some guidance on what tests to use on what patients. I think preferably by an algorithm. In the passage about multivariable diagnostic scores you write "Of note, the majority of studies assessed scores that had been developed before the widespread availability of CT and US imaging." I do not understand what implication this have? I miss a conclusion on the studies that have compared CT, US and early laparoscopy versus conventional management. I think these randomised trials that test the primary issue on what tests beside conventional management should be presented somewhere in the Discussion section. You have bypassed one of the most important questions – do we need imaging at all? Should it be routine or selective?	We do not make clinical recommendations in this report. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines. We believe that we have addressed the conceptual issues identified by the reviewer to the extent it is possible to do so.
Peer Reviewer #1	Discussion/ Conclusion	Future research needs. I miss randomised trials comparing selective and routine imaging. There is a need of defining optimal algorithms for a structured management, which can define the role of both clinical scores, imaging and laparoscopy. We need all diagnostic modalities but should define their place.	We have provided more extensive and specific guidance for future research in the revised report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	I guess you are right – there is limited data to draw any conclusions. More and better research is needed.	Thank you. No action required.
Peer Reviewer #2	Discussion/ Conclusion	Please see my general comments. I am worried that the report tried to artificially portray diagnostic strategies in isolation. Furthermore, the biases noted in prior studies are really a reflection of the heterogeneity of the true population for which diagnostic strategies are intended.	Unfortunately, it is often not possible to meaningfully assess the test performance of combinations of diagnostic tests because such information is either not reported or is reported inconsistently (e.g. different combinations of tests are assessed in each study or conditional test performance is not reported). That said, we agree that clinical signs and symptoms are almost always used in combination (with other signs and symptoms and with other tests). We think that it would be very cumbersome for primary studies to report all possible conditional test performance measures – and this probably explains the poor and inconsistent reporting. However, we think that studies could fit multivariable diagnostic models and report their results (e.g., using standard logistic regression methods). We have suggested this as a possibility in the Future Research Needs section of the report.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	Discussion/ Conclusion	The implications are not clearly stated. The authors leave the reader with the main conclusion that the quality of the evidence is, in general, not strong. That is a correct statement. However, there is also the conclusion that radiologic evaluation is more accurate than clinical evaluation. It should be stressed that this is not clearly associated with improved outcomes. I am concerned that this study may lead to even higher reliance on radiologic testing that is clearly not required for the majority of patients.	We agree that these are important issues, but we think that the reviewer is taking a somewhat strong position ("higher reliance on radiologic testing that is clearly not required"), which (as he admits) cannot be supported (or refuted by current evidence). We have addressed these issues in the Discussion section of the revised evidence report.
TEP Reviewer #1	Discussion/ Conclusion	Discussion and conclusions are appropriate and future research clearly defined and justified.	Thank you. No action required.
TEP Reviewer #2	Discussion/ Conclusion	Implications of major findings are not presented in clear manner.	We have tried to explicate the implications of our results, both in terms of the summary of the currently available evidence and the planning of future research. We do not make clinical recommendations in this report. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #2	Discussion/ Conclusion	Limitations of included reviews are presented adequately.	Thank you. No action required.
TEP Reviewer #2	Discussion/ Conclusion	In my opinion the authors have ignored important literature, especially as it applies to the pediatric population (specifically studies that examine the utility of a Low PAS score and the Low Risk appendicitis rule). Furthermore, the authors completely ignore the recent literature on implementation of clinical care pathways/guidelines into the Adult and Pediatric ED. These studies are useful as they describe the method of risk stratification of patients using clinical scores and the subsequent decreased reliance on diagnostic imaging without significant impact on safety outcomes.	Please see our responses to related comments by the same reviewer, regarding the inclusion/exclusion of specific studies. In brief, the studies mentioned have been included and analyzed properly. Their results are presented in a way that we think balances parsimony with comprehensiveness. Regarding studies of clinical pathways, we included only those that provided information on diagnostic performance and those that compared alternative test-and-treat strategies with respect to clinical outcomes (benefits and harms). We did not consider single group (non-comparative) studies of pathways that do not report test performance information to be within the scope of the report.
TEP Reviewer #3	Discussion/ Conclusion	Yes. No. I would think. The suggestions are appropriate. Trying to get these studies done is not easy and given the litigious nature of USA work. They will be hard to get by IRB or patient acceptance for enrollment in an emergency situation. I suspect this is part of the reason the literature doesn't reflect better comparative analysis.	Thank you for this comment. We have noted this practical difficulty in the Future Research Needs section, as suggested.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #4	Discussion/ Conclusion	Discussion: the authors do an excellent job of summarizing the large amount of data presented in the Results section. I especially like the section on "Future Research Needs" because the authors state explicitly what is lacking and furthermore give a prescription for study type (ex: paired test designs), how to conduct them, what to report, etc.	Thank you. No action required.
TEP Reviewer #4	Discussion/ Conclusion	Table F and G – Do these tables belong in the Discussion or the Results section?	We find that strength of evidence tables are best placed in the Discussion section, because strength of evidence assessment is based on unavoidably subjective judgments (which we have made every effort to explicate), as well as the empirical evidence. For that reason we have retained Tables F and G in their original location.
TEP Reviewer #4	Discussion/ Conclusion	Conclusions: no major changes necessary.	Thank you. No action required.
TEP Reviewer #4	Discussion/ Conclusion	Implications are clearly stated Limitations are described very well. I think this has been a very comprehensive and critical review of what has been done and what has not been done, but the future research section is what needs more work and there need to be more specific details of what should be done to fill in the gaps and answer the questions.	Thank you for these comments. We have provided additional specific details in the Future Research Needs section of the report.
Peer Reviewer #1	Clarity and Usability	Clarity and Usability: Is the report well structured and organized? In view of the enormous amount of literature and the large number of factors and variables treated I think it is as good as it can be.	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #2	Clarity and Usability	The report is extremely well organized around the key questions but I am not sure the key questions are the most important questions to be addressed to move forward with management strategies in patients with right lower quadrant pain.	Thank you for your comments regarding the report's organization. No action is required in response to that comment. Regarding the Key Questions, we note that they were developed through a stakeholder driven process, involving patients, frontline clinicians, and representatives of professional societies. At this point in the review process, it is not possible to modify the Key Questions. We encourage the reviewer to visit AHRQ's nominations website (http://effectivehealthcare.ahrq.gov/index.cfm/submit-a-suggestion-for-research/) to propose any additional questions that they would deem more clinically relevant. Details about the topic nomination process are provided on the Web site.
Peer Reviewer #3	Clarity and Usability	As commented above, I highly doubt that the average clinician will be able to use this report to inform practice decisions. For that purpose, a more clear summary has to be presented and a more decisive interpretation of the data. As far as policy, the report may aid in the issuance of guidelines for imaging, which would be useful. I find the executive summary and the main body of the report extremely duplicative. I think the executive summary should be significantly shortened and should stress the results and interpretation.	Should AHRQ decide to produce a clinical or patient guide on the basis of the report, we would be happy to offer our help. We have tried to streamline the Executive Summary, but this was a large project and we want to be sure that someone who only reads the Executive Summary will get a good sense for the conclusions of the report.
TEP Reviewer #1	Clarity and Usability	Report is well structured and organized.	Thank you. No action required.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #2	Clarity and Usability	Clarity and Usability: I do not believe that the report in its current form will lead to a significant impact on policy or practice decisions. What the authors have done is develop a very complete systematic review but have not made concrete conclusions or provided a clear path for future research initiatives. The future research section is well written, however at the same time too vague.	We agree the current systematic review – and any systematic review for that matter – cannot automatically produce clinical recommendations. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines. We share the reviewer's hope that the report can form the basis for future research. We have tried to make the future research needs section more concrete.
TEP Reviewer #3	Clarity and Usability	As to the first question- yes. Policy or practice decisions- no. There is no clear clinically relevant finding that would help a surgeon decide on a specific clinical pre-operative workup and operative strategy. The only conclusions I draw are that clinical impression plus selective imaging remains the best approach at present and that observation with delayed imagining is not dangerous overall. Among the problems out of the scope of the intentions of these authors is the cost versus benefit. Even the suggestion of universal MRI or CT scanning for this extremely common problem represents an enormous financial cost to a patient or the health care system. Studies balancing those factors need to be addressed, but obviously in a different forum	We agree about the importance of considering costs alongside clinical impacts and have recommended this as a natural next step for future research. We also agree that systematic reviews do not automatically produce clinical recommendations. The goal of the EHC program is to present evidence syntheses to clinicians and other stakeholders, so that they can use this evidence to inform practice. Professional societies and others may use EHC evidence syntheses to form the basis of practice guidelines.

Commentator & Affiliation	Section	Comment	Response
TEP Reviewer #4	Clarity and Usability	Very well organized and presented.	Thank you. No action required.