Comparative Effectiveness Research Review Disposition of Comments Report

Research Review Title: *Antinuclear Antibody, Rheumatoid Factor, and Cyclic-Citrullinated Peptide Tests for Evaluating Musculoskeletal Complaints in Children*

Draft review available for public comment from February 14, 2011 to March 14, 2011.

Research Review Citation: Wong KO, Bond K, Homik J, Ellsworth JE, Karkhaneh M, Ha C, Dryden DM. Antinuclear Antibody, Rheumatoid Factor, and Cyclic-Citrullinated Peptide Tests for Evaluating Musculoskeletal Complaints in Children. Comparative Effectiveness Review No. 50 (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. HHSA 290-2007-10021-I). AHRQ Publication No. 12-EHC015-EF. Rockville, MD: Agency for Healthcare Research and Quality. March 2012. Available at: 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 comparative effectiveness 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 email. At the conclusion of the public comment period, authors use the commentators' submissions and comments to revise the draft comparative effectiveness 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.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	General Comments	This report addresses an important topic - the utility of ANA, RF, CCP in children with MSK pain. These tests are not well-understood by general practicioners and this report therefore has the potential to significantly improve the use of these tests. The intended audience is presumed to be general pediatricians, but it is not explicitly stated and this information should be added.	We have incorporated this suggestion into the Introduction (p 1): "The report is intended for a broad audience including: primary care physicians who may consider ordering ANA, RF, or CCP tests in a child with MSK pain; health payers who provide coverage for these tests; and parents or caregivers who would like to know whether these tests can determine if their child does or does not have a particular disease."
Peer Reviewer #1	General Comments	The Key Questions appear to be appropriate. However, it appears that this review primarily addresses undiagnosed MSK pain - perhaps swelling can be removed from the wording of the questions?	The wording of the key questions was determined after extensive discussion with the technical expert panel and AHRQ (prior to and following public comment). The wording of the questions cannot be changed. We have added a statement in the summary of the evidence that none of the studies specifically looked at joint swelling.(p. 13)
Peer Reviewer #1	General Comments	ES-3 - Key Question 3.1 - would suggest that Key Question 3.1 include positive definition (or range) as ANA titer does impact the test properties. Would suggest that ANA titer be addressed in each Key Question response.	There is no good evidence to support the reviewer's suggestion that titer matters – and the titer chosen by labs as being positive was not consistent– and most papers do not look at this. We have reported the positive threshold for ANA titers that were used in the relevant studies.
Peer Reviewer #1	Introduction	Would suggest highlighting that many labs have moved away from the standard Hep-2 ANA assay and are using ELISA and other modalities that do not provide interchangeable results. It is imporant to highlight that these other assays have not been rigorously evaluated and physicians should ask for the Hep-2 assay specifically.	We have indicated that IIF is the gold standard for ANA testing and have referenced the ACR consensus statement (p 2). We have also indicated that IIF and EIA may not be interchangeable (p 3).
Peer Reviewer #1	Introduction	Would suggest that the authors site the ACR Consensus Statement on the different types of ANA assays.	We have referenced this statement.
Peer Reviewer #1	Introduction	Would also suggest that the authors site some of the most common causes of false positive ANAs (thyroid disease, intercurrent illness) and RF (viral infection, etc) here as well.	It is speculative to list reasons for false positive ANA, particular in children. Similarly there is insufficient data to conclude that the reasons for false positive RFs in children are the same as for adults. We have not added this information to the background section.
Peer Reviewer #1	Introduction	Estimates/ranges of costs of these assays should be addressed as well.	We agree this information may be of interest to some decisionmakers; however, costs of the tests were beyond the scope of this report.
Peer Reviewer #1	Methods	The methods are appropriate.	Thanks

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #1	Results	The results are overall clear and concise.	Thanks
Peer Reviewer #1	Results	Page 18 - 3rd paragraph states that 2 of the studies assessed different ANA assays and that these are indicated in Appendix F - however I do not see the results in the Appendix	These results appear in Appendix D. We have made this correction.
Peer Reviewer #1	Results	Page 25 / F-2 - a positive RF currently excludes the diagnosis based on the current ILAR criteria - so the tables as labelled do not make sense. This change in JIA criteria should be discussed / clarified to be consistent.	We have changed the section in the introduction (p 2) to indicate how RF is used in the different classifications and we have re-labeled the diagnostic criteria in the tables to reflect this.
Peer Reviewer #1	Discussion	The implications for the target audience should be more explicitly summarized - ie these tests should not be ordered by generalists in the setting of undiagnosed MSK pain?	EPC reports are intended to summarize the evidence and provide context for clinicians and other decisionmakers to understand the strength and limitations of the evidence. The systematic review is not intended to make clinical guidance recommendations which may require additional regional or individual considerations, such as patient values and available resources.
Peer Reviewer #1	Discussion	Although there were no cost data found by the authors, the costs of these tests and of the work-up of false positive results should be evaluated in the results. This should also be included as an important part of the Future Research Agenda.	Costs of the tests and related costs of work-up of false positive results were beyond the scope of this report. The report provides information on false positive rates, which could be used by researchers interested in determining the incremental costs of the tests.
Peer Reviewer #1	Discussion	Would suggest that rather than scheduling a short duration between the index test /reference standard, various durations should be evaluated as well (given the data in adults looking at duration between positive ANA and diagnosis of clinical ANA) as there is interest in the future significance of these tests and this could be part of their utility (eg if a child has an ANA of 1:320 at age 7 what does this mean for them at age 20, 30, etc).	We have removed this recommendation from our report. In terms of whether a positive RF or ANA in a healthy child today will mean anything in the future, we have incorporated this into our recommendation for future research (long-term prospective cohort studies).
Peer Reviewer 1	General	This report is well-structured.	Thanks.
Peer Reviewer #1	General	Would suggest a more explicit discussion of the results and their implications. With these changes I believe this report will provide very useful information for generalists evaluating children with undiagnosed MSK pain.	The systematic review is not intended to make clinical guidance recommendations which may require additional regional or individual considerations, such as patient values and available resources.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #2	General	This is an extremely useful report. The target audience is implied but not formally stated. It would include clinicians (who wish to use the ANA, RF or CCP tests in specific clinical situations), payers (who wish to determine appropriate utilization of resources), patients (who wish to determine the strength of evidence that they do or do not have a particular disease, researchers (who wish to see where better information is needed to improve utilization of these common tests). The Key questions are well thought out and clearly stated.	Thank you. We have incorporated this suggestion into the introduction (p 1).
Peer Reviewer #2	Introduction	The introduction is clear serving to familiarize the reader with the clinical issues, the population being studied and the tests that will be used.	Thank you.
Peer Reviewer #2	Introduction	It would be helpful if a few more details are provided for the testing in this section. For instance, on page 23, lines 23-31, the authors describe the indirect immunofluorescence test for ANA. While this is still a standard test, it is highly subjective with results varying from one laboratory to another.	We have incorporated this suggestion: "The assessment of fluorescence is based on a visual inspection and as a result may be somewhat subjective and vary from one laboratory to another." (p.2).
Peer Reviewer #2	Introduction	The studies that are included in this report also include cases where the ANA test was performed by Enzyme Immunoassay (EIA). Indeed, in many reference laboratories, EIA is the main test that is used to screen. The EIA methods vary from one manufacturer to another because there is no standard for which antigen preparations needed to be included, nor is there a standard as to what concentration(s) of the relevant antigen reparations should be used. Nonetheless, the EIA method should also be noted in general terms in the introduction.	We have reworded this section (p.3, top) to address the reviewer's comments.
Peer Reviewer #2	Introduction	Lastly, newer multiplexing assays are available and are also used for ANA testing. However, since the latter method was not used in the studies included in this report, it is not necessary to mention it in the introduction.	No change.
Peer Reviewer #2	Introduction	Finally there is a trivial error noted on line 29, an extra space is present in IgG (Ig G).	We have made this correction.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #2	Methods	The authors used a two-stage screening method for study selection. The initial broad screening had each article screened by two independent reviewers. They used the title and abstract to include (including the unsure articles) or exclude articles. I assumed they used the same criteria that were described in Table 1 on page 29. However, Table 1 is not referred to until the next paragraph where they discussed the second level of screening. Because it is not stated in so many words, and I wondered if other criteria were used by them.	The initial broad screen was conducted using a smaller number of and appropriately broader criteria. The following amendment has been made (p. 8): "…two independent reviewers who assessed the relevance of the study based on its title and abstract using prespecified broad screening criteria. Articles were excluded if they were judged clearly 1) not primary studies reporting on prevalence of conditions or diagnostic test accuracy or impact, 2) not on ANA, RF or CCP tests, or, 3) did not include a pediatric population."
Peer Reviewer #2	Methods	The authors also state on page 29 that discrepancies in decisions were resolved through discussion or third party adjudication. I wondered if this was a rare event, or if such adjudication were needed often. For the second level of screening, the full text of each article was examined. The inclusion and exclusion criteria in table 1 and their detailed form in Appendix C are appropriate for this study.	Inter-rater agreement is calculated during the pilot phase of the selection process as part of our procedure to improve consistency and reliability. Because complete agreement is required for all final decisions, this calculation is not made for the selection process. The need for third party adjudication was rare once the pilot phase of level 2 screening was completed.
Peer Reviewer #2	Methods	The definition and diagnostic criteria for the outcome measures were appropriate.	No change.
Peer Reviewer #2	Methods	However, it was disappointing that so little could be gleaned about their questions in this regard. The statistical methods are appropriate.	We agree. No change.
Peer Reviewer #2	Results	The tables are excellent. I appreciated the footnotes and legends. The figures were clear and I benefited from the hierarchical summary ROC curves.	Thank you.
Peer Reviewer #2	Results	However, I recommend expanding the information in tables 2 and 4 about papers where an ANA ELISA (EIA) method is used. The cut-off used for that test should been mentioned in the column for the positive threshold. It is not listed in the following cases: Table 2, page 38 Fawcett; Table 4, page 40 Fawcette (as above), and Nordal (here a "titer" is noted, but ELISA results are almost alwasy expressed as a cut-off of >1.0 etc, the value listed here "titer >1:101" is incorrect. There is no such titer and it is unlikely that a cut-off of 1.101 would be used since the cut-offs are usually derived from the performance of a standard control).	We have used the IIF in our analyses. Therefore, we have removed references to ELISA in the tables as we do not feel it adds additional useful information.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #2	Discussion	One thing I would like to know more about is if there is a breakdown by ages. The age range from 0 to 18 years is extreme. Clearly newborns barely make immunoglobulin whereas 18 year olds have adult levels. As I read the data, I couldn't help thinking that the remarkable range of some information reflected this fact. For instance, among healthy children, the range of data was that from 0 to 18% (with a median of only 3) had positive ANA results. This also seemed an odd figure to me. For instance, Kavanaugh et al. published a review of the ANA literature in Arch Pathol Lab Med 2000;124:71, they reported that at a cut-off titer of 1:40, 20-30% of "healthy" individuals were positive and at 1:80 the figure was about 10-12%. More recently in a study of "healthy" women, Sjowall et al. in J Rheumatoat 2008;35:1994-2000 reported number of greater than 40% at a 1:40 titer. I understand the variability and that one finds vastly different numbers, but zero is a huge surprise. I suspect that the lower numbers are from children under the age of 5 or so. But I couldn't figure this out from the data. If there could be a table showing a breakdown by age or age groups the titer data might be more meaningful.	We examined the studies that reported 'zero' and provided a possible explanation where we could. The potential effect of age on test accuracy is addressed in KQ4. Unfortunately, no studies were identified that reported test results by age. Our recommendations for future research include the need to examine potential modifiers of test accuracy including age.
Peer Reviewer #2	Discussion	Another concern I had was whether there may have been a bias with a positive RF or CCP result that might have increased the chances of the authors of the original publication placing that patient in the category of JIA.	As all the children were diagnosed based on accepted clinical criteria and not on the test positivity we think this is a low risk of index and reference test review bias.
Peer Reviewer #2	Discussion	The future research clearly lays out the major issues that need to be dealt with. In addition, I recommend studies that use both a classic HEp-2 FANA and compare that directly to ELISA and the newer multiplex assays in a large prospective trial would be very useful. In the future, methods will likely evolve to the newer platforms so the significance of an "ANA" may depend on the type of ANA test used.	While this would be an interesting line of research, it is not one of the recommendations that we would made based on the report.
Peer Reviewer #2	General	The report is quite well structured and its main points are clearly presented. The conclusions are important to discourage the more routine use of RF and CCP to screen for JIA in particular.	Thanks.
Peer Reviewer #2	General	Some caution about the variability of the types of ANA testing available may be helpful.	We have made changes to the introduction to address the reviewer's concerns.
Peer Reviewer #3	General	According to the "structured abstract", the objectives of this review were (1)"To assess the test performance of ANA, RF, and 'CCP' tests in children and adolescents with	Thank you for these observations. We agree with many of them and have reworded the conclusions of the report to better reflect the paucity of evidence on the

Commentator & Affiliation	Section	Comment	Response
		undiagnosed musculoskeletal pain and/or joint swelling compared with clinical diagnoses of pediatric SLE and JIA." A second objective was "to explore the difference in test performance for accuracy modifiers including age, sex, race/ethnicity, co-morbidities, and recent infections". The third objective was "to evaluate the impact of test results on clinical decision-making and clinically important outcomes such as referrals, ordering of additional tests, clinical management, and anxiety experienced by children and parents". I appreciate the desire on the part of pediatricians to determine the utility of these tests for the diagnosis of children with musculoskeletal symptoms. Unfortunately, as this review proves, there are no reliable data to address the question. Although the prevalence of musculoskeletal pain in pediatric patients may be estimated or guessed, there are no data as to how often the tests are ordered by clinicians and for which indications. There are also no data presented concerning what is done in response to the test results. These tests may be over-ordered and cause more distress than they are worth. Alternatively, they may be under-ordered and lead to missed diagnoses of SLE or RA. It is even possible that the tests are ordered completely appropriately, as determined by each clinician's ability and prior experience, and thereby help them decide on the next step in the evaluation of these difficult-to-diagnose patients. There are no studies presented in this manuscript that distinguish between these three possibilities. The researchers identified 10,512 citations and were able to include only 29 in the review based on their selected criteria. Of these 29 studies, the majority are published in journals that are electronically unobtainable from one of the world's largest medical libraries. In addition, nine of the manuscripts were written before 1990, long before many of the present techniques to detect ANAs, RF or anti-CCP antibodies came into use. After reviewing more than 10,000 citations, the authors did not find a single, high quality manuscript that addresses the objectives of this review.	usefulness of these diagnostic tests in healthy children with undiagnosed MSK pain.

Commentator & Affiliation	Section	Comment	Response
		The authors conclude: "Based on the available evidence, these (serological) tests should not be considered diagnostic tools by themselves." The authors ***should*** have concluded that "because of the absolute lack of any reliable evidence whatsoever, ***no*** conclusions can be drawn about the performance of these laboratory tests." This review has tremendous potential to be misinterpreted, misquoted and misunderstood. If it is used by insurers as justification to deny re-imbursement for these tests, then this review will cause significant harm to our patients.	
Peer Reviewer #3	General	Please consider the following issues: 1. The problem starts with the first objective: "To assess the test performance of ANA, RF and anti-CCP antibody tests in children and adolescents with undiagnosed musculoskeletal pain…" **The results of these tests, when obtained to evaluate a patient whose only complaint is musculoskeletal pain, can not be used to either establish or exclude any diagnosis. No child can be said to have (or not have) any disease (SLE, JIA or any other disease) with *only* the combination of musculoskeletal pain and a positive laboratory result. Therefore the sensitivity and specificity of these tests, used by themselves to evaluate musculoskeletal symptoms, is 0. This fact is certainly known to all rheumatologists.**	Perhaps this is known to all rheumatologists; however, this is not the target audience.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	General	2. If the laboratory results can not by themselves establish or exclude a diagnosis, why do clinicians order the tests? **<u>The tests are being used by health care providers (who have widely varying ability and experience in the field of pediatric musculoskeletal diseases) as markers for autoimmunity.</u>** The results are being used to help determine the next course of action: Should this patient be referred to a more experienced clinician or a specialist in the field? Should this patient be sent for radiological studies looking for evidence of synovitis that was either missed or not detectable on examination? Because each clinician's experience and expertise is different, the indications for ordering these tests may vary from one clinician to the next.	This is the reviewer's opinion.
Peer Reviewer #3	General	3. Why is the prevalence of musculoskeletal pain in the pediatric population relevant to this review? The real question is: "How often is an ANA, RF or anti-CCP test ordered on a pediatric patient with musculoskeletal pain?" The authors have no idea and seem to assume that the tests are ordered "commonly". On page 2 of the manuscript, the authors write: "An ANA test is commonly used to screen for autoimmune conditions…" and refer to reference 4. Ironically, reference 4 is a magnificent review of "Chronic musculoskeletal pain in children: assessment and management", by Clinch and Eccleston. The referenced manuscript does not even mention the use of serological tests in the evaluation of children with pain syndromes.	Thanks for pointing out the incorrect reference; we have made the correction. We believe that an understanding the prevalence of MSK pain in the pediatric population and prevalence of pSLE and JIA provides important background for potential role that ANA, RF, and CCP tests might play.
Peer Reviewer #3	General	4. I am not sure that the authors have a clear understanding of ANA testing. The ANA test is performed by hundreds of different laboratories across the United States. However, antinuclear antibody measurements are not the same as, for example, a sodium determination, for which a machine can measure an absolute concentration in the serum. The results of an ANA test depend on many different factors, including the dozens of different kits that are commercially available, as well as the expertise of the laboratory that performs the test. A task force established by the American College of Rheumatology in the late 1980s, led by Eng Tan, attempted to develop guidelines to permit standardization of ANA testing. The task force	Regardless of the task force recommendations, the labs in our studies have not adopted the standard of 30% positive at 1;40 dilution. The objective of the report was not to determine the utility of ANA testing in pSLE; it was to determine the utility of ANA testing in children with MSK pain. We have revised our tables and conclusion to reflect this. We have reworded the introduction to address changes in ANA testing (p 2-3).

Commentator & Affiliation	Section	Comment	Response
		involved many different laboratories, testing the same serum samples. The task force concluded that the ANA test should be standardized such that at a 1:40 dilution, 30% of a "normal" adult population should have a positive test. At a 1:160 dilution, 5% of this population should have a positive test. These results are not "false-positive" results. They are an inherent part of the ANA test procedure. To my knowledge, a similar standardization has not been performed for the pediatric population. Given the inherent limitations of ANA testing by indirect immunofluorescence, the manuscripts cited in this paper, reporting the prevalence of ANAs in the pediatric SLE population seem really, really good. For example, in the study by Wananukul et al, 15% of healthy children had ANAs detected at 1:40 and 3% had ANAs at 1:160. In contrast, 91% of children with lupus had ANAs $\geq$ 1:40. In the world of pediatric rheumatology, life probably does not get any better than this! Surprisingly, this bit of good news concerning the utility of ANA testing in lupus patients is down-played by the authors. Although this manuscript considers ANA testing by indirect immunofluorescence, the fact is that probably fewer than 50% of pediatricians in the United States can even order an ANA by indirect immunofluorescence. Many of the national laboratories screen for ANAs using solid phase assays. According to a recent report by an ACR Ad Hoc committee the performance of these assays for the diagnosis of autoimmune diseases in adults (no less children) is unknown. The authors do not address the issue of changes in ANA testing at all.	

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	General	5. Similarly, I am not sure that the authors are familiar with modern tests for rheumatoid factor. They unfortunately seem to lump all tests together: "The presence of RF's is typically determined by agglutination assays, nephelometry, or ELISA. Agglutination tests detecting the IgM-RF are commonly used in laboratory diagnosis in RA in adults. This assay method employs latex, charcoal, or human erythrocytes as carrier molecules to which human or rabbit IgG is bound." Some hard numbers and a few references would be very useful here. How many labs really still use charcoal or human red blood cells in their assays? The manuscripts that use some of the earliest tests for rheumatoid factors have little to add to the objectives of this manuscript.	We have reworded this section to address the reviewer's comments.
Peer Reviewer #3	General	6. The authors state that the major implication of the present manuscript is: "Based on the available evidence, the tests (ANA, RF, and anti-CCP) should not be considered as diagnostic by themselves." This statement is well-known to all rheumatologists and probably all clinicians. Musculoskeletal symptoms and a positive serological test do not establish or exclude any diagnosis. The authors can not pretend that the 29 papers used in this manuscript are the basis of this conclusion.	Perhaps this is well known to all rheumatologists, but perhaps it is not well known to all clinicians.
Peer Reviewer #3	General	7. I do not agree with the authors' suggestions for future research. There is no reason to conduct prospective cohort studies or diagnostic randomized trials in children with "musculoskeletal pain". It is absolutely clear that these tests will perform poorly in a population that has a low prevalence of systemic autoimmune disease. In addition, the suggestion that "studies examining ANA to diagnose SLE should explicitly describe whether a positive ANA is part of the diagnostic criteria" is non-sensical. Where would you find a cohort of "ANA-unknown", pediatric SLE patients? If you did find such a population of patients, do the authors really think that the probability of having a positive ANA will be ***lower*** in a cohort of children with four ***other*** SLE criteria?	This is the reviewer's opinion. We can infer this conclusion from the data but until someone looks critically at how the tests perform in this clinical situation, we do not know for sure. We have removed this from the future research section.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #3	General	**I think that the most important, potentially answerable, question for the future is: Did the health care provider get the information he or she wanted when the serological test was ordered and the results came back? The health care provider orders these tests, trying to distinguish autoimmune diseases from other causes of musculoskeletal pain. Did the test result lead to the correct additional tests? The correct referral? The correct final diagnosis? The correct treatment? The tests themselves are relatively inexpensive. The real question is: Did the results provide the clinician with useful information and was the patient well-served?**	We agree that these are some of the most important questions of the review and this was the purpose of KQ5 (to assess the impact of the tests on diagnostic decisionmaking including referral). Unfortunately, we did not find any research to address the question.
Peer Reviewer #4	General	a. General Comments: The report is clinically meaningful, especially in guiding general practitioners in the evaluation of musculoskeletal complaints in children, clarifying the lack of evidence to be using these tests as diagnostic screening tools.	Thank you.
Peer Reviewer #4		Overall, the target population is defined, but there is confusion for this reviewer how many studies looked at the question of joint swelling (as most of the literature reviewed seemed to be about MSK pain). The use of the term "undiagnosed" MSK pain and/or joint swelling is a bit misleading (as it seems most MSK pain never are formally diagnosed or explained and likely go into the idiopathic/non-specific cateogry).	None of the studies looked specifically at children with joint swelling; we added this to the description section of the included studies (p 13). In the one cohort study (Eichenfield et al) that looked at the prevalence of MSK pain in children, most children had a specific diagnosis – but most did not have a rheumatic disease (76%).
Peer Reviewer #4		Lastly, the audience for this report may get the impression to justify the use of these tests to confirm the diagnosis of rheumatic disorders of JIA and jSLE, which seems appropriate if the person is understanding of the clinical picture associated with these disorders but not if a practitioner is not familiar.	We have re-stated our results and conclusions to make it clear that there is insufficient evidence to support the use of these tests in children with MSK pain but no other symptoms suggestive of an inflammatory arthritis or connective tissue disease.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4		The key questions seem appropriate, although there is considerable overlap in answering the questions (i.e. explaining inflammatory and non-inflammatory etiologies of MSK pain are within the same category). Key question 1 and 2 are self-explanatory in the need to answer. For key question 3 with its different elements, the only concern is the context of the broader target population- is it any and all MSK complaints (which includes joint swelling alone) in children or just MSK pain? (See intro comments). Key question 4 would have been nice to have answered but I think that the lack of literature did not support any conclusion as noted by the authors. Also, key question 5 presents a legitimate concern of all screening tests, but again as the report shows, it has not been studied at all.	General observations of the reviewer that don not require any changes.
Peer Reviewer #4	Introduction	The introduction is good. As noted above, the issue arises when the focus of this section is about MSK pain (and not necessarily joint swelling). While I see that the search terms had included joint swelling, most of the literature evaluated in this report was about MSK pain. I suspect that no study has ever focused use of these diagnostic tests in children with joint swelling alone. If that is the case, it may need to be mentioned in the introduction (or results) and this may change some of the terms used in the key questions (i.e. may need to use joint pain with or without swelling, not the current terms of 'MSK pain and/or joint swelling').	No studies specifically examined children with joint swelling. We have added this to the results section on p. 13.
Peer Reviewer #4	Introduction	Also, the statistic that 15% of patients with JIA do not report pain (page 1, lines 25-27) is referenced with a jSLE study (#62). This is in contrast with the information provided in results section (page 12, lines 32-35), which states 16% with JIA present with pain, the exact opposite of the introduction information. Also, this same study (#71) does report that joint swelling is most predictive of JIA. This argues for my change in the definition of the target population as MSK pain with or without swelling, excluding the joint swelling alone.	We have added a sentence in the results section to address the differences in the reported statistics (p. 12): “In addition, the same study observed that among the 76 children diagnosed with JIA, only 12 (16 percent) included pain as part of their main complaints. Earlier observations by Sherry et al. stated that 14 percent of patients with a confirmed diagnosis of JIA reported no pain. Although the numbers differ, both studies confirm that the absence of MSK pain does not rule out a diagnosis of JIA.”
Peer Reviewer #4	Introduction	The background information on the different tests was well done and thorough.	Thanks.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Methods	The methodology section is well written and logical. The explanation of the search strategy is clear. From my understanding of performing systematic reviews of this nature, the statistical methods are appropriate. I think that it was a good idea to have done the extra subanalyses for the JIA categories when data were available.	Thanks
		The only complaint about the methodology is the definition of JIA used in attempts to provide a more unified definition than found in the literature. In doing this, I do think that sensitivity and specificity of these tests for JIA within the report are less than accurate. I understand was trying to be addressed using the subanalyses and the authors tried to explain it within their discussion.	We have reworded the introduction to explain how the clinical criteria used for diagnosis are similar and where they differ.
Peer Reviewer #4	Results	Given the lack of direct evidence, the information presented from the indirect evidence was thorough. The format of using tables to go over the different study characteristics helped to explain the methodology used and how well a study met the criteria for this review. The results overall answered the key questions as thoroughly as possible, but did show the inadequacies of studies in viewing the target population.	Thank you.
	Results	Additionally, the report does give a good job to explain the heterogeneity of the disease specific populations and the control groups used. They do try to address this as an issue within the results and the discussion, especially with regard to the variability of the control groups and how they may not be representative of the "target population".	Thanks
Peer Reviewer #4	Results	For the figures of the hierarchical summary receiver-operating characteristic curve for different tests and diagnoses could be improved with a detailed figure description. Without reading the body of the report, the figures are not intuitive to understand.	We have removed these figures. We agree with the reviewer that they are not intuitive and, on reflection, we do not think they add useful information to the results.
	Results	After looking over the references included and excluded, I cannot see how these can be changed given the clear stringent criteria they were using.	Thank you.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Discussion/Conclusion	The implications of the report are well stated. The major finding is clear for the target population (children with MSK pain) that the use of these tests is inappropriate. As for the explanations for the use of these tests in JIA and jSLE populations, the discussion and conclusion do point out that these tests alone should not be diagnostic, especially given the heterogenity of the populations (both disease-specific and control) used in reviewed studies.	Thank you.
Peer Reviewer #4	Discussion/Conclusion	The explanation of the report's limitations, especially when using case-control studies, was very good and should be highlighted. The authors gave a very clear, concrete explanation in the applicability section and in table 13. The summary and discussion section does go through a set of recommendations based on this report that are appropriate and reflect the literature to date. I assume that most pediatric rheumatologists also would agree that it coincides with their sentiments about using these tests in general for screening and confirmatory purposes for rheumatic conditions.	Thank you.
Peer Reviewer #4	Discussion/Conclusion	The future research section is clear, but it would be helpful if outline concrete examples for each suggestions to help guide future researchers in the field with their study designs. It would be interesting to have these recommendations be addressed in the many disease-based registries being done within pediatric rheumatology.	We have reworded this section to provide better guidance, but we have not provided concrete examples. The disease based registries likely would not address any of our key questions other than the incidence of +RF and ANA in JIA. The registries do not include children with MSK pain, which was our target audience.
Peer Reviewer #4	Conclusion	In terms of the conclusions, I think they are appropriate given the available literature they could review.	Thank you
Peer Reviewer #4		I do think that as far as these tests being used for confirmation of the specific diseases, I would say the evidence is less sound than the conclusions they present. This is especially the case for JIA patients, where there is so much heterogeneity when using the ILAR classification that the use of these tests may not be all that useful for diagnosis.	We agree with the reviewer and we have reworded our grading of the evidence and our conclusions

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #4	Clarity and usability	f. Clarity and Usability: Overall, the report is written well, but at times seemed like there was a bit of inconsistency in presentation of the results (i.e. as if multiple authors did various sections, the use of medium versus moderate, etc.). For example, the format regarding not having direct evidence for a specific test and the target population would be better if the same wording was used. For the one study of RF in JIA among children with MSK complaints which included the target population, this should be clearly stated at the beginning of that section to potentially set it apart.	Thank you for these comments. We have revised the appropriate results sections.
Peer Reviewer #4	References	Additionally, some references were missing in the report (page 21 and page 23).	We have added these references.
Peer Reviewer #5	General	a. General Comments: I found this a generally well written and clearly presented report. It was somewhat repetitive in places but this may be a feature of the way in which these reports are structured.	We have tried to reduce the repetitiveness as much as we felt was appropriate.
Peer Reviewer #5	Introduction	b. Introduction: This was a clear introduction to the review but I think it could be improved by adding some more details on the intended use of these tests. Presumably the aim is that they may allow an earlier diagnosis of the target condition than is possible based on current clinical criteria?	We are not sure why these tests are requested for children with MSK pain and no other symptoms of inflammatory arthritis or connective tissue disease. Therefore we cannot provide details.
Peer Reviewer #5	Methods	Search strategy: I have some concerns that the search appeared to include a diagnostic filter. This may have resulted in some studies being missed. Other than this the search appeared extensive and appropriate.	We have addressed this in the limitations section of the report (p 36). Our electronic searches identified over 11,000 records; we also searched for grey literature. We acknowledge that we may have missed some studies.
Peer Reviewer #5	Methods	Inclusion criteria: Participants: I am unclear about what you mean with "index test results". Study design: I think it would be clearer to just state studies of any design that included at least 2 participants - the info on sens and spec is covered under outcomes. Index tests: Do you mean studies of ANA published before 1980 were excluded? Outcomes: I found the bit about key question 5 confusing, suggest deleting bit about numerical data as presumably studies without numerical data would not have been excluded?	Thank you for these suggestions. We have incorporated them into Table 1, p. 8. By index test results we mean results of ANA, RF, and CCP tests. We have clarified this in Table 1. We reworded the study design section. Re pre-1980 ANA studies: yes, these were excluded. We have stated this in the table. We have reworded the outcome description for KQ5.
Peer Reviewer #5	Methods	Quality Assessment: QUADAS is the name of the tool not an acronym and so does not need to be spelled out. It would be helpful to include details of the criteria used to judge each QUADAS item as an appendix.	We have made this change to the way we refer to QUADAS. We have added the tool and the criteria to the appendix.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Methods	Data Analysis. I have concerns about presenting positive and negative predictive values. These statistics are almost meaningless without accompanying data on prevalence. I would suggest instead reporting likelihood ratios and for certain examples using these to modify pre-test probability of disease for different settings to give estimate of the post-test probability of disease.	We agree that PPV & NPV need to be interpreted in the context of disease prevalence and we have addressed this in the applicability section (p. 27 and the ES). We do not believe that reporting likelihood ratios would add clarity.
Peer Reviewer #5		I think that an important potential accuracy modifier which does not appear to have been considered in the review is symptom/disease duration. There is strong evidence that this affects estimates of accuracy of CCP for RA in adult populations and so I would imagine that this may act in a similar way in this review.	There is evidence to suggest that patients with a positive ANA and no clinical disease do not progress to the disease at least 10 years after the initial referral (Wijeyesinghe & Russell. Outcome of high titer ANA positivity in individuals w/o connective tissue disease: 10 year followup. *Clin Rheumatol* 2009 27:1399). Therefore, we do not believe that duration of symptoms is an important accuracy modifier in the context of this study.
Peer Reviewer #5	Methods	Rating the Body of Evidence How did you use the QUADAS assessment to obtain an overall judgement of the quality of the evidence? Could you add some more explicit details of this? I did not find the classification of direct and indirect evidence helpful. I am not sure if this is standard practice for AHRQ reported or whether this was how this grading was applied in this review. I think there is strong evidence that data from case control studies overestimtate estimates of accuracy. Therefore rather than providing indirect evidence I would suggest that these studies are at high risk of bias. This classification is used throughout the report and I would suggest revising it.	We agree with the reviewer's assessment of our use of direct/indirect evidence and have made the appropriate changes. Case control studies are at high risk of bias (changed from 'moderate'). The directness of the evidence is indirect because the measures of test accuracy are surrogate measures for clinically important outcomes.
Peer Reviewer #5	Results	General comments The results were a bit repetitive. I would suggest not starting each bit saying "no direct evidence". If you incorporate my earlier suggestion of removing this classification then you can just state how many studies of each design were included. Similarly, the sections on quality were repetitive and could be covered by a single section (see comments under quality).	We have reduced some of the repetition as per the reviewer's comments. In the results section, we have removed references to the 'indirect' evidence provided by case control studies.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Results	Flow diagram I think the direction of the arrow from the box on "studies identified from conference proceedings" is pointing the wrong way. The distinction between headings and subheadings in the boxes is not as clear as it could be e.g. pending and then language/reports pending retrieval.	We have confirmed the direction of the conference proceedings is correct (i.e., 109 proceedings screened in for review). We have reworded the box containing reports pending retrieval.
Peer Reviewer #5	Results	Suggest replacing direct and indirect evidence with details of study design (CC vs Cohort).	This has been done.
Peer Reviewer #5	Results	Methodological Quality As the quality of the studies is similar across key questions I think it would be sufficient to have a single section on quality rather than effectively repeating the same paragraph for each key question. A single table would also suffice if headings are added for target condition and index test.	We have made this change.
Peer Reviewer #5	Results	You state that selection bias is a potential problem in these studies but this is not specifically assessed by any of the QUADAS items - do you mean quality of reporting of selection criteria?	As is suggested by the AHRQ Methods guide for medical tests, we used the 14 items QUADAS tool, which includes "selection bias". We have referenced the methods guide in the methods section. Further, we have indicated in the narrative that few studies reported their selection criteria.
Peer Reviewer #5	Results	You appear to have scored case-control studies as unclear for the patient spectrum item. Could you explain why you have done this? I would expect a rating of "no" for these studies as in practice you would presumably be using these tests to help reach a diagnosis rather than in groups of patients with known disease status.	We agree that case control studies are at high risk of bias and have revised our assessment and the strength of evidence table to reflect this. We have changed our ratings for patient spectrum bias to 'No" in the individual quality assessment tables for each key question.
Peer Reviewer #5	Results	Key question 3.1, quantitative results. I find the term "overall" confusing as this suggests a summary estimate. You refer to three studies in this section but results are only presented for two and earlier you state that two studies met inclusion criteria.	We have removed "overall" from this section. We have corrected the reference to 3 studies; there were only 2 studies addressing this question.
Peer Reviewer #5	Results	Key question 3.2, study characteristics. The sentences about age are confusing.	We have reworded the sentence on age. "*For four studies the mean or median age of participants was under 18 years; two studies included a small number of young adults."*
Peer Reviewer #5	Results	Key question 3.2, Why did you report medians rather than pooling results?	A priori we decided not to pool the results due to anticipated heterogeneity in patient characteristics and test positive threshold values. We stated this in our methods section.
Peer Reviewer #5	Results	Key question 3.6, methodological quality - you use medium quality here, how does this differ from moderate quality?	We have changed this to moderate quality.

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #5	Results	Applicability You state that sensitivity and specificity does not vary across populations. Although in theory from a statistical point of view this may be the case this is not correct in practice. There is strong evidence evidence that various factors such as patient demographics, disease prevelance/severity and study design affect estimate of accuracy. This paragraph therefore needs to be reconsidered. I think that Table 13 would be more helpful if you consider different baseline prevalences based on different settings e.g. primary care, specialist clinic etc.	We have reworded this section. We have indicated that <1% prevalence of JIA or pSLE reflects the prevalence in a primary care setting. This figure (<1%) is from KQ1 and we have included the study reference.
Peer Reviewer #5	Discussion/Conclusion	I found the summary and discussion very repetitive of the results section. Other than that it was clearly presented.	We have reworded the summary and discussion to remove the repetition.
Peer Reviewer #5	Discussion/Conclusion	Future research. Could you give a bit more detail on what a diagnostic randomised trial would assess?	We have removed this study design from our future recommendations. It is unlikely such a trial would be conducted.
Peer Reviewer #5	Discussion/Conclusion	I query the recommendatin that future studies should include a short duration between the index test and reference standard. If the aim of these tests is to contribute to early diagnosis which I assume they are then in fact a longer time period between the index test and reference standard may be required. To evaluate whether these tests can be used early in the disease process they index needs to be applied in children with early symptoms of disease who are then followed up to determine whether they go on to fulfil clinical criteria for the target condition.	We have removed this recommendation. There is evidence to suggest that patients with a positive ANA and no clinical disease do not progress to the disease at least 10 years after the initial referral (Wijeyesinghe & Russell. Outcome of high titer ANA positivity in individuals w/o connective tissue disease: 10 year followup. *Clin Rheumatol* 2009 27:1399)
Peer Reviewer #5	Discussion/Conclusion	The final sentence of the future research section talks about assessing clinically important outcomes. How do you suggest this is done in the context of a diagnostic accuracy study? Or are you referring to other types of study at this point?	We are suggesting prospective cohort studies. This section has been reworded.
Peer Reviewer #5	Clarity and usability	This was generally a clearly presented report. My only criticism would be that it is a little repetitive and so by making the report more concise and removing some of the repetitive sections the structure and usability could be improved.	We have tried to address the repetitiveness where appropriate.
Peer Reviewer #7	General Comments	Dryden et al have done a systematic review of the performance and utility of ANA, RF and CCP testing in children with musculoskeletal complaints as screening tests for SLE and JIA. The key questions are appropriate and explicitly stated. The targed audience and population well defined.	Thank you

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer #7	Introduction	Good overview of MSK Pain, SLE and JIA. Appropriate mention of the different classification systems with references for JIA/JRA/JCA. minor: typo on line 54, page 1; .... effective immunosuppressive has>>> to improved outcomes> lead to? resulted in?	Thank you. We have made the correction on p.1
Peer Reviewer #7	Methods	Inclusion and exclusion criteria are justifiable. Search strategies appropriate and logical and comprehensive. There were pending studies listed in figure 2, and studies that needed to be translated. Given the limited number of studies that answered the study questions, it is possible that these include important information. Will their inclusion change the results?	We re-screened the foreign language titles/abstracts and have excluded those that clearly do not meet our inclusion criteria (literature reviews, case series, treatment). For the remaining 5 studies, they appear to be case-control studies. We do not believe that their inclusion would change the results of this review in any substantial way. We have revised Figure 2.
Peer Reviewer #7	Results	Results are presented in sufficient detail. The studies are well described. Figures are descriptive and helpful. I am not aware of overlooked studies. The appendix F shows subtypes of JIA, but has combined polyarticular JIA into a single entity. The ILAR criteria separate polyarthritis into two: RF negative polyarthritis and RF positive polyarthritis. Combining the two likely results in lower sensitivity. I suspect most studies do not separately present results for the RF negative and RF positive poly JIA, but this should be adequately addressed.	None of the studies provided subgroup data based on RF positivity. We have added this statement in Appendix F.
Peer Reviewer #7	Discussion	Discussion/ Conclusion: The major findings and implications are clearly stated. The limitations of the review are described. Future research directions are identified and are appropriate.	Thanks.
Peer Reviewer #7	General	Clarity and Usability: Report is well structured and organized. The points are presented clearly.	Thanks.