PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	FujiLAM for the diagnosis of childhood tuberculosis: A systematic review
AUTHORS	Olbrich, Laura Khambati, Nisreen Bijker, Else Margreet Ruhwald, Morten Heinrich, Nobert Song, Rinn

VERSION 1 – REVIEW

REVIEWER	Reviewer name: Dr. Martina Casenghi Institution and Country: Elisabeth Glaser Paediatric AIDS Foundation, Switzerland Competing interests: None
REVIEW RETURNED	28-Feb-2022

GENERAL COMMENTS	Generating solid evidence on performance of urine-based tests such as FujiLAM and Alere LAM represents a critical priority because non sputum-based tests have the potential to improve access to laboratory-based diagnosis of TB for children and can therefore contribute to address the pediatric TB diagbnostic gap. This manuscript provides an overview of the available evidence on the performance of FujiLAM for the diagnosis of pediatric TB. Importantly, it also highlights the evidence and methodological gaps that need to be urgently addressed in order to improve our knowledge and inform practival implementtaion in countries. Please find below this reviewer's comments on the manuscript, categorized into major and minor comments MAJOR COMMENTS Methodological approach The methodological approach used for this systematic review is not clear and needs to be improved. The authors stated that the objective of the systematic review was to assess the diagnostic performance of FujiLAM for pediatric TB, using AlereLAM as a comparator. However the Methods section does not provide neither a description of the comparison strategy that was selected for this systematic review nor the inclusion and exclusion criteria that have been established to include or exclude studies from the systematic review (and that should be consistent with the comparison strategy. In order to perform a rigorous diagnostic test accuracy comparison, authors should have choosen either a direct comparison strategy (i.e only studies that performed paired or unpaired head-to-head comparisons of Fuji and Alere LAM in the pediatric population) or in alternative an indirect comparison strategy (i.e. including all studies that have evaluated at least one of the tests of interest , FujiLAM or AlereLAM , in the population of interest). The approach currently used by the authors introduces an important bias because the performance of Alere LAM is calculated only based on the 2 publications that have included a head to head comparison

	of FujiLAM and AlereLAM rather than on all the possibly eligible publications that have assessed AlereLAM performance in the pediatric population. In contrast, the FujiLAM performance (pooled sensitivity and pooled specificity) has been calculated based on all the available FujiLAM studies in the pediatric population, irrespective of whether AlereLAM was used or not as a comparator (i.e see Barrio et al 2021) There are additional studies that have evaluated AlereLAM alone in the pediatric population (i.e no head-to head comparison with other comparator tests) and that have been excluded from this systematic review (just as examples: Osorio et al 2021, Schramm et al 2021, Kroidl et al 2015). Why have those studies been excluded? It is critical that the authors clarify which test comparison strategy they want to apply to the systematic review and provide a detailed description of inclusion and exclusion criteria that are consistent with the tests comparison strategy that has been selected. The comparison between FujiLam and AlereLam performance needs to be revised based on the revised methodological approach. Alternatively, the authors may decide to only focus the systematic review on the performance of FujiLAM and drop all data regarding the comparison with AlereLAM. But the value of the systematic review may decrease as the comparison between FujiLAM and AlereLAM is scientifically and programmatically relevant. Reporting of accuracy data -It would be important to add one table that shows the tests performance disaggregated by the 3 different TB case classifications (Confirmed TB, Unconfirmed TB, Unlikely TB). It is currently not possible to assess how the tests performed in the "Unconfirmed TB category". This is important as children in this category generally represent a very significant proportion of the sample size -Table 3 Add number of indeterminate and/or invalid results for each type of tests If those data were not reported by the original studies, please clearly state this and highlight it as one of the reporting limitations Statistical approach to calculation of pooled sensitivity and specificity -This reviewer does not have the needed statistical expertise to assess this aspect accurately but wonder if the pooled sensitivity and specificity of Alere LAM can be calculated based only on data from two studies . A statistician with expertise in DTA systematic reviews should be consulted MINOR COMMENTS Table 2 -Last column (Barrio et al)- 4th and 5th cell from the top: replace "sputum samples" with respiratory samples as authors specify that also nasopharyngeal/nasogastric aspirates were used) -6th cell: for the children that tested negative with TST/QFT, please specify if meeting only one of the conditions listed (two clinical criteria, X-ray consistent with TB, signs and symptoms of TB, close TB exposure, positive response to TB treatment) was sufficient for the children to be classified as unconfirmed TB. Please clarify the difference between clinical criteria and TB signs and symptoms (add a note to the table) DISCUSSION

	-Line 240- the authors state that " using the MRS is likely to underestimate specificity because the number of children with TB and therefore the proportion of true negative is underestimated" The sentence should be amended and should read: the MRS will underestimate the number of children with TB and therefore overestimate the number of true negatives (as MRS can misclassify pediatric TB positive cases as negative cases) In addition, the data included in this review do not really show such a difference in pooled specificity calculated using a MRS or a CRS (for FujiLam this is 87% with both MRS and CRS; and for AlereLAM is 83% and 88% with MRS and CRS respectively) Given the concerns regarding the methodological approaches and statistical analysis highlighted under major comments, the conclusions of the systematic review need to be reassessed once the methodological approach and analysis have been revised

REVIEWER	Reviewer name: Ms. Emily MacLean Institution and Country: McGill University, Canada Competing interests: None
REVIEW RETURNED	09-Feb-2022

GENERAL COMMENTS	Summary In this study, Olbrich and colleagues conducted a systematic review of FujiLAM's diagnostic accuracy in children. Over 100 studies were identified in the search, and results were available from 3 studies. At this point, it seems that FujiLAM has improved sensitivity compared to AlereLAM in children, although further work is needed to confirm this trend. This paper is a nice summary of the work already done and makes good recommendations for where further research is needed. I think conclusions are justified and qualifications about the limited data are well-stated throughout the paper. Specific comments are below. Major comments My only major comment is that I am a bit unclear about the pooling and why it was done. I would recommend against pooling or conducting a meta-analysis in these circumstances (i.e., 3 included studies), as there are probably insufficient studies here to get a proper estimate of variance. Regardless, there is no mention in the methods section that pooling of sensitivity and specificity was going to be attempted, rather, these results appear in the results section without preface. Can you explain the method used to pool the sensitivity and specificity? The abstract also states that no meta-analysis was performed but then also presents pooled sens and spec. If there were no meta-analyses undertaken due to between-study heterogeneity, why would it be necessary to pool those point estimates at all? Pooling without properly adjusting for sample size is not going to lead to the 'most right' answer… I think presenting the sensitivity and specificity +95%CI ranges with forest plots would be totally adequate. Minor comments Throughout the paper, when values from specific papers are being presented, please include a citation for the paper being referred to. E.g., in the section describing FujiLAM versus CRS performance, 3 sensitivity measures are presented but it is not evident from which paper each value was taken Line 39-40: Kind of awkward wording/tense in "was found to be" – maybe change to "has been observed" or similar? Line 43: the difference was not statistically significant though, right? So maybe add a qualifier about CIs being overlapping Line 64: as mentioned above, it's not clear why pooled sensitivity and specificity results are given, when on line 69-70 it says that meta-analysis was not performed Line 81-2: can you specify the age you mean by "the very young"?

	Line 87: would maybe suggest specifying that these TPPs include characteristics for diagnostic tools' use in children - as written, it sounds a bit like the TPPs are child-specific Line 96-7: this is a bit vague – I think it would help if it was specified over what time period this reduction was modeled to occur? Line 120: since the evidence base is so limited, did you consider using results from conference proceedings? Since a MA wasn't going to be done anyways, it could potentially be of value to show all available evidence. This is just a suggestion, not a strong recommendation by any means Line 125: are the answers for the QUADAS-2 bias assessment included in the supplement? If so, please indicate in this section; if not, I would strongly suggest including them in the supplement Line 135: were figures on indeterminate results or errors extracted? I think this would be valuable information for readers if it's available Summary measures and data analysis: were 2x2 tables reconstructed to do the pooling? Need to say how the pooling was done here. Line 144: I think there maybe should be a statement here about ethical approval not being needed as this was a secondary analysis etc etc Line 162: what exactly is meant by "value" – accuracy? Please revise in text Line 170: I find "microbiological investigations" to be a little bit unclear – what variable(s) is this describing? Line 176-7: this is kind of knit-picky but just to be super clear, I'd suggest changing the wording from "confirmed… by CRS" to something else, since CRS-positives aren't only confirmed cases. Maybe use wording more like CRS-positive or similar Discussion: I would maybe suggest the authors comment on what questions re: FujiLAM in kids are of the highest priority to address. The gaps that remain are peppered throughout the discussion, but I'd be curious what their recommendations are for immediate next steps Tables and Figures – clear and comprehensive

REVIEWER	Reviewer name: Dr. Peter Flom Institution and Country: Peter Flom Consulting New York, United States Competing interests: None
REVIEW RETURNED	14-Feb-2022

GENERAL COMMENTS	I confine my remarks to statistical aspects of this paper. These were fairly simple, but appropriately so, and I have only a couple of comments. First, the authors should state how the CIs were calculated for sensitivity and specificity. There are several methods of doing this and they can give different results. Second, in figure 2, I would remove a lot of the text to make the graph bigger. The columns for sensitivity and specificity can certainly be removed (or put in a separate table, if desired) as they are exactly what is in the figure. Peter Flom

VERSION 1 – AUTHOR RESPONSE

Dear Prof Choonara,
Thank you for considering our manuscript entitled
"The use of FujiLAM for the diagnosis of childhood tuberculosis: A systematic review"
We have revised our manuscript following the reviewers' comments, please find a point-by-point response below.

Reviewer #1:
Major comments
My only major comment is that I am a bit unclear about the pooling and why it was done. I would recommend against pooling or conducting a meta-analysis in these circumstances (i.e., 3 included studies), as there are probably insufficient studies here to get a proper estimate of variance. Regardless, there is no mention in the methods section that pooling of sensitivity and specificity was going to be attempted, rather, these results appear in the results section without preface. Can you explain the method used to pool the sensitivity and specificity? The abstract also states that no meta-analysis was performed but then also presents pooled sens and spec. If there were no meta-analyses undertaken due to between-study heterogeneity, why would it be necessary to pool those point estimates at all? Pooling without properly adjusting for sample size is not going to lead to the 'most right' answer… I think presenting the sensitivity and specificity +95%CI ranges with forest plots would be totally adequate.
¬ We thank the reviewer for this comment. Following the suggestion, we decided against presenting pooled estimates and report sensitivity and specificity for the individual studies.
Minor comments
Throughout the paper, when values from specific papers are being presented, please include a citation for the paper being referred to. E.g., in the section describing FujiLAM versus CRS performance, 3 sensitivity measures are presented but it is not evident from which paper each value was taken
¬ We have included the citations throughout to facilitate easy referencing.
Line 39-40: Kind of awkward wording/tense in "was found to be" – maybe change to "has been observed" or similar?
¬ We have changed the wording accordingly (line 39-40).
Line 43: the difference was not statistically significant though, right? So maybe add a qualifier about CIs being overlapping
¬ Following another reviewer's comment, we have decided against performing a head-to-head comparison between AlereLAM and FujiLAM as the main focus of the systematic review and have subsequently changed the wording of the principal findings of our paper. We do discuss the performance of the two tests and state that confidence intervals between the two tests overlapped, suggesting a lack of evidence for test superiority (line 259-270)
Line 64: as mentioned above, it's not clear why pooled sensitivity and specificity results are given, when on line 69-70 it says that meta-analysis was not performed
¬ As mentioned, we agree with this major concern and have now only reported the individual estimates with 95% confidence intervals.
Line 81-2: can you specify the age you mean by "the very young"?
¬ We were referring to children younger than five years. To make this clearer we have included this in the text (line 83).
Line 87: would maybe suggest specifying that these TPPs include characteristics for diagnostic tools' use in children - as written, it sounds a bit like the TPPs are child-specific
¬ We have changed the wording (line 88).
Line 96-7: this is a bit vague – I think it would help if it was specified over what time period this reduction was modelled to occur?
¬ We have included the time period as suggested (line 99).
Line 120: since the evidence base is so limited, did you consider using results from conference proceedings? Since a MA wasn't going to be done anyways, it could potentially be of value to show all available evidence. This is just a suggestion, not a strong recommendation by any means
¬ We appreciate this comment. Within the team, we had initially discussed and decided to exclude conference proceedings in our search strategy, mainly due to concerns that conference information may be incomplete and not enough for a ROB assessment. We still tried to be comprehensive in other ways by e.g. consulting experts in the TB diagnostic field.
Line 125: are the answers for the QUADAS-2 bias assessment included in the supplement? If so, please indicate in this section; if not, I would strongly suggest including them in the supplement
¬ We agree with the reviewer and have uploaded this as a supplement (please see Supplemental Table 1).
Line 135: were figures on indeterminate results or errors extracted? I think this would be valuable information for readers if it's available
¬ This information can now be found in the results section (line 180-182), as we agree these are important information for the reader.
Summary measures and data analysis: were 2x2 tables reconstructed to do the pooling? Need to say how the pooling was done here.

¬ We have now only reported the individual estimates.
Line 144: I think there maybe should be a statement here about ethical approval not being needed as this was a secondary analysis etc etc
¬ We have included a statement (line 145-146).
Line 162: what exactly is meant by "value" – accuracy? Please revise in text
¬ We have removed this line
Line 170: I find "microbiological investigations" to be a little bit unclear – what variable(s) is this describing?
¬ We have revised the text to make this more clear and added "microbiological investigations (specimen collected and tests performed)" (line 172-173).
Line 176-7: this is kind of knit-picky but just to be super clear, I'd suggest changing the wording from "confirmed… by CRS" to something else, since CRS-positives aren't only confirmed cases. Maybe use wording more like CRS-positive or similar
¬ We agree to use precise language. As we have decided not to report the pooled estimates, this specific phrase has been deleted.
Discussion: I would maybe suggest the authors comment on what questions re: FujiLAM in kids are of the highest priority to address. The gaps that remain are peppered throughout the discussion, but I'd be curious what their recommendations are for immediate next steps
¬ We believe the highest priorities for future studies include prospective evaluations on fresh specimens, direct head-to-head comparisons of FujiLAM to alereLAM within the same study population, specific subgroup analysis in children living with HIV and extrapulmonary disease, and recruitment from different geographical regions. We have summarized this in the "what this study adds" section (line 43-48), the conclusion of the abstract (line 74-80), and the final section of the discussion (line 278-284).
Tables and Figures – clear and comprehensive

Reviewer #2:
First, the authors should state how the CIs were calculated for sensitivity and specificity. There are several methods of doing this and they can give different results.
¬ We appreciate this important point. We have performed all analyses based on the raw data presented in the publications and used RevMan for both analyses and visualisation. To clarify this, we now included the following statement: "Point estimates and confidence intervals were calculated using the raw data provided by the original publication with the statistical software of RevMan (version 5, The Cochrane Collaboration, 2020)" (line 139-143)
Second, in figure 2, I would remove a lot of the text to make the graph bigger. The columns for sensitivity and specificity can certainly be removed (or put in a separate table, if desired) as they are exactly what is in the figure.
¬ We thank the reviewer for this important point. As we have omitted the comparison to AlereLAM, the figures should be much easier to read now.

Reviewer #3:
Major comments
Methodological approach
The methodological approach used for this systematic review is not clear and needs to be improved. The authors stated that the objective of the systematic review was to assess the diagnostic performance of FujiLAM for pediatric TB, using AlereLAM as a comparator. However the Methods section does not provide neither a description of the comparison strategy that was selected for this systematic review nor the inclusion and exclusion criteria that have been established to include or exclude studies from the systematic review (and that should be consistent with the comparison strategy. In order to perform a rigorous diagnostic test accuracy comparison, authors should have choosen either a direct comparison strategy (i.e only studies that performed paired or unpaired head-to-head comparisons of Fuji and Alere LAM in the pediatric population) or in alternative an indirect comparison strategy (i.e. including all studies that have evaluated at least one of the tests of interest , FujiLAM or AlereLAM , in the population of interest). The approach currently used by the authors introduces an important bias because the performance of Alere LAM is calculated only based on the 2 publications that have included a head to head comparison of FujiLAM and AlereLAM rather than on all the possibly eligible publications that have assessed AlereLAM performance in the pediatric population. In contrast, the FujiLAM performance (pooled sensitivity and pooled specificity) has been calculated based on all the available FujiLAM studies in the pediatric population, irrespective of whether AlereLAM was used or not as a comparator (i.e see Barrio et al 2021). There are additional studies that have evaluated AlereLAM alone in the pediatric population (i.e no head-to head comparison with other comparator tests) and that have been excluded from this systematic review (just as examples: Osorio et al 2021, Schramm et al 2021, Kroidl et al 2015). Why have those studies been excluded? It is critical that the authors clarify which test comparison strategy they want to apply to the systematic review and provide a detailed description of inclusion and exclusion criteria that are consistent with the tests comparison strategy that has been selected. The comparison between FujiLam and AlereLam performance needs to be revised based on the revised methodological approach.
Alternatively, the authors may decide to only focus the systematic review on the performance of FujiLAM

and drop all data regarding the comparison with AlereLAM. But the value of the systematic review may decrease as the comparison between FujiLAM and AlereLAM is scientifically and programmatically relevant.
¬ We appreciate this thorough feedback and detailed discussion of the methodological approach chosen in our review. We agree with the reviewer that the introduced bias in only presenting data on AlereLAM performance generated in the limited number of studies included here should be avoided. We have therefore decided to focus the review on summarising the available literature on FujiLAM and not make the comparison with AlereLAM as a main outcome of the review.

¬ However, as the reviewer stressed, it remains crucial to interpret FujiLAM's performance in relation to AlereLAM. Therefore, we do mention that two studies in the systematic review did a head-to-head comparison (line 188-192) and discussed the limitations of their findings in the Discussion (line 259-270). We also compared the estimates of FujiLAM performance from our systematic review to published evidence on AlereLAM in a Cochrane systematic review in the Discussion. Going forward, a key recommendation is the need for more direct comparisons between the two tests within the same study population (line 278 ff).

Reporting of accuracy data: It would be important to add one table that shows the tests performance disaggregated by the 3 different TB case classifications (Confirmed TB, Unconfirmed TB, Unlikely TB). It is currently not possible to assess how the tests performed in the "Unconfirmed TB category". This is important as children in this category generally represent a very significant proportion of the sample size

¬ Many thanks for pointing this out, we agree and have included a table outlining FujiLAM performance disaggregated by the diagnostic classifications (please see Table 4).
Table 3: Add number of indeterminate and/or invalid results for each type of tests aII those data were not reported by the original studies, please clearly state this and highlight it as one of the reporting limitations
¬ We agree with the reviewer and have included this information in the results section (line 180-182).
Statistical approach to calculation of pooled sensitivity and specificity -This reviewer does not have the needed statistical expertise to assess this aspect accurately but wonder if the pooled sensitivity and specificity of Alere LAM can be calculated based only on data from two studies . A statistician with expertise in DTA systematic reviews should be consulted
¬ This comment is duly noted and was raised by reviewer #1 as well. Following their suggestion, we decided to not report pooled estimates due to the insufficient number of studies and study heterogeneity. We instead report the individual study point estimates with 95% confidence intervals as stated in the original publications.

Minor comments
Table 2
-Last column (Barrio et al)- 4th and 5th cell from the top: replace "sputum samples" with respiratory samples as authors specify that also nasopharyngeal/nasogastric aspirates were used)
¬ We have adjusted accordingly (line 425).
-6th cell: for the children that tested negative with TST/QFT, please specify if meeting only one of the conditions listed (two clinical criteria, X-ray consistent with TB, signs and symptoms of TB, close TB exposure, positive response to TB treatment) was sufficient for the children to be classified as unconfirmed TB. Please clarify the difference between clinical criteria and TB signs and symptoms (add a note to the table)
¬ We included the information as requested (line 425).
DISCUSSION
-Line 240- the authors state that " using the MRS is likely to underestimate specificity because the number of children with TB and therefore the proportion of true negative is underestimated". The sentence should be amended and should read: the MRS will underestimate the number of children with TB and therefore overestimate the number of true negatives (as MRS can misclassify pediatric TB positive cases as negative cases)
¬ We thank the reviewer for suggesting a clearer and more concise phrasing and have changed the section accordingly (line 230).
In addition, the data included in this review do not really show such a difference in pooled specificity calculated using a MRS or a CRS (for FujiLam this is 87% with both MRS and CRS; and for AlereLAM is 83% and 88% with MRS and CRS respectively)
¬ We have decided not to report pooled estimates and have therefore deleted the respective line. Now, we discuss the individual specificity estimates of both tests for those studies that performed head-to-head comparisons (line 259-270).
Given the concerns regarding the methodological approaches and statistical analysis highlighted under major comments, the conclusions of the systematic review need to be reassessed once the methodological approach and analysis have been revised
¬ We thank the reviewer for the thorough review and suggestions, and have adapted our conclusions and the abstract accordingly. In particular, our conclusions now focus on the high specificity of FujiLAM

ranging from 84 to 93% in included studies, indicating its potential as a rule-in test and the need for more studies in children which perform prospective testing of fresh specimens, subgroup analyses for children living with HIV, and conduct direct paired comparison of FujiLAM with AlereLAM

¬ We thank the reviewer for suggesting a clearer and more concise phrasing and changed the section accordingly (line 74 ff): "The high specificity of FujiLAM demonstrates its potential as a point-of-care (POC) rule-in test for diagnosing paediatric TB. As an instrument-free point-of-carePOC test that uses an easy-to-obtain specimen, FujiLAM could significantly improve TB diagnosis in children in low-resource settings, however the small number of studies available highlight that further data is needed. Key priorities to be addressed in forthcoming paediatric evaluations include prospective head-to head comparisons with AlereLAM using fresh specimens, specific subgroup analysis in CLHIV, and extrapulmonary disease and studies in different geographical locations."

¬ Line 278 ff: "This review summarises the current evidence of FujiLAM, with the high specificity demonstrating its potential as a POC rule-in test for diagnosing paediatric TB. It reflects the current state of knowledge, highlighting that more data on FujiLAM in children are needed to understand the diagnostic value of this test in different groups at scale and suggests the priorities to be addressed in forthcoming evaluations. In particular, the need for prospective assessments that directly compare FujiLAM to AlereLAM in real-life settings, recruitment from several geographical regions, and subgroup analyses focusing on CLHIV and EPTB."

Yours sincerely,

Dr Laura Olbrich