Author's response to reviews

Title: Evaluation of chemiluminescence, toluidine blue and histopathology for detection of high risk oral precancerous lesions: a cross-sectional study

Authors:

- Shweta Ujaoney (shweta2503@gmail.com)
- Mukta B Motwani (motwanimukta@yahoo.co.in)
- Shirish Degwekar (drssdegwekar@yahoo.com)
- Vijay Wadhwan (pathogen@rediffmail.com)
- Prajakta Zade (drprajaktazade@yahoo.co.in)
- Minal Chaudhary (minal53@yahoo.com)
- Vinay Hazarey (vinay.hazarey@gmail.com)
- Tushar P Thakre (tpthakre@gmail.com)
- Manju Mamtani (manjuhemant@gmail.com)

Version: 3 Date: 11 January 2012

Author’s response to reviews: see over
January 10, 2012

To,

The Editor

BMC Clinical Pathology

Re: Submission of revised manuscript MS: 2259755885718795

Dear Sir/Madam,

Thank you very much for the critical and exhaustive review of our manuscript no. MS: 2259755885718795 titled “EVALUATION OF CHEMILUMINESCENCE, TOLUIDINE BLUE AND HISTOPATHOLOGY FOR DETECTION OF HIGH RISK ORAL PRECANCEROUS LESIONS: A CROSS-SECTIONAL STUDY” which we had submitted to your esteemed Journal for consideration of publication as a Research Article.

We wanted to place on record our sincere thanks to all the reviewers for a careful review and constructive suggestions which we believe have helped us improve the manuscript. We have redone some analyses, rewritten parts of the manuscript and have described these changes in a point-by-point detailed response letter at the end of this letter. In each response, we have also indicated where the corresponding change in the revised manuscript can be found.

We thank you very much for the opportunity to revise and resubmit our work to your esteemed Journal.

Sincerely

Manju Mamtani, MD
12023 Waterway Rdg
San Antonio, TX 78249, USA
Phone: (1) 210 602 5199
Fax: (1) 210 558 3177
E-mail: manjuhemant@gmail.com
RESPONSES TO REVIEWERS’ COMMENTS

Reviewer #1

Comment #1:
Frankly, the authors have done a lot and presented a not-very-well-known statistical analysis method (Hui and Walter’s multinomial latent class model). Unfortunately, the manuscript had some shortcomings, which are mentioned below:

Our response
We thank the reviewer for appreciating the work.

Comment #2
ABSTRACT
1.Background: Please clarify what you mean with “... technical...” in the second sentence.

Our response
We actually meant to address the issues related to the technical skill required to perform some of the screening procedures. However, we agree with the reviewer that the word ‘technical’ has other confusing connotations as well. For this reason we have now changed the word to ‘procedural requirements’.

Page 2, line 5

Comment #3
2.Methods:
a) Histopathology is not a screening method or tool. The authors should rephrase this sentence, i.e., “...the screening performance of chemiluminescence and toluidine blue, using histopathology as the reference standard”.

Our response
We fully agree with the Reviewer that, as currently practiced, histopathology is not a screening tool. However, in the latent class model analysis one does not consider any test as a reference standard. What LCA returns is the sensitivity and specificity (which we refer to as screening performance) of each imperfect test. Therefore, we had used the word ‘screening performance’. However, we appreciate the reviewer’s point of view and the fact that the sentence implies that histopathology is a screening tool. Therefore we have removed the word ‘screening’ from the sentence.

Page 2, line 7

Comment #4
b) In the 5th line, please correct the term “histopathological”.

Our response
We thank the reviewer for pointing this out. We have made the change.

Page 2, line 19

Comment #5
3.Results
Please state the sensitivity and specificity of the screening methods clearly.

Our response
Again, we are thankful to the Reviewer for this comment. We have made the suggested changes. (Page 2, lines 17-19)

Comment #6
4.Conclusion
This part needs to be written in a more academic language and style.

*Our response*
Thanks. We have now reworded this section. (Page 2, lines 21-22; page 3, lines 1-2)

Comment #7
INTRODUCTION
The main reason to conduct this study is not reasonable. The authors state that “…if the histopathological evaluation is itself subject to errors then the estimates of the sensitivity and specificity of the light-based protocols can be expected to be biased.” If we consider this statement true, then this approach can be proposed for every conventional or novel screening method. Unfortunately, in order to have scientific validity, we need to have a reference standard in such studies –even though the reference itself has pitfalls. In order to overcome this problem, a panel of pathologists which would provide an agreed histopathological decision for each lesion is suggested. So, this “purpose” is not suitable or appropriate to conduct the present research.

*Our response*
We respectfully beg to differ with this comment by the reviewer. There exists a substantial body of literature that deals with evaluating the screening/diagnostic tests when the “gold” standard is “alloyed” or when there is no reference standard. In a continued quest for more and more accurate tests for disease detection, we almost always encounter scenarios that use imperfect reference standards and the reviewer is right that each of such scenarios need to be subjected to more objective and unbiased assessment of the test accuracy without considering any test as the reference standard. While the choice of a test as reference standard can be understood to be a practical measure aimed at unifying test protocols, such measures may not be accurate even if they are precise. We therefore believe that it is important to shed off the bias implicit in treating a test as a reference standard and conduct a comparison of all available tests. LCA does exactly that and therefore we have persisted with our original reason to conduct the study.

Comment #8
PATIENTS AND METHODS
As far as I understand, the authors have used 2 screening methods: chemiluminescence and toluidine blue. On the other hand, on the last paragraph of this section, they have written that “..Therefore this model can be used only if there are at least three tests (number of parameters to be estimated = 7 and degrees of freedom = 7).” The authors have classified the lesions according to their toluidine blue staining degree. Dark staining lesions were considered positive, faint were considered equivocal, and unstained ones were negative. 1) TBLU group: positive if the lesions were dark stained, negative if the result was either equivocal or negative. 2) TBEQ group: positive if dark or equivocal, negative if unstained. 3) CHTB group: positive if the lesion was both CHEM-positive and TBLU-positive, negative if otherwise. Here, we have 2 screening tests, no matter how the authors classified the results.

*Our response*
We are extremely grateful to the reviewer for bringing up this point. Indeed, we agree that there are two tests – chemiluminescence and toluidine blue retention. However since these tests did not have binary outcomes, alternative binarization strategies using combined results yielded the four test protocols described. We agree that the TBEQ protocol was just another way to categorize the test results and therefore would be like correlated with the TBLU protocol. In the light of these comments, we decided to remove the TBEQ protocol from the analyses. Hence we have compared only four protocols in the revised manuscript – CHEM, TBLU, CHTB and HPMV. Using these four test protocols, we reran the LCA analysis and have edited Table 2 as well as the section titled “Comparison of diagnostic performance” to show the new results of LCA. We further agree that our original usage of the terms “tests” and “test protocols” was rather loose. In the revised manuscript we have corrected the language throughout the manuscript. (Page 6, last 4 lines; Table 2; Page 12, first paragraph)

Comment #9
Also, what had happened to “CHEM-positive and TBEQ positive lesions” in the authors’ subgrouping process?

Our response
Since we decided to remove the TBEQ protocol, in the revised manuscript we have also not reported the potential combination protocol resulting from CHEM and TBEQ.

Comment #10
The authors state that “…Histopathologic evaluation was done by two senior Oral Pathologists blinded to the clinical findings. The first pathologist evaluated each specimen at two time points.” However, there is not information about the second pathologist. In cases when there is a disagreement, what have the pathologists done to reach to a consensus?

Our response
The results of evaluations by the second histopathologist and the agreement of these results with those given by the first histopathologist are shown in Figure 2. How the results from these three evaluations were combined into a single histopathology status (HPMV) is shown in Figure 2F and described in section titled “Composite histopathological evaluation”. Regarding consensus, for the purpose of this research, a consensus was not deemed necessary.

Comment #11
The interevaluation agreement between 2 pathologists for each category of the classification was not clearly explained in the manuscript, and this part needs further clarification. The authors have presented the data analysis results of one pathologist in page 10: “..there was neither a bias in the histopathologist’s two evaluations nor a significant departure from variability at each time point as indicated by the Pitman’s test. Despite this, however, the Pearson’s correlation coefficient for scores at two time points by the same histopathologist was only 0.28.” Unfortunately, I have not received any information regarding the second histopathologist. These should be mentioned not only within the figure legends of Figure 2, but also within the text, as well.

Our response
We presented both inter- and intra-histopathologist agreement in our paper. For inter-histopathological evaluations, we had presented the data from the second pathologist as well in Figure 2B and 2D.
(columns titled H3 in yellow boxes). For intra-histopathologist agreement we presented Figure 2A and the portion of the text that the reviewer is here alluding to.

**Comment #12**
**DISCUSSION**
The authors declared that “Additionally, there is a need for an objective and more accurate diagnostic method when evaluating the histopathological specimens”. This is not one of the purposes of this study, and actually, I do not understand what the authors mean by “accurate diagnostic method for evaluation of histopathological specimens”.

*Our response*
We apologize for confusing the reviewer with this statement. Our statement was contingent upon the results shown in Figure 2 which shows intra- and inter-observer variability in histopathological evaluation and it was one of the study objectives to examine if such a variation exists. However, considering the comments of the reviewer, we have now removed that sentence in the revised manuscript.
Reviewer #2

Comment #1
The manuscript describes a comparison between several oral cancer diagnosis methods in subjects with clinically identified premalignant lesions in a tertiary-care setting. The manuscript attempts to deal with an extremely important aspect of emerging diagnostic techniques, namely the repeatability and accuracy of histopathological diagnosis, especially when used as a gold standard against which other methods are compared. Although there are several significant limitations of the study, including a patient population from which results may generalize poorly to a true screening setting and a small number of pathologists (2) used for comparison, the results are still likely to be of significant interest to researchers and physicians studying oral cancer diagnostics and the manuscript provides a methodological framework for future, larger studies. Additionally, these limitations are clearly and appropriately stated in the text. The possibility of using the single more cost effective test (TBLU) instead of the CHEM test is also of significance if it can be corroborated in the future. There are several statistical issues that must be addressed before publication. Additionally, if more pathologists can be included in this analysis the manuscript would be greatly strengthened.

Our response
We are extremely grateful to the reviewer for laying out the relevance of our work. We also agree with the reviewer’s comment that a larger number of histopathologists would enhance the value of the study. Unfortunately, due to feasibility reasons, generating those data is not possible at the present time.

Comment #2
Discretionary Revisions:
PATIENTS AND METHODS, Histopathological evaluation, in reference to figure 1H, in this image of a severe dysplasia, the basement membrane is not visible. I recommend replacing this image with an example where the full thickness of the epithelium is visible in order to demonstrate the histopathological diagnostic procedure.

Our response
We thank the reviewer for this suggestion. We have now substituted the figure with a new one.

Comment #3
Major Compulsory Revisions:
RESULTS, Composite histopathological evaluation: The authors should justify the choice of threshold of their binary separation of diagnostic categories. At this institution, are there significant differences in how patient with 0 and 1 scores are treated compared to patients with 2 and 3? Is there a more logical threshold based on clinical practice?

Our response
This binarization scheme was according to the recommended method by the WHO Working Group (Warnakulasuriya S et al, Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement.J Oral Pathol Med (2008) 37: 127–133). In our original submission this reference was missing and we thank the reviewer for pointing this out. We have now added that reference to justify the binarizing scheme. (Reference 18).
Comment #4
RESULTs, Composite histopathological evaluation: In figure 1 E, it appears that the “total category score” is a summation of the 0-3 risk category codes for the three evaluations by pathologists. The score only ranges between 1 and 7 however instead of between 0 and 9. Please clarify this point and alter the figure if necessary.

Our response
Thanks for pointing this out. We have added a statement in the legend to Figure 2E to clarify this point. (Page 21, line 11)

Comment #5
ABSTRACT, Conclusion: The statement: “There exists substantial variability in the histopathologically evaluation of oral precancerous lesions” is too strongly worded and too generalized based on the fact that only two pathologists were included in the study. This sentence, along with paragraph 2, sentence 4, should be modified to indicate that substantial variability existed within the context of this study. This will allow the reader to draw his or her own conclusions based on the presented data.

Our response
We are grateful to the reviewer for suggesting this change. We have made the suggested changes in the revised abstract. (Page 2, last paragraph)

Comment #6
RESULTs, Variability in reference standard evaluation: It appears the Kappa statistic was performed on repeated observations from one pathologist + a single observation from a second pathologist. This is potentially confounding two aspects of the data, the intra and inter-pathologist correlation. The authors should justify this point.

Our response
This is also a very important point raised by the reviewer. Using the Bland-Altman plots we found that there was no bias associated with the paired observations and the intra-class correlation coefficient for this dataset was close to 0. This indicated that for statistical purposes the two observations reflected two effectively independent evaluations (even if they were from the same histopathologist). Therefore, we conducted the kappa assessments on these effectively independent evaluations. We agree that this point was not clear in our original submission and thank the reviewer for pointing it out. In the revised manuscript we have now added this justification. (Page 10, lines 5-9) Also, we have restructured this section as well as Figure 2 to ensure better flow of thought and text.

Comment #7
RESULTs, Comparison of diagnostic performance: The authors should better define the HPMV as it is unclear. Is this the majority vote among the 3 readings (2 from the same pathologist and 1 from a different pathologist)? As mentioned before, it seems atypical to combine repeated observations with singular observations for a single consensus. The authors should justify this point.

Our response
Yes the majority vote was from three observations for the reasons explained above. We have added more explanation to the methods regarding this point in the revised manuscript. (Page 11, lines 15-19)
Reviewer #3

Comment #1
In their manuscript, the authors describe the results of a cross-sectional study evaluating the performance of histopathology reporting as well as chemiluminescence and/or toluidine blue staining on differentiating oral precancerous lesions. The study seems well constructed and performed, and a publication seems to be of interest for the readers of Clinical Pathology.

*Our response*
We are highly grateful to the reviewer for appreciating the importance of the work.

Comment #2
There are some minor revisions that needed attention: - In the results section of the abstract, chemiluminescence (and its performance) needs mentioning

*Our response*
Thanks. This was also suggested by Reviewer #1. We have now added this information. (Page 2, line 17)

Comment #3
- In the results section of the manuscript, the text often refers to Figure 1, which should be Figure 2 in reality

*Our response*
We apologize for this error. We have now corrected the manuscript accordingly.

Comment #4
- The discussion should also discuss the findings in relation to other screening methods (AF,...) as well as the results of "pioneers" of TB staining like Epstein et al

*Our response*
We thank the reviewer for this immensely important suggestion. In the revised manuscript we have now added this discussion. (Page 13, last paragraph)
Reviewer #4

Comment #1
While Bland-Altman plots are a good way of visually assessing agreement, Pearson’s correlation coefficient is not a valid method for quantifying agreement. There have been multiple papers on this topic. Valid approaches include quantifying agreement using an intraclass correlation coefficient from an appropriately specified model or the concordance correlation coefficient.

Our response
We agree with and thank the reviewer for this suggestion. In the revised manuscript we have reported the intra-class correlation coefficient. (Page 2, line 15; Page 10, lines 4-5).

Comment #2
There are multiple lesions per patient included in the analysis, but no mention is made of how this clustered data is handled in the analysis. Methods that take into account this correlation should be used, but it’s not clear that this was done. This comment pertains to all of the analyses done including the limits of agreement on the Bland-Altman plot.

Our response
In all our analyses, the unit of analysis is lesions and not persons. Indeed, in studies of similar disposition it is customary to estimate the performance measures for the lesions. We have used the same methods as are customary. Conceptually, clustering of lesions by individuals is unlikely to have a measure impact on histopathological evaluation since the histopathologists were provided with the biopsy slides (one per lesion) so they were blinded to individual identifiers.

To demonstrate this concept, we ran the Bland-Altman plot analyses again using the adjustment for clustered effects as described by Carstensen B, Simpson J, Gurrin LC. Statistical models for assessing agreement in method comparison studies with replicate measurements. Int J Biostat 2008; 4(1): 16. We found that the limits of agreement estimated using clustered data did not differ substantially from unclustered approach (Please compare the orange and purple lines in the adjoining figure). The full Stata output (based on notations used in the Carstensen et al paper is shown below. These analyses indicated that clustering had minimal, if any, effect on our interpretation in this study. Therefore, for simplicity of presentation we have persisted with lesions as our unit of analyses.

. xi: xtmixed y i.meth1 i.item || MI: || MIR: meth1, nocons var
  i.meth1    _Imeth1_0-1   (naturally coded; _Imeth1_0 omitted)
  i.item    _Iitem_1-55   (naturally coded; _Iitem_1 omitted)

Performing EM optimization:
Performing gradient-based optimization:
Iteration 0:  log likelihood = -633.72265
Iteration 1: log likelihood = -630.67615
Iteration 2: log likelihood = -630.51423
Iteration 3: log likelihood = -630.514
Iteration 4: log likelihood = -630.514

Computing standard errors:

Mixed-effects ML regression  Number of obs      =       198
                           -------------------
                              |   No. of       Observations per Group
Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
MI |      110          1        1.8          3
MIR |      198          1        1.0          1
-------------------

Wald chi2(55)      =    371.37
Log likelihood =   -630.514                     Prob > chi2        =    0.0000
------------------------------------------------------------------------------

y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imeth1_1 |  -2.353535   .9592464    -2.45   0.014    -4.233624    -.473447
   _Iitem_2 |   5.069246    4.38299     1.16   0.247    -3.521258    13.65975
   _Iitem_3 |  23.25384    4.38299     5.31   0.000     14.66334    31.84435
   _Iitem_4 |  20.38465   3.578697     5.70   0.000     13.37053    27.39877
   _Iitem_5 |   5.15386   3.578697     1.44   0.150    -1.860257    12.16798
   _Iitem_6 |   6.084614   3.578697     1.70   0.089    -.9295025    13.09873
   _Iitem_7 |   3.915386   3.578697     1.09   0.274    -3.182802     10.9295
   _Iitem_8 |   8.5       3.578697     2.38   0.018     1.485883    15.51412
   _Iitem_9 |   9.838456   3.578697     2.75   0.006      2.82434    16.85237
   _Iitem_10 |  21.02314   3.578697     5.87   0.000     14.00902    28.03726
   _Iitem_11 |   10.14618   3.578697     2.84   0.005     13.22011    17.16029
   _Iitem_12 |  21.02314   3.578697     5.87   0.000     14.00902    28.03726
   _Iitem_13 |   11.47695   3.578697     2.13   0.033    -3.937187    18.85237
   _Iitem_14 |   7.638491   3.578697     2.13   0.033     .6243745    14.65261
   _Iitem_15 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_16 |   13.059     3.266888     4.00   0.000     6.656017    19.46198
   _Iitem_17 |   7.30789    4.38299     1.66   0.097     16.85237
   _Iitem_18 |   7.73078    4.38299     1.76   0.078    -.8597139    16.32129
   _Iitem_19 |   8.976947   3.578697     2.51   0.012     1.962831    15.99106
   _Iitem_20 |  10.8154     3.578697     3.02   0.003     3.801287    17.82952
   _Iitem_21 |  11.38465   3.578697     3.21   0.001     4.462831    18.49106
   _Iitem_22 |   6.561561   3.578697     1.83   0.068    -2.028942    11.15206
   _Iitem_23 |   11.12314   3.578697     3.12   0.002      4.570532    17.69314
   _Iitem_24 |   9.415386   3.578697     2.63   0.009     2.401269    16.4295
   _Iitem_25 |   7.638491   3.578697     2.13   0.033    -3.937187    18.85237
   _Iitem_26 |   1.476955   3.578697     0.21   0.833    -6.656017    8.609019
   _Iitem_27 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_28 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_29 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_30 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_31 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_32 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_33 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_34 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_35 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_36 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_37 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_38 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106
   _Iitem_39 |   9.138491   3.578697     2.55   0.012     1.962831    15.99106

| _Iitem_40 | 14.20774 | 4.38299 | 3.24 | 0.001 | 5.617233 | 22.79824 |
| _Iitem_41 | 10.38465 | 3.578697 | 2.90 | 0.004 | 3.370532 | 17.39877 |
| _Iitem_42 | 1.730789 | 4.38299 | 0.39 | 0.693 | -6.859714 | 10.32129 |
| _Iitem_43 | 14.3616 | 4.38299 | 3.28 | 0.001 | 5.771093 | 22.9521 |
| _Iitem_44 | 27.41539 | 3.578697 | 7.66 | 0.000 | 20.40127 | 34.4295 |
| _Iitem_45 | 6.900017 | 4.38299 | 1.57 | 0.115 | -1.690486 | 15.49052 |
| _Iitem_46 | 9.392333 | 4.38299 | 2.14 | 0.032 | 0.8018299 | 17.98284 |
| _Iitem_47 | 32.57702 | 3.578697 | 9.10 | 0.000 | 25.5629 | 39.59113 |
| _Iitem_48 | 27.677 | 3.578697 | 7.73 | 0.000 | 20.66288 | 34.69112 |
| _Iitem_49 | 21.66932 | 4.38299 | 4.94 | 0.000 | 13.07881 | 30.25982 |
| _Iitem_50 | 15.37696 | 3.578697 | 4.30 | 0.000 | 8.362848 | 22.39108 |
| _Iitem_51 | 5.15386 | 3.578697 | 1.44 | 0.150 | -1.860257 | 12.16798 |
| _Iitem_52 | 7.57693 | 4.38299 | 1.73 | 0.056 | -1.013573 | 16.16743 |
| _Iitem_53 | 6.830772 | 3.578697 | 1.91 | 0.056 | -1.833446 | 13.48489 |
| _Iitem_54 | 18.20774 | 3.578697 | 5.09 | 0.000 | 11.19362 | 25.22185 |
| _Iitem_55 | 5.407702 | 4.38299 | 1.51 | 0.011 | -1.606415 | 12.42182 |
| _cons | 9.329039 | 2.535722 | 3.68 | 0.000 | 4.359115 | 14.29896 |

Random-effects Parameters | Estimate | Std. Err. | [95% Conf. Interval]
MI: Identity |
var(_cons) | 9.25e-22 | 5.07e-21 | 2.02e-26 | 4.24e-17 |
MIR: Identity |
var(meth1) | 60.26347 | 14.26908 | 37.88856 | 95.85177 |
var(Residual) | 15.41587 | 2.931447 | 10.61953 | 22.37848 |
LR test vs. linear regression: chi2(2) = 19.28 Prob > chi2 = 0.0001
Note: LR test is conservative and provided only for reference.

**Comment #3**
Generally a kappa statistic is calculated to show agreement among readers across categories of a variable. The presentation of kappa statistics within categories of the reference standard is therefore confusing and has me concerned that it was not calculated correctly. What is the kappa statistic for agreement between pathologists for the reference standard overall?

**Our response**
The reviewer is absolutely right that the kappa statistic summarizes the agreement but when the Siegel and Castellan’s method is used to compute the kappa, then for multi-category outcomes it is possible to estimate the per-category as well as overall kappa. We used the Stata program kap which reports the per-category as well as the overall kappa. The Stata output for analyses is pasted here:
```
.kap h1-h3
There are 3 raters per subject:
Outcome | Kappa | Z | Prob>Z
---------+-------+---+------
0 | 0.0350 | 0.60 | 0.2731
1 | 0.1222 | 2.11 | 0.0176
2 | 0.1328 | 2.29 | 0.0110
3 | 0.1008 | 1.74 | 0.0412
---------+-------+---+------
combined | 0.1126 | 2.57 | 0.0050
```
In our original submission, we had not reported the combined kappa. We thank the reviewer for asking us the combined kappa which we have now included in the revised manuscript. (Page 10, lines 14-15)

Comment #4
The interpretation of the AUC in the discussion (“a 9%-11% improvement over a single evaluation in the accuracy of diagnosis of a high risk lesion”) is not quite correct. The AUC can be interpreted as the probability that if 2 lesions are being considered, one high risk and one low risk, the high risk one gets the higher score. But it should not be interpreted as quantifying the accuracy of diagnosing a high risk lesion.

Our response
We have deleted this part from the manuscript as well as Figure 3 consequent to some other comments by other reviewers. However, we beg to respectfully differ with the reviewer. A very common interpretation of the area under an ROC curve is indeed the overall predictive accuracy of a test.

Comment #5
It seems as though 2 different LCA analyses were done, although this was a little confusing. If it is the case, the rationale was not clear.

Our response
The reviewer is right that we indeed conducted two different LCAs for two different purposes. First, we wanted to justify the use of a histopathological majority vote. Second, we estimated the performance characteristics of the test protocols as well as HPMV using another independent LCA analysis.