OPEN PEER REVIEW

Open Peer Review (OPR)

As a new feature for progressing towards transparency, we decided to open a new window for all of our editors as well as authors titled "**Open Peer Review**". We hope by this new facility, our reviewers will be more motivated and authors will be more satisfied with the review process. We believe that publishing our peer review reports could make a transparent and clear environment for all our efforts within a journal, but not all reviewers tend to publish their comments.

What is "Open Peer Review" process?

An "Open Peer Review" process is making the details of all review process (including reviewers, associate editors, and EICs comments) as "Public" as it is agreed by EIC, Authors, and reviewers.

Advantages of "Open Peer Review" process

- More transparency, constructiveness and tactful comments of the peer review process: leads to an increase in the quality of reviews
- More motivations for all involved roles in the review process
- Authors' satisfactions from the review process: Increases honesty between authors and reviewers
- Education of both authors and new students
- Prevents reviewers from following their individual agendas and leading to the detection of reviewers' conflicts of interests

You can find out more at:
https://sites.kowsarpub.com/kowsar/knowledgebase/category/tree.html#opr.html

Iranian Journal of Radiology
Article DOI: 10.5812/iranjradiol.86898
Published in: Iranian Journal of Radiology: 16(4); e86898

Peer Review Report for "Reliability Studies of Different Readers Using Computer Aided Diagnosis System in Ultrasound for Determining Identical Lesions in a Breast Phantom"

Author(s): Nari Choi, Bong Joo Kang, Na Young Jung, Sung Hun Kim, Jung Min Lee, Ga Eun Park

Review Timeline:

Submit Date: 29 Nov 2018
Revised Date: 19 Jun 2019
Accepted Date: 23 Jun 2019

Revision (0)

Here, you can see the **Reviewers**, **Associate Editors** and **EICs'** comments from the beginning to the end of the revision process.

REFEREE: Reviewer | **Revision** (0)

12 Jan 2019

Dear AE,
The authors have presented a well structured study with very good analysis. However, there are some basic questions that need explanation:
1- It is not clear why the kappa value of US (0.609) was significantly higher than CAD (0.411). In the method section it has been mentioned" The readers first detected the breast phantom lesion by US and then applied the CAD system. " I mean when the readers used CAD, they had analysed the US feature before, so kappa value of CAD should be higher. Please explain about this by detail or remove the CAD only analysis.
2- The sample volume is very low which decrease the accuracy of the study.
3- The legend of Fig 1 is" Diagram of Breast phantom of this study". Please correct this sentence because it seems you used only the Customized type.
4- It is not mentioned how many lesions of the phantom were evaluated by each reader. All 14 lesions? And how many lesions were evaluated totally? Have you evaluated the difference of kappa value between benign and malignant lesions in readers and between US and CAD?

5- Please clarify in the method section what is your mean about probably benign or malignant lesions based on BIRADS lexicon?
Thanks.

REFEREE: Reviewer | **Revision (0)** 2 Feb 2019

Dear AE,
There are some comments which should be addressed by the author:
1- This part is unclear: "…. in multiple studies conducted recently in order to breast ultrasound (US)" or "To investigate the reliability of computer aided diagnosis (CAD) system for US using breast phantom with multi-reader analysis for identical lesions." should be: "To investigate the reliability of different readers using computer aided diagnosis (CAD) system in US for determining identical lesions in a breast phantom." Or "….reliability study" should be "reliability studies." please correct the English writing style in all papers.
2- Can the reader draw his/her own boundary on the lesions? [in CAD] Isn't this a limitation that the boundary determination is not flexible and fully under reader control?
3- Are the features determined by CAD or by reader? [shape, margin, …..]
4- Why was the gold standard determined by two expert radiologists? We know about the nature of manufactured lesions [benign vs. malignant] embedded in the phantom. [fig 2 C note]
5- What conditions are possibly benign and possibly malignant in S detect? Plese define and clarify.
6- This sentence seems unclear: "As in the US, we analyzed two or more areas through ROI on the touch screen of CAD and then selected representative image." I recommend to write this sentence in clearer form.
7- Please add the full data of statistical program used in the paper.
8- Kappa is good for pairwise reliability and agreement assessment [between two raters]. We can assess all raters together in the same time by ICC [for both inter-rater and intra-rater assessments, this analysis could be used]. Why did not authors use ICC instead of Kappa? This question is related to tables 1 and 2.
9- It is not known which measurement is the target in table 1? Determining the characteristics of lesions? determining the nature of lesions? or determining their presence? please clarify in detail in the paper.
10- When we want to determine the inter-rater reliability using kappa, we can choose different possible combinations of raters [for example, senior rater 1 vs. junior rater 1, senior rater 1 vs junior rater 2., etc…]. the authors have determined only one value for each condition that is unclear. Please clarify this point. this is another why I recommended ICC seems better statistics than kappa for both inter-rater and intra-rater assessments. In fact, the authors could calculate ICC for all raters in one session [for UC and CAD and combination separately]. In addition, as the senior group are similar together and junior group are similar together, they can calculate ICC for these subgroups separately and also calculate the ICC for whole group.
11- It is unknown what is the P-values in table 1? Are they for Kappa or they are for comparison? In fact, each kappa is presented with a P-value, then this is questionable that what are P values in table 1?
12- The authors assessed the junior residents that were trained for 2 hours only. It is

questionable that what is the clinical benefit of enrolling such a group? In fact, it is expectable that a group of junior residents with only 2-hour education will show a worse efficacy in their imaging assessment. It seems better evaluation would be enrolling and comparing senior residents with their professors and attending radiologists and we will not gain a benefit from comparing the junior with senior residents. Please explain and rationalize the study design in this regard.
13- For subjective combination, did all readers have a consensus on the method for subjective combination or it was totally a task related to their personal knowledge and experience?
14- If the P-values in tables 1 and 2 are related to comparing the kappas, which test was used for comparing Kappa coefficients?
15- There are P-values in table 3 for comparing the index versus ultrasound as a reference based on a footnote in Table 3. It is unknown which ultrasound data was used for such comparison? Did they compare each index with its corresponding sonography index? Please clarify this point and it is better to say which test(s) were used for such comparisons?
16- The authors mentioned the performance of seniors is higher than juniors while they did not perform any statistical comparison for this comparison. It is better to document comparisons by statistical analyses.
17- The standard method of diagnostic efficacy papers is the STARD guideline. However, this guideline is used for papers performed on real patients and here, the study has been performed on phantoms. I recommend the authors consider this guideline and use relevant points of last version of this checklist to complete their paper.
18- Considering the use of phantom here, what is the prediction of authors for real patients? Please add and discuss.
19- Major part of discussion is on re-explaining the results that are repetitive. In addition, there are not reference in discussion and the authors is better to compare the results with other similar papers and discuss these comparisons.
Thanks.

REFEREE: Associate Editor | **Revision** (0) 5 Feb 2019

Dear EIC,
The article bneeds major revision.
Thanks.

REFEREE: EIC | **Revision** (0) 5 Feb 2019

Dear Author,
The reviewers' comments are mentioned. Please correct the details as mentioned above carefully according to reviewers' request.
Thanks.

OPEN PEER REVIEW

Revision (1)

Reply to Reviewers

Ideally, the reviewing process can significantly improve the submitted manuscripts by allowing the authors to take into account the advice of reviewers. Author(s) must reply to all reviewers' comments in a separate Word file, point by point. A "**Reply to Reviewers**" document is submitted along with revised manuscript during submission of revised files, summarizing the changes that the authors made in response to the reviewers' comments. The responses to reviewers' comments specifies how the authors addressed each comment the reviewers made.

You can read the authors' responses to the reviewers' comments in the next page.

Authors' Response to the reviewer`s Comments
Reviewer 1:

1-this part is unclear: "…. in multiple studies conducted recently in order to breast ultrasound (US)" or "To investigate the reliability of computer aided diagnosis (CAD) system for US using breast phantom with multi-reader analysis for identical lesions." should be: "To investigate the reliability of different readers using computer aided diagnosis (CAD) system in US for determining identical lesions in a breast phantom." Or "….reliability study" should be "reliability studies." please correct the English writing style in all paper.
→ **Authors` response:** As you commented, we corrected it.

2-can the reader draw his/her own boundary on the lesions? [in CAD] Isn't this a limitation that the boundary determination is not flexible and fully under reader control?
→ **Authors` response:**
When the reader identified the lesion and touched the center of the lesion in the monitor, a region-of-interest (ROI) was drawn along the border of the mass either automatically by the CAD program. And, BI-RADS lexicons are automatically analyzed and displayed by the CAD system.
However, the readers touched the center of lesion, which step could make different effect on CAD result based on the readers.
As you commented, we added it to limitation.

3-are the features determined by CAD or by reader? [shape, margin, …..]
→ **Authors` response:** No, it's not.
When the reader identified the lesion and touched the center of the lesion in the monitor, a region-of-interest (ROI) was drawn along the border of the mass either automatically by the CAD program. And, BI-RADS lexicons (shape, margin…) are automatically analyzed and displayed by the CAD system. We clarified it.

4-why was the gold standard determined by two expert radiologists? We know about the nature of manufactured lesions [benign vs. malignant] embedded in the phantom. [fig 2 C note]
→ **Authors` response:** Since there may be a discordance of judgment on the lesion of the breast phantom we actually designed, two dedicated breast radiologists have reaffirmed to reduce the discordance.

5-what conditions are possibly benign and possibly malignant in S detect? Please define and clarify.
→ **Authors` response:** The S-Detect adopted a deep learning algorithm in the processes of lesion segmentation. The S-Detect for breast module utilizes large data sets collected from numerous breast exam cases and provides the characteristics of displayed lesion as well as a suggestion as to whether a selected lesion is benign or malignant. We added it.

6-This sentence seems unclear: "As in US, we analyzed two or more areas through ROI on the touch screen of CAD and then selected representative image." I recommend to write this

sentence in clearer form.
→ **Authors` response:** As you commented, we corrected it.

7-please add the full data of statistical program used in the paper.
→ **Authors` response:** As you commented, we added full data.
'Statistical analyses were performed by SAS version 9.3 (Cary, NC, USA) and p-value below 0.05 was considered statistically significant.'

8-Kappa is good for pairwise reliability and agreement assessment [between two raters]. We can assess all raters together in the same time by ICC [for both inter-rater and intra-rater assessments, this analysis could be used]. Why did not authors use ICC instead of Kappa? This question is related to tables 1 and 2.
→ **Authors` response:** I agree with you. It was a matter of concern for us, and we also discussed with the statisticians.
However, in nominal (or categorical) variables, kappa statistics was used for analyzing inter/intra-rater reliability. The basis for this is the GRRAS guideline.
(Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. 2011;48:661-671)

9-it is not known which measurement is the target in table 1? Determining the characteristics of lesions? determining the nature of lesions? or determining their presence? please clarify in detail in the paper.
→ **Authors` response:** As you commented, we clarified it.
In Table 1, the Inter-reader reliability is for comparison the final assessment in CAD of senior or junior to gold standard.

And, the meaning of P value: 1) Inter-reader: comparison senior or junior to gold standard. 2) Intra-reader: comparison between the primary and secondary assessments.

10-when we want to determine the inter-rater reliability using kappa, we can choose different possible combinations of raters [for example, senior rater 1 vs. junior rater 1, senior rater 1 vs junior rater 2.,etc…]. the authors have determined only one value for each condition that is unclear. Please clarify this point. this is another why I recommended ICC seems better statistics than kappa for both inter-rater and intra-rater assessments. In fact, the authors could calculate ICC for all raters in one session [for UC and CAD and combination separately]. In addition, as the senior group are similar together and junior group are similar together, they can calculate ICC for these subgroups separately and also calculate the ICC for whole group.
→ **Authors` response:** I agree with you. It was a matter of concern for us, and we also discussed with the statisticians.
In Table 1, the Inter-reader reliability is for comparison the final assessment in CAD of senior or junior to gold standard.
However, in nominal (or categorical) variables, kappa statistics was used for analyzing inter/intra-rater reliability.

And, the meaning of P value: 1) Inter-reader: comparison senior or junior to gold standard. 2)

Intra-reader: comparison between the primary and secondary assessments.

As you commented, we clarified it.

11-it is unknown what is the P-values in table 1? Are they for Kappa or they are for comparison? In fact, each kappa is presented with a P-value, then this is questionable that what are P values in table 1?
→ **Authors` response:** As you commented, we clarified it.

P value meaning: 1) Inter-reader: comparison senior or junior to gold standard. 2) Intra-reader: comparison between the primary and secondary assessments.

12-the authors assessed the junior residents that were trained for 2 hours only. It is questionable that what is the clinical benefit of enrolling such group? In fact, it is expectable that a group of junior residents with only 2-hour education will show a worse efficacy in their imaging assessment. It seems better evaluation would be enrolling and comparing senior residents with their professors and attending radiologists and we will not gain a benefit from comparing the junior with senior residents. Please explain and rationalize the study design in this regard.
→ **Authors` response:** Our study is a study using breast phantom, not real patients. This is the greatest strength of our research. Junior residents that were trained for 2 hours only cannot perform ultrasound on the real patients. Our study is to demonstrate that breast phantom enables breast ultrasound education and training for the starters. If our study was a study comparing the performance of a senior resident and a professor, we did not have to use a breast phantom. And, there are several published studies comparing the performance of a senior resident and a professor.

13-for subjective combination, did all readers have a consensus on the method for subjective combination or it was totally a task related to their personal knowledge and experience?
→ **Authors` response:** As you commented, we clarified it.
The proper combined decision was chosen subjectively based on the US with CAD results, which was defined as the subjective combination. Readers subjectively choose based on the US (totally a task related to their personal knowledge and experience) with CAD results.

14-if the P-values in tables 1 and 2 are related to comparing the kappas, which test was used for comparing Kappa coefficients?
→ **Authors` response:** As you commented, we added the kappa test (Cohen kappa) and statistical package.

15-there are P-values in table 3 for comparing the index versus ultrasound as reference based on footnote in table 3. It is unknown which ultrasound data was used for such comparison? Did they compare each index with its corresponding sonography index? Please clarify this point and it is better to say which test(s) were used for such comparisons?
→ **Authors` response:** I understand your confusion. As you commented, we deleted the P-

values in table 3 for comparing the index versus ultrasound as reference based on footnote in table 3.
Instead of that, we added the P-values between senior and junior groups. However, the sample size is too small to show the significant difference of diagnostic performances between senior and junior groups. And, we added it to limitation also.
Comparison of AUC was performed by Dejong method. For Sensitivity, specificity, PPV, and NPV comparison, Chi-square exact test was used.

16-the authors mentioned the performance of seniors are higher than juniors while they did not perform any statistical comparison for this comparison. It is better to document comparisons by statistical analyses.
→ **Authors` response:** As you commented, we added it to table 3.
However, the sample size is too small to show the significant difference of diagnostic performances between senior and junior groups. And, we added it to limitation also.

17-the standard method of diagnostic efficacy papers is STARD guideline. However, this guideline is used for papers performed on real patients and here, the study has been performed on phantoms. I recommend the authors consider this guideline and use relevant points of last version of this checklist to complete their paper.
→ **Authors` response:** As you commented, we reviewed and followed the STARD guideline.

18-considering the use of phantom here, what is the prediction of authors for real patients? Please add and discuss.
→ **Authors` response:** As you commented, we mentioned it in discussion. In the case of a real patient, one reader may not be able to perform ultrasound multiple times. We used breast phantom, which enables multi reader analysis for the same lesion.

19-major part of discussion is on re-explaining the results that are repetitive. In addition, there are not reference in discussion and the authors is better to compare the results with other similar papers and discuss these comparisons.
→ **Authors` response:** As you commented, we corrected it. And. we added the comparisons with other papers.

Reviewer 2:
The authors have presented a well structured study with very good analysis. However, there are some basic questions that need explanation:
1- It is not clear why the kappa value of US (0.609) was significantly higher than CAD (0.411). In the method section it has been mentioned" The readers first detected the breast phantom lesion by US and then applied the CAD system. " I mean when the readers used CAD, they had analysed the US feature before, so kappa value of CAD should be higher. Please explain about this by detail or remove the CAD only analysis.
→ **Authors` response:** I understand your confusion.
When the reader identified the lesion and touched the center of the lesion in the monitor, a region-of-interest (ROI) was drawn along the border of the mass either automatically by the

CAD program. And, BI-RADS lexicons are automatically analyzed and displayed by the CAD system. So, the CAD result is not based on the US feature analyzed before. CAD analysis is separated from US analysis.
However, the readers touched the center of lesion, which step could make different effect on CAD result based on the readers.
As you commented, we added it to limitation.

2- The sample volume is very low which decrease the accuracy of the study.
→ **Authors` response:** I agree that it's the main limitation of our research. We added it to limitation.

3- The legend of Fig 1 is" Diagram of Breast phantom of this study". Please correct this sentence because it seems you used only the Customized type.
→ **Authors` response:** As you commented, we corrected it.

4- It is not mentioned how many lesions of the phantom were evaluated by each reader. All 14 lesions? And how many lesions were evaluated totally? Have you evaluated the difference of kappa value between benign and malignant lesions in readers and between US and CAD?
→ **Authors` response:** Each reader evaluated all 14 lesions. Because of the small number of cases, we did not evaluate the difference of kappa value between benign and malignant lesions in readers and between US and CAD.

5- Please clarify in the method section what is your mean about probably benign or malignant lesions based on BIRADS lexicon?
→ **Authors` response:** As you commented, we clarified it.

OPEN PEER REVIEW

Revision (1)

Here, you can see the **Reviewers**, **Associate Editors** and **EICs'** comments from the beginning to the end of the revision process.

REFEREE: Reviewer | **Revision** (1) — 23 Mar 2019

Dear AE,
The authors corrected all the requested comments and the article is now suitable for publication in IJR.
Thanks.

REFEREE: Reviewer | **Revision** (1) — 3 Apr 2019

Dear AE,
It seems the authors have corrected the paper.
Thanks.

REFEREE: Associate Editor | **Revision** (1) — 9 Apr 2019

Dear EIC,
The authors corrected the requested changes and the article is now acceptable for publication in IJR.
Thanks.

REFEREE: Associate Editor | **Revision** (1) — 3 Jun 2019

Dear EIC,
The authors should respond to all comments one by one.
1-In the Materials and Methods, “It has the same anatomy as the normal breast US and the lesions are embedded in.” is not correct.
Do you mean, “It has the same anatomy as the normal breast US and the lesions are embedded in it.”?
2- In the Discussion, “Junior residents that were trained for 2 hours only cannot perform ultrasound on the real patients.”
This sentence is wrong. Do you mean, “Junior residents who are trained for 2 hours cannot perform ultrasound on real patients.”?
Thank you.

REFEREE: EIC | **Revision** (1) — 8 Jun 2019

Dear Author,

Your article is reviewed by the reviewers carefully. According to their comments, this document needs minor revision. Please follow their comments as requested.

Thanks.

OPEN PEER REVIEW

Revision (2)

Reply to Reviewers

Ideally, the reviewing process can significantly improve the submitted manuscripts by allowing the authors to take into account the advice of reviewers. Author(s) must reply to all reviewers' comments in a separate Word file, point by point. A "**Reply to Reviewers**" document is submitted along with revised manuscript during submission of revised files, summarizing the changes that the authors made in response to the reviewers' comments. The responses to reviewers' comments specifies how the authors addressed each comment the reviewers made.

You can read the authors' responses to the reviewers' comments in the next page.

Authors' Response to the reviewer`s Comments

Associate Editor 1:
Dear editor in chief,
The authors corrected the requested changes and the article is now acceptable for publication in IJR.
Associate Editor 2:
Dear author,
Please respond to all comments one by one. Then, highlight any changes you make in the manuscript. Please do not make any changes in the manuscript other than responding to the comments, because they will not be considered.
1-In the Materials and Methods, "It has the same anatomy as the normal breast US and the lesions are embedded in." is not correct.
Do you mean, "It has the same anatomy as the normal breast US and the lesions are embedded in it."?
→ **Authors` response:** As you commented, we corrected it.

2- In the Discussion, "Junior residents that were trained for 2 hours only cannot perform ultrasound on the real patients."
This sentence is wrong. Do you mean, "Junior residents who are trained for 2 hours cannot perform ultrasound on real patients."?
Thank you.
→ **Authors` response:** As you commented, we corrected it.

OPEN PEER REVIEW

Revision (2)

Here, you can see the **Reviewers**, **Associate Editors** and **EICs'** comments from the beginning to the end of the revision process.

REFEREE: Associate Editor | **Revision** (2) 23 Jun 2019

Dear EIC
Based on the authors' answers to the reviewers' comments, the reviewer(s) have recommended publication.
Kind regards.

REFEREE: EIC | **Revision** (2) 23 Jun 2019

Dear Author,
The document is reviewed carefully by the reviewers. According to their comments and your revisions as requested, this manuscript is accepted.
Thanks.