Peer Review File

Article information: https://dx.doi.org/10.21037/atm-21-3449

Reply to Reviewer Comments

This paper studied the feasibility of a self-supervised learning assistant system for helping diagnosis of severity of mitral regurgitation (MR). This algorism could help physician in diagnosis of MR. Although the research was interesting, several major concerns arose for the application in clinical practice.

Major comments:

Comment 1: Using self-supervised learning (SSL) in this study for avoiding time-consuming and laborious manual annotation of videos by experts to train the model was an advantage of this study, but also was a major limitation. The outputs from this SSL algorithm was limited.

Reply 1: Indeed, the use of model pre-training of the SSL model helps reduce the number of required annotated images for model development. However, the main factor that limited the outputs in our study was the fact that echocardiographic images/videos were reliably obtained from only one view: 2D color echocardiographic apical 4-chamber view. This limitation was therefore not inherent to the SSL algorithm itself. As detailed, a major consideration in the design of our algorithm was for it to be practical for use in clinical practice. While limiting the analysis to segmentation of LA and MR contours in the training dataset may have some limitations, as we showed, have excellent correspondence with five metrics of MR severity. To clarify this, we expanded the Limitations section of the Discussion.

Changes in the text:

We have added the following to the limitations section （page 19 line 412-429）

> "First, grading of MR severity based on society guidelines requires analyses based on multiple 2D color, continuous and pulse wave Doppler signals. However, in clinical practice, as in our overall database, images from all required views of sufficient quality are not always available. Accordingly, by design, our algorithm relied only on the apical 4-chamber view, which was reliably obtained in the overall dataset. While this limited our analysis to metrics of MR derivable from segmentation of LA and MR contours, we demonstrated that this approach yielded excellent results when compared with thorough analyses by experts based on a larger number of metrics obtained from multiple views which are required by society guidelines. Notably, all views required for such thorough

> analysis were available in only a ~5% (148/2766) of echocardiograms obtained in our real-world retrospective analysis. Overall, the results showing the excellent performance of our segmentation model and index quantification algorithms demonstrate the effectiveness of this approach. Nevertheless, the DL architecture developed for the present study could readily be trained and applied to other views if so desired.
>
> Further related to model training, our algorithm employed SSL pre-training to overcome the need for a large number of labeled images, and thus markedly reduced the burden of manual annotation."

Comment 2: Grading the severity of MR is a complex process involving both qualitative and quantitative from both 2D and Doppler echocardiography. This study only used apical four chamber color Doppler image in the model. Although the results were satisfied, the clinical usefulness was limited. Why not using all major views included apical 4 chamber, apical 2 chamber, and parasternal long axis, short axis view in the training process? Since our evaluation of echocardiography was not only apical 4 chamber view, a helping system for each view will be more useful.

Reply 2: We agree with the comment. Indeed, ground truth was determined from all views and analyses required by guideline standards. But, as noted, such comprehensive sets of views were available in only 5% of all studies in our real-world database. Therefore, as detailed in response to comment 1, we took advantage of the fact that apical 2D color 4-chamber Doppler videos were reliably available. We demonstrate that our model yields information that assists physicians in making the correct diagnosis based on this single view. In future research, the algorithm architecture can be applied to additional views for automated analysis of additional parameters. To the best of our knowledge, this is the first study applying self-supervised learning in analyzing color Doppler echocardiographic videos.

Changes in the text:

Please refer to "Changes in the text" in response to Comment 1.

Comment 3. Six candidate indexes of MR severity were developed from the model. These indexes can be easily measured manually. Were the AI derived indexes better than manually measured?

Reply 3: The ground truth for training the model on disease severity was provided by expert clinician assessments derived from their manual analyses. Therefore, it is not a matter of whether the AI metrics are "better", it is just a question of whether the metrics provided by the algorithm lead to the correct final diagnosis. We are therefore addressing the reality that, in actual clinical practice, physicians do not typically measure all these indexes, but only make a qualitative judgment based on visual assessment of the color Doppler videos. Therefore, rather than directly measuring these six indexes, physicians merely segment the outlines of left atrium and MR jet area. Therefore, with our dataset, it was not possible to compare manual and AI measurements. Although comparing the values of those indexes could not be performed, we compared the final outcome - diagnostic accuracy improved by providing these indexes to clinicians. Our results showed that the sensitivity increased when physicians were provided with six indexes deriving from the algorithm output, compared those without the AI assistant. So, we believe that these indexes shown in Figure 3 are very helpful to physicians.

Changes in the text:

We have added the following to the introduction (page 7 line 130-134):

> "The required quantitative measurements can be helpful, but in actual clinical practice are time-consuming to obtain and are not in widespread use. Therefore, automated algorithms which can provide quantitative indexes have the potential to improve accuracy for grading MR severity in clinical practice."

Comment 4. The task for the deep machine learning in grading of the MR severity would be the automatic grading of MR. This study was only for assisting diagnosis of significant MR.

Reply 4: Correct. This is because from among the 2766 patients included in this study, only 148 had high-quality labels required for the final diagnosis of MR severity based on society guidelines. In order to ensure the validity of the results, we used these 148 patients as the testing dataset. We did not train a MR grading model on this dataset, because such a small sample may lead to over-fitting of the model. On the contrary, the acquisition of segmentation labels is relatively easy and has few limitations, which is more suitable for training machine learning network. Therefore, the algorithm is designed to provide quantitative indexes to assistant clinicians, rather than replacing physicians by making the MR classification on its own.

Changes in the text: No changes made in the text. It is explained at several places that the algorithm determines indexes of MR to assist physicians to make the diagnosis.

Comment 5. Apical 4 chamber color Doppler images were selected then manually select the biggest MR color jet frame for annotation. Did the view and the frame could be automatic selection from the testing, training and validation? This mean that did the whole echocardiography video clip as an input in the process? I did not see the algorithm for the view and flame selection.

Reply 5: We assume the word “view” you mentioned refers to the identification of apical 4 chamber from all chamber section. If so, the algorithm for automatic view selection was developed and presented in prior study from our group, which is accepted for publication in JACC: Cardiovascular Imaging and is currently *in press*. So, when the algorithm proposed in this study will be implemented, the view selection will be automatic. In terms of the identification of the max MR jet frame, we classify the max MR jet frame among a series of echocardiography video clip in an end-to-end manner with the algorithm. The process has been explained in the Supplementary section entitled **MR jet recognition and segmentation** (page 3). The blue arrows in **Figure 2** also illustrate the process of frame identification.

Changes to the text: The following as been added on page 10 line 202-204 in Methods:

> “These video clips were identified automatically from each study using a previously developed and validated view classification algorithm having an accuracy of >90% (10).”

Comment 6. Study number was probably enough for testing the sensitivity and specificity for the model in the helping physician diagnosis, but the number was too low for training machine learning.

Reply 6: The number of training samples does affect the training efficiency, but some algorithms have been developed to solve the small sample problem such as SSL [r1], zero-shot learning [r2], few-shot learning [r3], and so on. Our algorithm employs SSL pretraining to obviate the requirement for large-labeled training samples, and thus saves labor for annotation labor; this is one of the advantages of this work. The results demonstrated the effectiveness of this approach.

[r1] Kolesnikov A, Zhai X, Beyer L. Revisiting Self-Supervised Visual Representation Learning. IEEE Conference on Computer Vision and Pattern Recognition 2019; pp. 1920-1929. https://doi.org/10.1109/CVPR.2019.00202

[r1] Li Y, Zhang J, Zhang J, Huang K. Discriminative learning of latent features for zero-shot recognition. IEEE Conference on Computer Vision and Pattern Recognition 2018; pp. 7463-7471. https://doi.org/10.1109/CVPR.2018.00779.

[r2] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. International Conference on Machine Learning 2015. Volume 37.

Changes in the text: We have added the following to Discussion (page 19 line 427-429):

> **“**Further related to model training, our algorithm employed SSL pre-training to overcome the need for a large number of labeled images, and thus markedly reduced the burden of manual annotation.**”**

Minor comments:

Comment 7. There were several abbreviation without mention of full name in the abstract and also in the text, such as AI, LA, DICE, ResNet-Unet, ASE, ESC…etc.

Reply 7: Thank you for your suggestion.

Changes in the text: We have added the **Abbreviation** page (page 4 line55-71)

“Abbreviations

A4C = Apical four-chamber

ASE = American Society of Echocardiography

AI = Artificial intelligence

DICE = Dice similarity coefficient

EROA = Effective regurgitant orifice area

ESC = European Society of Cardiology

GEE = Generalized estimating equation

LA = Left atrium

LSTM = Long short-term memory

MR = Mitral Regurgitation

PISA = Proximal isovelocity surface area

ResNet-UNet = Residual U-shape Network

SV = Stroke volume

SSL = Self-supervised learning

STARD = Standards for Reporting Diagnostic Accuracy

FVCD = Full-volume color Doppler transthoracic echocardiography”

Comment 8. Totally 2766 studies were include. There were 148 for testing, 592 for training, and 148 for validation (totally 888). How were the remaining 1878 studies (figure 1) ? How to select 888 studies from 2766 studies?

Reply 8: Patient selection for testing, training, and validation was described in Methods (Figure 1). 148 studies with sufficient views to calculate all quantitative indexes according to the guideline as the reference standard were selected as the test dataset for the segmentation algorithm. Then a stratified random sampling was performed to select patients for training and validation datasets. The remaining 1879 patients not selected in the stratified random sampling process were excluded from the analyses. We have added the “stratified random sampling” process in the flow chart, to make the patients selection and exclusion more clear.

Changes in the text: We have updated the Figure 1 showing the flow chart of patients selection.