Peer Review File

Article information: http://dx.doi.org/10.21037/atm-20-1519

Response to the comments of reviewer A

Comment 1: The Methods of this paper are not suited to address the question being asked. The rationale for choosing the various types of model is not clear. The nature if the training data is not clear, because data replacement occurred. This compromises the interpretation of Results and severely undermines the reproducibility of the experiment. The data do not appear to be well balanced.

The models do not seem to perform well until month 1, so cannot really be regarded as predicting outcomes based on pre-treatment features. The performance only improved once the model incorporates the improvements see at 1-month post treatment.

Reply 1: Thank you for your advice. To predict the patient prognosis with CSC after laser treatments, we have applied different models, including deep learning models. Deep learning is probably the most convenient way to process image data by reducing the research workload. Compared with machine learning, the output of deep learning is more intuitive. First, we applied the generative adversarial net (GAN) to predict OCT characteristics (such as subretinal fluid absorption [SFA]) at 1-, 3-, and 6-month after laser treatment. However, the results were disappointing. The prediction results of deep learning could not reflect the changes in SFA accurately after laser treatment. This failure probably occurred because data were collected from fewer than 500 patients and limited. In addition, the synthetic images could not really show us the ellipsoid zone [EZ] and sensory retina, which is important to the ophthalmologist. Therefore, we chose machine learning models that performed better on our prediction tasks. The collection and allocation of our data sets were based on TRIPOD Statement: the development of prediction model using one dataset and the evaluation of its performance on separate data.

Common retinal diseases such as age-related macular degeneration (AMD) and diabetic retinopathy (DR) as well as central serous retinopathy (CSC) have many prognostic factors. In addition to the characteristics of the sensory retina and the therapy, other factors such as the patient profession, physical and mental state, and

even genetic differences can influence the effectiveness of the treatment. There are two possible reasons for the unsatisfactory results of the 1-month prediction. First, patients have different sensitivities to the therapy due to the complexity of the disease and its influence factors. However, due to the limitations of the existing cognition, the feature extraction is also limited, which may be the primary factor affecting the accuracy of the short-term prediction. At 6 months, due to the treatability of CSC and because some patients might be treated more than once, the long-term recovery is better. Therefore, we obtained a better long-term prediction result. We will continue to improve our study and hope to accumulate more data in the future to obtain better results with deep learning, so as to serve the patients and clinicians.

Response to the comments of reviewer B

Comment 1: In this paper, authors have utilized machine learning to predict sub-retinal fluid absorption (SFA) in central serous chorioretinopathy (CSC) subjects. Using Random Forest, they were able to achieve the accuracy of 0.651 ± 0.068, 0.753 ± 0.065 and 0.818 ± 0.058 for predicting SFA for CSC cases of 1 month, 3 months and 6 months, respectively. On external validation, they achieved the accuracy of 0.734, 0.727, and 0.900 for predicting SFA on 1 month, 3 months and 6 months CSC cases, respectively using XGBoost. The dataset used in their research was locally acquired from Zhongshan Ophthalmic Center and Xiamen Eye Center.

My prime concern is that why the authors have not explored modern deep learning architectures for predicting SFA? Many researchers have proposed promising solutions encompassing deep learning for predicting (and grading) CSC subjects, and validated their frameworks using publicly available datasets. Furthermore, there are many deep learning systems which robustly extract retinal lesions for the lesion-influenced grading of CSC (especially as acute and chronic). Considering all of these frameworks, what this study adds new in the body of knowledge?

Reply 1: Thank you very much for your recognition! We sincerely appreciate your scrupulous and constructive suggestions, which were really valuable for improving the quality of our manuscript.

We have explained the reasons for using machine learning in the reply to reviewer A.

In fact, before manually extracting the data for machine learning, we applied deep learning to complete our prediction task. However, the deep learning results were disappointing. Most of the existing literature reports are on cross-sectional data. For example, Narendra and his colleagues segmented OCT images and identified the SRF of CSC patients with deep learning (1); Yuki and Yi Zhen et al. analyzed fundus photographs of CSC patients using deep learning to detect the choroidal thickness and SRF. (2,3) They are all great studies and help tremendously with image parsing. However, these studies used cross-sectional data but no follow-up data. Therefore, to predict the SFA after laser treatments, we used the machine learning models that performed better on prediction tasks. Although the workload increased greatly, a good long-term prediction effect was achieved, which could assist clinicians in choosing laser therapy in the future.

Comment 2: Although, clinical validation of machine learning frameworks on locally acquired datasets is important. But, testing them on publicly available datasets using (standardized ground truths) is also essential so that the proposed study can be really beneficial for the targeted research community.

Reply 2: Thank you for your advice. The clinical validation and the dataset allocation were performed according to Type 3 of the TRIPOD Statement: the development of the prediction model using one dataset and the evaluation of its performance on separate data. (4) In the studies to come, we will test the accuracy of our prediction models in external datasets from different sources. We did not find the datasets suitable for our prediction model. We would test our prediction models on the publicly available datasets if there is one.

Comment 3: Apart from this, the training details in the paper are hardly vague, making the whole framework extremely non-reproducible. Either the authors release the code or they should exactly mention how the training (of all the models) is performed. What is the criteria for training/testing data split? How the cross-validation is performed? How the overfitting is avoided especially for the Random Forest and Decision Trees?

Reply 3: Thank you for your advice. While selecting the optimal algorithms, we

randomly divided the clinical data from the Zhongshan Ophthalmology Center (ZOC) into 10 parts, and we calculated the decision tree, AdaBoost.R2, gradient boosting, XGBoost, random forest and extra trees by means of 10-fold cross-validation. We then selected the three best algorithms for the ensemble according to the prediction accuracy. After determining the selected algorithms, to make the most of the existing clinical data, we used all the data from the ZOC to retrain the selected algorithms (Page 7, line 157-163). We supplemented the criteria for training and testing the data split (Page 6, line 139-141), and we also supplemented the details of the machine learning algorithms in the supplementary material (Supplements Page 2-3). The code is not available for public access because of privacy concerns, but it is available from the corresponding author upon reasonable request. To avoid the overfitting in random forest, we adjusted the hyper parameters of the algorithm, such as the numbers of samples in the leaf node. For the decision tree, our experiment is based on the Cart algorithm, by selecting the Gini coefficient as the feature selection criterion. To avoid overfitting, we used the grid search method for the maximum depth in decision tree, and decision tree is optimized by a post-pruning method.

Changes in the text: See Manuscript, Page 6, line 139-141; Page 7, line 157-163; Supplements Page 2-3.

Comment 4: Also, authors should provide valid justification for choosing Random Forest (and the other old models)? They should also present a thorough comparison with state-of-the-art deep learning solutions on standardized publicly available datasets and state how (and why) their proposed scheme is better than already published schemes?

Reply 4: Thank you for your advice. Deep learning may be the best way to predict disease prognoses with sufficient data. In our study, random forest performed the best in predicting SFA during the internal validation and performed well in the external validation, which was far better than the deep learning prediction models. We did not find the publicly available datasets suitable for our prediction model. The most recent applications of deep learning in CSC are image structure recognition and detecting SRF in cross-sectional data, not follow-up data for predictions (1-3). After many

attempts to use different data types and algorithm selection, random forest and XGBoost performed best in our prediction models. Consequently, we chose machine learning.

Comment 5: In addition to this, the ROC curves reported in Figure 2 (D, E, F, K, L) and Figure 3 (D, E, F, K, L) are not convincing. First of all, they are generated using very few samples. Secondly, the performance is not satisfactory for some classes. Why? Authors should thoroughly discuss the reason for such low performance in the manuscript.

Reply 5: First, the most important reason for the unsatisfactory results at 1-month may be the insufficient follow-up data. The second is the complexity and uncertainty of the prognosis in fundus diseases. Although the overall therapeutic effect in patients with CSC was relatively good compared with other fundus diseases, even experienced ophthalmologists cannot predict recovery after laser treatment in a particular patient. Therefore, we tried to help clinicians to predict the prognosis of patients with CSC, and we achieved a satisfactory result, with 80-90% accuracy, in the 6-month predictions. In paragraph 1 of the discussion, we analyzed the reasons for the unsatisfactory 1-month prediction results after laser treatment (Page 11 line 232-238). Thanks to the suggestions of reviewer B, we will continue to collect follow-up data for patients with CSC and continue to improve our models.

Changes in the text: We added the reason for the low performance of the 1-month predictions in paragraph 1 of the discussion. Last but not least, for the prognosis predictions of complex fundus disease, we still need to accumulate follow-up data to improve our models. (see Page 11, line 237-238).

Reference

1. Narendra Rao TJ, Girish GN, Kothari AR, et al. Deep Learning Based Sub-Retinal Fluid Segmentation in Central Serous Chorioretinopathy Optical Coherence Tomography Scans. *Conf Proc IEEE Eng Med Biol Soc.* 2019;2019:978-981.
2. Komuku Y, Ide A, Fukuyama H, et al. Choroidal thickness estimation from colour fundus photographs by adaptive binarisation and deep learning,

according to central serous chorioretinopathy status. *Sci Rep.* 2020;10(1):5640.

3. Zhen Y, Chen H, Zhang X, et al. Assessment of Central Serous Chorioretinopathy Depicted on Color Fundus Photographs Using Deep Learning. *Retina.* 2019.

4. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594.