JULY 12–16
2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

# Deep Learning based Treatment Plan Evaluation

Xinran Zhong, Dan Nguyen, Rafe McBeth, Azar Balagopal, Maryam Mashayekhi, Mu-Han Lin, Steve Jiang
Medical Artificial Intelligence and Automation (MAIA) Laboratory,
Department of Radiation Oncology,
UT Southwestern Medical Center, Dallas, Texas
Xinran.Zhong@UTSouthwestern.edu

MAIA Laboratory
UT Southwestern Medical Center
Radiation Oncology

## INTRODUCTION

Reliable automatic treatment plan evaluation can help accelerate the plan optimization and approval process, especially for adaptive treatment re-planning. Plan quality quantification is subjective to individual physicians' preference and can hardly be done with a simple algorithm such as PlanIQ.

Although there are several attempts to evaluate treatment plans quantitatively, the score could not represent fully physician's opinion. Deep learning models have the potential to automatically evaluate the plan, and the main challenge is that the clinical treatment plan evaluation label is limited and qualitative.

## AIM

Here we propose a deep learning model that can automatically rank treatment plans for each patient based on specific physician's preference.

## METHODS

The general workflow of the proposed method was shown in Figure 1. A Siamese model [1] is trained using pairwise comparison label between plans within a patient, and the predicted scores for each plan gives the rank of all plans for one patient.

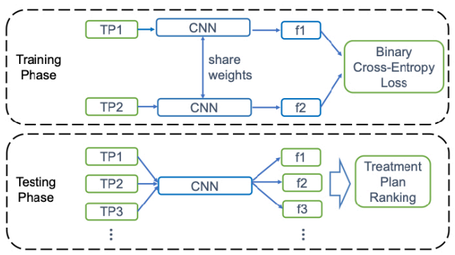


*Figure 1. CNN based automatic plan evaluation model. In training phase it's trained via pair-wise ranking while in testing phase it can rank multiple plans based on score f*

## METHODS

- The input of the model is the dose and contours for the PTV and OARs and the output is that 1) the first plan is better; 2) the second plan is better; or 3) two plans are similar.
- The model is first trained using the PlanIQ scores [2] for all the plans in training dataset so it works for an idealized virtual physician called "Dr. PlanIQ". Then the trained model is adapted to the real physician's preference through transfer learning.

## DATASETS

- We have in total 66 head and neck patients treated by the same physician
  - 1-29 plans per patient
  - 738 plans in total
- We randomly picked 62 cases for training and 4 cases for validation. We plan to acquire another 10 cases for separate testing.
- Plan resolution is 5 x 5 x 5 mm

## RESULTS

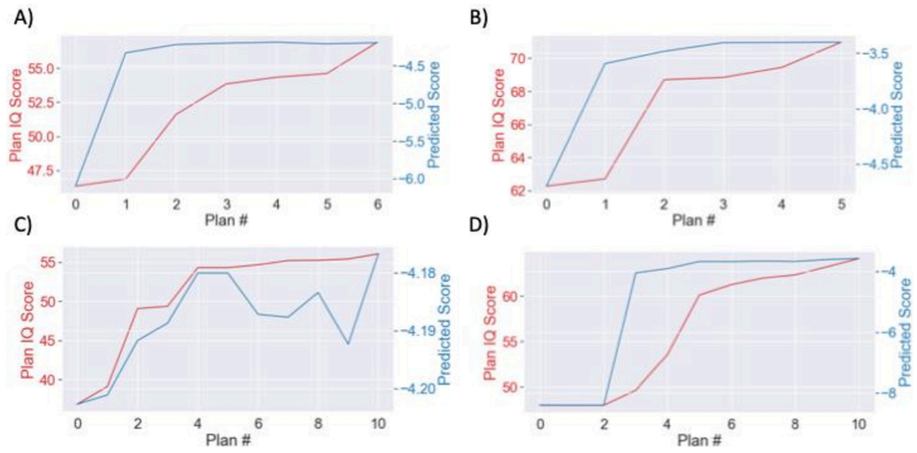


*Figure 3. The predicted score ranking vs. Plan IQ score ranking for all four validation cases.*

The model ranked all the plans correctly. Notice that considering we don't have a score for each plan from our physician, plan-IQ score is not directly put as the label during the training process. Instead, the ranking of the plan-IQ scores is the label, so it is a more challenging task compared to regression.

To evaluate the transfer learning feasibility, we have compared the similarity between the Plan-IQ scoring and physician's preference. We calculated the approved plan Plan-IQ ranking among all the plans for each patient, and the median ranking for approved plan is 83.3%. This shows that the score and physician's opinion are mostly consistent, indicating the transfer learning can possibly achieve high accuracy in this task.

The trained model for "Dr. PlanIQ" can achieve 0.9 binary accuracy on the validation data. Minor overfitting was observed with data augmentation implemented.

Among the 4 validation cases, the number of plans per patient ranges from 6 to 11, and the model predicts the ranking correctly except for one case when the predicted score were very similar across all plans.

## CONCLUSIONS

- We proposed a model to use Siamese network to utilize the qualitative preference label and transfer learning to deal with limited training data.
- We have shown the feasibility of evaluating and ranking a set of plans for one patient trained with paired-wise plan preference label.
- We have shown the correlation between plan IQ score and physician's approval decision for head and neck patients.
- We expect to see similar results after transfer learning based on each physician's preference.

## REFERENCES

[1] Doughty, H., Damen, D. and Mayol-Cuevas, W., 2018. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6057-6066).

[2] Ouyang, Z., Liu Shen, Z., Murray, E., Kolar, M., LaHurd, D., Yu, N., Joshi, N., Koyfman, S., Bzdusek, K. and Xia, P., 2019. Evaluation of auto-planning in IMRT and VMAT for head and neck cancer. Journal of applied clinical medical physics, 20(7), pp.39-47.