2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

Quality-Driven Deep Learning Dose Prediction for Head and Neck Cancer Using Limited High-Quality Training Data

Ivan Vazquez[1]; Alexandra Olivia Leone[2]; Christine V. Chung[1]; Skylar Gay[2]; Ramon Maurilio Salazar[1]; Carlos Sjogreen[1]; Thomas J. Whitaker[1,2]; Laurence Edward Court[1,2]; Ming Yang[1,2]

[1]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center; [2]Medical Physics Program, The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Science

ABSTRACT

Purpose: To evaluate whether deep learning models trained on a small number of high-quality plans (e.g., ≤30) can predict dose distributions of comparable quality, and whether the predicted quality improvements are achievable.

Methods: This study includes 138 oropharyngeal VMAT patients, whose clinical treatment plans were scored using a validated percentile-based plan quality scoring system developed in-house (PQS; higher scores indicate superior plan quality). Three training strategies were compared: (1) the baseline model trained with 84 cases spanning all quality levels, (2) the model trained with only 30 top-scoring cases (PQS≥65), and (3) the model trained with only 30 top-scoring cases after transfer learning from 110 external automated-planning cases to address limited training data. The baseline model was evaluated on its 26-patient test set; the high-quality models were applied to the 88 patients outside the high-quality subset. Two cases underwent dosimetrist replanning guided by predicted dose distributions and structure-specific quality metrics derived from them.

Results: The baseline and high-quality-only strategies predicted negative PQS changes (−6.64±16.95 and −9.46±18.84 percentile points), reflecting regression toward average quality. The high-quality with transfer learning strategy predicted positive changes (+8.20±21.34 points; p<0.001), with 67% of patients showing predicted improvement. In a paired comparison of 16 overlapping test patients, the transfer learning approach produced higher predicted PQS in 14/16 cases (87.5%), with a mean advantage of 9.38±8.87 percentile points (p<0.001). Dosimetrist replanning confirmed predictions were achievable, improving PQS from 7.97 to 81.2 and from 44.20 to 70.3.

Conclusion: Despite training on only 30 high-quality cases, models using transfer learning predicted achievable quality improvements rather than reproducing average practice, providing actionable optimization targets for clinical planners.

INTRODUCTION

- Standard deep learning dose prediction models are typically trained on clinical plans of potentially highly variable quality
- These models learn the average practice, reproducing routine outcomes instead of guiding planners toward optimal results
- Isolating only top-tier plans restricts the dataset (e.g., 30 of 138 cases), leaving sample sizes below typical numbers used for deep learning-based dose prediction applications
- ***We hypothesized that a model trained on a small, curated set of high-quality plans, with transfer learning to offset the reduced sample size, can predict dose distributions representing achievable quality improvements over clinical plans***
- We compared three training strategies evaluating dose accuracy, predicted PQS gains, and whether those gains were achievable in dosimetrist replanning

METHODS AND MATERIALS

- **Data:** 138 oropharyngeal VMAT patients treated at our institution; an external automated-planning dataset (110 H&N VMAT plans) served as the transfer-learning source
- **Quality metric:** Plan Quality Score (PQS) based on dosimetric quantities scored as percentiles and combined into a single 0–100 score
- **Model:** 3D Dense Dilated U-Net with 21 input channels (planning CT + 20 Volumes of interest)
- **Three training configurations:**
  - **Baseline:** 84 mixed-quality clinical plans
  - **High-quality:** 30 top-scoring plans (PQS ≥ 65)
  - **High-quality + TL:** pre-trained with automated plans and fine-tuned with top 30 scoring
- **Evaluation:**
  - Accuracy: Dose Score, DVH Score, and Paddick Conformity Index on held-out test sets
  - Quality gain: ΔPQS (predicted − original) for the 88 patients outside the high-quality subset
  - Achievability: a dosimetrist replanned two cases using the predicted dose and PQS summaries as guidance

RESULTS

Dose prediction accuracy

- High-quality + TL had the lowest error: Dose Score 1.839 ± 0.042 Gy vs 2.076 ± 0.059 (baseline) and 2.035 ± 0.062 (high-quality); $p < 0.05$ vs both
- DVH Score also lowest with transfer learning (1.658 ± 0.025 Gy)
- High-quality model matched baseline accuracy using 36% of the training data; conformality was preserved with TL (ΔPCI −0.005 ± 0.009)

Predicted plan quality change

- **Baseline:** ΔPQS = −6.64 ± 16.95
- **High-quality only:** −9.46 ± 18.84
- **High-quality + TL:** +8.20 ± 21.34 ($p < 0.001$ vs both)
  - 67% of patients showed predicted improvement
  - 45.5% gained more than 10 percentile points
  - Largest gains in mid-range plans (PQS ≈ 15–60)
- Head-to-head in 16 overlapping patients: TL better in 14/16 (87.5%), mean advantage 9.38 ± 8.87 points ($p < 0.001$)

Table 1. Dosimetrist replanning

Patient	Original PQS	Predicted PQS	Replanned PQS
A	7.97	65.04 ± 6.27	81.2
B	44.20	84.60 ± 5.53	70.3

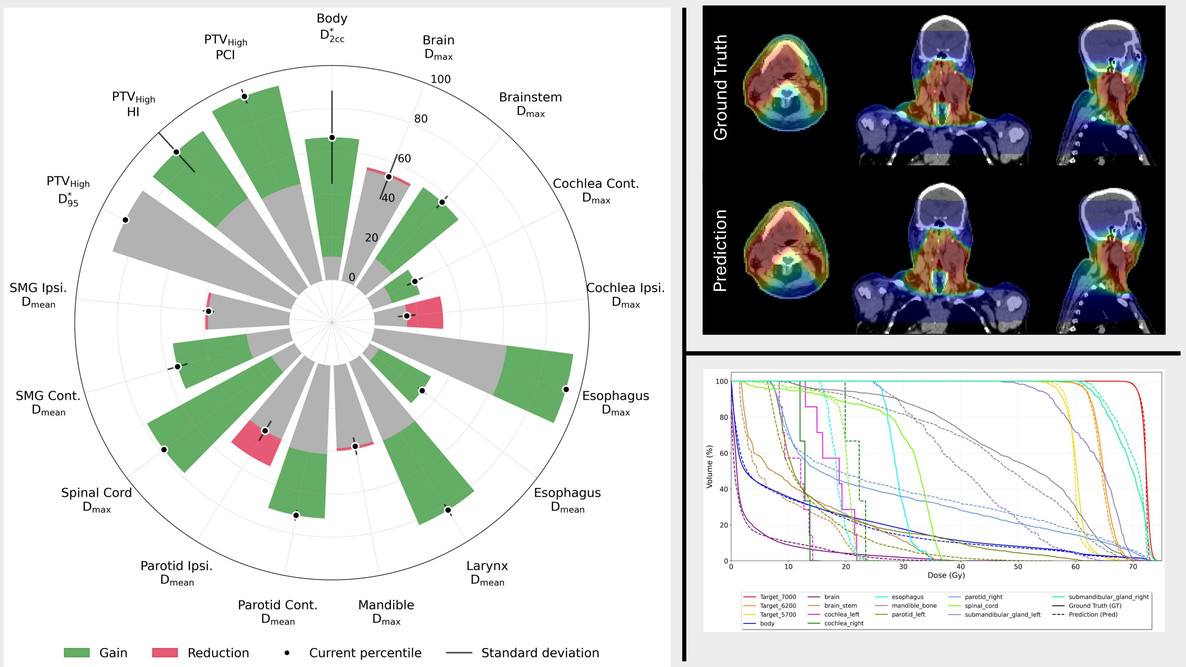

Figure 1. Representative model output. Ground-truth vs. predicted dose distributions (top right), corresponding DVHs (bottom right, solid = clinical, dashed = predicted), and the PQS radial plot (left) showing per-structure percentile with predicted gains (green) and reductions (red).

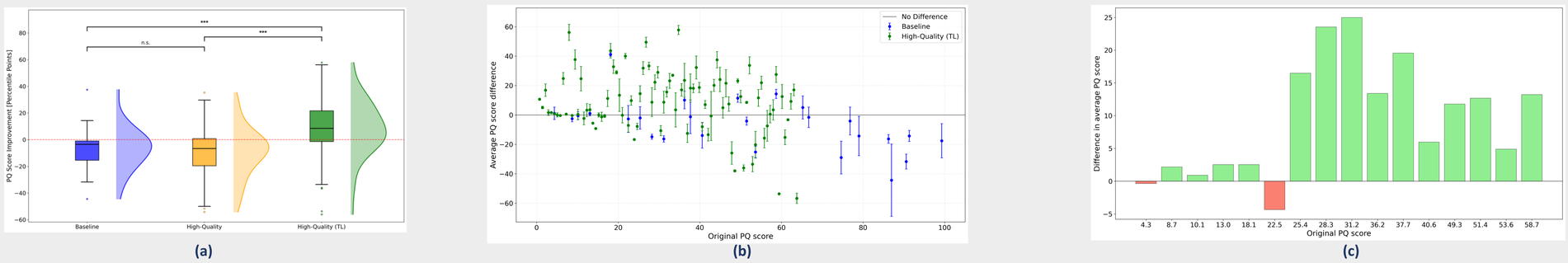

Figure 2. Predicted plan quality improvement across models. (a) Distribution of PQS change (Predicted − Original, percentile points) for the baseline model (n = 26), the high-quality model, and the high-quality transfer-learning model (both applied to the 88 clinical plans outside the high-quality training subset). Box plots show median and interquartile range with overlaid kernel density estimates; the dashed line marks zero change. The baseline and high-quality distributions are centered below zero, whereas the transfer-learning distribution is shifted positive. n.s. = not significant; *** $p < 0.001$. (b) Mean PQS change per patient plotted against original PQS for the baseline (blue) and high-quality transfer-learning (green) models. Error bars denote the standard deviation across the four-fold ensemble; the horizontal line indicates no change. (c) Mean PQS change grouped by bins of original PQS. Green bars indicate a net predicted gain, red bars a net predicted loss.

DISCUSSION AND CONCLUSIONS

- Training on 30 high-quality plans alone was not sufficient; the model predicted lower-quality plans than the originals (ΔPQS −9.46). Transfer learning offset the reduced sample size and reversed decreased in PQS (+8.20)
- The high-quality model matched baseline accuracy with 36% of the training data, which could indicate less variability in training and evaluation data due to PQS-based data curation
- The baseline model predicted reduced PQS for most patients, consistent with a model reproducing the average quality of its heterogeneous training data
- Gains peaked for mid-range plans: the lowest-scoring cases may be constrained by intrinsic case difficulty the model cannot overcome, while the highest-scoring plans approach the model's own predictive ability
- In two cases, a dosimetrist achieved PQS improvements comparable to those predicted, suggesting the predictions are realistic optimization targets
- Limitations: single institution, oropharyngeal VMAT only; PQS is an imperfect surrogate for clinical benefit; only two cases were replanned

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com