UCLA Health

David Geffen School of Medicine

Generalizability and Domain Adaptation of a Deep Learning Dose Prediction Model for Head-and-Neck Radiotherapy

Yasin Abdulkadir, Jennifer Kadowaki, Jack Neylon, James Lamb

Department of Radiation Oncology, University of California, Los Angeles, CA

ROSA-ML LABORATORY

Local institutional domain shift substantially degraded dose prediction accuracy; targeted fine-tuning reduced MAE from 7.39 to 3.68 Gy and improved protocol compliance.

STUDY DESIGN

Pretrained MedNeXt model trained on GDP-HMM challenge data

↓

Local cohort split: 350 fine-tuning / 40 validation / 37 test

↓

Compare direct deployment vs institutional fine-tuning

↓

Secondary extension: patient-specific plan quality assessment

Key Definitions

- **MAE**: Mean Absolute Error (voxel-wise)
- **QA-gap**: Difference between clinical and AI predicted benchmark
  - OAR metrics: Clinical – Predicted
  - Target metrics: Predicted - Clinical
  - Positive QA gap: Clinical plan is worse
- **QA-gap flag**: Exceeding the metric-specific threshold derived from the validation cohort (currently the 95th percentile, q95)
- **Protocol opportunity**: AI-predicted dose meets constraint but clinical plan does not
- **Review rule**: ≥ 1 QA-gap flag and ≥ protocol opportunity

INSTITUTIONAL DOMAIN SHIFT AND RECOVERY WITH FINE-TUNING

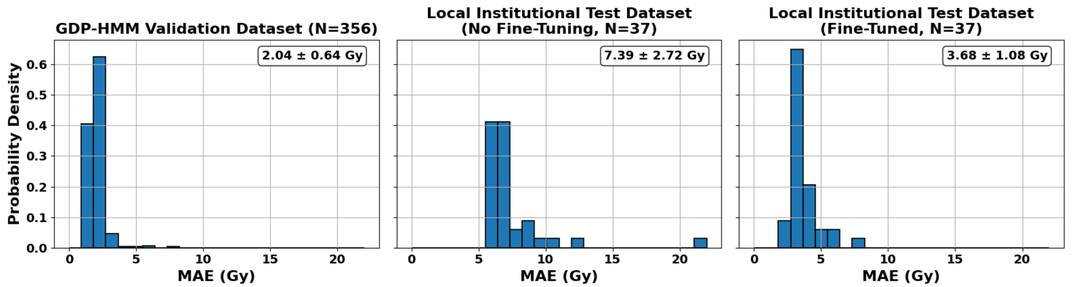


Figure 1. Institutional domain shift and recovery after fine-tuning. Normalized distributions of per-case voxel-wise mean absolute error within the patient body. The GDP-HMM–trained model performed well on its original validation cohort but showed substantial error inflation when directly applied to the independent institutional test set. Fine-tuning on local cases reduced test-set MAE from **7.39 ± 2.72 Gy to 3.68 ± 1.08 Gy**, an approximately 50% reduction.

REPRESENTATIVE LOCAL CASE: FINE-TUNING REDUCES SPATIAL DOSE ERROR

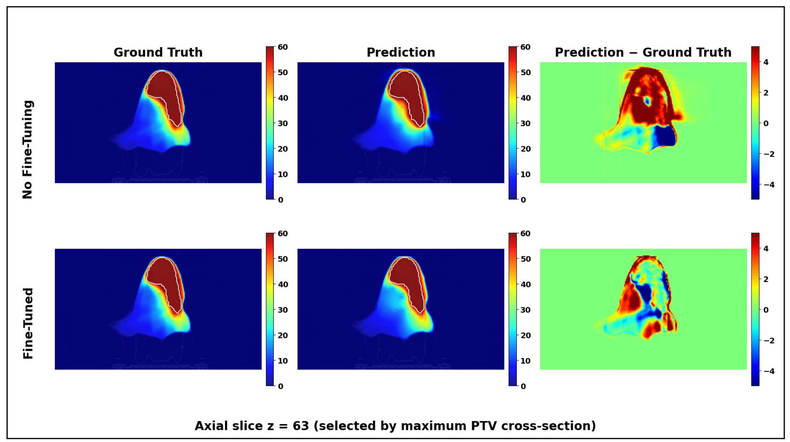
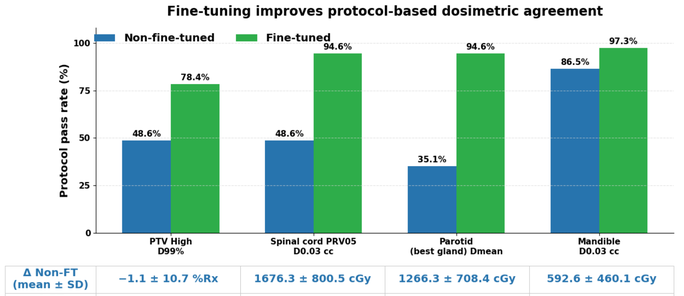


Figure 2. Spatial dose-prediction improvement following institutional fine-tuning. Representative axial slice from a local head-and-neck case with non-fine-tuned MAE near the test-cohort mean. Fine-tuning reduced systematic spatial differences between predicted and ground-truth dose. The white contour denotes the PTV; difference maps show prediction minus ground truth and are clipped at ±5 Gy.

↓ ~50%

reduction in local test MAE



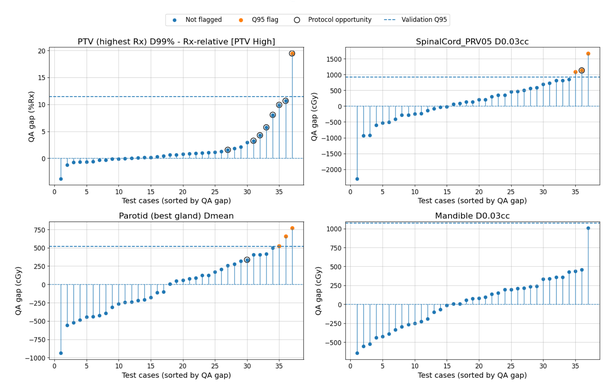
	PTV High D99%	Spinal cord PRV05 D0.03 cc	Parotid (best gland) Dmean	Mandible D0.03 cc
Δ Non-FT (mean ± SD)	−1.1 ± 10.7 %Rx	1676.3 ± 800.5 cGy	1266.3 ± 708.4 cGy	592.6 ± 460.1 cGy
Δ FT (mean ± SD)	−0.4 ± 5.7 %Rx	−195.5 ± 711.4 cGy	−25.1 ± 381.4 cGy	−131.4 ± 330.3 cGy

Predicted doses were normalized to the highest-dose PTV D95%.

Figure 3. Fine-tuning improves protocol-based dosimetric agreement. Pass rates for selected NRG head-and-neck dose metrics before and after institutional fine-tuning. Predicted doses were normalized to the highest-dose PTV D95%. Fine-tuning improved compliance across all four evaluated endpoints, with particularly large gains for spinal cord PRV and parotid dose.

Figure 4. Patient-specific plan-quality screening on the independent test cohort. QA gaps were calculated by comparing clinical dose metrics with the fine-tuned model's patient-specific predictions. Cases are sorted by QA gap; orange markers denote validation-derived q95 flags, open circles denote protocol opportunities, and dashed lines indicate q95 thresholds. Six of 37 cases had a QA-gap flag, 10 had a protocol opportunity, and 3 met the combined review criterion.

PLAN QA EXTENSION



- Benchmark-level performance did not directly generalize to the local clinical environment.
- Institutional fine-tuning substantially restored voxel-wise accuracy and improved clinically relevant dose metrics.
- Adapted dose prediction may support automated, patient-specific plan quality screening.