

Exploring Uncertainty-Aware Deep Learning for Segmentation of Primary Tumor and Pathologic Lymph Nodes in Oropharyngeal Cancer: Insights from a Multi-Center Cohort

Alessia De Biase, PhD Candidate, Department of Radiation Oncology, University Medical Centre Groningen
Nanna Sijtsema; Lisanne Dijk; Johannes Langendijk; Peter Ooijen

Introduction

Within the medical field, there is a growing demand for deep learning (DL) models which convey model certainty to the end-user, while maintaining alignment between model accuracy and certainty. For oropharyngeal cancer (OPC) primary tumor segmentation in PET/CT images, tumor probability maps (TPM) that show voxel-level predicted probabilities were introduced as possible output of DL models [1]. This study extended the network to generate TPMs for both primary tumor and pathologic lymph nodes, and explored whether quantified uncertainty in TPMs can predict segmentation model accuracy in an external cohort.

Hypothesis

DL generated TPMs can serve as a tool for uncertainty-aware primary tumor and pathologic lymph nodes segmentation in OPC patients across internal and external datasets.

Methods

We gathered PET/CT images and manual delineations of gross tumor volume of the primary tumor (GTVp) and pathologic lymph nodes (GTVln) of 405 OPC patients treated with (chemo)radiation in our institute in 2010-2022. The publicly available 2022 HECKTOR challenge dataset served as external test set (Figure1). The existing ensemble-based DL tumor segmentation method [1] was adapted to perform multi-label segmentation. 15 models were trained, and the ensemble-average of the predicted TPMs was considered per patient. Surface dice similarity coefficient was computed for the predicted contours at different probability thresholds. The model's performance on the patient was assessed using the average (Average Surface DSC). Uncertainty was quantified using coefficient of variation (CV) of the multiple predictions. Calibration was evaluated by correlation between the two metrics.

Results

A moderate negative correlation (around -0.6) was observed in both internal and external test sets between Average Surface DSC and CV for both GTVp and GTVln (Figure2), indicating a significant calibration.

Conclusion

Significant accuracy versus uncertainty calibration was achieved for TPMs in both internal and external test sets, indicating the potential use of CV to identify cases with lower segmentation accuracy, independently of the dataset.

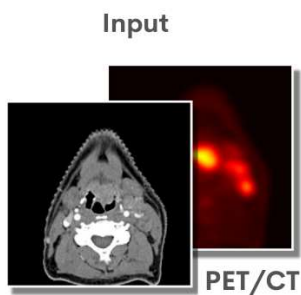
Synopsis

This study addresses uncertainty-aware DL in oropharyngeal cancer segmentation in PET/CT images. Results show that quantified uncertainty in ensemble-averaged primary tumor and lymph nodes probability maps aligns with segmentation accuracy across internal and external datasets.

Figure(s)

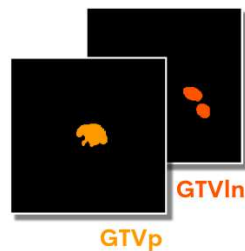
Training / Validation

301 cases



[1]

Groundtruth



5-folds CV

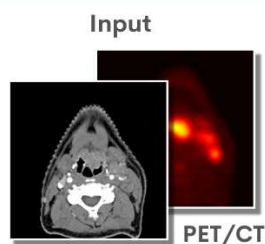
[1] Alessia De Biase et al 2023 Phys. Med. Biol. 68 055013.

Testing

Internal Set:
104 cases



External Set:
519 cases



Model 1



Model 2



....

Model 14



Model 15



Output

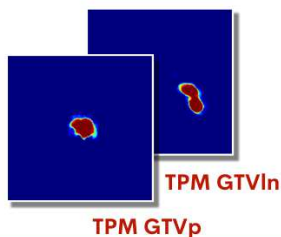


Figure 1. Illustration of Training/Validation and Testing Process.

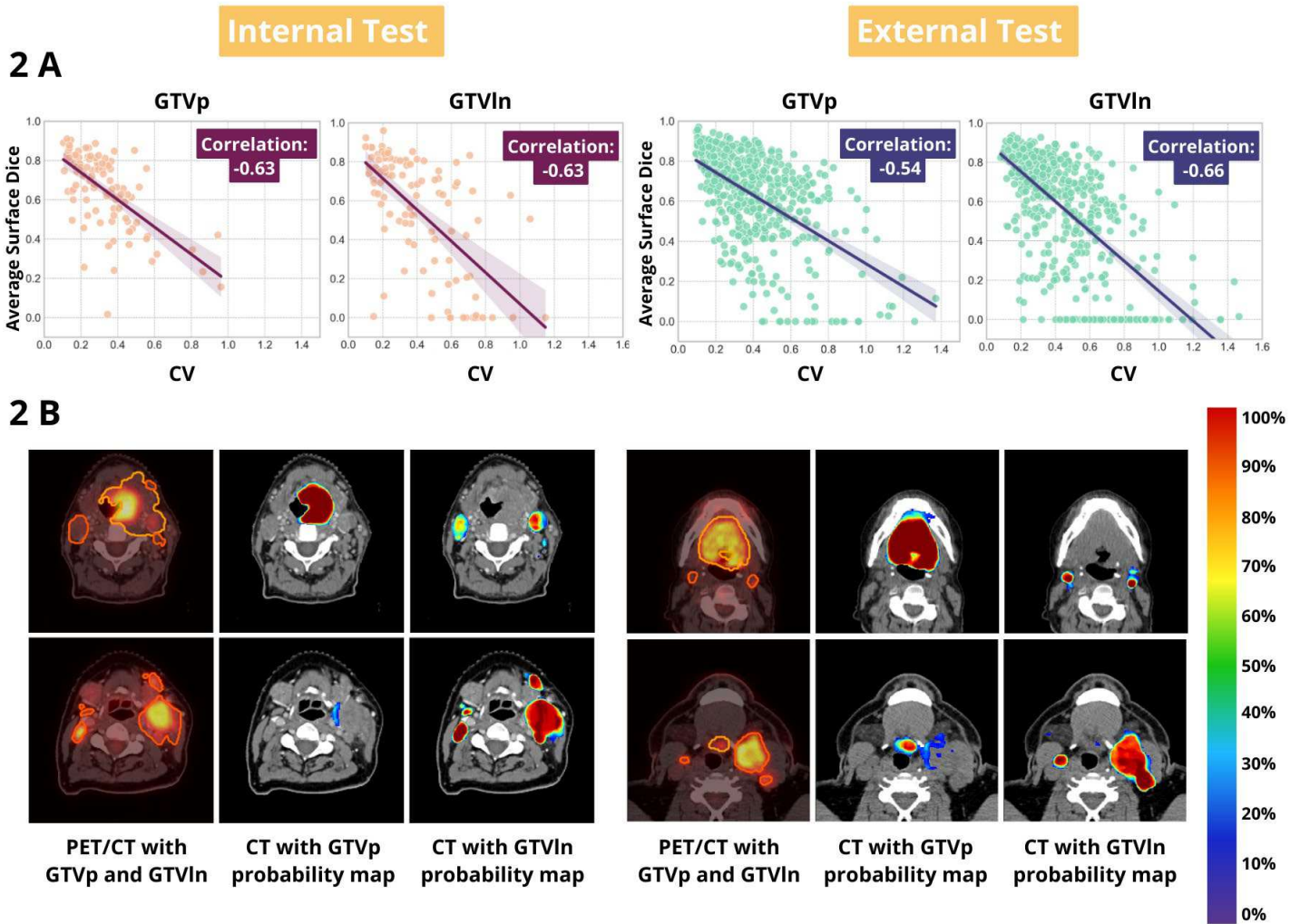


Figure 2. Quantitative and qualitative results. In row A) correlations plots showing coefficient of variation (CV) versus Average Surface Dice for GTVp and GTVIn segmentation results in the internal and external test sets. In row B), examples of predicted tumor probability maps of GTVp and GTVIn for four patients, two in the internal (left side) and two in the external (right side) test sets. The GTVp manual contour is in light orange, the GTVIn in dark orange.

Keywords

Artificial Intelligence / Machine Learning; Clinical Workflow & Productivity; Imaging Research; Quality Improvement & Quality Assurance