Modality-Aware Cross-Modal Transformer for Joint Fusion of PET/CT and Clinical Data for Relapse-Free Survival Prediction in Multi-Centric HPV-Associated Oropharyngeal Cancer

Yujing Zou[1,2,4]; Sébastien Quetin[1,2,4]; Juan Duran[1,2,4],Shirin Abbasinejad Enger[1,2,3,4]

[1]Medical Physics Unit, Department of Oncology, Faculty of Medicine, McGill University, Montréal, Québec, Canada

[2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada

[3]Research Institute of the McGill University Health Centre (RI-MUHC), Montréal, Québec, Canada

[4]Montreal Institute for Learning Algorithms – Quebec AI Institute, Montréal, Canada

2026 JULY 19–22 VANCOUVER, BC

JOINT AAPM | COMP MEETING

ABSTRACT

Purpose: To develop and validate a modality-aware cross-modal transformer for joint fusion in relapse-free survival prediction in HPV-associated oropharyngeal cancer (OPSCC), integrating PET/CT imaging and clinical data for risk stratification.

Methods: We used the HECKTOR 2025 multi-center dataset (N = 678; 542 training, 136 test). The proposed Cross-Modal Transformer architecture employs:(1) 3D DenseNet imaging encoders for PET/CT and FTTransformer encoder for clinical variables (i.e., demographics, stage, HPV status), and radiomic features derived from nnUNet-segmented primary/nodal tumour volumes. (2) linear tokenizers projecting modalities to shared latent space, (3) learnable positional encodings for modality identity, (4) shallow 3-layer transformer encoder with 8-head self-attention enabling cross-modal interactions, and (5) mask-aware pooling for missing data robustness. We compared against unimodal baselines (clinical-only, imaging-only, radiomic-only) and late fusion variants (attention, weighted, stacking). Models were trained with DeepHit and Contrastive losses using 5-fold cross-validation. Discrimination was assessed via concordance index (C-index) with 95% 1000-resample bootstrap CIs; calibration via Integrated Brier Score (IBS), and risk stratification used training-derived risk cutoffs on HPV+ test patients, with hazard ratios (HR) comparing high vs. low-risk groups.

Results: The Cross-Attention model achieved a holdout C-index of 0.675 (95% CI: 0.563-0.770) with the best calibration (IBS=0.160). It successfully stratified HPV+ patients (HR=4.46; 95%CI: 1.38-14.40; p=0.012) between high- and low-risk groups. Late Fusion with attention achieved the highest discrimination (C-index=0.701, HR=7.59, p=0.009), with wider confidence intervals, indicating less precise stratification. All unimodal models demonstrated reasonable cross-validation discrimination (C-index=0.659-0.714) yet failed to produce clinically meaningful risk stratification.

Conclusion: Multimodal transformer joint fusion of PET/CT imaging and clinical data improves clinically actionable risk stratification beyond unimodal and late-fusion approaches in HPV-associated OPSCC. Modelling cross-modal interactions reveals biologically distinct risk groups within HPV-associated disease, achieving a balanced combination of discrimination, calibration, and stratification precision that may guide treatment intensification and future de-escalation strategies. External validation of an independent cohort will be required.

CONTACT

Yujing Zou, PhD candidate

Medical Physics Unit, McGill University

yujing.zou@mail.mcgill.ca

INTRODUCTION

HPV+ OPSCC generally has favorable outcomes, but identifying patients at higher relapse risk remains clinically important as de-escalation may compromise control in some patients. We developed and validated a modality-aware cross-modal transformer joint fusion model for Relapse-Free-Survival (RFS) prediction, using multi-centric patient data, that integrates pre-treatment PET/CT imaging, clinical variables, and Gross-Tumour-Volume (GTV)-derived features.

METHODS

Cohort: HECKTOR 2025 [1] multi-center data (N=678; 542 training, 136 test).

Encoders: 3D DenseNet for PET/CT [2]; FTTransformer for clinical variables [3] and radiomic features derived from nnUNet-segmented primary and nodal GTVs [4].

Joint fusion: Linear modality tokenizers, learnable positional encodings for missing modalities , and a shallow 3-layer, 8-head transformer enabled cross-modal interactions; mask-aware pooling supported missing modalities (**Figure 1**).

Comparators: Clinical-, imaging-, and radiomic-only models plus attention, weighted, and stacking late-fusion variants.

Training: DeepHit and contrastive losses with 5-fold cross-validation.

Evaluation: C-index with 1,000-resample bootstrap 95% CIs, IBS, and training-derived risk cutoffs with high-vs-low HRs in HPV+ test patients.

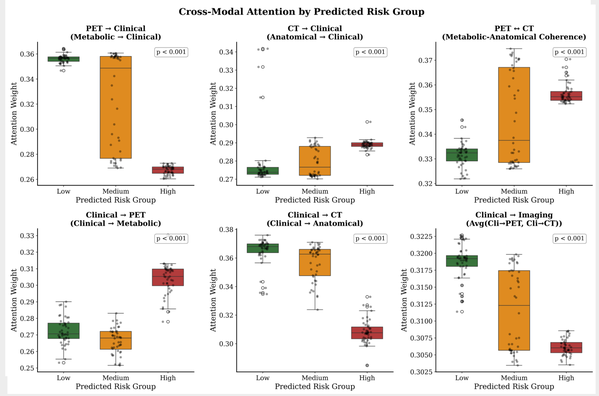

Figure 2. Cross-modal attention weights differed across predicted risk groups (all p<0.001).

RESULTS

Cross-attention transformer : holdout C-index 0.675 (95% CI 0.563–0.770), best calibration (IBS=0.160), and the most precise statistically significant HPV+ risk stratification (HR=4.46; 95% CI 1.38–14.40; p=0.012) (**Table 1, Figure 3E**) in HPV+ internal holdout test set. Cross-modal self-attention enables PET/CT/clinical representations to be unified and to inform one another, integrating complementary metabolic, anatomical, and clinical signals for improved risk stratification (**Figure 2**).

Late-fusion attention: highest discrimination (C-index=0.701) and HR=7.59 (p=0.009) in the internal holdout test set, but unable to yield significant risk stratification using train set risk score median threshold cutoff for HPV+ OPSCC patients (**Table 1, Figure F-H**).

Unimodal models: All unimodal models achieved reasonable cross-validation concordance indices (0.659-0.714) but failed to stratify HPV-positive patients into significant risk groups (**Table 1, Figure 3 A-C**).

Clinically, the identified HPV+ low-risk group represents potential candidates for future treatment de-escalation trials, while the high-risk group may benefit from intensified therapy or closer surveillance despite favourable HPV status. These findings underscore that explicit cross-modal integration improves not only model performance metrics but also the reliability and interpretability of risk stratification needed for personalized radiotherapy strategies.

Table 1: Evaluation of unimodal and multimodal survival models for relapse-free survival prediction: reporting discrimination (C-index) for train cross-validation (CV) mean±sd and internal holdout test set, calibration (Integrated Brier Score, IBS), and risk stratification (Hazard Ratio, HR, 95% CIs) between high (H)- and low (L)-risk groups of the tertile risk stratification with Wald p-values.

Method	Train	Internal Holdout		Hazard Ratio (H/L)		
	CV C-index	C-index	IBS	Holdout (All)	Train (HPV+)	Holdout (HPV+)
Unimodal						
Clinical	0.659 ±0.057	0.430 [0.304–0.554]	0.164 [0.124–0.196]	–	0.63 [0.36–1.11] $p = 0.108$	–
Radiomic (GTV)	0.673 ±0.036	0.354 [0.243–0.471]	0.163 [0.124–0.196]	–	1.51 [0.89–2.59] $p = 0.130$	–
Imaging (PT/CT)	**0.714 ±0.033**	0.525 [0.409–0.637]	0.163 [0.124–0.195]	–	1.37 [0.77–2.42] $p = 0.283$	–
Multimodal Fusion						
Variational Autoencoder	0.701 ±0.028	0.610 [0.491–0.723]	0.163 [0.124–0.195]	3.44 [1.08–10.98] $p = 0.037$	1.45 [0.84–2.50] $p = 0.182$	2.61 [0.76–8.92] $p = 0.127$
Cross-Attention Transformer	0.693 ±0.051	0.675 [0.563–0.770]	**0.160 [0.123–0.198]**	**4.41 [1.68–11.59]** $p = 0.003$	10.15 [4.35–23.66] $p < 0.001$	**4.46 [1.38–14.40]** $p = 0.012$
Late Fusion						
Weighted	0.664 ±0.020	0.697 [0.584–0.787]	0.163 [0.122–0.196]	7.17 [1.62–31.65] $p = 0.009$	3.90 [2.02–7.55] $p < 0.001$	7.25 [1.55–33.81] $p = 0.012$
Attention	0.662 ±0.021	**0.701 [0.592–0.790]**	0.163 [0.122–0.197]	7.56 [1.72–33.22] $p = 0.007$	**3.87 [2.01–7.44]** $p < 0.001$	7.59 [1.65–34.92] $p = 0.009$
Stacking	0.668 ±0.027	0.641 [0.527–0.741]	0.161 [0.122–0.194]	3.01 [1.08–8.42] $p = 0.035$	3.63 [1.93–6.83] $p < 0.001$	2.90 [0.96–8.74] $p = 0.059$

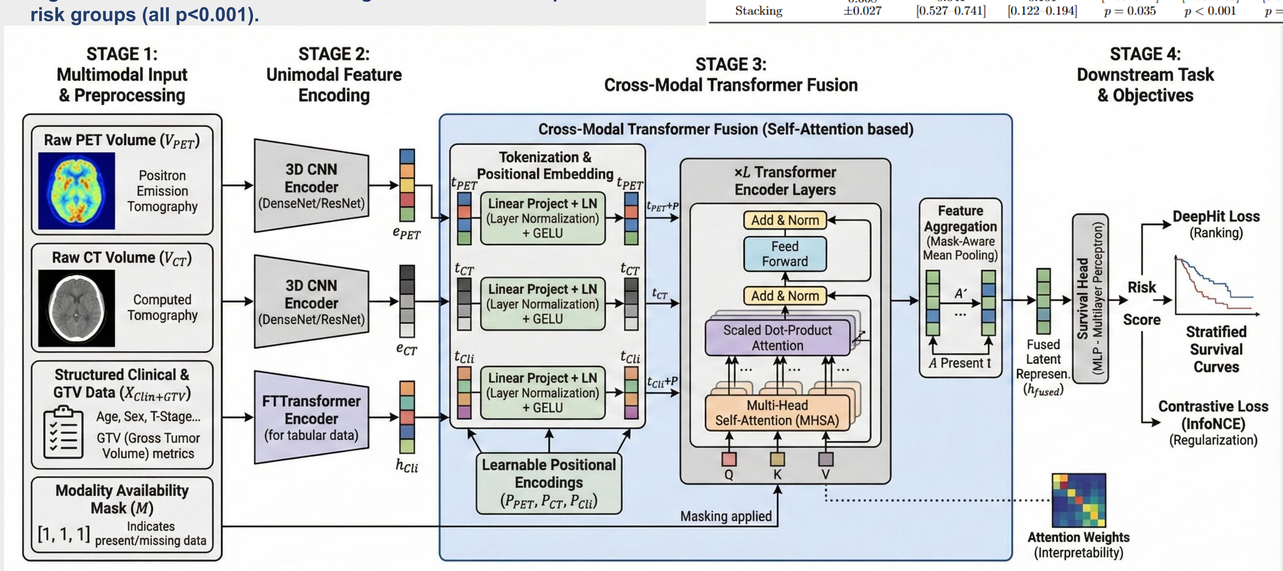

Figure 1: Architecture of the Multimodal Model with Cross-Modal Transformer Joint Fusion for Survival Prediction in HPV-Associated Oropharyngeal Head and Neck Cancers using PET/CT and clinical information. A) Multimodal inputs, including Positron Emission Tomography (PET), Computed Tomography (CT), and Clinical / Gross Tumour Volume (radiomic) data, are acquired. B) Unimodal input encoders, 3D Convolutional Neural Networks for diagnostic images, and a Feature Tokenizer Transformer for structured clinical variables. C) The multimodal fusion module tokenizes these features, adds positional encodings, and processes them through L Transformer Encoder layers. Each layer employs Multi-Head Self-Attention (MHSA) for all-to-all cross-model modelling, with masking for possible patient-specific missing modalities. Features are aggregated via a mask-aware mean pooling into a fused representation (hfused). D) A Multilayer Perceptron (MLP) survival head predicts a risk score, optimized with DeepHit and Contrastive losses. MHSA weights provide interpretability.

DISCUSSION

Risk stratification. Multimodal cross-modal transformer joint fusion models separated HPV+ risk groups that unimodal models did not.

Performance trade-off. Cross-attention provided the best calibration and a balanced combination of discrimination and stratification. Late-fusion attention achieved a higher C-index but less precise risk estimates.

Cross-modal interactions. Attention patterns varied across predicted risk groups (all p<0.001), uncovering risk-dependent relationships among metabolic, anatomic, and clinical information (**Figure 2**).

Clinical implication. These interactions may help identify higher-risk patients within otherwise favorable HPV-associated disease.

Unimodal limitations. Higher model discrimination did not generalize to better risk stratification in test sets. As HPV-associated OPSCC prognosis is inherently multimodal. Clinical variables capture treatment context but miss biological aggressiveness, imaging captures tumour phenotype but lacks patient-level context, and tumour volume features were derived from radiomics.

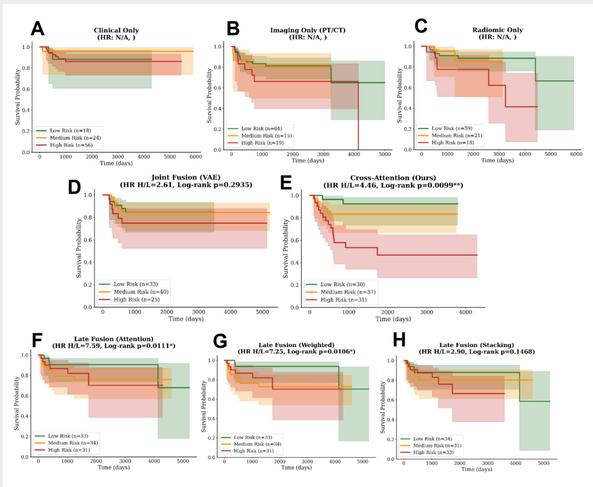

Figure 3: Kaplan-Meier curves for HPV+ test cohort risk stratification (N=105) for each method. Risk groups (low/intermediate/high) defined by training-set-derived tertile cutoffs. (A) Clinical-only, (B) PET/CT Imaging Only, (C) Radiomic-only, (D) Joint Fusion Variational Autoencoder, (E) Joint Fusion Cross-Attention Transformer, (F) Late Fusion Attention, (G) Late Fusion Weighted, (H) Late Fusion Stacking. Multimodal methods (E, F, G) significantly stratified patients into three risk groups with a log-rank test p-value < 0.05; unimodal methods (A-C) showed insufficient stratification.

CONCLUSIONS

Joint PET/CT–clinical fusion transformer improves clinically actionable risk stratification beyond unimodal approaches in HPV-associated OPSCC. Explicit modelling of cross-modal interactions reveals biologically distinct risk groups and may support future treatment-intensification and carefully selected de-escalation studies. External validation on an independent cohort will be required.

REFERENCES

1. Saeed N, et al. A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. arXiv:2509.00367 (2025).
2. Ma B, et al. PET and CT-based DenseNet outperforms advanced deep learning models for outcome prediction of oropharyngeal cancer. Radiother Oncol. 2025;207:110852.
3. Quetin S, Enger SA. Automatic lesion and lymph node segmentation from PET and CT scans of the head and neck region: a HECKTOR 2025 challenge report.
4. Duran J, Zou Y, Vallières M, Enger SA. Beyond single-run metrics with CP-fuse: a rigorous multi-cohort evaluation of clinico-pathological fusion. Mach Learn Appl. 2025:100789.