Precision Medicine and Imaging

Deep learning for fully automated prediction of overall survival in patients with oropharyngeal cancer using FDG-PET imaging: an international retrospective study

Nai-Ming Cheng[1], Jiawen Yao[2], Jinzheng Cai[2], Xianghua Ye[3], Shilin Zhao[4], Kui Zhao[5], Wenlan Zhou[6], Isabella Nogues[7], Yuankai Huo[8], Chun-Ta Liao[9], Hung-Ming Wang[10], Chien-Yu Lin[11], Li-Yu Lee[12], Jing Xiao[13], Le Lu[2], Ling Zhang[2], and Tzu-Chen Yen[1]

N.-M. Cheng and J. Yao contributed equally to this work.

[1]Departments of Nuclear Medicine and Molecular Imaging Center, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC

[2]PAII Inc., Bethesda, MD, USA

[3]Department of Radiotherapy, The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, China

[4]Departments of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[5]Department of PET center, The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, China

[6]NanFang PET Center, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China

[7]Departments of Biostatistics, Harvard University T.H. Chan School of Public Health, Boston, MA, USA

[8]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

[9]Departments of Otorhinolaryngology, Head and Neck Surgery, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC

[10]Departments of Medical Oncology, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC

[11]Departments of Radiation Oncology, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC

[12]Departments of Pathology, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC

[13]Ping An Technology Co., Ltd., Shenzhen, China

Corresponding Authors:

Tzu-Chen Yen,

Department of Nuclear Medicine, Chang Gung Memorial Hospital at Linkou, Taiwan, ROC

5 Fu-Shin Street, Kueishan, Taoyuan 333, Taiwan, ROC

Tel: +886-3-328-1200 ext. 2673

Email: yentc1110@gmail.com

Ling Zhang

PAII Inc., Bethesda, MD

6720b Rockledge Dr, Bethesda, MD 20817

Tel: +1-3195126453

Email: zhangling300@paii-labs.com

Running title: Automated Prediction of Survival in Oropharyngeal Cancer

Word count: 4682

Total number of figures: 3

Total number of tables: 3

Translational Relevance

While rapid technical advances are furthering the application of deep learning in cancer prognostication based on imaging data, the reliance on manually selected slices and segmentation, the failure to account for traditional risk factors, and the limited sample sizes without ethnic diversity are major obstacles for translation into the clinic. Using data from FDG-PET imaging, we devised the first deep learning-based fully-automated tool for predicting overall survival in patients with oropharyngeal squamous cell carcinoma. Our tool revealed a robust performance across different geographic regions, PET scanners, and treatment protocols in a large, international study. On the one hand, such an approach enables an objective, unbiased, and rapid assessment that is suitable for clinical prognostication. On the other hand, the use of our biomarker has the potential to tailor treatment at the individual level.

Abstract

Purpose: Accurate prognostic stratification of patients with oropharyngeal squamous cell carcinoma (OPSCC) is crucial. We developed an objective and robust deep learning-based fully-automated tool called the DeepPET-OPSCC biomarker for predicting overall survival (OS) in OPSCC using [^{18}F]fluorodeoxyglucose PET imaging.

Experimental Design: The DeepPET-OPSCC prediction model was built and tested internally on a discovery cohort (n = 268) by integrating five convolutional neural network models for volumetric segmentation and ten models for OS prognostication. Two external test cohorts were enrolled – the first based on the Cancer Imaging Archive (TCIA) database (n = 353) and the second being a clinical deployment cohort (n = 31) – to assess the DeepPET-OPSCC performance and goodness of fit.

Results: After adjustment for potential confounders, DeepPET-OPSCC was found to be an independent predictor of OS in both discovery and TCIA test cohorts (HR = 2.07; 95% CI 1.31–3.28 and HR = 2.39; 1.38–4.16; both P = 0.002). The tool also revealed good predictive performance, with a c-index of 0.707 (95% CI 0.658–0.757) in the discovery cohort, 0.689 (0.621–0.757) in the TCIA test cohort, and 0.787 (0.675–0.899) in the clinical deployment test cohort; the average time taken was 2 min for calculation per exam. The integrated nomogram of DeepPET-OPSCC and clinical risk factors significantly outperformed the clinical model (AUC at 5 years: 0.801 [95% CI 0.727–0.874] *versus* 0.749 [0.649–0.842]; P = 0.031) in the TCIA test cohort.

Conclusions: DeepPET-OPSCC achieved an accurate OS prediction in patients with OPSCC and enabled an objective, unbiased, and rapid assessment for OPSCC prognostication.

Introduction

Oropharyngeal squamous cell carcinoma (OPSCC) is frequently associated with human papillomaviruses (HPV) infection (1). However, there are significant differences in five-year overall survival (OS) rates between HPV-related (HPV+) and tobacco- and alcohol-related (HPV–) cases (75%−80% *versus* 45%−50%, respectively) (2). Recent years have witnessed a growing interest in less-intensive treatment approaches for HPV+ OPSCC, with the main goal of reducing toxicity while maintaining comparable disease control rates (3, 4). However, there is still insufficient evidence to recommend de-intensified treatment protocols owing to the risk of less favorable outcomes (5, 6). Moreover, these de-escalation therapies depend on patient response to induction chemotherapy (7, 8) which remains unpredictable, particularly in the pretreatment phase (9). More worryingly, there remains a paucity of effective therapies for patients with HPV– OPSCC (3), although a few enhanced therapies for such patients have been investigated (10). In this scenario, novel operator-independent risk stratification tools are eagerly awaited to facilitate and optimize clinical trials by identifying specific patient subgroups who are more likely to benefit from novel therapeutic approaches. This would ultimately make the treatment of OPSCC more personalized and reduce unnecessary morbidity (11, 12).

As for HPV+ OPSCC, PIK3CA mutations have been associated with less favorable disease control in de-escalation trials (13). On the contrary, TRAF3 and CYLD losses have been reported to portend a favorable prognosis (14). With regard to HPV– cases, mutations in p53 have been associated with poor outcomes (15). Moreover, a measure of intratumor genetic heterogeneity (termed quantitative mutant allele tumor heterogeneity) has been linked to unfavorable outcomes (16). Despite intense research on the ability of these tools to comprehensively capture the molecular underpinnings of head and neck malignancies, these

biomarkers have not yet been implemented in clinical practice. Compared with tissue-based biomarker testing, algorithm-guided medical imaging features have inherent advantages in terms of being real-time, non-invasive, independent of sampling bias, and not limited to the portion of tested tissue (17). While radiomics – defined as high throughput extraction of quantitative imaging features – has been successfully used for predicting prognosis in OPSCC (18-21), its reproducible application in everyday practice is limited because of its dependence on manual segmentation and handcrafted features (17).

Deep learning-based artificial neural networks comprise algorithms and techniques that enable computers to identify complex patterns in large data sets without resorting to handcrafted feature extraction. In human cancer imaging, deep learning approaches have increasingly been applied to different steps of the entire workflow (22-24). While rapid technical advances are furthering the application of deep learning in cancer prognostication based on image data (25-28), their implementation in clinic practice remains a major hurdle. Among the methodological barriers, the reliance on manually selected two-dimensional slices and manual segmentation (which have a significant adverse impact in terms of reproducibility), the failure to account for traditional risk factors, and the limited sample sizes without ethnic diversity are major obstacles for translation.

The objective of this study was to develop a scalable, objective, and robust deep learning-based fully-automated tool – termed DeepPET-OPSCC biomarker – for predicting OS in patients with OPSCC using [^{18}F]fluorodeoxyglucose (FDG) PET imaging. DeepPET-OPSCC – which integrates an automated three-dimensional (3D) deep segmentation model with a deep learning Cox model – was subsequently tested in an international multicenter study to validate its applicability and generalizability regardless of potential confounders.

Materials and Methods

Study design

This international retrospective study included three patient cohorts – a discovery cohort, on which the best-fitting prediction models were built and tested internally, and two external test cohorts, on which performance and goodness of fit were assessed. Inclusion criteria were as follows: 1) non-metastatic (M0) OPSCC and absence of other concomitant malignancies; 2) availability of baseline pretreatment PET images covering the head and neck region; 3) treatment with curative intent; and 4) follow-up continued for at least 18 months or until death. Patients without identifiable tumors on PET/CT scans were excluded.

All patients in the three cohorts were staged according to the seventh edition of the American Joint Committee on Cancer (AJCC) staging system. Details are available in the **Supplementary Protocol** (**Sections 1** and **4)**. OS – which was defined as the time from cancer diagnosis to the last follow-up or death from any cause – served as the main outcome measure. Ethics approval for the retrospective review of imaging and clinical data was received from the local ethics committees for the discovery and validation cohorts. The need for informed consent was waived.

Discovery cohort

The discovery cohort included 268 patients who had been treated between June 2006 and December 2017 at the Linkou Chang Gung Memorial Hospital (CGMH; Taiwan, ROC). The CGMH database contained complete information on demographics, clinical characteristics, and therapeutic procedures of each patient and was, thus, selected for model development. FDG-

PET/CT images were acquired using either GE or Siemens scanners, within a median of 9 (IQR 3–14) days from the pathological diagnosis. HPV status was ascertained using p16 immunohistochemistry. According to the CGMH treatment policy, OPSCC patients were treated with concurrent chemoradiotherapy (CCRT), whereas those in T1–T2 stages with no nodal metastasis received radiotherapy or surgery. Patients with advanced-stage OPSCC in a prospective clinical trial received induction chemotherapy, followed by CCRT (IC + CCRT).

External test cohorts

The first test cohort consisted of 353 patients with OPSCC from Western countries. The Cancer Imaging Archive (TCIA) public database was thoroughly queried for PET image data and clinical information of patients who had been treated between October 2003 and November 2014 at six centers (Hôpital Général Juif, Centre Hospitalier Universitaire de Sherbrooke, Hôpital Maisonneuve-Rosemont, and Centre Hospitalier de l'Université de Montréal, Canada; University of Texas MD Anderson Cancer Center, USA; MAASTRO Clinic, the Netherlands). The HPV status, which was available for 44% of the cases, was ascertained by *in situ* hybridization or p16 immunohistochemistry. Most patients received CCRT treatment, whereas others were treated with either single or combined modalities—for example, surgery, radiotherapy, induction chemotherapy, or cetuximab.

The second test cohort included 31 patients with OPSCC from an Asian country. We enrolled patients who had been treated between April 2011 and March 2019 at two hospitals (First Affiliated Hospital of Zhejiang University [ZJU1] and Nanfang Hospital, China) with available baseline PET imaging. Except for one HPV case (based on the results of p16 immunohistochemistry), the HPV status was unknown for all patients. The study patients were

treated with surgery, CCRT, or both. The complete model was locked before deployment in ZJU1.

DeepPET-OPSCC discovery and internal testing

Nested cross-validation

Figure 1A summarizes the discovery and internal testing of the DeepPET-OPSCC prognostic biomarker, which comprises five PET image segmentation models and ten prognostic models (**Supplementary Protocol Sections 2** and **3**). All models were trained in the discovery cohort using nested five-fold cross-validation, with 64%, 16%, and 20% of the data as the training, validation, and test sets at each repeat time (one fold), respectively. The same data-splitting approach was used for segmentation and prognosis. This technique was implemented in order to avoid the overoptimistic issue inherent to conventional cross-validation, as individual DeepPET-OPSCC scores in the discovery cohort were obtained in the setting of internal testing (i.e., test sets in the nested cross-validation) with automated segmentation.

Segmentation models

All PET image volumes were converted to standardized uptake values (SUV) maps/volumes. For generating annotations of tumor and lymph nodes in the discovery cohort, volumetric delineation was performed semi-automatically by an experienced nuclear radiologist (NMC), with 14 years of experience in nuclear imaging and image processing. The segmentation models were built upon the 3D version of nnUNet (29), with extensive data augmentation for improving generalization performance (30). The full description is provided in **Supplementary Protocol** (**Section 2**).

Prognostic models

The prognostic models were trained on three types of 3D region-of-interests (i.e., SUV map, automatically segmented tumor mask, and node-to-tumor [N-T] distance map; **Figure 1C**) using OS time and patient status (alive *versus* dead) as labels. The N-T distance map was included as a region-of-interest type because nodal metastases to the lower neck reflect spread to more distant sites and are associated with reduced OS (2,31). The prognostic model consisted of 3D convolutional neural networks that relied on the Cox proportional hazards assumptions (ConvCox) (32). Nonlinear associations between 3D images and time-dependent censored OS were directly modeled. The architecture (**Figure 1C**) and implementation of our ConvCox network are detailed in **Supplementary Protocol** (**Section 3.2)**.

The following scheme was adopted to train the prognostic models in each of the five folds. For each fold, we separately trained (with extensive data augmentation) two distinct ConvCox models: 1) DeepPET-OPSCC-T with two input channels (SUV and tumor mask), and 2) DeepPET-OPSCC-TN with three input channels (SUV, tumor mask, and N-T distance map). Given that there is more variability in the appearance image (N-T distance map) as compared to that in the binary image (tumor mask), the deep learning model may not capture adequate information in the tumor mask. Therefore, in order to allocate sufficient network capacity for adequately and comprehensively capturing both tumor and lymph node information, we trained the two models separately. The optimal ConvCox models were selected in the validation set based on the highest Harrell's concordance index (c-index) (33) and subsequently tested in the test set. The predicted risk score reflecting the probability of less favorable OS in each test set was normalized by subtracting, for each fold, the mean risk score in the training set. The final continuous DeepPET-OPSCC score was calculated by averaging the DeepPET-OPSCC-T and

DeepPET-OPSCC-TN scores. The nested cross-validation process was repeated five times, thereby yielding five DeepPET-OPSCC-T and five DeepPET-OPSCC-TN models.

In order to determine the DeepPET-OPSCC risk category (i.e., dichotomized into high *versus* low risk), the median value of all DeepPET-OPSCC scores obtained in the test sets was used as the cut-off threshold. Further, the continuous DeepPET-OPSCC score was categorized into three, four, or five risk subgroups using tertiles, quartiles, and quintiles, respectively, of the total risk scores.

External testing

For external testing (**Figure 1B**), the models trained for segmentation and OS prediction were integrated into the UNet and ConvCox ensemble models, respectively. Further, the ten DeepPET-OPSCC-T/TN normalized prediction scores were averaged to obtain the final DeepPET-OPSCC score, which was subsequently dichotomized to obtain the DeepPET-OPSCC risk category based on the previously determined cutoff threshold from the discovery cohort.

Visualization

A renormalized class-activation heatmap was used to visualize/highlight tumor and nodal areas associated with unfavorable OS. Our heatmap represented risks at both the voxel and patient levels for facilitating the visual interpretation of the local and global risks. The heatmap value of each voxel directly reflected its predicted risk score. The heatmap values of all voxels were renormalized to [0, 1] based on the maximal and minimal values in the corresponding training set.

Comparison with other computational approaches

To compare our method to other computational prognostic approaches, we developed three distinct tools – lightweight 3D ResNet-OPSCC (designed for insufficient training data), 2D DeepPET-OPSCC (using the largest tumor and lymph node slices as network input), and a radiomics signature that reflected both tumor and nodal characteristics – which were trained and assessed as DeepPET-OPSCC (**Supplementary Protocol Section 3.5**).

Research reproducibility

The major components of our tool have been made available in open-source repositories and libraries – including PyTorch (https://pytorch.org/), nnUNet (https://github.com/MIC-DKFZ/nnUNet), and SALMON (https://github.com/huangzhii/SALMON). All experimental and implementation methods have been also described in sufficient detail (Supplementary Protocol) to enable independent replication by other researchers. The trained prognostic models, inference code, and an illustrative example of SUV image, tumor mask, and N-T distance map are publicly available through the DeepPET-OPSCC GitHub repository (https://github.com/deep-med/DeepPET-OPSCC-Example). All of the data in the TCIA test cohort can be accessed at The Cancer Imaging Archive (http://www.cancerimagingarchive.net/).

Statistical analysis

This study conforms to the REMARK guidelines (34) and the acceptance criteria set forth by the AJCC for the inclusion of risk models (**Table S1 and S2**) (35). The performance of the automated segmentation model was assessed as described in the **Supplementary Protocol Section 2.6**.

The c-index was used to investigate the predictive ability of the prognostic model. We carried out a time-dependent receiver operating characteristic (ROC) curve analysis and calculated the areas under the curves (AUCs) for OS at two and five years. The overall c-index and AUC in the discovery cohort were calculated by concatenating all normalized scores from the five test sets. In order to assess the improvements in the c-indexes between the compared models, the Student's *t*-test for dependent samples was used (36). A similar approach has been implemented in previous studies (37, 38). The 95% confidence intervals (CIs) for AUC were constructed from 1000 bootstrap replicates of the test sets of discovery cohort and external test cohorts. In addition, the *z*-test was used to compare the differences in bootstrapped AUCs from different models (39).

Univariable and multivariable Cox proportional hazards regression survival analyses were also conducted. The Wald χ^2 test was used to calculate *P* values in multivariable models. Because of missing HPV information (56%) in the TCIA test cohort, the HPV status was not entered into the multivariable model. Because only 26 patients who died had a known HPV status in the TCIA test cohort, cases in the discovery and TCIA test cohorts with existing HPV data were grouped in a unique cohort (i.e., the entire cohort) for multivariable analyses. Due to the limited number of patients (n = 31), events (n = 15), and HPV information (n = 1) in the clinical deployment test cohort, multivariable and subgroup analyses were not performed for this cohort. In addition, smoking information was missing in seven of the eight external centers;

therefore, this variable was investigated only in the discovery cohort. Kaplan-Meier estimate curves were generated for OS, and survival differences were compared with the log-rank test. Further, Spearman's correlation coefficients were calculated to investigate the associations between the DeepPET-OPSCC risk category and clinical variables. Following established criteria for developing nomograms in the field of oncology (40), we devised integrated nomograms from Cox regression coefficients using inputs from the DeepPET-OPSCC score and clinical risk factors. All calculations were performed in R, version 3.6.1. Statistical significance was determined by *P* value <0.05.

Results

Patient characteristics

Table 1 presents the general characteristics of the study participants. Patients in the external test cohorts (n = 384) underwent PET imaging with nine unseen scanners from three vendors (**Supplementary Protocol Table 1**). The HPV status was available for 424 (65%) cases (165 HPV+ and 259 HPV–). Among patients for whom the HPV status was known, there were 211 (79%) and 47 (30%) HPV– cases in the discovery and TCIA test cohorts, respectively. Primary radiotherapy, either with or without chemotherapy, was given to 258 (96%) patients in the discovery cohort as well as to 339 (97%) and 7 (23%) patients in the two external test cohorts, respectively. The remaining patients were treated with primary surgery, either with or without postoperative treatments. Chemotherapy was used for 252 (94%), 255 (72%), and 25 (81%) patients in the discovery and two external test cohorts, respectively. The clinical characteristics

of the entire cohort with known information of HPV status, cT, cN, and cTNM stages are summarized in **Table S3**.

DeepPET-OPSCC

Univariable and multivariable analyses

The distribution of the DeepPET-OPSCC score in the discovery and TCIA test cohorts is depicted in **Figure S1**. The median DeepPET-OPSCC score (–0.12) in the discovery cohort was used as the cutoff to obtain the DeepPET-OPSCC risk category (dichotomized into high risk *versus* low risk), which was a strong predictor of OS in all the three study cohorts (**Figures 2A and 2B**; **Table S4**). After adjustment for age, sex, HPV status, cT stage, cN stage, maximum SUV (SUV_{max}), metabolic tumor volume (MTV), and use of chemotherapy in multivariable analysis, the DeepPET-OPSCC risk category was retained as an independent predictor of OS (discovery cohort: HR 2.07, 95% CI 1.31–3.28; TCIA test cohort: HR 2.39, 95% CI 1.38–4.16; $P = 0.002$; **Table 2, Table S5**). The use of chemotherapy was associated with a reduced mortality in patients from the TCIA test cohort. However, after adjusting for the HPV status, this significance was no longer evident (**Table 2**).

On multivariable analysis, the components of DeepPET-OPSCC (i.e., -T and -TN models) were independent predictors of OS in the discovery and TCIA test cohorts (**Tables S6 and S7**). The continuous DeepPET-OPSCC score was also retained as a strong predictor in the multivariable model (**Table S8**). Validation with additional clinical variables (e.g., smoking) as well as pathological (e.g., tumor grade) and immunohistochemistry-based (e.g., Cyclin D1) markers in the discovery cohort is provided in **Tables S9–S11**.

Prediction accuracy

The c-indices of the DeepPET-OPSCC score for OS were 0.707 (95% CI 0.658–0.757) and 0.689 (95% CI 0.621–0.757) in the discovery and TCIA test cohorts, respectively. The constituents of DeepPET-OPSCC (i.e., -T and -TN models) were also strongly associated with OS (**Table 3**). Nonetheless, ensemble models provided more robust and reliable performance (especially with respect of unseen data) than a single model both in terms of OS prediction and in univariable and multivariable analyses (**Tables 3, S4–S7**, and **S12**). In addition, prognostic markers generated by three other computational approaches of 3D ResNet-OPSCC, 2D DeepPET-OPSCC, and conventional radiomics all underperformed ($P < 0.01$) the DeepPET-OPSCC score in both the discovery and TCIA test cohorts, with the exception of 3D ResNet-OPSCC in the TCIA test cohort ($P = 0.21$; **Table S13**; **Supplementary Protocol, Section 3.5**).

Visualization

Our tool allowed obtaining a renormalized heatmap that can depict risk at both voxel and patient levels through a hot-cold color code (**Figure 3**; **Figure S2A-E**). We found that the DeepPET-OPSCC-T model focused mostly on the tumor's interior, while the DeepPET-OPSCC-TN model tended to fixate on the interface between the tumor and lymph nodes.

Subgroup analyses

The DeepPET-OPSCC risk category retained its ability to predict OS when patients in the entire cohort with a known HPV status were stratified into different subgroups according to HPV, cT, cN, and cTNM stages, or the use of chemotherapy (**Figures S3–S10**). The majority of the study patients were staged as cTNM IVA (113 [70%] of the 161 HPV+ patients and 158 [61%] of the

258 HPV– patients), and the DeepPET-OPSCC risk category was capable of predicting OS in the two subgroups (for high *versus* low risk, HR 4.20, 95% CI 1.18–14.92; $P = 0.016$; **Figure 2C**; HR 2.64, 95% CI 1.65–4.21; $P < 0.001$; **Figure 2D**). We also investigated the relationship between the DeepPET-OPSCC risk category and the usage of induction chemotherapy before CCRT. For patients with HPV– and TNM stage IVB cancer, induction chemotherapy was associated with an inferior OS than CCRT alone in the DeepPET-OPSCC high-risk group (HR 2.44, 95% CI 1.03–5.79; $P = 0.037$; **Table S14**, **Figure S14**).

Correlations between DeepPET-OPSCC and clinical parameters

The DeepPET-OPSCC risk category was significantly correlated with a number of clinical parameters, including sex, HPV status, cT stage, cN stage, cTNM stage, SUV_{max}, and MTV – both in the entire cohort (**Table S15**) and TCIA test cohorts (**Table S16**). Scatter plots of the relationships between SUV_{max}/MTV and DeepPET-OPSCC scores are provided in **Figure S15**. A large proportion of HPV+ (e.g., among 104 HPV+ cases, DeepPET-OPSCC identified 82 patients being at low risk and 22 as being at high risk), cT1-cT3, cN0-cN2, and cTNM stage I-IVA diseases were classified as being at low risk by DeepPET-OPSCC in the validation cohort, thereby supporting the clinical utility of DeepPET-OPSCC in Western populations.

Nomograms

Finally, we devised integrated nomograms by combining DeepPET-OPSCC score and the clinical risk factors (i.e., age; sex; HPV status; and cT, cN, and cTNM stages). In the subgroup of patients with known HPV status, the five-year AUCs for the integrated nomogram were 0.793

(95% CI 0.749–0.834) and 0.801 (95% CI 0.727–0.874) in the discovery and TCIA test cohorts, respectively, thereby outperforming clinical models and each individual risk factor (e.g., clinical model: 0.749 [95% 0.649–0.842], clinical model plus MTV: 0.754 [0.659–0.843], HPV: 0.624 [0.530–0.729], and AJCC cTNM stages: 0.517 [95% CI 0.423–0.614] in the TCIA test cohort with known HPV status; $P < 0.05$; **Figure 2E**; **Table S17**). A full description—including c-indices and two-year AUCs—is provided in **Tables S18 and S19** and **Figures S16 and S17** as well as in **Supplementary Protocol Section 5**. On analyzing all of these results, the single DeepPET-OPSCC score was never found to underperform ($P > 0.05$) the clinical model when the HPV status was missing in both the discovery and TCIA test cohorts.

Clinical deployment: fully-automated prediction

Different procedures – including SUV conversion, segmentation, and prognostic prediction – were assembled into a unique fully automated processing pipeline, whose performance was analyzed in the clinical deployment test cohort. The mean processing time for the complete automated process was 2 min 6 sec per PET exam on an NVIDIA Titan RTX-6000 GPU. The fully-automated tool significantly predicted OS ($P = 0.002$; **Figure S18**) with a c-index of 0.787 (95% CI 0.675–0.899), thereby indicating a robust performance across different geographic regions, PET scanners, and treatment protocols. In this cohort, the DeepPET-OPSCC outperformed the clinical model and each individual risk factor when the HPV was missing (**Figure S19**).

Discussion

Using data from FDG-PET imaging, we devised a deep learning-based fully-automated tool – based on deep segmentation and prognostication models – for predicting OS in patients with OPSCC. The system, which captured PET information from both the primary tumor and lymph nodes, offered a rapid (calculation time: ~2 min) prediction of OS and performed satisfactorily in an international multicenter study. Notably, the DeepPET-OPSCC risk category was retained in the multivariable analysis as an independent predictor of OS in all cohorts, with an approximately two-fold increased risk for mortality in the high-risk *versus* low-risk group. Further, the nomogram combining the DeepPET-OPSCC score, age, sex, HPV status, cT, cN, and cTNM stage significantly improved the prediction accuracy of OS.

Our work is currently the largest computational imaging-based prognostic study conducted in patients with OPSCC (18-21). The DeepPET-OPSCC score had c-indices of 0.689–0.787 for the prediction of OS from baseline imaging—these values being substantially higher than those previously reported (0.59–0.63) for radiomics markers (19, 20). In addition, our tool showed a robust performance on PET data from different geographic regions, scanners, and treatment protocols. While the discovery cohort consisted of patients treated primarily with combined radiotherapy and chemotherapy, the DeepPET-OPSCC biomarker is applicable to patients primarily treated with surgery or who did not receive chemotherapy. Given that the AJCC principle requires a staging system that must be applicable to any treatment approach that meets accepted guidelines (2), the DeepPET-OPSCC score – which remained an independent predictor after adjustment for different treatments – might have the potential to complement the future staging system. In addition, our automated tool is highly objective and reproducible.

Recent years have witnessed a growing interest in the development of deep learning-based prognostic systems based on imaging findings for patients with malignancies (25-28).

However, published approaches have inherent limitations, which include the need for manual segmentation and the inability to extract the three-dimensional tumor characteristics from two-dimensional slices. Moreover, fully automated prediction systems may improve the objectiveness and are currently gaining traction (41).

Our prognostic tool was implemented on FDG-PET images, which exhibit high image contrast and small variation among various acquisitions and reconstructions (42), thereby making fully-automated image analysis a more promising task. While the segmentation model (nnUNet) is clinically applicable for distinct segmentation tasks (23,29), extensive data augmentation enabled the generalization of this model to unseen domains (30). The ConvCox prognostic model developed in our study is a regression network that has the capacity to learn time-dependent events directly from all the available data. This is a highly desirable feature for prognostic applications, where the number of patients with complete baseline imaging data tends to be limited. Moreover, the ConvCox network is designed with consideration of several architectural modifications, optimized training and inference configurations, incorporation of domain knowledge (e.g., N-T distance map), and the model ensemble of -T and -TN constituents (focusing on the tumor itself and its relationship with lymph nodes, respectively), thereby improving its robustness and generalization. In deep learning practice, assembling models trained from several training-validation data splits (e.g., five models trained from nested five-fold cross-validation in the current study) is a commonly utilized solution that is efficient and effective in improving model robustness on unseen data (23, 29, 41).

The DeepPET-OPSCC outperformed all other clinical variables for OS prediction at two and five years. In addition, it was found to correlate with known clinical and PET-derived prognostic parameters. Taken together, these observations indicate an association between the

prognostic features captured by deep learning and established prognostic markers in OPSCC, including the HPV and AJCC stages. These interrelationships may also explain why DeepPET-OPSCC performed similarly well in Asian and Western populations, despite different disease characteristics (e.g., different proportions of HPV+ cases and five-year OS). Moreover, DeepPET-OPSCC followed the path of the eighth AJCC staging system, which downstaged stage IV to stages I–III for HPV+ OPSCC. Accordingly, 182 (67%) of the 271 cases with stage IV disease in the TCIA test cohort (70% HPV+) were classified as being at low risk by the DeepPET-OPSCC.

Further, our tool enabled us to obtain a renormalized heatmap that can depict risk at both patient and voxel levels through a hot-cold color encoding. While we hypothesize that personalized radiation plans with higher tumoricidal doses could potentially target the identified high-risk regions (12, 43), this requires further investigation.

The application of DeepPET-OPSCC enabled the identification of different prognostic subgroups even when current classification approaches (i.e., HPV and AJCC stages) were applied. For example, we were able to show that certain subgroups of HPV+ patients with AJCC stage IVA or N2 disease (**Figures 2C**, **S5, and S11–S13**) have favorable outcomes and may benefit from less intensive treatment protocols (e.g., de-intensified radiotherapy or chemoradiotherapy, which have been shown to achieve clinically favorable results for HPV+ patients with respect to induction chemotherapy response (7, 8), without evidence of hypoxia on baseline or inter-treatment PET imaging (9), or T0-T2, N0-N2c OPSCC [AJCC seventh edition]) (44). Conversely, certain subgroups of HPV– patients with cT1-3, T4a, N1, N2, AJCC III, or IVA stages (**Figures 2D**, **S7–S9, and S11–S13**) had a dismal prognosis and, thus, may be candidates for more aggressive treatment strategies (e.g., the combination of an antagonist of the

multiple inhibitor-of-apoptosis protein [Debio 1143] with chemoradiotherapy outperformed high-dose chemoradiotherapy in patients with stages III, IVA, and IVB [AJCC seventh edition] head and neck cancer [58% are HPV– OPSCC]) (10). Interestingly, CCRT was associated with a better OS compared to induction chemotherapy and CCRT in patients with the most advanced disease stage (HPV– and stage IVB) and a high-risk DeepPET-OPSCC category. This can be explained by the observation that higher toxicity delays or even prevents patients from completing subsequent CCRT, which is critical for maximizing OS (3).

Several caveats of our study must be considered. First, the performance of the DeepPET-OPSCC prognostic biomarker needs to be tested in larger longitudinal investigations. Second, unavailable data on HPV status for several patients in the TCIA and clinical deployment test cohorts pose a limitation regarding the ability to generalize our conclusions with regard to the presence or absence of HPV infections. Third, the retrospective nature of the study did not permit the application of the more recent (eight edition) AJCC staging system, although this is likely non-influential on our main conclusions. Fourth, the automated tool may be unsuitable to segment a minor percentage (1−3%) of early-stage tumors, which will ultimately require manual segmentation. Finally, we selected cut-off values for risk categorization based on the Asian population – with most patients being HPV–. In future prospectively designed studies with larger sample sizes, it might be reasonable to select more suitable cut-off values separately for HPV+ and HPV– patients.

In summary, the primary novelty of this large international study lies in the possibility of obtaining an accurate prediction of OS in patients with OPSCC through a fully-automated deep learning-based tool. On the one hand, such an approach enables an objective, unbiased, and rapid

assessment that is suitable for clinical prognostication. On the other hand, the use of our biomarker has the potential to tailor treatment at the individual level.

Disclosure of Potential Conflicts of Interest

The authors declare no conflicts of interest.

Authors' contributions

Conception and design: Nai-Ming Cheng, Jiawen Yao, Ling Zhang, and Tzu-Chen Yen

Development of methodology: Jiawen Yao, Ling Zhang, and Jinzheng Cai

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): Nai-Ming Cheng, Xianghua Ye, Kui Zhao, Wenlan Zhou, Chun-Ta Liao, Hung-Ming Wang, Chien-Yu Lin, Li-Yu Lee, Ling Zhang, and Tzu-Chen Yen

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Nai-Ming Cheng, Jiawen Yao, Jinzheng Cai, Xianghua Ye, Shilin Zhao, Hung-Ming Wang, Jing Xiao, Le Lu, Ling Zhang, and Tzu-Chen Yen

Writing, review, and/or revision of the manuscript: Ling Zhang, Jiawen Yao, Nai-Ming Cheng, Hung-Ming Wang, Isabella Nogues, Le Lu, and Tzu-Chen Yen

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Nai-Ming Cheng, Jiawen Yao, Jinzheng Cai, Xianghua Ye, Ling Zhang, and Tzu-Chen Yen

Study supervision: Le Lu and Tzu-Chen Yen

Acknowledgments

The authors are indebted to The Cancer Imaging Archive for data availability.

References

1. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer N Engl J Med **2010**;363(1):24-35.
2. O'Sullivan B, Huang SH, Su J, Garden AS, Sturgis EM, Dahlstrom K, *et al.* Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study Lancet Oncol **2016**;17(4):440-51.
3. Chow LQM. Head and Neck Cancer. N Engl J Med **2020**;382(1):60-72.
4. Cramer JD, Burtness B, Le QT, Ferris RL. The changing therapeutic landscape of head and neck cancer. Nat Rev Clin Oncol **2019**;16(11):669-83.
5. Gillison ML, Trotti AM, Harris J, Eisbruch A, Harari PM, Adelstein DJ, *et al.* Radiotherapy plus cetuximab or cisplatin in human papillomavirus-positive oropharyngeal cancer (NRG Oncology RTOG 1016): a randomised, multicentre, non-inferiority trial. Lancet (London, England) **2019**;393(10166):40-50 doi 10.1016/s0140-6736(18)32779-x.
6. Mehanna H, Robinson M, Hartley A, Kong A, Foran B, Fulton-Lieuw T, *et al.* Radiotherapy plus cisplatin or cetuximab in low-risk human papillomavirus-positive oropharyngeal cancer (De-ESCALaTE HPV): an open-label randomised controlled phase 3 trial. Lancet (London, England) **2019**;393(10166):51-60 doi 10.1016/s0140-6736(18)32752-1.
7. Chen AM, Felix C, Wang PC, Hsu S, Basehart V, Garst J, *et al.* Reduced-dose radiotherapy for human papillomavirus-associated squamous-cell carcinoma of the oropharynx: a single-arm, phase 2 study. The Lancet Oncology **2017**;18(6):803-11 doi 10.1016/s1470-2045(17)30246-2.
8. Marur S, Li S, Cmelak AJ, Gillison ML, Zhao WJ, Ferris RL, *et al.* E1308: Phase II Trial of Induction Chemotherapy Followed by Reduced-Dose Radiation and Weekly Cetuximab in Patients With HPV-Associated Resectable Squamous Cell Carcinoma of the Oropharynx- ECOG-ACRIN Cancer Research Group. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2017**;35(5):490-7 doi 10.1200/jco.2016.68.3300.
9. Riaz N, Sherman E, Pei X, Schöder H, Grkovski M, Paudyal R, *et al.* Precision Radiotherapy: Reduction in Radiation for Oropharyngeal Cancer in the 30 ROC Trial. Journal of the National Cancer Institute **2021** doi 10.1093/jnci/djaa184.
10. Sun XS, Tao Y, Le Tourneau C, Pointreau Y, Sire C, Kaminsky MC, *et al.* Debio 1143 and high-dose cisplatin chemoradiotherapy in high-risk locoregionally advanced squamous cell carcinoma of the head and neck: a double-blind, multicentre, randomised, phase 2 study. The Lancet Oncology **2020**;21(9):1173-87 doi 10.1016/s1470-2045(20)30327-2.

11. Budach V, Tinhofer I. Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer: a systematic review. The Lancet Oncology **2019**;20(6):e313-e26 doi 10.1016/s1470-2045(19)30177-9.
12. Caudell JJ, Torres-Roca JF, Gillies RJ, Enderling H, Kim S, Rishi A, *et al.* The future of personalised radiotherapy for head and neck cancer. The Lancet Oncology **2017**;18(5):e266-e73 doi 10.1016/s1470-2045(17)30252-8.
13. Beaty BT, Moon DH, Shen CJ, Amdur RJ, Weiss J, Grilley-Olson J, *et al.* PIK3CA Mutation in HPV-Associated OPSCC Patients Receiving Deintensified Chemoradiation. Journal of the National Cancer Institute **2020**;112(8):855-8 doi 10.1093/jnci/djz224.
14. Hajek M, Sewell A, Kaech S, Burtness B, Yarbrough WG, Issaeva N. TRAF3/CYLD mutations identify a distinct subset of human papillomavirus-associated head and neck squamous cell carcinoma. Cancer **2017**;123(10):1778-90 doi 10.1002/cncr.30570.
15. Carlos de Vicente J, Junquera Gutiérrez LM, Zapatero AH, Fresno Forcelledo MF, Hernández-Vallejo G, López Arranz JS. Prognostic significance of p53 expression in oral squamous cell carcinoma without neck node metastases. Head & neck **2004**;26(1):22-30 doi 10.1002/hed.10339.
16. Rosenberg AJ, Vokes EE. Optimizing Treatment De-Escalation in Head and Neck Cancer: Current and Future Perspectives. The oncologist **2021**;26(1):40-8 doi 10.1634/theoncologist.2020-0303.
17. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, *et al.* Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin **2019**;69(2):127-57 doi 10.3322/caac.21552.
18. Cheng NM, Fang YD, Tsan DL, Lee LY, Chang JT, Wang HM, *et al.* Heterogeneity and irregularity of pretreatment (18)F-fluorodeoxyglucose positron emission tomography improved prognostic stratification of p16-negative high-risk squamous cell carcinoma of the oropharynx. Oral Oncol **2018**;78:156-62.
19. Haider SP, Zeevi T, Baumeister P, Reichel C, Sharaf K, Forghani R, *et al.* Potential Added Value of PET/CT Radiomics for Survival Prognostication beyond AJCC 8th Edition Staging in Oropharyngeal Squamous Cell Carcinoma. Cancers (Basel) **2020**;12(7) doi 10.3390/cancers12071778.
20. Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, *et al.* External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta oncologica (Stockholm, Sweden) **2015**;54(9):1423-9 doi 10.3109/0284186x.2015.1061214.
21. Wu J, Gensheimer MF, Zhang N, Guo M, Liang R, Zhang C, *et al.* Tumor Subregion Evolution-Based Imaging Features to Assess Early Response and Predict Prognosis in Oropharyngeal Cancer. J Nucl Med **2020**;61(3):327-36 doi 10.2967/jnumed.119.230037.
22. Kann BH, Hicks DF, Payabvash S, Mahajan A, Du J, Gupta V, *et al.* Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2020**;38(12):1304-11 doi 10.1200/jco.19.02031.
23. Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, *et al.* Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. The Lancet Oncology **2019**;20(5):728-40 doi 10.1016/s1470-2045(19)30098-1.
24. Luo H, Xu G, Li C, He L, Luo L, Wang Z, *et al.* Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. The Lancet Oncology **2019**;20(12):1645-54 doi 10.1016/s1470-2045(19)30637-0.
25. Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep **2019**;9(1):2764 doi 10.1038/s41598-019-39206-1.

26. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med **2018**;15(11):e1002711 doi 10.1371/journal.pmed.1002711.
27. Jiang Y, Jin C, Yu H, Wu J, Chen C, Yuan Q, *et al.* Development and Validation of a Deep Learning CT Signature to Predict Survival and Chemotherapy Benefit in Gastric Cancer: A Multicenter, Retrospective Study. Ann Surg **2020** doi 10.1097/sla.0000000000003778.
28. Peng H, Dong D, Fang MJ, Li L, Tang LL, Chen L, *et al.* Prognostic Value of Deep Learning PET/CT-Based Radiomics: Potential Role for Future Individual Induction Chemotherapy in Advanced Nasopharyngeal Carcinoma. Clin Cancer Res **2019**;25(14):4271-9 doi 10.1158/1078-0432.Ccr-18-3065.
29. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **2021**;18(2):203-11 doi 10.1038/s41592-020-01008-z.
30. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. 2017 International Conference on Learning Representations; Toulon, France.
31. Xing Y, Zhang J, Lin H, Gold KA, Sturgis EM, Garden AS, *et al.* Relation between the level of lymph node metastasis and survival in locally advanced head and neck squamous cell carcinoma. Cancer **2016**;122(4):534-45 doi 10.1002/cncr.29780.
32. Yao J, Shi Y, Lu L, Xiao J, Zhang L. DeepPrognosis: Preoperative Prediction of Pancreatic Cancer Survival and Surgical Margin via Contrast-Enhanced CT Imaging. 2020 International Conference on Medical Image Computing and Computer Assisted Intervention; San Miguel, Peru.
33. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama **1982**;247(18):2543-6.
34. Sauerbrei W, Taube SE, McShane LM, Cavenagh MM, Altman DG. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): An Abridged Explanation and Elaboration. Journal of the National Cancer Institute **2018**;110(8):803-11 doi 10.1093/jnci/djy088.
35. Kattan MW, Hess KR, Amin MB, Lu Y, Moons KG, Gershenwald JE, *et al.* American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA Cancer J Clin **2016**;66(5):370-4 doi 10.3322/caac.21339.
36. Haibe-Kains B, Desmedt C, Sotiriou C, Bontempi G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics (Oxford, England) **2008**;24(19):2200-8 doi 10.1093/bioinformatics/btn374.
37. Qiang M, Li C, Sun Y, Sun Y, Ke L, Xie C, *et al.* A Prognostic Predictive System Based on Deep Learning for Locoregionally Advanced Nasopharyngeal Carcinoma. Journal of the National Cancer Institute **2020** doi 10.1093/jnci/djaa149.
38. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based Deep Learning Model for Predicting Disease-Free Survival in Patients with Lung Adenocarcinomas. Radiology **2020**;296(1):216-24 doi 10.1148/radiol.2020192764.
39. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **2011**;12(1):77 doi 10.1186/1471-2105-12-77.
40. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2008**;26(8):1364-70 doi 10.1200/jco.2007.12.9791.
41. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, *et al.* Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet (London, England) **2020**;395(10221):350-60 doi 10.1016/s0140-6736(19)32998-8.

42. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer **2012**;48(4):441-6 doi 10.1016/j.ejca.2011.11.036.
43. Horsman MR, Mortensen LS, Petersen JB, Busk M, Overgaard J. Imaging hypoxia to improve radiotherapy outcome. Nature reviews Clinical oncology **2012**;9(12):674-87 doi 10.1038/nrclinonc.2012.171.
44. Chera BS, Amdur RJ, Green R, Shen C, Gupta G, Tan X, *et al.* Phase II Trial of De-Intensified Chemoradiotherapy for Human Papillomavirus-Associated Oropharyngeal Squamous Cell Carcinoma. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2019**;37(29):2661-9 doi 10.1200/jco.19.01007.

Figure legends

Figure 1. Flowchart for discovery and external testing of the DeepPET-OPSCC prognostic biomarker.

(A) The DeepPET-OPSCC biomarker consists of five UNet segmentation models and ten convolutional Cox (ConvCox) prognostic models. All models were trained by nested five-fold cross-validation in the discovery cohort, with 64%, 16%, and 20% of all data considered as training, validation, and test sets for each repeat time (one fold), respectively. For each fold, 3-dimensional (3D) SUV images and the corresponding manual masks were used to train and validate a UNet model, which was subsequently applied to the test set to segment the tumor and lymph nodes. Based on these results, the node-to-tumor (N-T) distance maps were generated. Thereafter, 3D regions-of-interest and the corresponding OS time and status were used to train and tune two distinct ConvCox models: 1) a DeepPET-OPSCC-T model with two input channels (SUV and tumor mask), and 2) a DeepPET-OPSCC-TN model with three input channels (SUV, tumor mask, and N-T distance map). The optimal ConvCox models were subsequently tested in the test set to predict risk scores, thereby reflecting the probabilities of less favorable OS. Upon completion of five folds, DeepPET-OPSCC scores were obtained for all data in the discovery cohort for the purpose of the internal test setting. (B) For external testing, the five UNet and ten

ConvCox models were integrated to generate the DeepPET-OPSCC score. The median value of all DeepPET-OPSCC scores in the discovery cohort was used as the cutoff threshold to classify patients as being at high *versus* low risk. (C) Architecture, input, and output of the 3D ConvCox network in the DeepPET-OPSCC-T/-TN prognostic models. CGMH, Chang Gung Memorial Hospital; SUV, standardized uptake values; -T1, DeepPET-OPSCC-T model 1; -TN1, DeepPET-OPSCC-TN model 1; TCIA, The Cancer Imaging Archive.

Figure 2. Kaplan-Meier plots and time-dependent ROC curves for the DeepPET-OPSCC biomarker.

Patients in the discovery cohort (A) and TCIA test cohort (B) stratified according to DeepPET-OPSCC risk category. HPV+ patients and cTNM stage IVA disease (C) stratified according to DeepPET-OPSCC risk category. HPV– patients and cTNM stage IVA disease (D) stratified according to DeepPET-OPSCC risk category. **Figures S11-S13** depict Kaplan-Meier plots using DeepPET-OPSCC risk categories (with three, four, or five groups defined by tertiles, quartiles, and quintiles, respectively, of the risk scores in the discovery cohort) similar as **Figure 2 (A)–(D)**. (E) AUCs at five years were used to assess the prognostic accuracy of the integrated nomogram (combining the DeepPET-OPSCC score with clinical risk factors), clinical model, DeepPET-OPSCC score, and individual clinical risk factors (full description provided in **Supplementary Protocol Section 5**). HR, hazard ratio; HPV, human papillomaviruses; TCIA, The Cancer Imaging Archive; ROC, receiver operating characteristic; AUC, area under the curve.

Figure 3. Examples of 3D PET images (consecutive image slices), corresponding activation maps (heatmaps), and two enlarged images with heatmaps for better visual observation.

In this illustrative example, auto-segmented tumors and lymph node boundaries are indicated by red and green curves, respectively. The PET images are anonymized by blocking the eye region with black boxes.

Table 1. Clinical characteristics in the discovery, TCIA test, and clinical deployment test cohorts

Characteristic	Discovery cohort (n=268)	TCIA test cohort (n=353)	TCIA test cohort with known HPV status (n=155)	Clinical deployment test cohort (n=31)
Age, years	53 (47–60)	61 (54–67)	61 (55–65)	59 (55–65)
Age, years				
<55	154 (58%)	99 (28%)	39 (25%)	8 (26%)
≥55	114 (43%)	254 (72%)	116 (75%)	23 (74%)
Sex				
Female	22 (8%)	75 (21%)	30 (19%)	8 (26%)
Male	246 (92%)	278 (79%)	125 (81%)	23 (74%)
HPV status				
+	57 (21%)	108 (31%)	108 (70%)	0
–	211 (79%)	47 (13%)	47 (30%)	1 (3%)
Missing	0	198 (56%)	0	30 (97%)
cT stage (AJCC seventh edition)				
cT1	14 (5%)	54 (15%)	26 (17%)	5 (16%)
cT2	85 (32%)	149 (42%)	63 (41%)	9 (29%)
cT3	55 (21%)	87 (25%)	38 (25%)	1 (3%)
cT4a	79 (30%)	51 (14%)	21 (14%)	12 (39%)
cT4b	35 (13%)	7 (2%)	3 (2%)	2 (6%)
cT4 (substage missing)	0	5 (1%)	4 (3%)	2 (6%)
cN stage (AJCC seventh edition)				
cN0	57 (21%)	50 (14%)	23 (15%)	10 (32%)
cN1	25 (9%)	35 (10%)	17 (11%)	6 (19%)
cN2	168 (63%)	247 (70%)	108 (70%)	12 (39%)
cN3	18 (7%)	21 (6%)	7 (5%)	3 (10%)
cTNM stage (AJCC seventh edition)				
I	4 (2%)	6 (2%)	5 (3%)	3 (10%)
II	23 (9%)	23 (7%)	10 (7%)	2 (6%)
III	32 (12%)	48 (14%)	20 (13%)	3 (10%)
IVA	163 (61%)	244 (69%)	108 (70%)	19 (61%)
IVB	46 (17%)	28 (8%)	9 (6%)	4 (13%)
IV (missing substage)	0	4 (1%)	3 (2%)	0
Primary treatment				
Surgery	10 (4%)	14 (3%)	5 (3%)	24 (77%)
Radiotherapy	258 (96%)	339 (97%)	150 (97%)	7 (23%)
Chemotherapy				
Yes	252 (94%)	255 (72%)	104 (67%)	25 (81%)
No	16 (6%)	98 (28%)	51 (33%)	6 (19%)
Follow-up time, years	2.8 (1.5–5.6)	4.3 (2.9–6.6)	3.9 (2.8–5.5)	2.3 (1.3–2.8)
Event				
Death	127 (53%)	70 (20%)	27 (17%)	15 (48%)
Overall survival rate (95% CI)				
2 years	67.4% (62.0–73.3)	91.4% (88.6–94.4)	91.6% (87.3–96.1)	63.6% (48.5–83.4)
5 years	50.0% (44.0–56.8)	79.9% (75.2–84.9)	79.0% (71.1–87.7)	44.8% (28.2–71.2)

Note: Data are expressed as medians (interquartile ranges) or counts (percentages) unless otherwise specified.
TCIA, The Cancer Imaging Archive; HPV, human papillomavirus; AJCC, American Joint Committee on Cancer; CI, confidence interval.

Table 2. Multivariable Cox regression analysis of overall survival in the discovery, TCIA test, and entire (with known HPV status) cohorts

Variable	Discovery cohort (n=268, events=127)		TCIA test cohort (n=348, events=70)		Entire cohort with known HPV status (n=419, events=153)	
	HR (95% CI)	*P*	HR (95% CI)	*P*	HR (95% CI)	*P*
DeepPET-OPSCC risk category						
Low risk	Reference	..	Reference	..	Reference	..
High risk	2.07 (1.31–3.28)	0.002	2.39 (1.38–4.16)	0.002	2.24 (1.50–3.39)	<0.001
Age, years						
<55	Reference	..	Reference	..	Reference	..
≥55	0.95 (0.65–1.40)	0.804	2.21 (1.18–4.11)	0.013	0.86 (0.61–1.21)	0.388
Sex						
Female	Reference	..	Reference	..	Reference	..
Male	1.38 (0.54–3.52)	0.506	1.96 (0.91–4.19)	0.084	1.37 (0.65–2.89)	0.408
HPV						
–	Reference	..	..	..	Reference	..
+	0.19 (0.09–0.41)	<0.001	..	..	0.24 (0.14–0.41)	<0.001
cT stage	..	0.012	..	0.019	..	0.003
cT1	0.88 (0.24–3.15)	0.839	0.54 (0.21–1.35)	0.185	1.18 (0.48–2.96)	0.714
cT2	Reference	..	Reference	..	Reference	..
cT3	1.75 (0.91–3.52)	0.093	1.61 (0.81–3.18)	0.171	2.05 (1.17–3.60)	0.012
cT4a	2.96 (1.53–5.73)	0.001	3.43 (1.42–8.29)	0.006	3.27 (1.82–5.88)	<0.001
cT4b	2.09 (0.96–4.53)	0.064	3.54 (0.98–12.76)	0.054	2.70 (1.34–5.44)	0.005
cN stage	..	<0.001	..	0.006	..	0.004
cN0	Reference	..	Reference	..	Reference	..
cN1	2.41 (1.12–5.22)	0.025	1.60 (0.61–4.20)	0.341	2.18 (1.10–4.33)	0.026
cN2	2.41 (1.38–4.20)	0.002	1.08 (0.50–2.31)	0.851	2.29 (1.37–3.82)	0.002
cN3	4.96 (2.28–10.80)	<0.001	4.27 (1.51–12.08)	0.006	3.36 (1.63–6.90)	0.001
SUV_{max}[a]						
<14.65	Reference	..	Reference	..	Reference	..
≥14.65	0.60 (0.40–0.88)	0.010	1.50 (0.85–2.65)	0.163	0.75 (0.53–1.07)	0.113
MTV[a]						
<22.66 cm^3	Reference	..	Reference	..	Reference	..
≥22.66 cm^3	1.18 (0.72–1.95)	0.509	0.40 (0.19–0.83)	0.014	0.88 (0.56–1.37)	0.571
Chemotherapy						
No	Reference		Reference			
Yes	0.50 (0.23-1.09)	0.080	0.45 (0.26-0.80)	0.006	0.76 (0.48-1.21)	0.245

TCIA, The Cancer Imaging Archive; HPV, human papillomavirus; SUV_{max}, maximum standardized uptake value; MTV, metabolic tumor volume.
[a] Cutoff threshold was the median value in the discovery cohort.

Table 3. Harrell's concordance index (c-index), hazard ratio (HR), and area under the curve (AUC) at 5 years, all with 95% confidence intervals (CIs), of different deep learning and radiomics approaches evaluated on the discovery and TCIA test cohorts

Methods	**Discovery cohort (n=268)**				**TCIA test cohort (n=353)**			
	c-index	**HR**	***P***	**5yr AUC**	**c-index**	**HR**	***P***	**5yr AUC**
DeepPET-OPSCC	0.707 (0.658-0.757)	3.17 (2.18-4.63)	<0.001	0.728 (0.677-0.777)	0.689 (0.621-0.757)	3.15 (1.97-5.05)	<0.001	0.669 (0.600-0.743)
DeepPET-OPSCC-T	0.702 (0.652-0.752)	3.07 (2.11-4.46)	<0.001	0.723 (0.670-0.774)	0.672 (0.604-0.739)	2.89 (1.81-4.63)	<0.001	0.682 (0.623-0.743)
DeepPET-OPSCC-TN	0.682 (0.632-0.733)	2.82 (1.95-4.09)	<0.001	0.705 (0.663-0.754)	0.692 (0.625-0.760)	2.71 (1.68-4.35)	<0.001	0.664 (0.595-0.738)
3D ResNet-OPSCC	0.646 (0.595-0.697)	1.95 (1.36-2.79)	<0.001	0.638 (0.584-0.699)	0.665 (0.599-0.731)	1.68 (1.05-2.69)	0.031	0.662 (0.604-0.719)
3D ResNet-OPSCC-T	0.633 (0.583-0.683)	1.87 (1.31-2.68)	<0.001	0.612 (0.547-0.674)	0.676 (0.616-0.736)	1.98 (1.24-3.17)	0.005	0.656 (0.598-0.715)
3D ResNet-OPSCC-TN	0.627 (0.575-0.678)	1.88 (1.32-2.69)	<0.001	0.623 (0.566-0.677)	0.657 (0.591-0.724)	2.10 (1.30-3.38)	0.002	0.661 (0.602-0.719)
2D DeepPET-OPSCC	0.605 (0.552-0.658)	1.92 (1.35-2.73)	<0.001	0.600 (0.542-0.657)	0.591 (0.519-0.663)	1.61 (1.00-2.60)	0.051	0.550 (0.478-0.621)
2D DeepPET-OPSCC-T	0.616 (0.564-0.668)	2.01 (1.41-2.88)	<0.001	0.621 (0.566-0.678)	0.572 (0.498-0.647)	1.37 (0.84-2.22)	0.21	0.541 (0.469-0.615)
2D DeepPET-OPSCC-TN	0.586 (0.533-0.638)	1.49 (1.05-2.12)	0.026	0.575 (0.520-0.631)	0.596 (0.526-0.667)	1.77 (1.08-2.89)	0.024	0.563 (0.501-0.629)
Radiomics signature	0.621 (0.570-0.672)	1.85 (1.30-2.65)	<0.001	0.619 (0.560-0.676)	0.608 (0.538-0.677)	1.81 (1.13-2.90)	0.014	0.564 (0.488-0.642)

-T, prognosis model uses SUV map/image and tumor mask as input; -TN=prognosis model uses SUV map/image, tumor mask, and nodes-to-tumor (N-T) distance map as input; TCIA, The Cancer Imaging Archive.

Figure 1

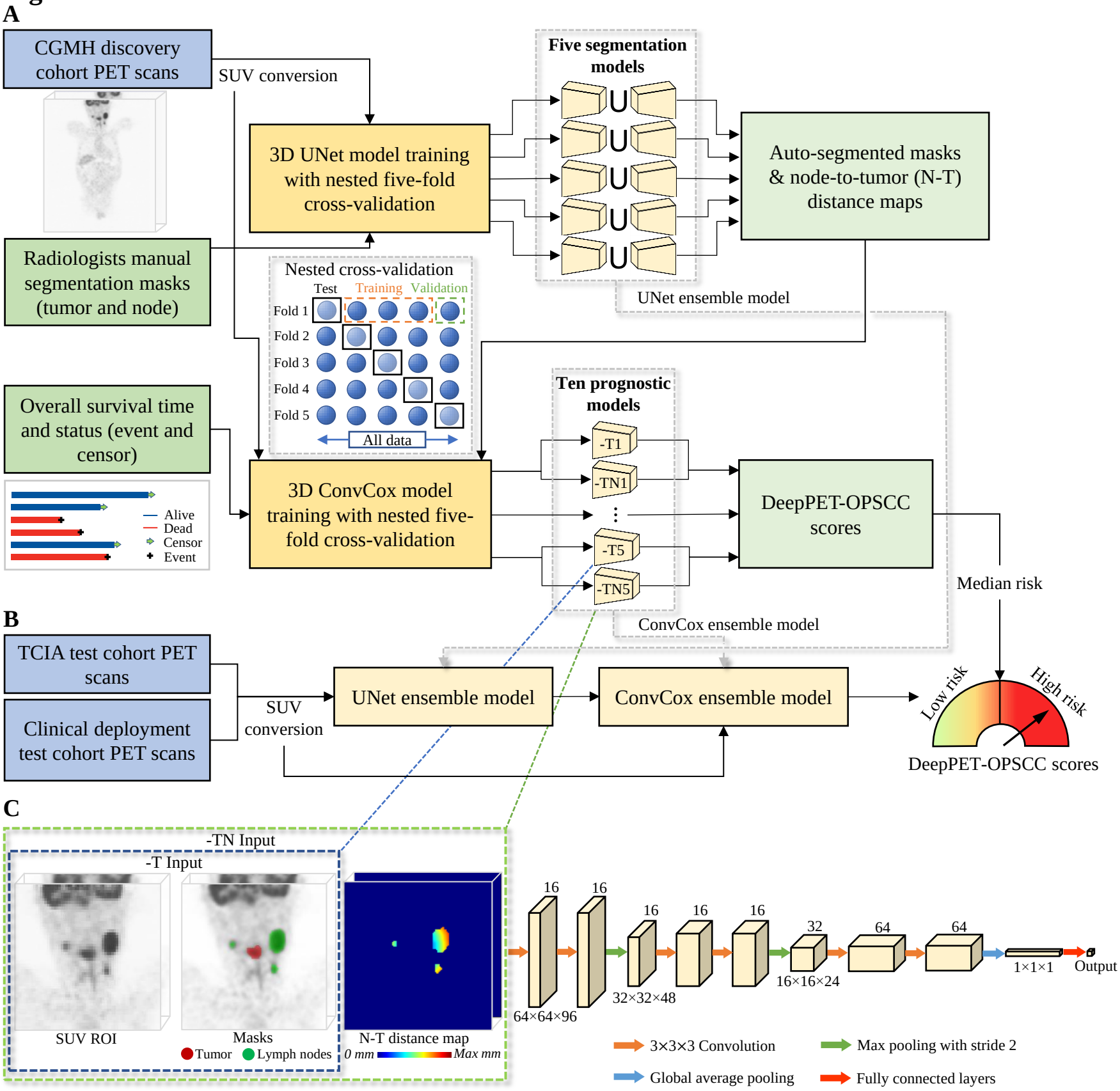

A
CGMH discovery cohort PET scans
SUV conversion
3D UNet model training with nested five-fold cross-validation
Five segmentation models
U
U
U
U
U
Auto-segmented masks & node-to-tumor (N-T) distance maps
UNet ensemble model
Radiologists manual segmentation masks (tumor and node)
Nested cross-validation
Test
Training
Validation
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
All data
Overall survival time and status (event and censor)
Alive
Dead
Censor
Event
3D ConvCox model training with nested five-fold cross-validation
Ten prognostic models
-T1
-TN1
-T5
-TN5
DeepPET-OPSCC scores
Median risk
ConvCox ensemble model
B
TCIA test cohort PET scans
Clinical deployment test cohort PET scans
SUV conversion
UNet ensemble model
ConvCox ensemble model
Low risk
High risk
DeepPET-OPSCC scores
C
-TN Input
-T Input
SUV ROI
Masks
Tumor
Lymph nodes
N-T distance map
0 mm
Max mm
16
16
16
16
16
32
64
64
64×64×96
32×32×48
16×16×24
1×1×1
Output
3×3×3 Convolution
Max pooling with stride 2
Global average pooling
Fully connected layers

Figure 2

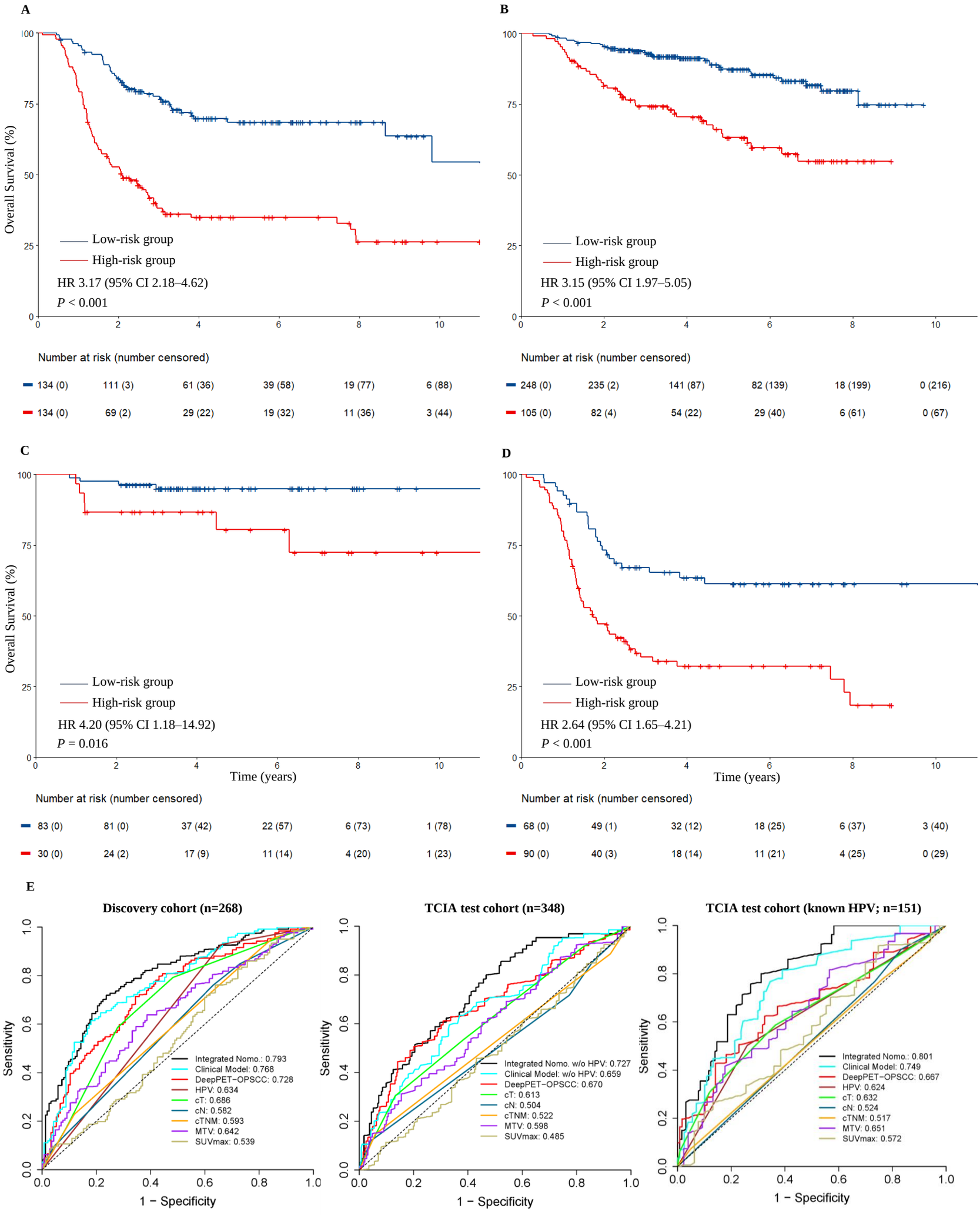
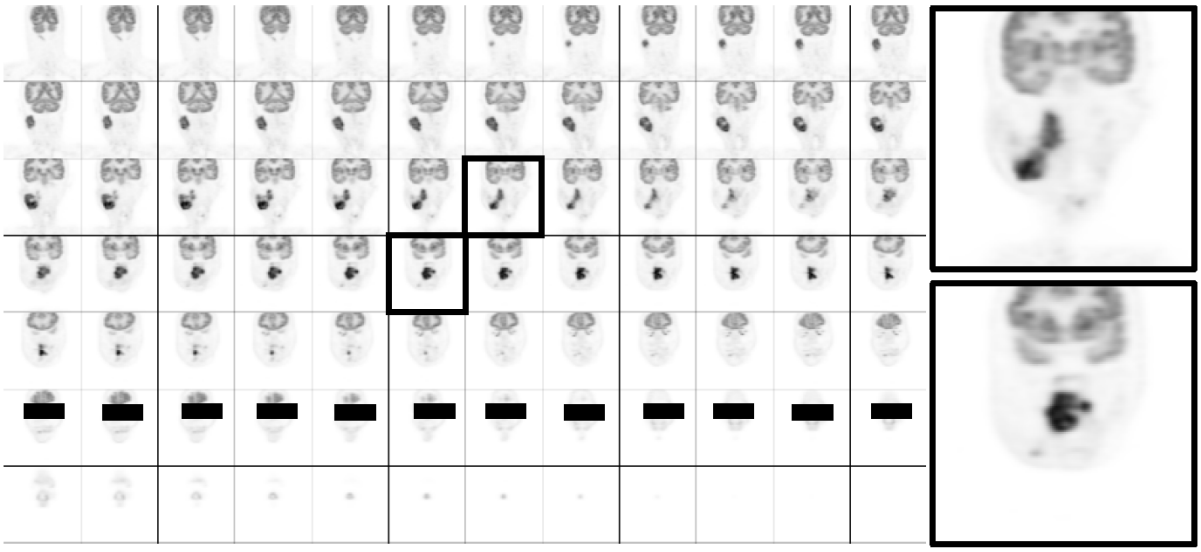
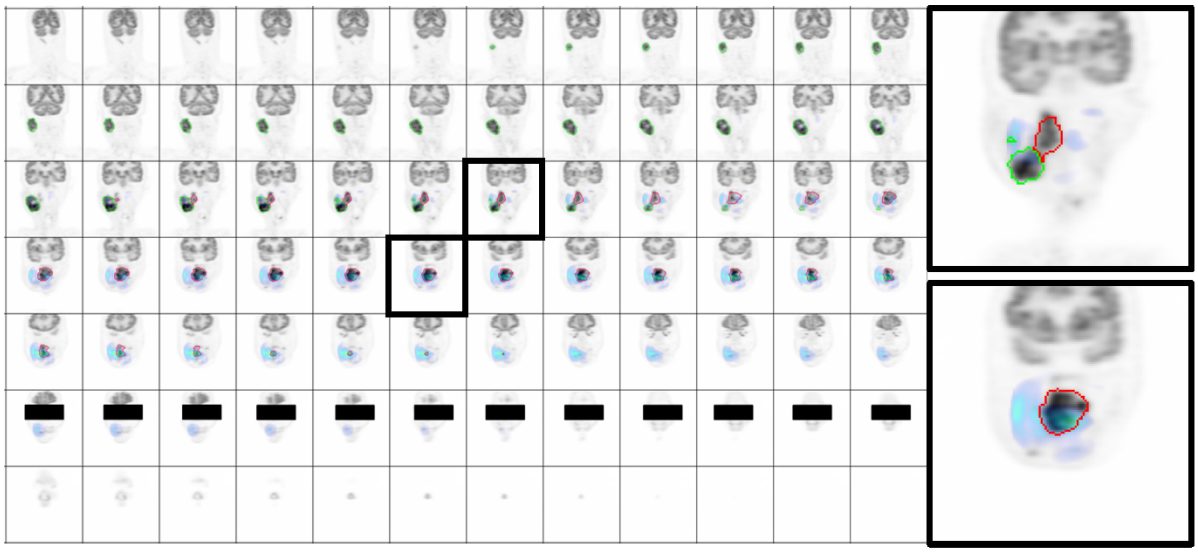

A
Overall Survival (%)
Low-risk group
High-risk group
HR 3.17 (95% CI 2.18–4.62)
P < 0.001
Number at risk (number censored)
134 (0) 111 (3) 61 (36) 39 (58) 19 (77) 6 (88)
134 (0) 69 (2) 29 (22) 19 (32) 11 (36) 3 (44)
B
Low-risk group
High-risk group
HR 3.15 (95% CI 1.97–5.05)
P < 0.001
Number at risk (number censored)
248 (0) 235 (2) 141 (87) 82 (139) 18 (199) 0 (216)
105 (0) 82 (4) 54 (22) 29 (40) 6 (61) 0 (67)
C
Overall Survival (%)
Low-risk group
High-risk group
HR 4.20 (95% CI 1.18–14.92)
P = 0.016
Time (years)
Number at risk (number censored)
83 (0) 81 (0) 37 (42) 22 (57) 6 (73) 1 (78)
30 (0) 24 (2) 17 (9) 11 (14) 4 (20) 1 (23)
D
Low-risk group
High-risk group
HR 2.64 (95% CI 1.65–4.21)
P < 0.001
Time (years)
Number at risk (number censored)
68 (0) 49 (1) 32 (12) 18 (25) 6 (37) 3 (40)
90 (0) 40 (3) 18 (14) 11 (21) 4 (25) 0 (29)
E
Discovery cohort (n=268)
Integrated Nomo.: 0.793
Clinical Model: 0.768
DeepPET−OPSCC: 0.728
HPV: 0.634
cT: 0.686
cN: 0.582
cTNM: 0.593
MTV: 0.642
SUVmax: 0.539
Sensitivity
1 − Specificity
TCIA test cohort (n=348)
Integrated Nomo. w/o HPV: 0.727
Clinical Model: w/o HPV: 0.659
DeepPET−OPSCC: 0.670
cT: 0.613
cN: 0.504
cTNM: 0.522
MTV: 0.598
SUVmax: 0.485
Sensitivity
1 − Specificity
TCIA test cohort (known HPV; n=151)
Integrated Nomo.: 0.801
Clinical Model: 0.749
DeepPET−OPSCC: 0.667
HPV: 0.624
cT: 0.632
cN: 0.524
cTNM: 0.517
MTV: 0.651
SUVmax: 0.572
Sensitivity
1 − Specificity

Figure 3

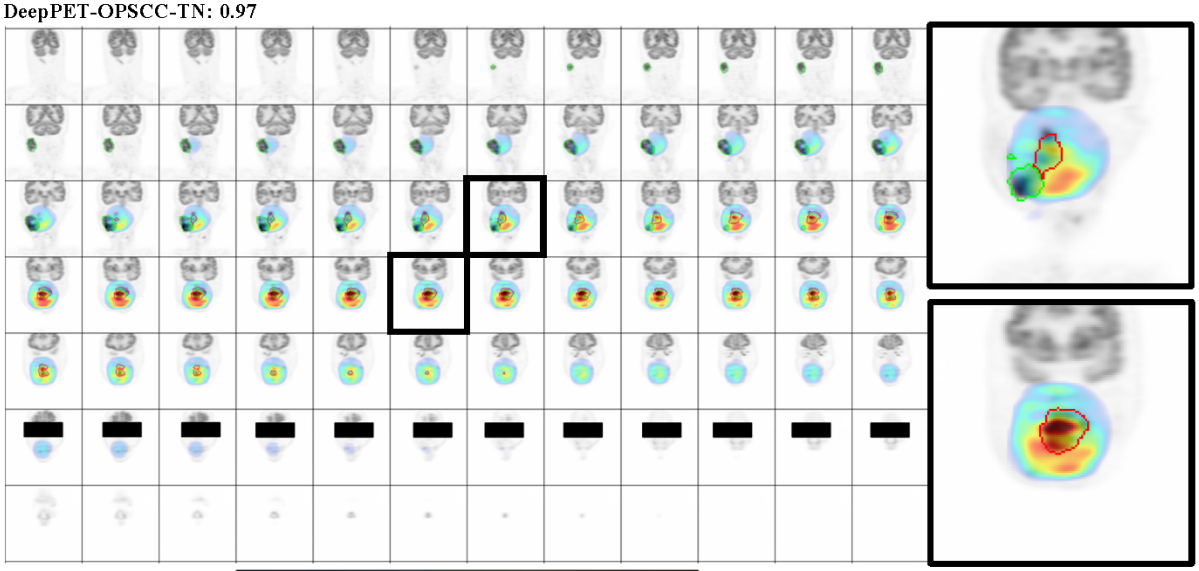
Male, 59 years, HPV–, cT4b-cN3-cM0, Stage: IVB, SUVmax: 19.12 ; Died at 15 months



DeepPET-OPSCC-T: 0.36



DeepPET-OPSCC-TN: 0.97



Supplementary Protocol

Deep learning for fully-automated prediction of overall survival in patients with oropharyngeal cancer using FDG PET imaging: an international retrospective study

Table of Contents

Evaluation of a deep learning model for automated prediction of overall survival in patients with oropharyngeal cancer from pretreatment ^{18}F-fluorodeoxyglucose positron emission tomography imaging

1 Discovery cohort

We formed the discovery cohort as an initial single-center dataset with 273 histologically confirmed non-metastatic (M0) oropharyngeal squamous cell carcinoma (OPSCC) patients treated between June 2006 and December 2017, at the Linkou Chang Gung Memorial Hospital (CGMH), Taiwan, ROC. FDG PET/CT images were acquired using GE or Siemens scanners (Protocol Table 1), within a median of 9 (3–14) days from the pathological diagnosis. The conversions from PET to standardized uptake value (SUV) map were performed using the Chang-Gung Image Texture Analysis (CGITA) toolbox (https://code.google.com/archive/p/cgita/)[1] and were saved in NifTI format (https://nifti.nimh.nih.gov/). For development and validation of segmentation models, 273 patients were used. For prognosis: one patient was excluded from the study as the lymph node (stage = N3) was resected before PET imaging; four patients were excluded, as the automated segmentation model identified no tumor in their SUV images – the remaining 268 patients comprised the CGMH discovery cohort (Protocol Fig. 1). Human papillomavirus (HPV) status was assessed by a pathologist (LYL) blinded to the clinical data on the p16 staining of biopsy sections. Patients were staged and treated based on the 7th edition TNM staging system.

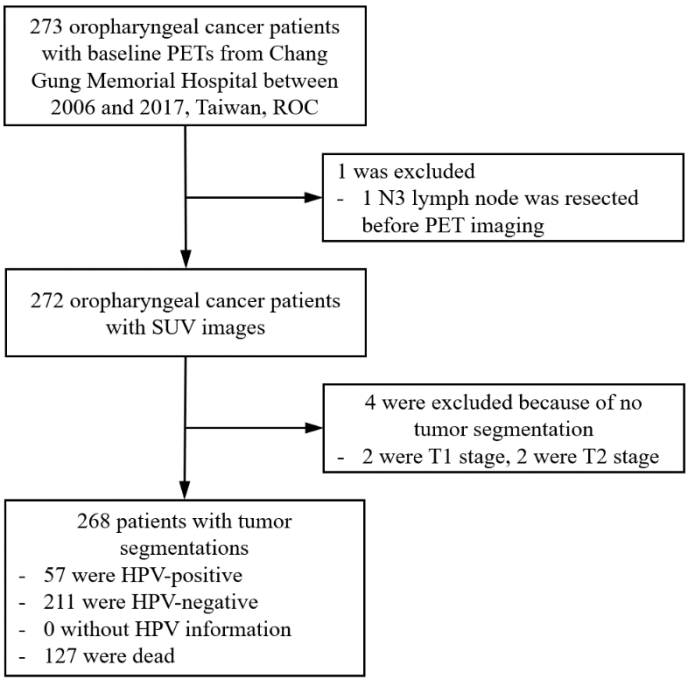
Protocol Table 1. PET scanner characteristics of the discovery, TCIA test, and clinical deployment test cohorts

	Discovery cohort	**TCIA test cohort**						**Clinical deployment test cohort**	
Institutions	CGMH	HN-PET-CT: HGJ	HN-PET-CT: CHUS	HN-PET-CT: HMR	HN-PET-CT: CHUM	HNSCC	Head-Neck-Radiomics-HN1	ZJU1	NFH
Patient (n)	268	54	72	18	53	121	35	22	10
PET scanners									
GE	149 (56%)	54 (100%)	-	18 (100%)	15 (28%)	120 (99%)	-	-	2 (20%)
Discovery ST	146 (55%)	54 (100%)	-	-	14 (26%)	32 (26%)	-	-	-
Discovery STE	3 (1%)	-	-	17 (94%)	1 (2%)	39 (32%)	-	-	-
Discovery RX	-	-	-	-	-	41 (34%)	-	-	-
Discovery HR	-	-	-	-	-	8 (7%)	-	-	-
Discovery LS	-	-	-	-	-	-	-	-	2 (20%)
Unknown	-	-	-	1 (6%)	-	-	-	-	-
Philips	-	-	72 (100%)	-	38 (72%)	-	-	-	-
Guardian Body(C)	-	-	72 (100%)	-	38 (72%)	-	-	-	-
Siemens	119 (44%)	-	-	-	-	1 (1%)	35 (100%)	22 (100%)	8 (80%)
Biograph40	42 (16%)	-	-	-	-	-	-	-	-
Biograph40 mCT	74 (28%)	-	-	-	-	-	-	-	-
Biograph64 mCT	1 (0.3%)	-	-	-	-	-	-	-	-
Biograph128 mCT	-	-	-	-	-	-	-	-	8 (80%)
Definition AS	2 (1%)	-	-	-	-	-	-	-	-
CPS 1080	-	-	-	-	-	1 (1%)	-	-	-
CERR	-	-	-	-	-	-	35 (100%)	-	-
1080	-	-	-	-	-	-	-	22 (100%)	-
Voxel spacing (mm)									
Median [IQR]	4.69 [3.13, 4.69]	3.52 [3.52, 3.52]	4.0 [4.0, 4.0]	3.52 [3.52, 3.52]	4.0 [4.0, 4.0]	5.47 [5.47, 5.47]	2.67 [2.67, 2.67]	4.06 [4.06, 4.06]	4.07 [4.07, 4.07]
Range	3.13-5.47	3.52-4.69	4.0-4.0	3.52-5.47	3.52-5.47	3.91-5.47	2.67-2.67	4.06-4.06	3.91-4.07
Slice thickness (mm)									
Median [IQR]	3.27 [2.03, 3.27]	3.27 [3.27, 3.27]	4.0 [4.0, 4.0]	3.27 [3.27, 3.27]	4.0 [3.27, 4.0]	3.27 [3.27, 3.27]	3.0 [3.0, 3.2]	5.0 [5.0, 5.0]	3.0 [3.0, 3.0]
Range	2.0-3.27	3.27-3.27	4.0-4.0	3.27-3.27	3.27-4.0	2.00-3.27	3.0-3.0	5.0-5.0	3.0-4.25

TCIA=The Cancer Imaging Archive. CGMH=Chang Gung Memorial Hospital. HGJ=Hôpital Général Juif. CHUS=Centre Hospitalier Universitaire de Sherbrooke. HMR=Hôpital Maisonneuve-Rosemont. CHUM=Centre Hospitalier de l'Université de Montréal. HNSCC=University of Texas MD Anderson Cancer Center. Head-Neck-Radiomics-HN1=MAASTRO Clinic. ZJU1=First Affiliated Hospital of Zhejiang University. NFH=Nanfang Hospital.

According to the treatment policy of our institute, OPSCC patients were treated with platinum-based concurrent chemoradiotherapy (CCRT). Patients who participated in the open-label, prospective clinical trial to assess the survival benefits of induction chemotherapy on advanced stage OPSCC, received induction chemotherapy followed by platinum-based CCRT (IC + CCRT). Patients were followed-up for at least 36 months or until death. All patients received radiation therapy with an intensity-modulated technique and completed FDG PET staging before therapy.

Two hundred and one patients (75%) underwent CCRT: 187 patients received chemotherapy with cisplatin (50 mg/m2 of body surface area, day 1), tegafur (200 mg/capsule, 1# qid on days 1–14), and



Protocol Fig. 1 A diagram specifying inclusion and exclusion information for patients and PET images from the discovery cohort. HPV=human papillomavirus.

leucovorin (15 mg/capsule 1# qid on days 1–14) biweekly; 12 received cisplatin (40 mg/m2) weekly; two with cisplatin (100 mg/m2) triweekly. IC + CCRT was administered to 44 patients (16.4%). Thirty cases completed IC of cisplatin (50 mg/m2 on day 1), tegafur (200 mg/capsule, 1# qid on days 1–14), and leucovorin (15 mg/capsule 1# qid on days 1–14) biweekly for 4 cycles and continued this regimen during radiotherapy. Fourteen patients received taxotere (75 mg/m2, day 1), cisplatin (75 mg/m2, day 1), and 5-fluorouracil (750 mg/m2 on days 1–5) triweekly for 3 cycles, followed by cisplatin (40 mg/m2, weekly) during radiotherapy. Three patients with an early stage (one T1N0 and two T2N0 cases) received surgery only. Although different treatment modalities were utilized, intensive radiotherapies (median radiation dose: 72 Gy, range: 66–80 Gy; 2 Gy per day, 5 days per week) were performed for the rest 265 (99%) patients.

2 Segmentation

The segmentation method is built upon a state-of-the-art medical image segmentation backbone nnUNet[2], augmented using extensive data augmentation. The nnUNet is recognized for its high performance on several medical image segmentation tasks, especially tumor segmentation.[2,3] Extensive data augmentation is known to help networks maintain a high level of classification/segmentation accuracies when applied to unseen data.[4,5] The trained augmented nnUNet model can then robustly produce 3D masks of the primary tumor and the lymph nodes, given an input SUV image.

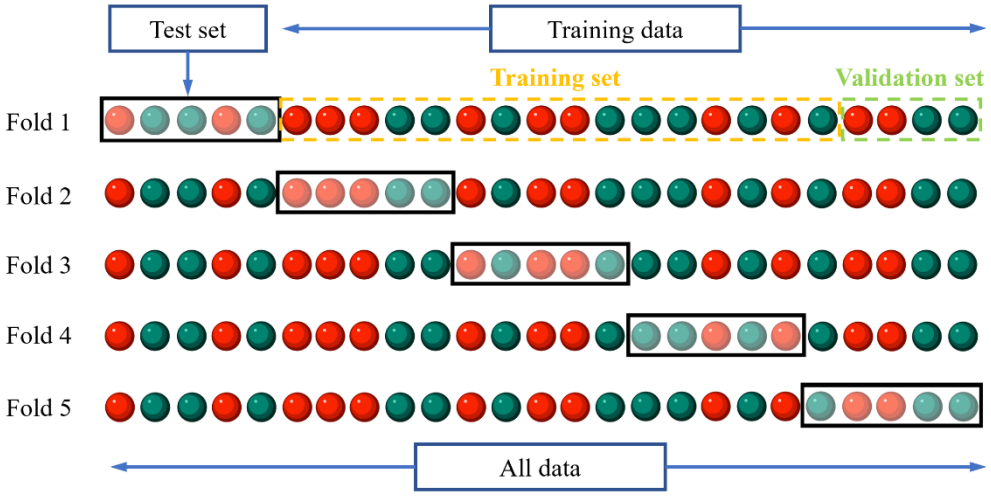
2.1. Training Images and Annotations

Ground truth tumor and lymph node segmentation masks were first generated for 168 patients in the CGMH discovery cohort (by NMC, a nuclear radiologist with 14 years of experience in nuclear imaging and image processing). Specifically, the PET images were converted to SUV maps – with the SUV value of 2.5 defined as the threshold in delineating tumors – using the CGITA toolbox (https://code.google.com/archive/p/cgita/)[1]. Lymph nodes in the SUV maps were delineated using ITK-SNAP (www.itksnap.org)[6]. To ensure consistency of masks with the true clinical data, the radiologist (NMC) carefully reviewed all corresponding CT images, radiology, and pathology reports to guide the manual delineation task. Next, to facilitate/accelerate the manual annotation of the remaining unannotated 105 patients, a nnUNet ensemble model – composed of five models trained from five-fold cross-validation on the 168 annotated patients – was applied to SUV images of the 105 remaining patients. The resulting tumor and lymph node auto-segmentations were checked (by NMC) by referring to their

corresponding CT images and reports if necessary. Inaccurate, false positive, and false negative segmentations were edited and corrected using ITK-SNAP. The 273 SUV images with (both manual and corrected) annotations/masks were used as the developing set of the automated segmentation model.

2.2. Training, Validation, and Test Sets Splitting

The final automated segmentation models were trained using nested five-fold cross-validation (Protocol Fig. 2) on the 273 patients with annotations described above. The following scheme was adopted for each of the five folds. 20% of the data was used as the test set and the remaining 80% as the training data. We ensured that there was no overlap between the test sets across folds. Of the training data, 80% was used for the optimization of network parameters (as a training set), and the remaining 20% for model selection (as a validation set). The final model, selected based on performance on the validation set, was then evaluated on the test set in the corresponding fold. After repeating the process five times, the testing results were generated for the full dataset. Due to this subdivision of the data into training, validation, and test splits in each fold, nested cross-validation can overcome the tendency of standard cross-validation to generate over-optimistic estimates, as the automated segmentation of every PET volume was obtained in the test sets. Note that the current data splitting scheme was identical to that in the prognosis experiments (Protocol section 3), where the distributions of events in the training and validation sets were similar in each fold.



Protocol Fig. 2 A diagram showing the nested five-fold cross-validation used in this study.

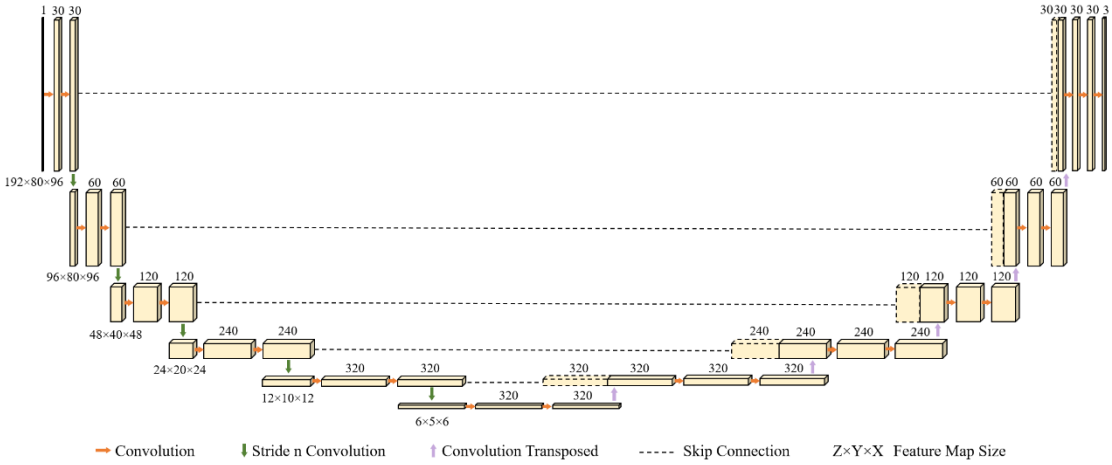
2.3. nnUNet

nnUNet is a self-configuring method for deep learning-based medical image segmentation, including 2D and 3D versions of UNet architectures.[2] Since 3D spatial semantics are crucial to predicting the full tumor and lymph node segmentation masks, a 3D UNet, trained on pairs of volumetric image patches (input: cropped ROI with a size of 192×80×96 [Z×Y×X] voxels) and corresponding segmentation masks (output: ROI with a same size of 192×80×96 voxels) at full resolution, is used as the network architecture (Protocol Fig. 3). These ROIs are cropped to the region of nonzero values with consideration of the GPU memory constraints. To properly learn the 3D spatial semantics, all SUV volumes are first resampled to the median voxel spacing (3.27×4.69×4.69 mm) of the developing dataset, using third-order spline interpolation for images and nearest-neighbor interpolation for segmentation masks. Next, to increase the robustness to variations in SUV intensity among different patients, image intensities for each patient are normalized using a z-score normalization method based on the mean and standard deviation of intensity values for the patient individually. The convolutional kernel size is 3×3×3. A combination of Dice and cross-entropy loss is utilized.[2] The loss function is calculated between the softmax output of the final feature map and the ground truth segmentation mask, in which the labels of background, tumor, and lymph node voxels are 0, 1, and 2, respectively.

For nnUNet training, most parameters are set to be default values. We find that increasing the batch size (setting as 2) or changing the optimizer from default Adam[7] to most recent RAdam[8] does not significantly affect performance, and thus use the Adam optimizer. The models are trained to optimize the average

loss of the primary tumor and lymph nodes. The models are trained with 200 epochs, with 250 training batches per epoch. Each training process takes one day on a NVIDIA Titan RTX-6000 GPU. The model that produces the best average Dice score of the primary tumor and lymph nodes on the validation set is selected as the best segmentation model.

In the testing phase, a sliding window of overlapping (by patch size / 2) volumetric image patches/ROIs (with size of 192×80×96 voxels) is applied to the whole 3D volume. In addition, testing-time data augmentation, i.e., flipping of all image patches along three axes, is used to further increase the model's robustness. All final predictions are aggregated to produce the final segmentations for each voxel.



Protocol Fig. 3 The 3D segmentation network architecture for automated tumor and lymph node segmentation in PET. Yellow boxes represent the 4D feature maps. The network processes the 3D input image (top-left) and outputs the tumor and lymph node segmentation masks (top-right).

2.4. Data Augmentation

One major obstacle to introducing deep learning models into clinical practice is their poor generalizability to unseen domains, such as different scanner vendors, hospitals, populations, etc[9,10]. For deep learning based predictive biomarkers in particular, it is usually difficult to obtain a large variety of imaging data and clinical outcomes to train highly generalizable models. Furthermore, it is infeasible to collect data from the target domain (e.g., every new hospital or vendor) to implement transfer learning or domain adaptation approaches to refine the trained models. Data augmentation is one of the most important techniques used to enhance the generalization performance of deep learning.[4] Extensive data augmentation transformations have demonstrated promising generalization performance on unseen domains for image segmentation tasks.[5] Specifically, for each 3D image patch randomly cropped during nnUNet training, a sequence of the following eight data augmentation transforms based on an open-source implementation (*batchgenerators*; https://github.com/MIC-DKFZ/batchgenerators) is applied: *deformation*, *rotation*, *scaling*, *sharpening*, *blurring*, *noise*, *brightness*, and *contrast*. Each transform function is associated with two parameters: 1) the probability p to apply the function, set to 0.5, and 2) the magnitude m of the function, which is set as the default settings as in *batchgenerators*.

2.5. Postprocessing

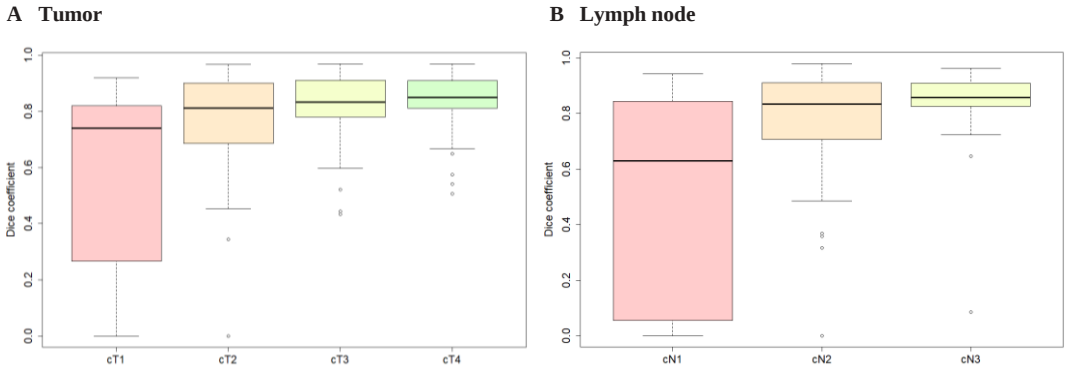
The resulting segmentations are further processed through two steps. First, only segmentations in the head-and-neck region are retained. We empirically find that most head-and-neck PET images contain 91 slices with a slice thickness of 3.27 mm in the z-direction, yielding a total of 298 mm sufficient to cover the entire head-and-neck region. Therefore, we use 300 mm starting from the uppermost slice as a distance cutoff. Next, volume cutoffs of the segmented tumor and lymph nodes are set. For the segmented tumor, we keep the largest connected component as a tumor to remove false positives. For the segmented lymph nodes, we remove segmentations (for both ground truth and auto-segmentations) with volume < 500 mm^3, since the RECIST 1.1 guidelines consider lymph nodes < 10 mm in the shortest diameter as normal[12]. We choose 500 mm^3 as our cutoff size, as it more or less corresponds to the volume of a 3D spherical object equal to $(4/3) \times \boldsymbol{\pi} \times 5 \times 5 \times 5 = 523$ mm^3.

2.6. Evaluation

Segmentation masks of tumors and lymph nodes are inputs to our deep learning-based prognosis models, that they may guide the extraction of deep features from the imaging data. We evaluate the segmentation accuracy using the Dice coefficients[11], which is the standard metric used to evaluate segmentation performance.[3] It measures the spatial overlap between the ground truth (*GT*) and automated segmentation masks (*Seg*) as

$$Dice = \frac{2|GT \cap Seg|}{|GT| + |Seg|}$$

Results of tumor (Protocol Fig. 4 A, Protocol Table 2) and lymph node (Protocol Fig. 4 B, Protocol Table 3) segmentation on the discovery test sets of the nested 5-fold cross-validation were median Dice coefficients of 0.84 (IQR 76–91) for tumors and 0.84 (70–91) for lymph nodes. The segmentation was more accurate for advanced than earlier cT and cN stages. It identified the tumor in 269 (99%) of 273 cases in the discovery cohort, and its performance was comparable in the external TCIA and clinical deployment test cohorts, i.e., identify the tumor for 384 (97%) of 397 cases. Note that for calculations of lymph node Dice coefficients, we did not include patients who had a manually segmented lymph node < 500 mm^3, according to the RECIST 1.1 criteria mentioned in section 2.5. For patients in N0 stage, a false positive result was defined as the auto-segmented lymph node ≥ 500 mm^3. 11 out of 58 (19.0%) patients in N0 stage were found to have false positive lymph node segmentations.



Protocol Fig. 4 Box plots show the agreement between the automated 3D segmentation and the radiologist-generated ground truth for tumor segmentation categorized by cT stages (A) and lymph nodes by cN stages (B). The values illustrated are Dice coefficients. The upper and lower edges of each box show the 25'th and 75'th percentiles. Whiskers above and below the boxes correspond to the 10'th and 90'th percentiles, respectively, horizontal central lines to the median values, and dots to outliers.

Protocol Table 2. Evaluation of tumor segmentation on 273 CGMH patients using nested 5-fold cross validation. Dice coefficients on the test sets are reported.

Dice	**cT1**	**cT2**	**cT3**	**cT4**	**cT1-cT4**
Median (IQR)	0.74 (0.27-0.82)	0.81 (0.69-0.90)	0.83 (0.78-0.91)	0.85 (0.81-0.91)	0.84 (0.76-0.91)

Protocol Table 3. Evaluation of lymph node segmentation on 273 CGMH patients using nested 5-fold cross validation. Dice coefficients on the test sets are reported.

Dice	**cN1**	**cN2**	**cN3**	**cN1-cN3**
Median (IQR)	0.63 (0.06-0.84)	0.83 (0.71-0.91)	0.86 (0.83-0.91)	0.84 (0.70-0.91)

For the outliers in Protocol Fig. 4 A for tumor segmentation, Dice scores of 0 are obtained for 7 out of 273 PET scans of patients in cT1 and cT2 stages. Four of these cases (2 / 16 (12.5%) of cT1 stage and 2 / 89 (2.2%) of cT2 stage) are false negatives, whereby our automated segmentation model does not identify any voxels as the tumor. The other 3 (2 / 16 (12.5%) of the cT1 stage and 1 / 89 (1.1%) of the cT2 stage), are false positives, where our model identifies some voxels that do not overlap with the ground truth tumor regions. These cases are not especially concerning, as the FDG avid regions of the tumors were very small. In fact, it can even difficult for radiologists to identify such tumors based only on PET imaging without clinical and histopathological information. In the prognosis experiments, the 4 false-negative cases are excluded since a tumor location is needed to perform the prognosis. For the

outliers in lymph node segmentation (Protocol Fig. 4 B), Dice scores of 0 are obtained in 15 cases (5 / 25 (20.0%) of cN1 stage and 10 / 172 (5.8%) of cN2 stage), as our model does not identify any lymph node with volume larger than the cutoff size (500 mm^3), thus yielding false negatives. Similar to the false negatives in tumor segmentation, some small lymph node lesions were difficult to identify partly due to the limited spatial resolution of PET images. The manual annotations for these cases usually require additional information from clinical, histopathological, CT or MRI imaging. Nevertheless, as shown in these quantitative results, our model produces higher Dice scores for patients in advanced T or N stages.

Any patient can have an arbitrary number of lymph nodes. Reliable detection of lymph nodes is also important to prognosis, e.g., the definition of the N stage relies not only on the size but also on the number and distribution of lymph nodes. 390 lymph nodes were manually detected and/or annotated in 273 patients. Lymph node detection performance was evaluated based on the number of true positives (TP), false negatives (FN), false positives (FP), the true positive rate (TPR), the positive predictive value (PPV), and the false positives per volume (Protocol Table 4). The Dice coefficient between the segmentation and ground truth is used as the criteria to define a TP detection. We define TPs, FNs, and FPs as follows:

- TP: a lymph node is successfully detected if there is an auto-segmentation with Dice coefficient ≥ threshold with its ground truth.
- FN: a lymph node is missed if there is no auto-segmentation with Dice coefficient ≥ threshold with its ground truth.
- FP: a lymph node is incorrectly detected if there is no ground truth lymph node that has Dice coefficient ≥ threshold with an auto-segmentation.

Protocol Table 4. Evaluation of Lymph Node detection (n=390) on 273 CGMH patients using nested 5-fold cross-validation. Results on the test sets are reported.

TP criteria	TP (n)	FN (n)	FP (n)	TPR (%)	PPV (%)	FP per volume
Dice ≥ 0.9	93	297	242	23.9	27.8	0.89
Dice ≥ 0.8	183	207	152	46.9	54.6	0.56
Dice ≥ 0.7	231	159	104	59.2	69.0	0.38
Dice ≥ 0.6	258	132	77	66.2	77.0	0.28
Dice ≥ 0.5	270	120	65	69.2	80.6	0.24
Dice ≥ 0.4	276	114	57	70.8	82.9	0.21
Dice ≥ 0.3	282	108	50	72.3	84.9	0.18
Dice ≥ 0.2	286	104	45	73.3	86.4	0.16
Dice ≥ 0.1	289	101	43	74.1	87.0	0.16
Dice > 0.0	291	99	20	74.6	93.6	0.07

For example, defining Dice coefficient ≥ 0.5 as a true positive of lymph node detection, the TPR and PPV were 69.2% and 80.6% at 0.24 FP per volume.

3 Prognosis

In this section, we describe the deep learning-based prognostic models that were developed for the prediction of OPSCC patients' survival. The models were implemented with Python 2.7.0 (https://www.python.org/) using Pytorch package version 1.4.0 (https://pytorch.org/). The CGMH discovery cohort was used to train, validate, and test the prognosis models (with the same nested five-fold cross-validation in the segmentation model; Protocol Fig. 2). The multicenter TCIA and clinical deployment test cohorts were used for independent external testing and application of the model.

We explicitly analyze both the primary tumor and involved lymph nodes by capturing not only their morphology and texture but also their location and distance information. Most previous studies focused exclusively on the primary tumor.[13,14] Several recent studies have shown that specific imaging features or characteristics of the involved lymph nodes could be more strongly associated with patients' outcomes than those derived from the primary tumor, especially on OPSCC patients with HPV+ disease.[15,16] However, those studies do not include a larger, multicenter validation of the imaging markers. Furthermore, they usually use 2D tumor slices and merely determine the maximum tumor occurrence in the image, in order to reduce the amount of computations required to extract hand-crafted features. Such an approach is limited, in that it could fail to fully capture the tumor's heterogeneity and the complexity of its underlying biological processes.

Despite the growing clinical interest in deep learning-based methods,[17–20] the limited sample size of cancer prognosis datasets (baseline imaging with follow-up outcomes) usually prevents deep neural networks from being adequately trained to generalize to unseen data effectively. Indeed, a recent study[21] in lung cancer outcome prediction had observed that when 60% (n≈400) of the training data were used for model training, deep learning underperformed even standard clinical risk factors, such as tumor volume. How to efficiently leverage a limited dataset to train a robust deep learning-based model still requires investigation in cancer prognosis tasks, including OPSCC.

Most previous deep learning methods used classification-based loss functions and could only be trained using a small number of patients (with a distinct outcome at specific time points, e.g., 2 years) out of all available data. Longer or shorter outcomes had to be treated as the same risk level, respectively. In our study, inspired by our recent method for pancreatic cancer survival prediction[22], we present a 3D convolutional neural network with a Cox proportional hazard assumptions (ConvCox) framework. Our ConvCox model, instead, is a regression network capable of leverage all available data to directly learn time-dependent events from the full tumor and lymph nodes in 3D. This is a highly desirable property for prognosis problems, where the number of patients with baseline imaging and outcomes, tends to be limited. We aim to maximize the generalization performance of the ConvCox network. To this end, we perform several preliminary experiments to optimize its architectural design and results. In particular, we evaluate several architectural modifications – such as skip connections, side supervision, and autoencoder reconstruction, and found that the proposed relatively shallower network architecture with optimized preprocessing (e.g., resampling and normalization), extensive data augmentation, and various training and inference strategies (namely nested cross-validation, model ensembling, and test-time augmentation respectively) produce the best generalization performance. In addition, we incorporate domain knowledge about cancer metastasis from the tumor to lymph nodes into the learning of ConvCox, further improving its robustness and generalizability. Since 3D ResNet with pretrained weights is deeper than our networks, the limited training data size can easily lead to overfitting.

3.1. Training Images and Labels

For each patient in the CGMH discovery cohort, three types of volumetric images (i.e., SUV, automatically segmented tumor mask, and nodes-to-tumor [N-T] distance map) related to metabolism uptake, tumor morphology, and clinical evidence were extracted, serving as inputs to the prognosis networks. We note that although the original SUV volumetric image could be used as the unique input, this could more likely induce overfitting of our network to irrelevant areas, due to the limited sample size of the 3D training dataset. The tumor mask was used as additional information, guiding the training process and granting the network the flexibility required to extract features mainly from the tumor regions (both inside and outside). For the N-T distance map, the underlying clinical evidence is that OPSCC patients with lower neck lymph node metastasis, i.e., farther cancer spread, are associated with increased risk of distant metastasis[23] and reduced overall survival (OS).[24,25] We explicitly represented the locations of all lymph node voxels by calculating the distance from each lymph node voxel to the nearest surface of the primary tumor. More specifically, the tumor mask was first converted to tumor mesh, which contained boundary points of the tumor. The mesh conversion process was done with VTK 8.2.0 (https://vtk.org/Wiki/VTK/Examples/Python). Next, distances from lymph nodes voxels to their nearest tumor boundary points in the mesh were calculated using NVIDIA's 3D deep learning open-source library Kaolin (https://github.com/NVIDIAGameWorks/kaolin).

Each patient was associated with two outcome labels, i.e., OS time and status (death or censor). The OS time was defined as the time from cancer diagnosis to the last follow-up or death from any cause.

3.2. 3D Convolutional Cox Model

The DeepPET-OPSCC prognosis model is trained to produce risk scores from input SUV images. The model is based on 3D ConvCox, which can generate imaging signature for survival outcome prediction from 3D imaging of the full tumor and all lymph nodes.

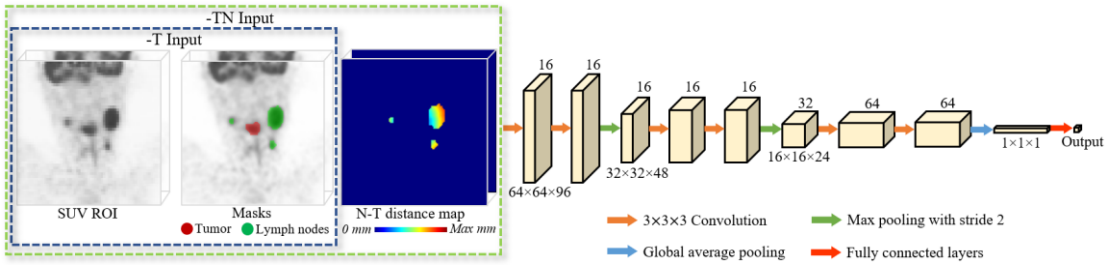
3.2.1. Sub-volume Extraction

When designing a 3D network, the amount of GPU memory available is the main hardware constraint. Since most of the GPU memory is usually occupied by feature map activations (as they need to be stored

for the backward pass), we address this issue by cropping the input sub-volumes of interest, that contain the entire tumor and relevant lymph nodes. Given that the maximum distance from the lymph nodes to the primary tumor is about 120 mm in our discovery cohort, the sub-volume size of 64×64×96 voxels (after resampling the original image resolution to 2×2×2 mm^3) could cover the full region of the tumor and lymph nodes for almost all patients. The SUV values are cut by a window of [0 26] and then normalized to [0 1] (by dividing all remaining values by 26).

3.2.2. Two Models: DeepPET-OPSCC-T and DeepPET-OPSCC-TN

Given that there is more variability in the appearance image (N-T distance map) than the binary image (tumor mask), the deep learning model may not capture adequate information in the tumor mask. To allocate sufficient network capacity to learn both tumor and lymph nodes information, we train two models separately: 1) DeepPET-OPSCC-T, which takes two-channels of input (SUV and tumor mask), and 2) DeepPET-OPSCC-TN, which takes three-channels of input (SUV, tumor mask, and T-N distance map) (Protocol Fig. 5). The tumor and lymph node masks used as network inputs are the automated segmentation instead of the manual segmentation masks, as our preliminary experiments show that the former yield slightly better results. This may be due to the fact that the network yields better predictions based on masks it has created in the past, as it is most familiar with its own segmentation style.



Protocol Fig. 5 The architecture, input, and output of the 3D ConvCox network in the DeepPET-OPSCC-T/-TN prognosis model.

3.2.3. Network Architecture

Both DeepPET-OPSCC-T and -TN models were implemented using the same 3D ConvCox network. The ConvCox network has 6 convolutional layers with filter size 3×3×3 and stride 1. Each convolutional layer is followed by a batch normalization layer and a rectified linear units (ReLUs) nonlinearity activation function. In the first layer, there are 16 filters; the filter size is doubled every two layers. Max-pooling operations with a pooling size of 2×2×2 are adopted after the second and fourth convolutional layers. The last convolutional layer is followed by one global average pooling layer that reduces feature maps to single channel-wise features. This reduces the feature dimension from 64×64×96 (input) to 64 (latent space). Afterward, a fully connected layer is used to obtain outcome predictions based on the latent features.

3.3. Training Procedure

3.3.1. Data Augmentation in Training

Deep neural networks tend to overfit, given a limited amount of training data. Various of data augmentation techniques were used on the fly to train the DeepPET-OPSCC network. At each training iteration, before being fed into the network, each 3D sub-volume undergoes several image transformations, including:

- All input sub-volumes are randomly cropped to the required size (64×64×96) but must contain the full tumor and all lymph nodes.
- Next, the selected 3D sub-volume is randomly flipped from left to right around its axial plane with a probability of 50% and randomly rotated by either 0, 90, 180, or 270 degrees in the axial or horizontal plane.

- 50% of sub-volumes are augmented with image appearance and quality transformations, including *contrast, sharpening, blurring,* and *noise*. The magnitudes of these transformations are set to the same values as those mentioned in Protocol section 2.4.

3.3.2. Loss Function

The deep neural network with Cox proportional hazard assumptions (DeepCox)[26] models nonlinear associations between covariates and outcomes and can handle censored data without the need for binarizing it at specific time points (e.g., at two years). Similar to the Cox model, no assumptions were made about the form of the baseline hazard function, which made DeepCox more generalizable to different applications in cancer outcome prediction. Let X_i denotes the input 3D image of the *i*-th patient with death (or censoring) time t_i. The predicted risk o_i of the *i*-th patient is generated when the input volume X_i is processed by the 3D ConvCox. Let $t_1 < t_2 < \cdots < t_N$ denote the ordered event times and $R(t_i)$ the risk set of all individuals who are still present in the study at time t_i ($j \in R(t_i)$ and $t_j >= t_i$). Assuming statistical independence of the patients, the loss function of the network is computed as a negative log partial likelihood of all samples in the training dataset:

$$L(o_i) = \sum_i \delta_i \left(-o_i + \log \sum_{j \in R(t_i)} \exp(o_j) \right).$$

where $\delta_i = 1$ if the *i*-th patient's death event occurs and $\delta_i = 0$ if the patient is censored.

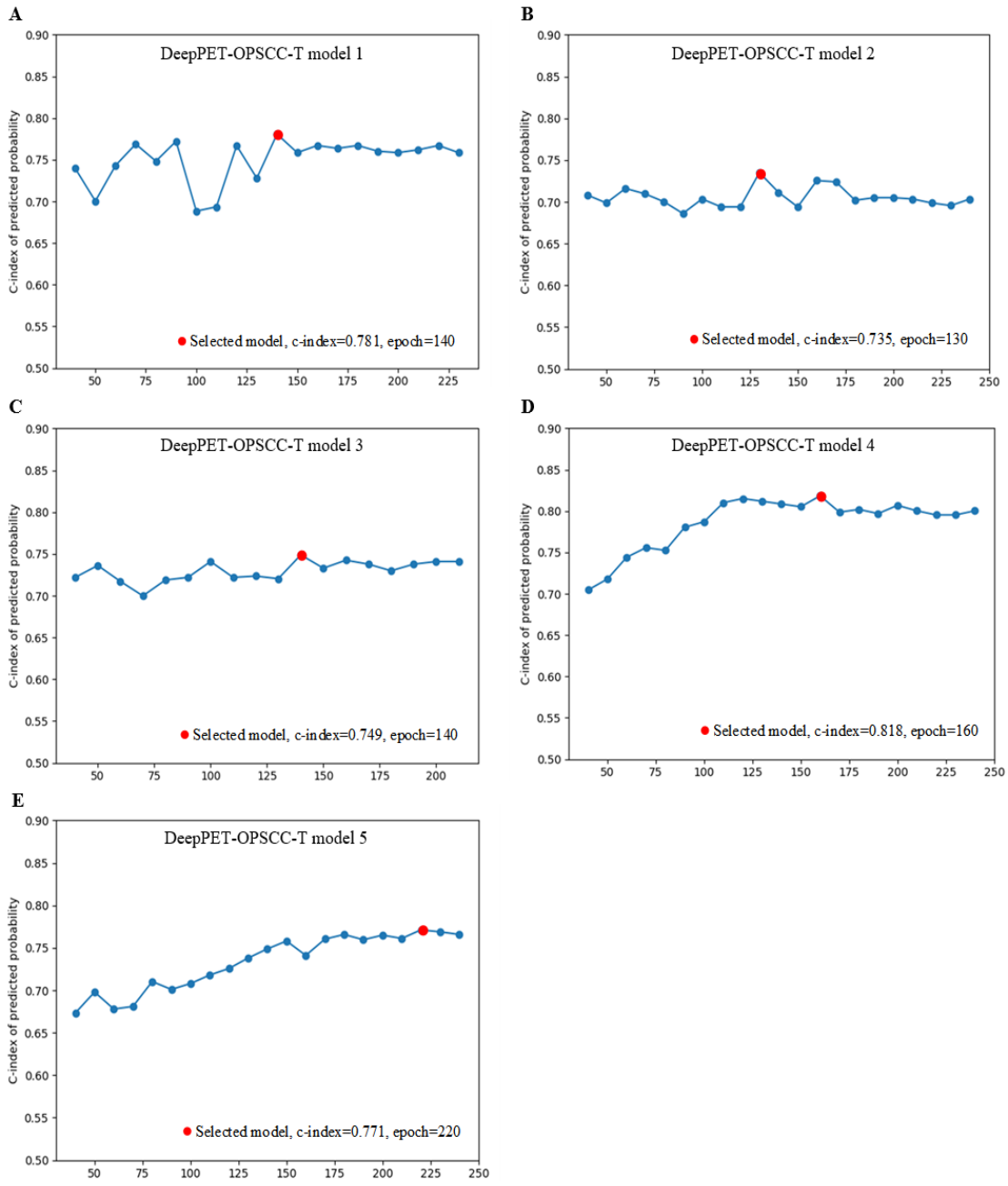
3.3.3. Model Training and Selection

Ten models, five for DeepPET-OPSCC-T and five for DeepPET-OPSCC-TN, were trained using nested five-fold cross-validation (Protocol Fig. 2), where 64%, 16%, and 20% of the data were used as the training, validation, and test sets in each fold. Data splits in each fold were identical to those created to train the segmentation models; the distributions of events in the training and validation sets were similar in each fold. L_2 regularization is applied to all network parameters, and the weight decay is set to 1×10^{-5}. The network was trained with a batch size of 8 sub-volumes, each with a size of 64×64×96 voxels at a spacing of 2×2×2 mm. The Adam optimizer, with an initial learning rate of 1×10^{-3}, is used. The learning rate is adjusted and reduced by a factor of 0.5 when the loss value in the validation set does not decrease for 20 epochs. The network was trained for 250 epochs, and the model with the highest Harrell's concordance index (c-index)[27] on the validation set in each fold was selected and saved as the best model. The c-index compared the observed time to death or censoring to a model's predicted score of poor prognosis. Indeed, the model with the highest c-index on the validation set could potentially generalize to unseen domains substantially well. The training encounters early stopping when no validation loss decreases after 160 epochs. Note that we checked the model's performance on the validation set every 10 epochs starting from the 40th epoch (~80 000 iterations). Protocol Fig. 6 and Fig. 7 show performances results at different training epochs across different nested cross-validation folds (folds 1, 2, 3, 4, and 5) for DeepPET-OPSCC-T and DeepPET-OPSCC-TN, respectively.

The network output was the predicted probability of poor prognosis for the input 3D volumes. The probability was normalized to [0,1], as we used the sigmoid activation function in the last output layer.

3.4. Inference

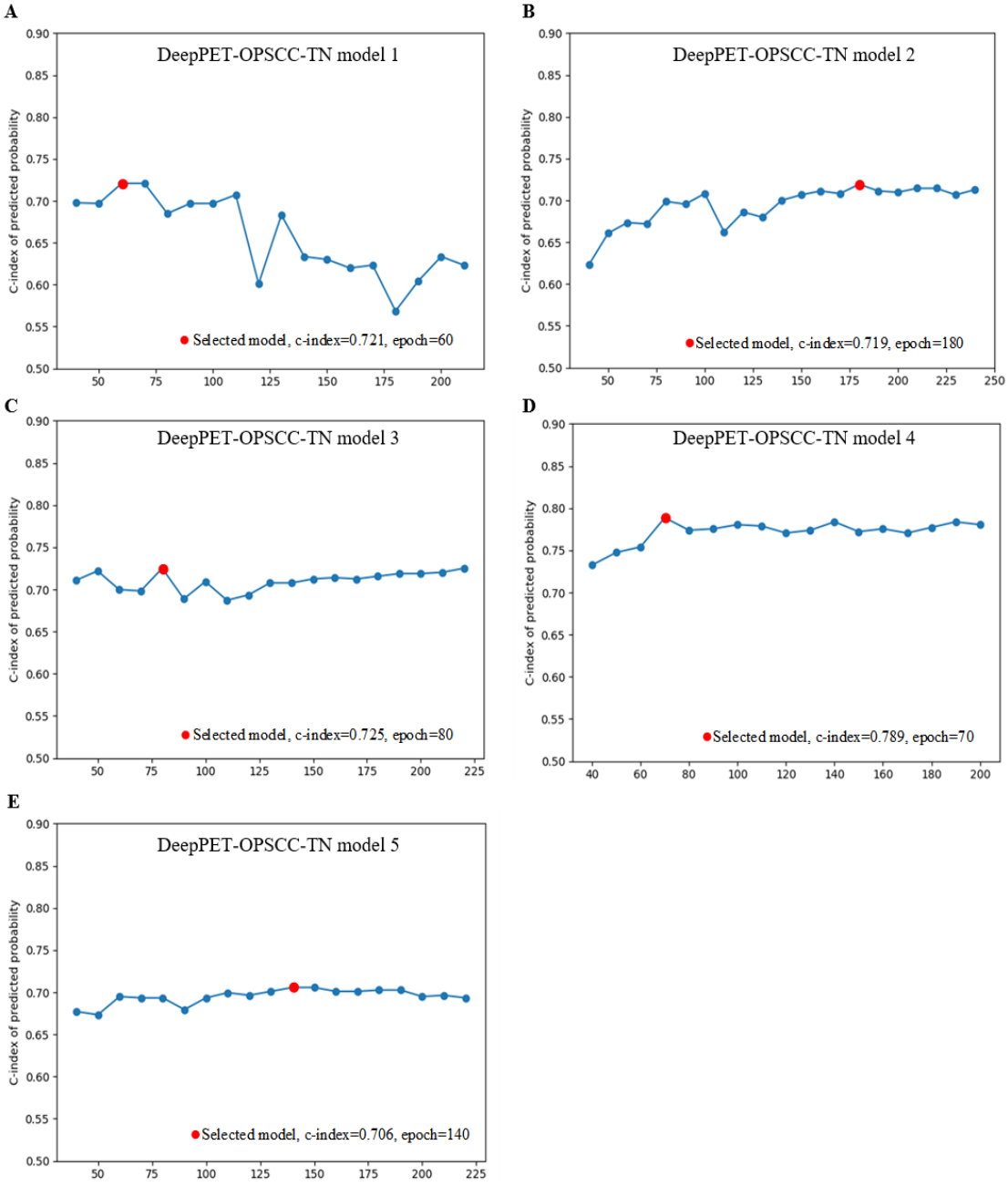
The best DeepPET-OPSCC-T and DeepPET-OPSCC-TN models selected above were applied to perform inference (i.e., prediction of survival). For the CGMH discovery cohort, individual models were applied to the test sets in the nested cross-validation setting – each patient's risk score was obtained by averaging results of one DeepPET-OPSCC-T model and the corresponding DeepPET-OPSCC-TN model from the same fold. Note that the test sets ensure that the predictions were performed in an independent testing scenario. For the TCIA and clinical deployment test cohorts, an ensemble model consisting of all ten models was applied.



Protocol Fig. 6 C-index of tumor only models (i.e., DeepPET-OPSCC-T) on the five validation sets in the nested cross-validation process. Subplots A-E show training and validation results on folds 1 to 5.

3.4.1. Data Augmentation in Testing

For robust and stable prediction, we performed testing-time augmentation by shifting the cropped 3D volumes by up to 5 voxels in the x, y, or z directions ten times. The original volume focused on the head and neck region. After undergoing such augmentation, 10 ROI volumes are shifted around the center and then cropped for testing. This testing-time augmentation scheme was applied to the validation and test sets in the CGMH discovery cohort, and also the TCIA and clinical deployment test cohorts.



Protocol Fig. 7 C-index of tumor with lymph nodes models (i.e., DeepPET-OPSCC-TN) on the five validation sets in the nested cross-validation process. Subplots A-E show training and validation results on folds 1 to 5.

3.4.2. Individual models

When applying each model to a new patient, the predicted score is individualized/normalized by subtracting the mean risk scores of all training patients, $O_n^i(x) = O_\beta^i(x) - O_{risk}^i$, i = 1, 2, ... , 5 for model i, where $O_n(x)$ is the normalized score, $O_\beta(x)$ is the output of the DeepPET-OPSCC-(T/TN) model, and O_{risk} is the mean risk scores of all training patients in fold i. The O_{risk} of the -T models (1st through 5th) are 0.308, 0.287, 0.290, 0.277, and 0.266 and of the -TN models (1st through 5th) are 0.491, 0.396, 0.266, 0.369, and 0.3870, respectively.

3.4.3. Ensemble models

For prediction in the external TCIA and clinical deployment test cohorts, the five DeepPET-OPSCC-T

and five DeepPET-OPSCC-TN models were combined into an ensemble model. The five normalized prediction scores of each T/TN ensemble were used to obtain the final prediction score of the T and TN ensemble models. These two scores were then averaged to generate the final DeepPET-OPSCC score. This score was then binarized, using the median value of DeepPET-OPSCC scores in the discovery cohort's test sets as a cutoff value to classify DeepPET-OPSCC scores of external test cohorts into high and low-risk patient groups.

3.4.4. Visualization with Activation Mapping

To generate visualization maps, we used an activation mapping method[28] to map important regions in the volume with respect to predictions made. The last convolutional block before the global average pooling layer was set to produce activation heatmaps during backpropagation. The visualization maps were enlarged by resampling to match the input volume size (64×64×96). The value of each voxel in the heatmaps directly reflects the given voxel's prediction risk score. This analysis allowed us to observe the most relevant regions, with the most impact on predictions, both within and beyond the primary tumor. To display the heatmaps (Fig. S13), we renormalized their values to [0,1] based on the max and min values from the corresponding training set. For better visualization, we transparently overlay the heatmaps on PET images with heatmap values lower than 0.1 not shown. We found that the -T model focused mostly on the tumor's interior, while the -TN model tended to fixate on the interface between the tumor and the lymph nodes.

3.5. Compared Methods

3.5.1. 3D ResNet-OPSCC

We built a 3D residual network for OPSCC (3D ResNet-OPSCC) as the comparison method. Though the classical 3D ResNet is widely used in 3D computer vision learning tasks, it will face training data scarcity issues when applied to the medical imaging field. Hence, we considered an effective lightweight 3D ResNet proposed for coronavirus disease 2019 (COVID-19) classification (DeCoVNet),[29] as it could potentially be more effective than classical ResNet models on training datasets of limited size. We validated this hypothesis on our data in a preliminary experiment. The 3D ResNet-OPSCC has three stages: the network stem, ResBlocks, and the classifier. The first two stages used the same settings as DeCoVNet. The classifier is changed to that used in DeepPET-OPSCC. The loss function is adapted to the negative partial likelihood function instead of the cross-entropy loss for classification, and the same training procedures (data augmentation, model selection, -T and -TN ensemble model, etc.) used for DeepPET-OPSCC were performed for 3D ResNet-OPSCC.

3.5.2. 2D DeepPET-OPSCC

We investigated the performance of 2D DeepPET-OPSCC models by using 2D key slices from the 3D tumor volume. The largest primary tumor mask in the axial view is used to select each patient's 2D slice. The 2D DeepPET-OPSCC-T model takes the SUV image with its corresponding tumor mask as inputs, while the -TN model takes the SUV image with its corresponding tumor mask and N-T distance map in the same 2D slice. The 2D DeepPET-OPSCC network's parameters shared the same parameter settings in the 3D DeepPET-OPSCC network, except the convolutional and pooling operations used 2D kernels. We followed the same training procedures (data augmentation, model selection, -T and -TN ensemble model, etc.) as DeepPET-OPSCC.

3.5.3. Radiomics Features Extraction and Signature Building

We built two radiomics signatures reflecting the phenotypic characteristics of the primary tumor and the lymph nodes in SUV images, respectively, as independent predictors of OS. The 3D radiomics features were extracted from both the primary tumor and the lymph nodes. There are two sets of handcrafted radiomics features: 482 extracted for primary tumors and 482 for lymph nodes, all of which were extracted using an open-source Python package, Pyradiomics (https://pyradiomics.readthedocs.io/en/latest/).[30]

Each set of 482 handcrafted features can be divided into four groups: 1) intensity, 2) geometry, 3) texture, and 4) wavelet features.

- The intensity features quantified the first-order statistical distribution of the voxel intensities within the volumes of interest (tumor or lymph nodes). The statistical measurements include Energy, Entropy, etc. This group has 18 features in total (1st-18th).
- The texture features measured the spatial distribution of the voxel intensities, thereby quantifying the intra-tumoral heterogeneity. There are 34 texture features in total (19th-52nd).
- The geometry feature group contains features that quantified 3D shape characteristics of the tumor or lymph nodes. It is composed of 14 features in total (53rd to 66th).
- Wavelet features were calculated by applying wavelet transformations to the original input images. There are 416 features in total (67th -482nd).

To accommodate the especially large number of radiomics features and to prevent overfitting, feature selection was performed with nested five-fold cross-validation, as follows. First, univariable analysis was completed for each feature in the training set. Features with p-value < 0.1 were considered to be potentially associated with OS and thus were selected into the following process. The least absolute shrinkage and selection operator (LASSO) Cox regression method[31] was then used to obtain the most useful prognostic features from candidate features. These selections were all performed in the training set. We evaluated the performance of all individual selected features using the c-index in the validation set and retained the best feature from each of the four groups (Protocol Table 5). These four chosen features were then introduced into a multivariable Cox model to predict OS. The radiomics signatures used in this study were calculated for each patient as a linear combination of selected features weighted by their respective coefficients. This building process was done for both the primary tumor and the lymph nodes, resulting in two signatures, the primary tumor and lymph nodes radiomics signatures, which were then averaged to generate the final radiomics marker. The performance on the discovery cohort was evaluated on the test sets. All other evaluation settings were the same as those used for DeepPET-OPSCC.

Protocol Table 5. Results of radiomics feature selection for tumor and lymph nodes.

	Tumor features	**Lymph nodes features**
1st Fold	Firstorder_TotalEnergy GLRLM_GrayLevelNonUniformity Shape_MeshVolume Wavelet.HHH_glrlm_GrayLevelVariance	Firstorder_Energy GLRLM_GrayLevelNonUniformity Shape_MajorAxisLength Wavelet.HHH_glcm_Contrast
2nd Fold	Firstorder_Uniformity GLCM_ClusterTendency Shape_MinorAxisLength Wavelet.HHH_firstorder_Kurtosis	Firstorder_Minimum GLCM_DifferenceEntropy Shape_MajorAxisLength Wavelet.HHL_glcm_Imc2
3rd Fold	Firstorder_Maximum GLCM_Correlation Shape_Sphericity HHH_firstorder_Maximum	Firstorder_RootMeanSquared GLCM_JointAverage Shape_Maximum3DDiameter Wavelet.HHH_firstorder_InterquartileRange
4th Fold	Firstorder_Minimum GLCM_Autocorrelation Shape_LeastAxisLength Wavelet.HHH_firstorder_TotalEnergy	Firstorder_Energy GLCM_ClusterShade Shape_Maximum2DDiameterColumn Wavelet.HHL_glcm_Imc1
5th Fold	Firstorder_TotalEnergy GLRLM_GrayLevelVariance Shape_MeshVolume Wavelet.HHH_glrlm_ShortRunLowGrayLevelEmphasis	Firstorder_Maximum GLCM_ClusterTendency Shape_MajorAxisLength Wavelet.HHL_glcm_Imc2

4 External TCIA and Clinical Deployment Test Cohorts

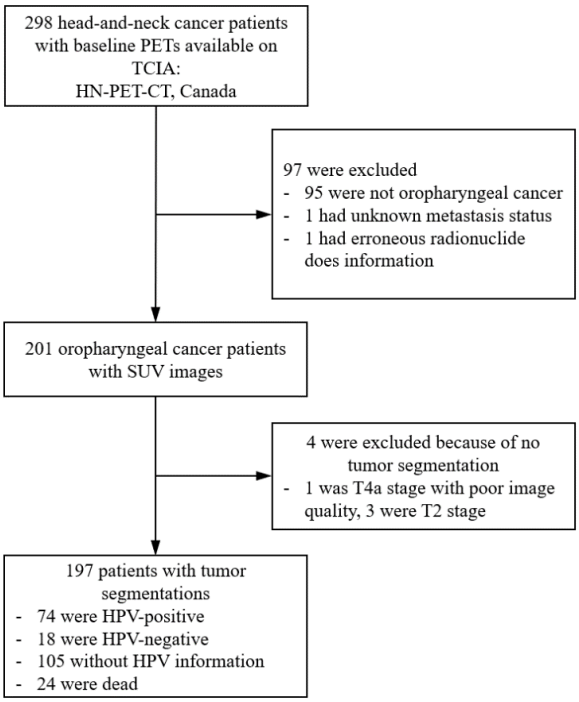
4.1. TCIA Test Cohort

Six external cohorts from six hospitals in North America and Europe (Canada, USA, and Netherlands) were used as the test cohort: the HN-PET-CT cohort (including four sub-cohorts), the HNSCC cohort, and the Head-Neck-Radiomics-HN1 cohort that are described in the following subsections. These data are publicly available on The Cancer Imaging Archive (TCIA). PET scans were obtained on 6 unseen different PET scanners from three manufacturers (Protocol Table 1).

4.1.1. HN-PET-CT cohort

The HN-PET-CT cohort included 298 patients with head-and-neck squamous cell carcinoma (HNSCC) treated between April 2006 and November 2014 at four separate institutions in Canada: Hôpital général

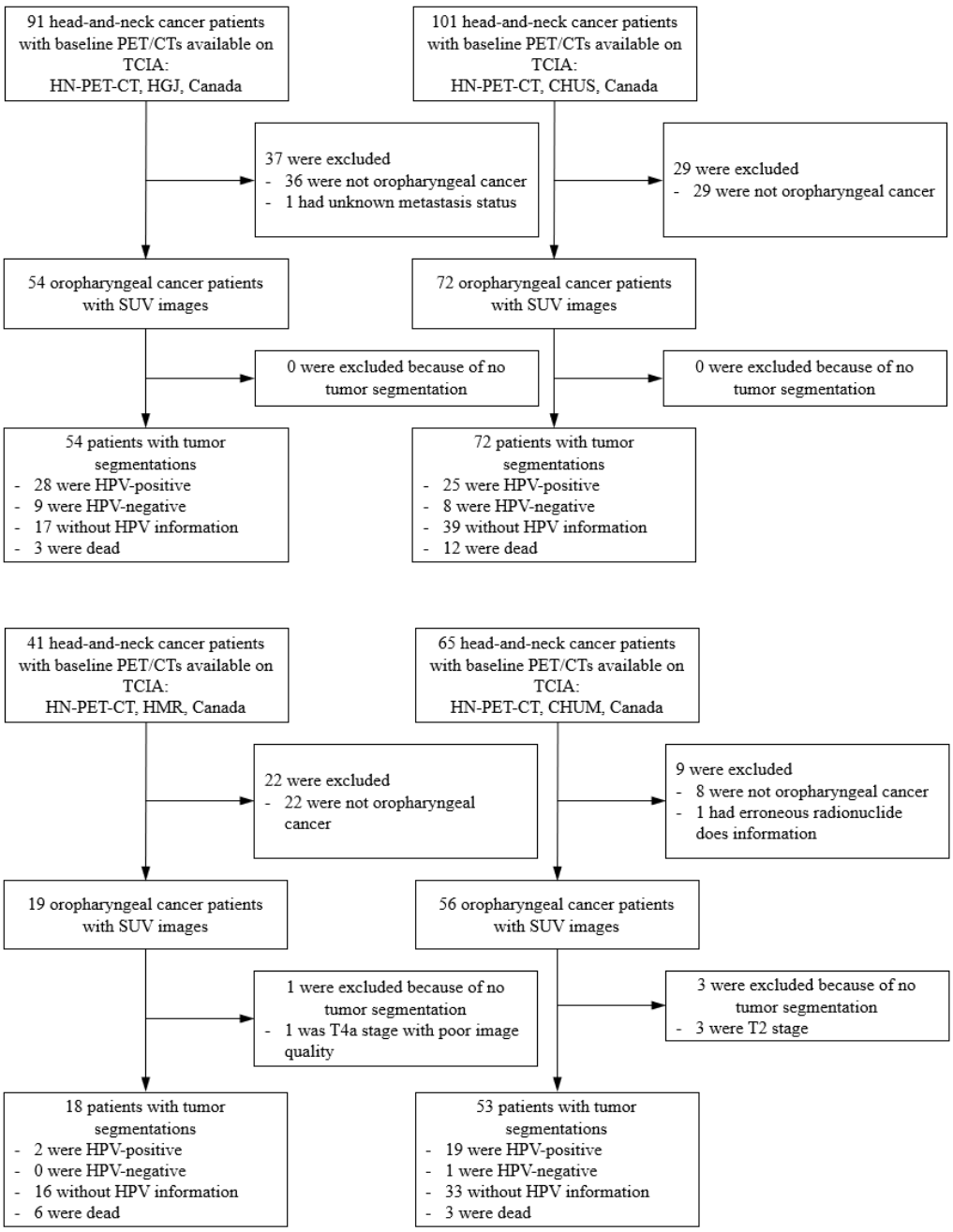
juif (HGJ), Centre hospitalier universitaire de Sherbrooke (CHUS), Hôpital Maisonneuve-Rosemont (HMR), and Centre hospitalier de l'Université de Montréal (CHUM).[32] All patients underwent FDG-PET/CT imaging scans within a median of 18 days (range: 6-66) before treatment. Ninety-seven patients were excluded from the study, due to location of tumors not in the oropharynx (95), unknown metastasis status (one), and erroneous radionuclide dose information (one) (Protocol Fig. 8). The remaining 201 oropharyngeal cancer patients' PET images were converted to standard uptake value (SUV) maps using a radiomics package (https://github.com/mvallieres/radiomics) provided by the HN-PET-CT cohort data provider, and the file formats were converted from DICOM to NifTI (https://nifti.nimh.nih.gov/). The automated segmentation model identified no tumor for four patients on the SUV images. The final 197 included patients with segmentations of primary tumor and lymph nodes were defined as the HN-PET-CT cohort. HPV status was available for 92 patients, among which 74 (80%) were HPV+ and 18 (20%) were HPV–. TNM staging for all patients were available, and no patient had distant metastases. Twenty-eight patients received radiation alone (14%), 158 received chemoradiation (80%) with curative intent, and 11 (6%) received chemoradiation and surgery. Twenty-four patients (12%) died during the follow-up period. Protocol Fig. 9 specifies the detailed inclusion and exclusion information for the four sub-cohorts (HGJ, CHUS, HMR, and CHUM) in the HN-PET-CT cohort.



Protocol Fig. 8 A diagram specifying inclusion and exclusion information for patients and PET images from the TCIA: HN-PET-CT cohort, and the HPV status and the number of events of the included patients. TCIA=The Cancer Imaging Archive (http://www.cancerimagingarchive.net/). HPV=human papillomavirus.

4.1.2. HNSCC cohort

The HNSCC cohort included 215 patients with HNSCC treated from October 2003 to August 2013 at the University of Texas MD Anderson Cancer Center in USA[33]. The majority of patients underwent PET/CT before and after treatment. Only PET imaging before treatment was considered in our prognosis study. The median time between initial diagnostic imaging and treatment planning was 0.87 months (interquartile range 0.37-2.27 months). Fifty-nine patients had tumors not in the oropharynx, 13 patients did not have PET imaging before treatment, 13 patients' PET imaging did not contain head-and-neck region, and one patient's lymph node (stage = N3) was resected before PET scanning. Hence, these patients could not be considered as pretreatment imaging. The PET-to-SUV image conversion of the remaining 129 oropharyngeal cancer patients was performed by using the radiomics package (https://github.com/mvallieres/radiomics) and converted to NifTI format (https://nifti.nimh.nih.gov/). However, the scanning time in the calculation of decay was changed to the series time instead of the default acquisition time, after double-checking the generated SUV images with the 3D Slicer PETDICOMExtension (https://www.slicer.org/wiki/Documentation/4.10/Extensions/PETDICOM). The

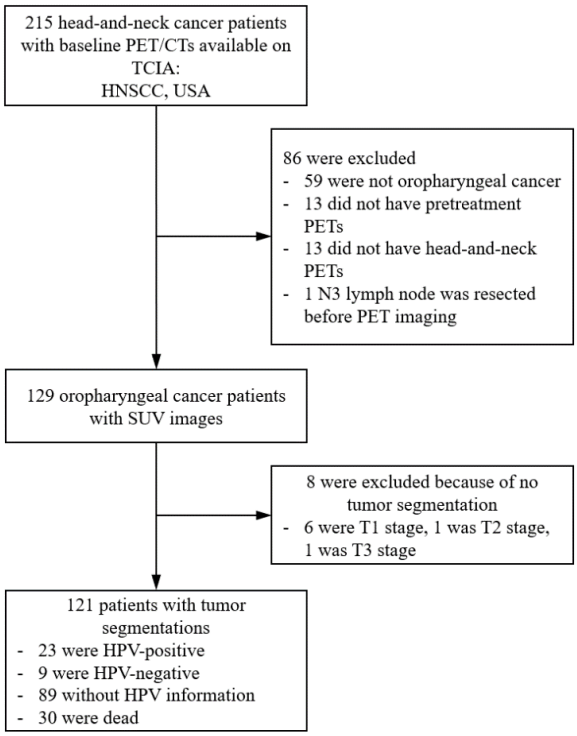


Protocol Fig. 9 Diagrams specifying inclusion and exclusion information for patients and PET images from the four sub-cohorts in TCIA: HN-PET-CT cohort. TCIA=The Cancer Imaging Archive (http://www.cancerimagingarchive.net/). HPV=human papillomavirus.

automated segmentation model identified no tumor for eight patients on SUV images. The remaining 121 patients comprised the HNSCC cohort (Protocol Fig. 10). HPV status was available for 32 patients, among which 23 (72%) were HPV+ and 9 (28%) were HPV–. Staging for all patients was assigned according to the AJCC 7th TNM staging system, and no patient had distant metastases. Treatment included RT, CCRT, IC, surgery, cetuximab, and various combinations of these. Thirty (25%) patients died during the follow-up period.

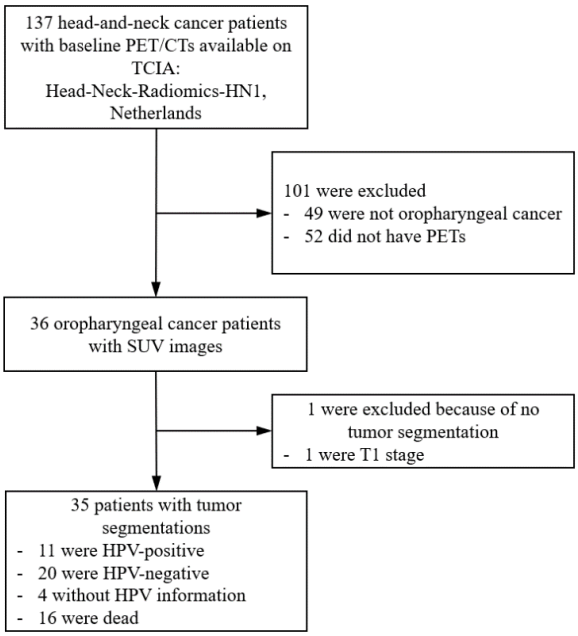
4.1.3. Head-Neck-Radiomics-HN1 cohort

The Head-Neck-Radiomics-HN1 cohort contained 137 HNSCC patients treated by radiotherapy at MAASTRO Clinic, in The Netherlands[34]. For these patients, PET/CT or CT scans before treatment were available. Forty-nine patients with tumors not in the oropharynx, and 53 with no PET imaging were



Protocol Fig. 10 A diagram specifying inclusion and exclusion information for patients and PET images from the TCIA: HNSCC cohort. TCIA=The Cancer Imaging Archive (http://www.cancerimagingarchive.net/). HPV=human papillomavirus.

excluded (Protocol Fig. 11), resulting in 36 oropharyngeal cancer patients with PET images, which were converted to SUV maps with NifTI (https://nifti.nimh.nih.gov/) format by the radiomics package (https://github.com/mvallieres/radiomics). Since the required patient body weight information was missing, we assumed that the weight of male patients was 84kg and that of female patients 70 kg according to the national statistical office, Statistics Netherlands (CBS) (https://www.cbs.nl/en-gb/news/2012/49/dutch-population-taller-and-heavier). The automated segmentation model identified no



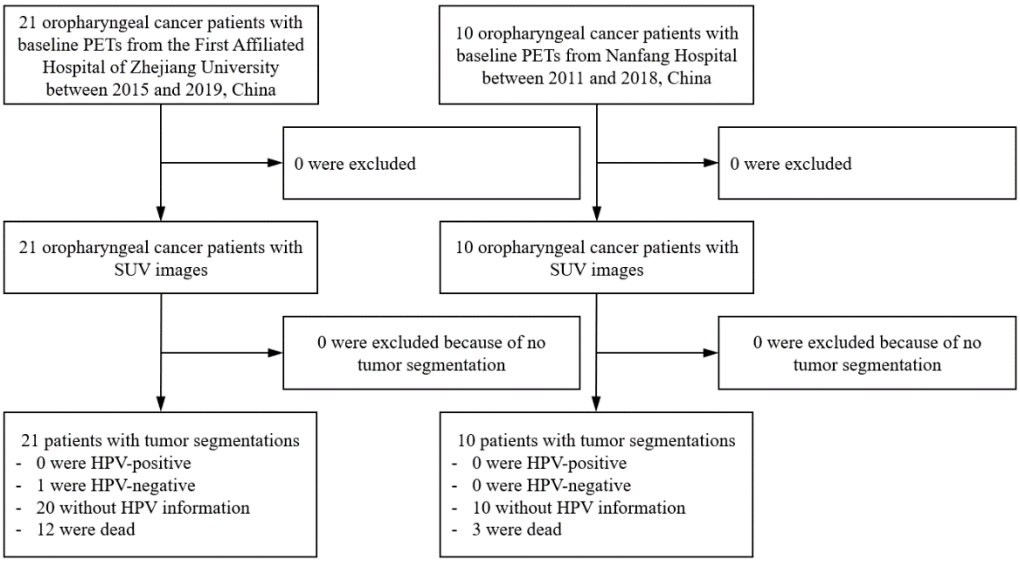
Protocol Fig. 11 A diagram specifying inclusion and exclusion for patients and PET images from the TCIA: Head-Neck-Radiomics-HN1cohort. TCIA=The Cancer Imaging Archive (http://www.cancerimagingarchive.net/). HPV=human papillomavirus.

tumor for one patient on SUV images. The remaining 35 included patients with the segmentation of the primary tumor and lymph nodes were defined as the Head-Neck-Radiomics-HN1 cohort. HPV status was available for 31 patients, with 11 (35%) HPV+ and 20 (65%) HPV–. TNM staging for all patients were available, and no patient had distant metastases. 16 (46%) patients died during the follow-up period.

4.2. Clinical Deployment Test Cohort

Two external cohorts from two hospitals in Asia (China) were used as the clinical deployment test cohort: the First Affiliated Hospital of Zhejiang University (ZJU1) and the Nanfang Hospital (NFH) cohorts. PET scans were obtained on 3 unseen different PET scanners from two manufacturers (Protocol Table 1). This cohort was used to test our automated DeepPET-OPSCC assay in the clinical environment (i.e., deployed the model and software in the First Affiliated Hospital of Zhejiang University, as detailed in Protocol Section 6). The investigators only needed to assign the PET DICOM data to the software, which will perform SUV conversion, segmentation, and prognosis automatically.

The clinical testing cohort included 31 patients with oropharyngeal cancer treated between April 2011 and March 2019 at ZJU1 and NFH. For these patients, whole body PET scans before treatment were available. The automated segmentation successfully identified tumors for all patients (Protocol Fig. 12). HPV status (p16 staining) was available for only 1 patient who was HPV–. The 7th edition TNM staging for all patients were available, and patients with distant metastases were preliminarily excluded.



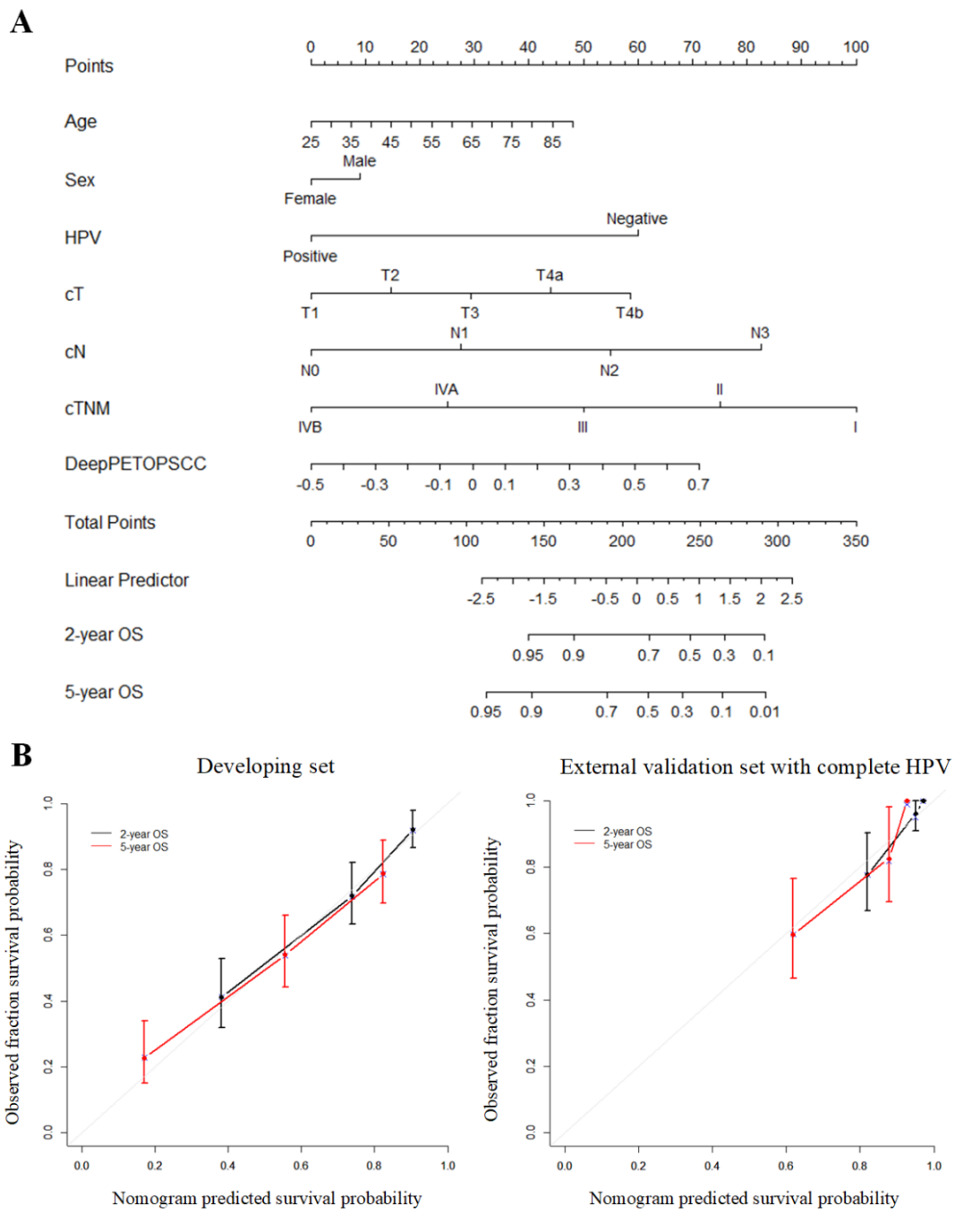
Protocol Fig. 12 A diagram specifying inclusion and exclusion information for patients and PET images from the two sub-cohorts in the clinical testing cohort, and the HPV status. HPV=human papillomavirus.

Twenty-five patients (81%) underwent CCRT: all received chemotherapy with cisplatin (75 mg/m2, high does) every 3 weeks. Twenty-three patients (74%) received surgery. Among them, four patients with an early stage (three T1N0 and one T2N0 cases) and one patient with decreased kidney function (T2N2b case with curative intended surgery) received surgery only. Although different treatment modalities were utilized, intensive radiotherapies (radiation dose range: 66-70 Gy; 2 Gy per day, 5 days per week) were performed for 26 (84%) patients.

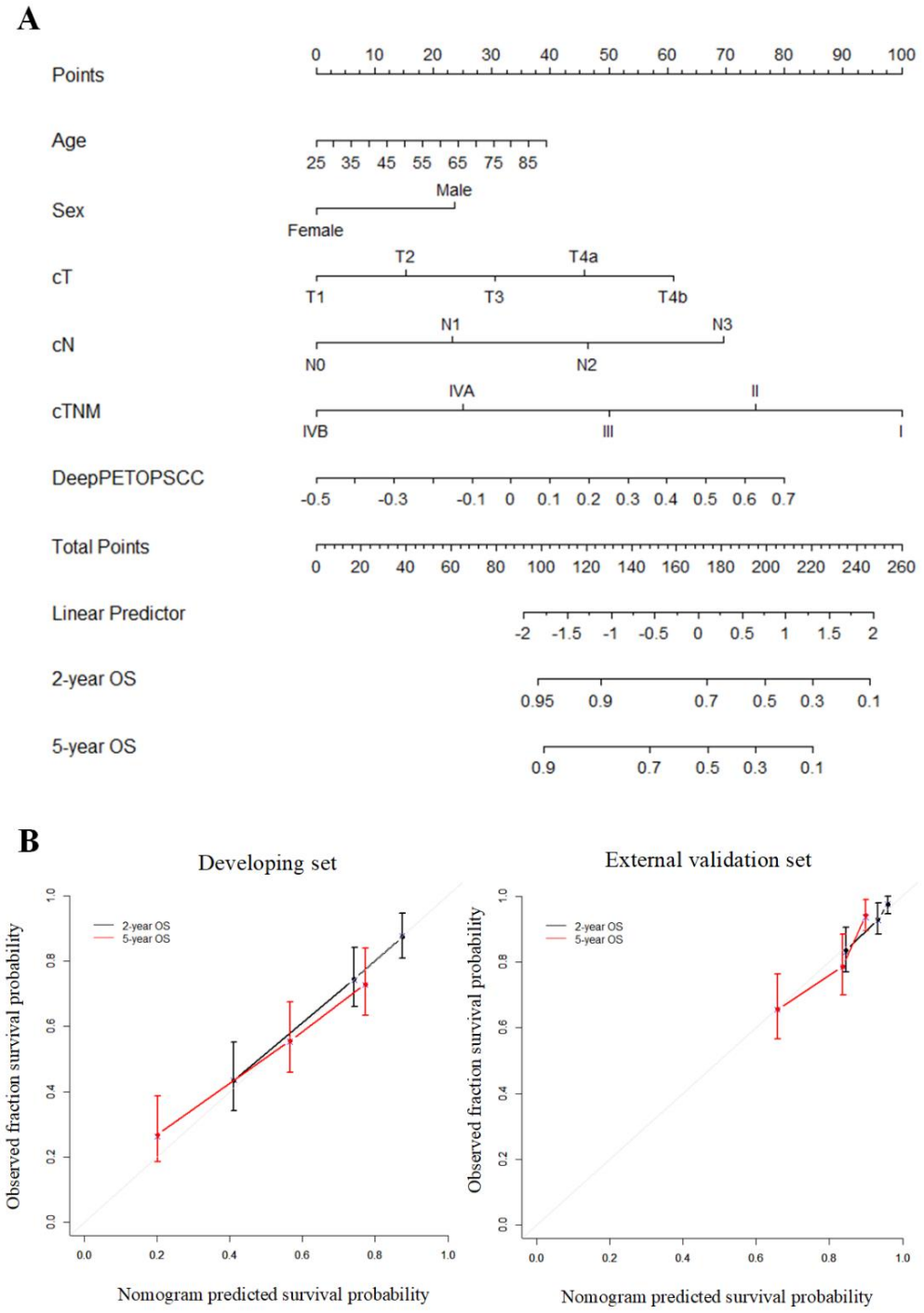
5 Nomogram

Integrated nomograms were built upon the DeepPET-OPSCC score and other clinical risk factors, i.e.,

- Integrated nomogram: combining DeepPET-OPSCC score, age, sex, HPV status, cT, cN, and cTNM stage.
- Integrated nomogram without HPV: combining DeepPET-OPSCC score, age, sex, cT, cN, and cTNM stage.



Protocol Fig. 13 Establishing and validating a nomogram which combines the DeepPET-OPSCC score and clinical factors to predict overall survival (OS). (A) Integrated nomogram combing the DeepPET-OPSCC score and age, sex, HPV, cT, cN, and cTNM stage. The points of DeepPET-OPSCC score, HPV status, cT, cN, and cTNM stages are obtained based on the top "points" bar with scale of 0-100. Then, the total point is calculated by summing the five points. By mapping the total point to the "Total Points" bar, the predicted risk score is obtained by checking the "Linear Predictor" bar, and the predicted n-year overall survival (OS) is obtained by checking the "n-year OS" bar. (B) Calibration curves of the integrated nomogram in prediction of 2-year and 5-year OS.

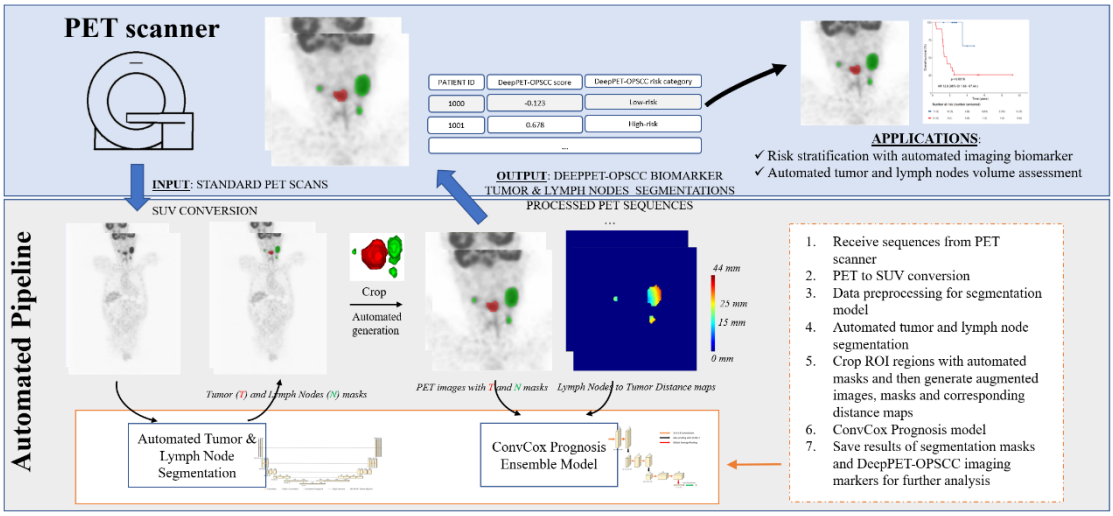


Protocol Fig. 14 Establishing and validating a nomogram which combines the DeepPET-OPSCC score and clinical factors without HPV status to predict overall survival (OS). (A) Integrated nomogram combing the DeepPET-OPSCC score and age, sex, cT, cN, and cTNM stage. The points of DeepPET-OPSCC score, HPV status, cT, cN, and cTNM stages are obtained based on the top "points" bar with scale of 0-100. Then, the total point is calculated by summing the five points. By mapping the total point to the "Total Points" bar, the predicted risk score is obtained by checking the "Linear Predictor" bar, and the predicted n-year overall survival (OS) is obtained by checking the "n-year OS" bar. (B) Calibration curves of the integrated nomogram in prediction of 2-year and 5-year OS.

Calibration curves were used to assess the consistency of predicted outcomes from the nomograms with true outcomes. The nomogram calibration curve demonstrated good agreement between the predicted and true survival probabilities, especially for 2-year OS (Protocol Fig. 13, Fig. 14). The integrated nomogram yielded a significantly higher c-index of 0.792 (95% CI 0.720–0.865) than the DeepPET-OPSCC score and any individual clinical factor in the TCIA test cohort with known HPV status. Similar improvements were observed for the integrated nomogram without HPV status (Supplementary Table S18). Time-dependent ROC analysis yielded AUCs at 2 years of the integrated nomograms with known HPV status of 0.804 (95% CI 0.758–0.848) and 0.867 (95% CI 0.797–0.931) in the discovery and TCIA test cohort, respectively; and without HPV of 0.754 (95% CI 0.659–0.841) in the TCIA test cohort. As such, the time-dependent ROC analysis significantly outperformed each individual factor (Supplementary Figure S16). AUCs at 5 years in the discovery cohort of the integrated nomogram without HPV status are provided in the Supplementary Figure S17.

6 Clinical Integration

We deployed an application-ready pipeline using the NVIDIA-Docker container (version 2.0, https://github.com/NVIDIA/nvidia-docker). We aimed to enable translation and application of our DeepPET-OPSCC in daily clinical practice and clinical trials. In routine clinical practice, automated processing of a PET scan begins directly after the images have been saved to the local workstation from the PET scanner. Processing of PET scans is completed by our pipeline in a fully automated fashion and requires no additional manual intervention or labor costs. The processed results (e.g., tumor and lymph nodes segmentation masks on individual PET sequences, tables that include DeepPET-OPSCC risk scores) are automatically saved in the local workstation. In current research, many prognosis studies require human manual input to annotate tumor (and lymph node) masks. Hence the appeal of our fully automated process, which could considerably help in deploying objective, reproducible, and scalable imaging biomarkers. We tested the developed pipeline in a simulated clinical environment (the First Affiliated Hospital of Zhejiang University) with automated processing of all PET scans from retrospectively enrolled patients. Each PET exam was processed by our pipeline in an average of 2 min 6 s on a machine with an Intel and an NVIDIA Titan RTX-6000 GPU. The processing pipeline can be scaled up linearly by adding additional processing power without changing and interrupting the existing workflow.



Protocol Fig. 15 The developed segmentation and prognosis model is part of a scalable and fully automated processing pipeline for PET exams. The pipeline is implemented in a fully automated fashion and does not require any additional human intervention and labors. The processed results (e.g. tumor and lymph node segmentation masks on individual PET sequences, tables that include DeepPET-OPSCC scores and risk categories) are automatically saved for interpretation. Overall, this enables objective, fully automated, high-throughput generation of imaging-based predictive biomarkers for OPSCC.

The Docker container that encapsulates our workflow executes the following fully automated steps:

1. Parallelized SUV conversion of DICOM images to NIfTI format.

2. Tumor and lymph node segmentation inference through the ensemble models described within this manuscript.
3. Parallelized data pre-processing (resampling both images and masks, distance map calculation and augmented inference samples generation)
4. Imaging marker generation through the ensemble DeepPET-OPSCC models.

The complete steps can be found in Protocol Fig. 15.

References

1. Fang YHD, Lin CY, Shih MJ, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. BioMed Res Int 2014:1-9, 2014.
2. Isensee F, Jäger PF, Kohl SA, Petersen J, Maier-Hein KH. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv 1904.08128, 2019.
3. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol 20:728–740, 2019.
4. Zhang C, Bengio S, Hardt M, Recht B, and Vinyals, O. Understanding deep learning requires rethinking generalization. 2017 International Conference on Learning Representations; Toulon, France; April 24–26, 2017.
5. Romera E, Bergasa LM, Alvarez JM, Trivedi M. Train here, deploy there: Robust segmentation in unseen domains. 2018 Institute of Electrical and Electronics Engineers (IEEE) Intelligent Vehicles Symposium; Changshu, China; June 26–30, 2018.
6. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006; 31: 1116-1128.
7. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2015 International Conference on Learning Representations; San Diego, USA; May 7–9, 2015.
8. Liu L, Jiang H, He P, et al. On the variance of the adaptive learning rate and beyond. 2020 International Conference on Learning Representations; Virtual Conference; Apr 26–May 1, 2020.
9. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24:1342-1350, 2018.
10. Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: Current applications and future directions. PLoS Med 15:1002707, 2018.
11. Dice LR. Measures of the amount of ecologic association between species. Ecology 26:297-302, 1945.
12. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247, 2009.
13. Leijenaar RT, Carvalho S, Hoebers FJ, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol 54:1423–1429, 2015.
14. Cheng NM, Fang YHD, Tsan DL, et al. Heterogeneity and irregularity of pretreatment 18F-fluorodeoxyglucose positron emission tomography improved prognostic stratification of p16-negative high-risk squamous cell carcinoma of the oropharynx. Oral Oncol 78:156–162, 2018.
15. Wu J, Gensheimer MF, Zhang N, et al. Tumor subregion evolution-based imaging features to assess early response and predict prognosis in oropharyngeal cancer. J Nucl Med 61:327–336, 2020.
16. Haider SP, Zeevi T, Baumeister P, et al. Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma. Cancers 12:1778–1790, 2020.
17. Peng H, Dong D, Fang MJ, et al. Prognostic value of deep learning PET/CT-based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. Clin Cancer Res 25:4271–4279, 2019.
18. Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 15:e1002711, 2018.
19. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 115:E2970-E2979, 2018.
20. Jiang Y, Jin C, Yu H, et al. Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. Ann Surg 2020.
21. Lou B, Doken S, Zhuang T, et al. An image-based deep learning framework for individualising

radiotherapy dose: a retrospective analysis of outcome prediction. Lancet Digit Health 1:e136–e147, 2019.
22. Yao, J., Shi, Y., Lu, L., Xiao, J., and Zhang, L. DeepPrognosis: preoperative prediction of pancreatic cancer survival and surgical margin via dynamic contrast-enhanced CT imaging. 2020 International Conference on Medical Image Computing and Computer Assisted Intervention; San Miguel, Peru; October 4–8, 2020.
23. Riaz N, Setton J, Tam M, et al. Patients with low lying lymph nodes are at high risk for distant metastasis in oropharyngeal cancer. Oral Oncol 50:863–868, 2014.
24. O'Sullivan B, Huang SH, Su J, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. Lancet Oncol 17:440–451, 2016.
25. Xing Y, Zhang J, Lin H, et al. Relation between the level of lymph node metastasis and survival in locally advanced head and neck squamous cell carcinoma. Cancer 122:534–545, 2016.
26. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18:24, 2018.
27. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 247:2543–2546, 1982.
28. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016 Institute of Electrical and Electronics Engineers (IEEE) Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA; June 27–30, 2016.
29. Wang X, Deng X, Fu Q, et al. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans Med Imaging 39:2615–2625, 2020.
30. Van Griethuysen JJ, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 77:e104–e107, 2017.
31. Tibshirani, R. The lasso method for variable selection in the Cox model. Stat Med 16:385–395, 1997.
32. Vallieres M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep 7:1–14, 2017.
33. Grossberg AJ, Mohamed AS, Elhalawani H, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci Data 5:180173, 2018.
34. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:1–9, 2014.

Supplementary appendix

Supplement to: Nai-Ming Cheng, Jiawen Yao, Jinzheng Cai, et al. Deep learning for fully-automated prediction of overall survival in patients with oropharyngeal cancer using FDG PET imaging: an international retrospective study.

Table of Contents

Table S1: REporting recommendations for tumour MARKer prognostic studies (REMARK)

	Item to be reported	Where reported
INTRODUCTION		
1	State the marker examined, the study objectives, and any pre-specified hypotheses.	Introduction, Methods, appendix
MATERIALS AND METHODS		
Patients		
2	Describe the characteristics (e.g., disease stage or co-morbidities) of the study patients, including their source and inclusion and exclusion criteria.	Methods, Results
3	Describe treatments received and how chosen (e.g., randomized or rule-based).	Results, Protocol
Specimen characteristics		
4	Describe type of biological material used (including control samples) and methods of preservation and storage.	Methods, Protocol
Assay methods		
5	Specify the assay method used and provide (or reference) a detailed protocol, including specific reagents or kits used, quality control procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols. Specify whether and how assays were performed blinded to the study endpoint.	Methods (FDG PET imaging)
Study design		
6	State the method of case selection, including whether prospective or retrospective and whether stratification or matching (e.g., by stage of disease or age) was used. Specify the time period from which cases were taken, the end of the follow-up period, and the median follow-up time.	Methods
7	Precisely define all clinical endpoints examined.	Methods
8	List all candidate variables initially examined or considered for inclusion in models.	Appendix
9	Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.	Include as many samples as possible
Statistical analysis methods		
10	Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.	Methods, appendix, Protocol
11	Clarify how marker values were handled in the analyses; if relevant, describe methods used for cutpoint determination.	Method, Protocol
RESULTS		
Data		
12	Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.	Protocol
13	Report distributions of basic demographic characteristics (at least age and sex), standard (disease-specific) prognostic variables, and tumor marker, including numbers of missing values.	Table 1, appendix
Analysis and presentation		
14	Show the relation of the marker to standard prognostic variables.	Table 2 and 3
15	Present univariable analyses showing the relation between the marker and outcome, with the estimated effect (e.g., hazard ratio and survival probability). Preferably provide similar analyses for all other variables being analyzed. For the effect of a tumor marker on a time-to-event outcome, a Kaplan-Meier plot is recommended.	Results, figure 2, appendix
16	For key multivariable analyses, report estimated effects (e.g., hazard ratio) with confidence intervals for the marker and, at least for the final model, all other variables in the model.	Table 2
17	Among reported results, provide estimated effects with confidence intervals from an analysis in which the marker and standard prognostic variables are included, regardless of their statistical significance.	Table 2, appendix
18	If done, report results of further investigations, such as checking assumptions, sensitivity analyses, and internal validation.	Internally tested by nested cross-validation, externally tested in two cohorts
DISCUSSION		
19	Interpret the results in the context of the pre-specified hypotheses and other relevant studies; include a discussion of limitations of the study.	Discussion
20	Discuss implications for future research and clinical value.	Discussion

Table S2: Checklist necessary for possible approval of an AJCC (American Joint Committee on Cancer) risk model proposed by Kattan et al

Item number	Checklist item	Comments
Inclusion Criteria (Model Must Have All of These Characteristics)		
1	The probability of overall survival, DSS, or DSM must be the outcome predicted.	OK, overall survival is predicted.
2	The model should address a clinically relevant question—a prediction someone cares about.	Yes, predict overall survival of non-metastasis oropharyngeal cancer with modern treatment.
3	At face value, the model should include the relevant predictors or explain why something relevant was not included.	Yes, included established predictive factors in the nomogram model.
4	The model validation study should specify precisely which patients were used to evaluate the model and the validation data set's inclusion/exclusion criteria.	OK, main manuscript (Methods) and protocol.
5	The model should be assessed for generalizability and external validation.	Yes, assessed in internal test sets and external validation cohorts from eight centers.
6	The model should have a well-defined prognostic time zero.	Yes, date of cancer diagnosis, Methods and protocol.
7	All predictors must be known at time zero and sufficiently defined for someone else to use.	OK, main manuscript and appendix.
8	Sufficient detail must be available to implement the model (need the equation, not a crippled version or simple, not yet validated score chart) OR the author must allow free access to the model.	OK, main manuscript (Methods) and protocol.
9	A measure of discrimination must have been reported. This is often measured as the concordance index and needs to be assessed on the validation data set(s).	OK, c-index and AUC are used and assessed on the validation cohorts.
10	Calibration in the small must be assessed (from the external validation data set) and provided. Calibration in the small is a plot of predicted probability versus observed proportion.	OK, appendix (risk distribution).
11	The model should be validated over a time frame and in a practice setting that are relevant to contemporary patients with disease.	OK, main manuscript, our marker is validated in patients treated between 2003 and 2019.
12	It should be clear which initial treatment(s), if any, were applied, and with what frequency.	OK, main manuscript, our marker is effective in patients treated primarily by radiation with or without chemotherapy, surgery with or without postoperative treatments, or initial induction chemotherapy.
13	The development and/or the validation of the prediction model must appear as a peer-reviewed journal article. The reference(s) needs to be provided.	In submission.
Exclusion Criteria (Any of These Exclude a Model From Consideration)		
1	A substantial proportion of patients had essentially no follow-up (either missing entirely or very short censored follow-up) in the validation data set.	No, main manuscript, the median follow-up time is 2.3-4.3 years in validation cohorts.
2	No information is available on the number of missing values in the validation data set.	No, main manuscript and appendix, information missing values is provided.
3	The number of events in the validation data set is small. This is a relatively unexplored literature to make firm suggestions on how small is small. However, 100 events may be the minimum needed.	No, main manuscript, there are totally 212 events in the internal test sets and validation cohorts.

Source: Kattan, M.W., Hess, K.R., Amin, M.B., Lu, Y., Moons, K.G., Gershenwald, J.E., Gimotty, P.A., Guinney, J.H., Halabi, S., Lazar, A.J. and Mahar, A.L. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA Cancer J Clin* 2016; 66: 370-374.

Table S3: Clinical characteristics in the entire cohort with known HPV status, cT, cN, and cTNM stage information (n=419)

	HPV+ (n=161)	HPV– (n = 258)
Age, years	58 (54-64)	53 (47-61)
Age, years		
<55	48 (30%)	144 (56%)
≧55	113 (70%)	114 (44%)
Sex		
Female	35 (22%)	16 (6%)
Male	126 (78%)	242 (94%)
cT stage		
cT1	25 (16%)	15 (6%)
cT2	74 (46%)	74 (29%)
cT3	39 (24%)	54 (21%)
cT4a	17 (11%)	83 (33%)
cT4b	6 (4%)	32 (12%)
cN stage		
cN0	23 (14%)	57 (22%)
cN1	15 (9%)	26 (10%)
cN2	116 (72%)	158 (61%)
cN3	7 (4%)	17 (7%)
cTNM stage		
I	6 (4%)	3 (1%)
II	11 (7%)	22 (9%)
III	19 (12%)	33 (13%)
IVA	113 (70%)	158 (61%)
IVB	12 (8%)	42 (16%)
Primary treatment		
Surgery	5 (3%)	10 (4%)
Radiotherapy	156 (97%)	248 (96%)
Chemotherapy		
Yes	129 (80%)	223 (86%)
No	32 (20%)	35 (14%)
Follow-up time, years	3.8 (2.8-5.7)	2.6 (1.3-5.4)
Event		
Death	18 (10.9%)	135 (52.3%)
Overall survival (95% CI)		
2 years	95.7% (92.7-98.9)	63.8% (58.1-69.9)
5 years	85.4% (78.7-92.8)	45.6% (39.6-52.5)

NOTE. The median age 58 of HPV+ is similar with the median age 57 in the ICON-S study (reference 2 in the main text); the median age 53 of HPV– is similar with the median age 52 in a study from Asian (reference 18 in the main text). Four patients with HPV+ were excluded due to the missing cT and cTNM IV sub-stage information. The only HPV– patient in the clinical deployment test cohort is not included in this table and the following K-M plot subgroup analysis.
Abbreviations: HPV=human papillomavirus.

Table S4: Univariable Cox regression for overall survival of the DeepPET-OPSCC risk category, its constituents, and established prognostic markers in the discovery, TCIA test, and entire cohorts with known HPV status

Variables	Discovery cohort (n=268)		TCIA test cohort (n=353)		Entire cohort with known HPV status (n=423)	
	HR (95%CI)	p value	HR (95%CI)	p value	HR (95%CI)	p value
DeepPET-OPSCC (high vs low)	3·17 (2·18-4·63)	<0·0001	3·15 (1·97-5·05)	<0·0001	3·79 (2·70-5·32)	<0·0001
DeepPET-OPSCC-T (high vs low)	3·07 (2·11-4·46)	<0·0001	2·89 (1·81-4·63)	<0·0001	3·89 (2·78-5·45)	<0·0001
DeepPET-OPSCC-TN (high vs low)	2·82 (1·95-4·09)	<0·0001	2·71 (1·68-4·35)	<0·0001	3·08 (2·20-4·32)	<0·0001
Age (≥55 years vs <55 years)	0·67 (0·46-0·95)	0·026	2·17 (1·19-3·98)	0·012	0·59 (0·43-0·82)	0·0014
Sex (male vs female)	2·51 (1·02-6·14)	0·048	2·23 (1·07-4·68)	0·032	2·93 (1·44-5·97)	0·0031
HPV (+ vs –)	0·18 (0·09-0·38)	<0·0001	..	..	0·17 (0·10-0·27)	<0·0001
cT stage (T4 vs T1-T3)	2·92 (2·04-4·17)	<0·0001	2·47 (1·49-4·10)	0·00045	3·51 (2·55-4·84)	<0·0001
cN stage (N2-N3 vs N0-N1)	1·81 (1·20-2·74)	0·0051	0·94 (0·55-1·62)	0·83	1·65 (1·13-2·41)	0·010
cTNM stage (IV vs I-III)	1·86 (1·16-3·00)	0·011	1·20 (0·67 -2·15)	0·55	1·80 (1·17-2·79)	0·0081
Radiomics signature (high vs low)	1·85 (1·30-2·65)	0·00070	1·81 (1·13-2·90)	0·014	2·21 (1·59-3·06)	<0·0001
SUVmax (≥14.65 vs <14.65)	1·22 (0·86-1·73)	0·26	1·17 (0·73-1·87)	0·52	1·31 (0·95-1·80)	0·095
MTV (≥22.66 cm3 vs <22.66 cm3)	2·41 (1·68-3·47)	<0·0001	1·50 (0·94 -2·39)	0·093	2·40 (1·73-3·34)	<0·0001
Primary treatment (surgery vs radiotherapy)	0·96 (0·35-2·61)	0·94	0·54 (0·13-2·20)	0·39	0·95 (0·39-2·31)	0·91
Chemotherapy (yes vs no)	0·80 (0·39-1·64)	0·55	0·61 (0·38-0·98)	0·041	1·06 (0·69-1·63)	0·79

NOTE. Cutoff thresholds for SUVmax and MTV were median values in the discovery cohort.
TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S5: Multivariable Cox regression for overall survival in the discovery, TCIA test, and entire cohort with known HPV status; the multivariable model included the DeepPET-OPSCC risk category and established prognostic markers

Variables	Discovery cohort (n=268)		TCIA test cohort (n=353)		Entire cohort with known HPV status (n=423)	
	HR (95%CI)	p value	HR (95%CI)	p value	HR (95%CI)	p value
DeepPET-OPSCC (high vs low)	2.09 (1.34-3.25)	0.0011	2.78 (1.62-4.76)	0.0002	2.34 (1.58-3.47)	<0.0001
Age (≥55 years vs <55 years)	0.88 (0.60-1.30)	0.53	2.20 (1.20-4.05)	0.011	0.86 (0.61-1.20)	0.37
Sex (male vs female)	1.39 (0.55-3.54)	0.49	2.00 (0.95-4.25)	0.068	1.47 (0.70-3.07)	0.31
HPV (+ vs –)	0.19 (0.09-0.41)	<0.0001	..	..	0.23 (0.14-0.39)	<0.0001
cT stage (T4 vs T1-T3)	1.69 (1.07-2.69)	0.026	2.17 (1.10-4.25)	0.024	1.75 (1.15-2.64)	0.0084
cN stage (N2-N3 vs N0-N1)	1.83 (1.19-2.81)	0.0058	1.02 (0.57-1.83)	0.95	1.78 (1.20-2.65)	0.0041
SUVmax (≥14.65 vs <14.65)	0.69 (0.47-1.02)	0.061	1.33 (0.78-2.30)	0.30	0.77 (0.54-1.08)	0.12
MTV (≥22.66 cm3 vs <22.66 cm3)	1.28 (0.77-2.13)	0.34	0.60 (0.30-1.17)	0.13	1.09 (0.70-1.70)	0.69
Chemotherapy (Yes vs No)	1.47 (0.98-2.20)	0.062	0.54 (0.32-0.92)	0.024	0.96 (0.66-1.39)	0.83

TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S6: **Multivariable Cox regression for overall survival in the discovery, TCIA test, and entire cohort with known HPV status; the multivariable model included the DeepPET-OPSCC-T risk category and established prognostic markers**

Variables	Discovery cohort (n=268)		TCIA test cohort (n=353)		Entire cohort with known HPV status (n=423)	
	HR (95%CI)	p value	HR (95%CI)	p value	HR (95%CI)	p value
DeepPET-OPSCC-T (high vs low)	1.64 (1.03-2.61)	0.037	2.41 (1.41-4.14)	0.0014	2.14 (1.44-3.18)	0.00016
Age (≥55 years vs <55 years)	0.94 (0.64-1.39)	0.77	2.21 (1.20-4.08)	0.011	0.91 (0.65-1.28)	0.59
Sex (male vs female)	1.44 (0.57-3.66)	0.44	1.96 (0.93-4.14)	0.079	1.45 (0.69-3.02)	0.33
HPV (+ vs –)	0.21 (0.10-0.45)	<0.0001	..	..	0.24 (0.14-0.41)	<0.0001
cT stage (T4 vs T1-T3)	1.79 (1.14-2.82)	0.012	2.24 (1.14-4.41)	0.020	1.71 (1.13-2.59)	0.011
cN stage (N2-N3 vs N0-N1)	1.94 (1.26-2.99)	0.0026	1.02 (0.57-1.84)	0.94	1.81 (1.22-2.69)	0.0032
SUVmax (≥14.65 vs <14.65)	0.73 (0.50-1.07)	0.11	1.28 (0.74-2.20)	0.38	0.80 (0.57-1.13)	0.21
MTV (≥22.66 cm3 vs <22.66 cm3)	1.36 (0.82-2.27)	0.23	0.64 (0.32-1.26)	0.20	1.13 (0.73-1.76)	0.59
Chemotherapy (Yes vs No)	0.54 (0.25-1.16)	0.11	0.55 (0.33−0.94)	0.028	0.95 (0.66-1.38)	0.79

TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S7: **Multivariable Cox regression for overall survival in the discovery, TCIA test, and entire cohort with known HPV status; the multivariable model included the DeepPET-OPSCC-TN risk category and established prognostic markers**

Variables	Discovery cohort (n=268)		TCIA test cohort (n=353)		Entire cohort with known HPV status (n=423)	
	HR (95%CI)	p value	HR (95%CI)	p value	HR (95%CI)	p value
DeepPET-OPSCC-TN (high vs low)	1.92 (1.22-3.01)	0.0045	2.53 (1.46-4.41)	0.0010	1.85 (1.24-2.77)	0.0028
Age (≥55 years vs <55 years)	0.86 (0.59-1.26)	0.44	2.24 (1.22-4.13)	0.0097	0.86 (0.62-1.21)	0.39
Sex (male vs female)	1.53 (0.60-3.90)	0.37	2.06 (0.97-4.34)	0.059	1.50 (0.72-3.13)	0.28
HPV (+ vs –)	0.19 (0.09-0.39)	<0.0001	..	..	0.22 (0.13-0.37)	<0.0001
cT stage (T4 vs T1-T3)	1.63 (1.02-2.61)	0.042	2.25 (1.14-4.46)	0.020	1.74 (1.14-2.66)	0.0010
cN stage (N2-N3 vs N0-N1)	1.97 (1.28-3.03)	0.002	1.07 (0.60-1.92)	0.82	1.82 (1.23-2.71)	0.0028
SUVmax (≥14.65 vs <14.65)	0.68 (0.46-1.00)	0.048	1.26 (0.74-2.14)	0.39	0.80 (0.57-1.13)	0.20
MTV (≥22.66 cm3 vs <22.66 cm3)	1.36 (0.83-2.25)	0.27	0.58 (0.29-1.16)	0.12	1.17 (0.75-1.84)	0.49
Chemotherapy (Yes vs No)	0.46 (0.22-0.98)	0.045	0.51 (0.30-0.86)	0.012	0.92 (0.63-1.33)	0.66

TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S8: **Multivariable Cox regression for overall survival in the discovery, TCIA test, and entire cohort with known HPV status; the multivariable model included the DeepPET-OPSCC score (continuous variable) and established prognostic markers**

Variables	Discovery cohort (n=268)		TCIA test cohort (n=353)		Entire cohort with known HPV status (n=423)	
	HR (95%CI)	p value	HR (95%CI)	p value	HR (95%CI)	p value
DeepPET-OPSCC (continuous variable)	4.59 (2.29-9.20)	<0.0001	8.93 (2.96-26.95)	0.0001	4.45 (2.49-7.96)	<0.0001
Age (≥55 years vs <55 years)	0.99 (0.67-1.46)	0.95	2.16 (1.17-4.00)	0.014	1.07 (0.75-1.53)	0.71
Sex (male vs female)	1.57 (0.62-3.98)	0.34	1.80 (0.85-3.80)	0.12	1.51 (0.72-3.14)	0.28
HPV (+ vs –)	0.22 (0.11-0.47)	<0.0001	..	..	0.25 (0.15-0.42)	<0.0001
cT stage (T4 vs T1-T3)	1.92 (1.21-3.07)	0.0061	1.72 (0.78-3.75)	0.18	1.63 (1.08-2.44)	0.019
cN stage (N2-N3 vs N0-N1)	1.94 (1.26-2.99)	0.0026	0.95 (0.52-1.71)	0.85	1.83 (1.23-2.72)	0.0027
SUVmax (continuous variable)	0.95 (0.88-1.03)	0.21	1.02 (0.92-1.13)	0.24	1.00 (0.93-1.07)	0.98
MTV (continuous variable)	1.00 (0.99-1.00)	0.15	1.00 (0.98-1.01)	0.55	0.99 (0.98-1.00)	0.019
Chemotherapy (Yes vs No)	0.49 (0.23-1.02)	0.056	0.53 (0.31-0.89)	0.017	0.52 (0.32-0.85)	0.0097

TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S9: **Complete clinical characteristics (to complement Table 1 in the main text) in the discovery cohort**

	Discovery cohort (n=268)
Smoking	
Yes	213 (79%)
No	55 (21%)
Pack-years	
> 10	183 (68%)
≤ 10	65 (24%)
Missing	20 (7%)
Alcohol	
Yes	192 (72%)
No	76 (28%)
Betel nut	
Yes	153 (57%)
No	115 (43%)
TP53 intensity	
0	40 (15%)
1	4 (1%)
2	8 (3%)
3	191 (71%)
Missing	25 (9%)
TP53 percentage	
0-30%	148 (55%)
31%-60%	83 (31%)
61%-90%	60 (22%)
90%+	1 (<1%)
Missing	24 (9%)
EGFR intensity	
0	1 (<1%)
1	17 (6%)
2	80 (30%)
3	147 (55%)
Missing	23 (9%)
EGFR percentage	
0-30%	54 (20%)
31%-60%	77 (29%)
61%-90%	95 (35%)
90%+	19 (7%)
Missing	23 (9%)
Cyclin D1 intensity	
0	12 (4%)
1	13 (5%)
2	39 (15%)
3	180 (67%)
Missing	24 (9%)
Cyclin D1 percentage	
0-30%	156 (58%)
31%-60%	47 (18%)
61%-90%	41 (15%)
90%+	0
Missing	24 (9%)
ERCC Ki intensity	
0	25 (9%)
1	48 (18%)
2	71 (26%)
3	98 (37%)
Missing	26 (10%)
ERCC Ki percentage	
0-30%	106 (40%)
31%-60%	59 (22%)
61-90%	71 (26%)
90%+	6 (2%)
Missing	26 (10%)
Cell differentiation	
Poor	70 (26%)
Moderate	152 (57%)
Well	10 (4%)
Missing	36 (13%)

Data are expressed as counts (percentages).

Table S10: **Univariable Cox regression for overall survival in the discovery cohort (with complete variables)**

Variables	Discovery cohort (n=268)	
	HR (95%CI)	p value
DeepPET-OPSCC (high vs low)	3.17 (2.18-4.63)	<0.0001
DeepPET-OPSCC-T (high vs low)	3.07 (2.11-4.46)	<0.0001
DeepPET-OPSCC-TN (high vs low)	2.82 (1.95-4.09)	<0.0001
Age (≥55 years vs <55 years)	0.67 (0.46-0.95)	0.026
Sex (male vs female)	2.51 (1.02-6.14)	0.048
HPV/p16 (+ vs –)	0.18 (0.09-0.38)	<0.0001
cT stage (T4 vs T1-T3)	2.92 (2.04-4.17)	<0.0001
cN stage (N2-N3 vs N0-N1)	1.81 (1.20-2.74)	0.0051
cTNM stage (IV vs I-III)	1.86 (1.16-3.00)	0.011
SUVmax (≥14.65 vs <14.65)	1.22 (0.86-1.73)	0.26
MTV (≥22.66 cm3 vs <22.66 cm3)	2.41 (1.68-3.47)	<0.0001
Smoking (Yes vs No)	3.72 (2.00-6.92)	<0.0001
Pack-years (>10 vs ≤10)	2.13 (1.20-3.81)	0.010
Alcohol (Y vs N)	2.97 (1.82-4.85)	<0.0001
Betel nut (Y vs N)	1.70 (1.18-2.46)	0.0046
TP53 intensity	0.89 (0.77-1.04)	0.14
TP53 percentage	1.01 (1.00-1.01)	0.019
Cyclin D1 intensity	1.50 (1.13-2.00)	0.0056
Cyclin D1 percentage	1.01 (1.00-1.02)	0.0021
EGFR intensity	1.18 (0.89-1.58)	0.25
EGFR percentage	1.01 (1.00-1.02)	0.015
ERCC Ki intensity	0.87 (0.73-1.03)	0.11
ERCC Ki percentage	1.00 (0.99-1.00)	0.43
Cell differentiation	1.28 (0.91-1.82)	0.15

HPV=human papillomavirus. MTV=metabolic tumor volume.

Table S11: **Multivariable Cox regression for overall survival in the discovery cohort (with complete variables)**

Variables	Discovery cohort (n=243*)	
	HR (95%CI)	p value
DeepPET-OPSCC (high vs low)	1.87 (1.19-2.92)	0.0063
Smoking (Yes vs No)	1.62 (0.68-3.89)	0.30
Alcohol (Y vs N)	1.69 (0.86-3.33)	0.13
BetNut (Y vs N)	0.73 (0.47-1.14)	0.16
HPV/p16 (+ vs –)	0.43 (0.19-0.99)	0.047
cT stage (T4 vs T1-T3)	1.84 (1.17-2.90)	0.0088
cN stage (N2-3 vs N0-1)	1.78 (1.15-2.77)	0.010
MTV (≥22.66 cm3 vs <22.66 cm3)	1.24 (0.77-1.99)	0.38
Cyclin D1 intensity	1.32 (0.97-1.79)	0.078
Cyclin D1 percentage	1.00 (0.99-1.01)	0.61
TP53 percentage	1.00 (0.99-1.01)	0.99
EGFR percentage	1.00 (0.99-1.01)	0.098

*25 patients excluded due to missing observations. Due to the limited number of death events (n=123), up to 12 variables that are significant ($p<0.05$) in the univariable analysis (Table S9) were chosen for the multivariable analysis to avoid overfitting in the model. Pack-years is correlated with Smoking so it is not included. HPV=human papillomavirus. MTV=metabolic tumor volume.

Table S12: Harrell's concordance index with 95% confidence intervals (CIs) between overall survival and a DeepPET-OPSCC individual or ensemble model scores, evaluated on the discovery cohort and TCIA test sub-cohorts. Bold highlights the best performing single model on the discovery cohort and the ensemble model

Variable	Discovery cohort	TCIA test cohort		
	CGMH	TCIA: HN-PET-CT	TCIA: HNSCC	TCIA: Head-Neck-Radiomics-HN1
DeepPET-OPSCC-T score model 1	0.725 (0.625-0.825)	0.616 (0.507-0.725)	0.649 (0.545-0.753)	0.724 (0.586-0.862)
DeepPET-OPSCC-T score model 2	0.718 (0.619-0.817)	0.643 (0.523-0.763)	0.675 (0.572-0.778)	0.667 (0.524-0.811)
DeepPET-OPSCC-T score model 3	**0.732 (0.638-0.826)**	**0.649 (0.527-0.772)**	**0.644 (0.536-0.752)**	**0.649 (0.520-0.777)**
DeepPET-OPSCC-T score model 4	0.659 (0.529-0.788)	0.729 (0.650-0.809)	0.662 (0.568-0.757)	0.653 (0.523-0.784)
DeepPET-OPSCC-T score model 5	0.644 (0.520-0.768)	0.663 (0.547-0.779)	0.653 (0.547-0.760)	0.642 (0.506-0.777)
DeepPET-OPSCC-T score combine	0.702 (0.652-0.752)	..	..	..
DeepPET-OPSCC-T score ensemble	**..**	**0.669 (0.559-0.778)**	**0.666 (0.560-0.768)**	**0.658 (0.517-0.799)**
DeepPET-OPSCC-TN score model 1	**0.729 (0.632-0.825)**	**0.638 (0.520-0.755)**	**0.676 (0.571-0.781)**	**0.677 (0.550-0.804)**
DeepPET-OPSCC-TN score model 2	0.684 (0.573-0.794)	0.642 (0.530-0.755)	0.701 (0.616-0.786)	0.717 (0.591-0.843)
DeepPET-OPSCC-TN score model 3	0.712 (0.622-0.803)	0.617 (0.495-0.739)	0.623 (0.513-0.734)	0.675 (0.560-0.789)
DeepPET-OPSCC-TN score model 4	0.683 (0.565-0.800)	0.630 (0.509-0.752)	0.692 (0.587-0.797)	0.677 (0.536-0.818)
DeepPET-OPSCC-TN score model 5	0.616 (0.506-0.726)	0.620 (0.473-0.766)	0.692 (0.604-0.779)	0.667 (0.544-0.791)
DeepPET-OPSCC-TN score combine	0.682 (0.632-0.733)	..	..	..
DeepPET-OPSCC-TN score ensemble	**..**	**0.648 (0.522-0.773)**	**0.694 (0.596-0.792)**	**0.672 (0.534-0.811)**

In the discovery cohort, each of five models was evaluated on the corresponding test sets in the nest five-fold cross-validation. When applying to the TCIA test cohort, each of five models was evaluated on all the patients. DeepPET-OPSCC-T=deep learning prognosis model uses SUV map/image and tumor mask as input. DeepPET-OPSCC-TN=deep learning prognosis model uses SUV map/image, tumor mask, and nodes-to-tumor distance map as input. CGMH=Chang Gung Memorial Hospital. TCIA=The Cancer Imaging Archive.

Table S13: Comparison between DeepPET-OPSCC and other computational models for overall survival prediction

Variable	Discovery cohort (n=268)		TCIA test cohort (n=353)	
	c-index	*P*	c-index	*P*
DeepPET-OPSCC	0.707 (0.658-0.757)	ref	0.689 (0.621-0.757)	ref
DeepPET-OPSCC-T	0.702 (0.652-0.752)	0.29	0.672 (0.604-0.739)	0.10
DeepPET-OPSCC-TN	0.682 (0.632-0.733)	0.012	0.692 (0.625-0.760)	0.63
3D ResNet-OPSCC	0.646 (0.595-0.697)	0.0022	0.665 (0.599-0.731)	0.21
2D DeepPET-OPSCC	0.605 (0.552-0.658)	<0.0001	0.591 (0.519-0.663)	0.0058
Radiomics signature	0.621 (0.570-0.672)	0.0001	0.608 (0.538-0.677)	0.0044

p-value is measured by the dependent Student *t* test.

Table S14: Clinical characteristics in the treatment analysis of using induction chemotherapy (IC) before concurrent chemoradiotherapy (CCRT) for patients with HPV– and TNM stage IVB cancer, in the entire cohort with known HPV status

	IC+CCRT (n=14)	CCRT (n = 26)	p value
Age, years	48 (44-54)	51 (44-58)	0.48
Age, years			0.48
<55	11 (79%)	17 (65%)	
≧55	3 (21%)	9 (35%)	
Sex			0.53
Female	0 (0%)	2 (8%)	
Male	14 (100%)	24 (92%)	
cT stage			0.45
cT1	0 (0%)	0 (0%)	
cT2	1 (7%)	4 (15%)	
cT3	0 (0%)	3 (12%)	
cT4a	1 (7%)	1 (4%)	
cT4b	12 (86%)	18 (69%)	
cN stage			0.61
cN0	1 (7%)	4 (15%)	
cN1	0	1 (4%)	
cN2	8 (57%)	10 (39%)	
cN3	5 (36%)	11 (42%)	
Follow-up time, years	1.8 (1.0-2.7)	3.1 (1.1-6.0)	
Event			
Death	11 (78.6%)	15 (57.7%)	
Overall survival (95% CI)			
2 years	50.0% (29.6-84.4%)	61.5% (45.4-83.4%)	
5 years	21.4% (7.86-58.4%)	43.8% (27.8-69.1%)	

*p values were calculated by the Kruskal-Wallis test for continuous variables and the Chi-square test or Fisher exact test for categorical variables, as appropriate.

Table S15: Associations between DeepPET-OPSCC risk category and the general characteristics of patients in the entire cohort with known HPV status (n=419)

Variable	DeepPET-OPSCC risk category		Spearman's correlation	
	High risk (n=179)	Low risk (n=240)	Rho (95% CI)	*P*
Age (continuous), years	55 (48–62)	56 (50–62)	-0.063 (-0.16 to 0.030)	0.198
Age (dichotomous), years	..	..	-0.068 (-0.16 to 0.028)	0.168
<55	89 (50%)	103 (43%)	..	..
≧55	90 (50%)	137 (57%)	..	..
Sex	..	..	0.14 (0.058 to 0.22)	0.003
Female	12 (7%)	39 (16%)	..	..
Male	167 (93%)	201 (84%)	..	..
HPV	..	..	-0.28 (-0.36 to -0.19)	<0.001
+	41 (23%)	120 (50%)	..	..
–	138 (77%)	120 (50%)	..	..
cT stage	..	..	0.48 (0.40 to 0.55)	<0.001
cT1	8 (4%)	32 (13%)	..	..
cT2	28 (16%)	120 (50%)	..	..
cT3	41 (23%)	52 (22%)	..	..
cT4a	69 (39%)	31 (13%)	..	..
cT4b	33 (18%)	5 (2%)	..	..
cN stage	..	..	0.18 (0.089 to 0.27)	<0.001
cN0	26 (15%)	54 (23%)	..	..
cN1	11 (6%)	30 (13%)	..	..
cN2	125 (70%)	149 (62%)	..	..
cN3	17 (9%)	7 (3%)	..	..
cTNM stage	..	..	0.35 (0.27 to 0.42)	<0.001
I	1 (<1%)	8 (3%)	..	..
II	4 (2%)	29 (12%)	..	..
III	12 (7%)	40 (17%)	..	..
IVA	120 (67%)	151 (63%)	..	..
IVB	42 (23%)	12 (5%)	..	..
SUV_{max} (continuous variable)	15.9 (11.8–21.8)	12.6 (9.5–17.6)	0.26 (0.16 to 0.34)	<0.001
MTV (continuous variable), cm^{3a}	41.6 (22.5–62.2)	14.2 (7.8–23.3)	0.54 (0.46 to 0.60)	<0.001
Death time	15.7 (11.8-25.2)	23.9 (16.1-37.0)	-0.22 (-0.36 to -0.06)	0.006
Follow-up time (Censored)	52.2 (31.6-86.7)	51.3 (37.6-77.9)	-0.02 (-0.15 to -0.10)	0.703

HPV, human papillomavirus; SUV_{max}, maximum standardized uptake value; MTV, metabolic tumor volume.

[a] Spearman's correlation calculated for an increase of 1 cm^3.

Table S16: Associations between the DeepPET-OPSCC risk category and the general characteristics of patients in the TCIA test cohort (n=348)

Variables	DeepPET-OPSCC risk category		Spearman's correlation	
	High risk (n=103)	Low risk (n=245)	Rho (95% CI)	p value
Age (continuous), years	62 (54-68)	61 (54-67)	0.002 (-0.11 to 0.11)	0.97
Age (dichotomous), years	..	..	-0.03 (-0.14 to 0.08)	0.60
<55	31 (30%)	67 (27%)	..	..
≧55	72 (70%)	178 (73%)	..	..
Sex	..	..	0.14 (0.04 to 0.22)	0.011
Female	13 (13%)	61 (25%)	..	..
Male	90 (87%)	184 (75%)	..	..
HPV*	..	..	-0.28 (-0.45 -0.12)	0.00047
+	22 (49%)	82 (77%)	..	..
–	23 (51%)	24 (23%)	..	..
cT stage	..	..	0.33 (0.21 to 0.43)	<0.0001
cT1	11 (11%)	43 (18%)	..	..
cT2	25 (24%)	124 (51%)	..	..
cT3	27 (26%)	60 (24%)	..	..
cT4a	37 (36%)	14 (6%)	..	..
cT4b	3 (3%)	4 (2%)	..	..
cN stage	..	..	0.11 (0.01 to 0.21)	0.035
cN0	8 (8%)	41 (17%)	..	..
cN1	11 (11%)	23 (9%)	..	..
cN2	75 (73%)	170 (69%)	..	..
cN3	9 (9%)	11 (4%)	..	..
cTNM stage	..	..	0.16 (0.06 to 0.25)	0.0025
I	0	6 (2%)	..	..
II	2 (2%)	21 (9%)	..	..
III	12 (12%)	36 (15%)	..	..
IVA	77 (75%)	167 (68%)	..	..
IVB	12 (12%)	15 (6%)	..	..
SUVmax (continuous)	15.3 (10.7-22.7)	12.9 (9.3-17.6)	0.18 (0.08 to 0.28)	0.0010
MTV (continuous), cm^3**	32.3 (14.9-57.1)	12.8 (6.6-22.6)	0.38 (0.27 to 0.48)	<0.0001
Death time	24.4 (14.5-44.9)	36.0 (19.2-57.4)	-0.17 (-0.41 to 0.07)	0.159
Follow-up time (Censored)	64.8 (42.3-86.0)	53.9 (39.9-82.4)	0.06 (-0.05 to 0.17)	0.325

*197 patients without HPV status. ** included as continuous variable (Spearman's correlation corresponds to an increase of 1 cm^3). Patients with substage missing were not included. TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S17: 5-year area under the curves (AUCs) of integrated nomograms, DeepPET-OPSCC score, and clinical factors in the discovery and TCIA test cohorts

Variables	Discovery cohort (n=268)		TCIA test cohort (n=348)		TCIA test cohort with known HPV status (n=151)	
	AUC (95% CI)	p value	AUC (95% CI)	p value	AUC (95% CI)	p value
Integrated nomogram	0.793 (0.749-0.834)	1 (ref)	..	..	0.801 (0.727-0.874)	1 (ref)
Integrated nomogram without HPV	0.755 (0.706-0.801)	0.019	0.727 (0.671-0.780)	1 (ref)	..	..
Clinical model + MTV	0.765 (0.717-0.810)	0.025	..	..	0.754 (0.659-0.843)	0.036
Clinical model	0.768 (0.722-0.813)	0.052	..	..	0.749 (0.649-0.842)	0.031
Clinical model (wo HPV) + MTV	0.715 (0.663-0.770)	0.0029	0.657 (0.594-0.721)	0.0035	..	..
Clinical model (wo HPV)	0.713 (0.662-0.767)	0.0029	0.659 (0.597-0.723)	0.0066	..	..
DeepPET-OPSCC score	0.728 (0.677-0.777)	0.0048	0.670 (0.594-0.745)	0.057	0.667 (0.549-0.794)	0.0052
HPV	0.634 (0.593-0.674)	<0.0001	..	..	0.624 (0.530-0.729)	<0.0001
cT	0.686 (0.632-0.738)	0.0003	0.613 (0.553-0.674)	0.0003	0.632 (0.522-0.745)	0.0014
cN	0.582 (0.529-0.634)	<0.0001	0.504 (0.443-0.564)	<0.0001	0.524 (0.441-0.610)	<0.0001
cTNM	0.593 (0.536-0.645)	<0.0001	0.522 (0.458-0.586)	<0.0001	0.517 (0.423-0.614)	<0.0001
SUVmax	0.539 (0.481-0.595)	<0.0001	0.485 (0.421-0.547)	<0.0001	0.572 (0.479-0.671)	0.0002
MTV	0.642 (0.584-0.697)	<0.0001	0.598 (0.534-0.662)	0.0003	0.651 (0.552-0.747)	0.0038

Integrated nomogram: combining DeepPET-OPSCC score, age, sex, HPV status, cT, cN, and cTNM stage.
Clinical model: combining age, sex, HPV status, cT, cN, and cTNM stage.
p-value is measured by the *z* test using 1000 bootstrap replicates. TCIA=The Cancer Imaging Archive. HPV=human papillomavirus.
Notably, for TCIA test cohort (n = 353), 5 patients were labelled as cT4 stage but no information of cT4a or cT4b was available. Four of those five cases had HPV data. Since we applied integrated nomogram in subsequent studies that included clinical information (age, sex, HPV status, cT, cN, and cTNM stage), those five patients lacked cT information did not participate in nomogram and ROC curves analyses. Therefore, 348 (353 minus 5) cases were enrolled in model performance studies.

Table S18: Harrell's concordance index (c-index) of integrated nomograms, clinical nomogram, clinical nomogram with MTV, radiomics, and PET-based markers in the discovery and TCIA test cohorts

Variables	Discovery cohort (n=268)		TCIA test cohort (n=348)		TCIA test cohort with known HPV status (n=151)	
	c-index (95% CI)	p value	c-index (95% CI)	p value	c-index (95% CI)	p value
Integrated nomogram	0·757 (0·714-0·800)	1 (ref)	..	..	0·792 (0·720-0·865)	1 (ref)
Integrated nomogram (without HPV)	0·731 (0·684-0·779)	0.024	0·712 (0·646-0·777)	1 (ref)	..	..
Clinical model + MTV	0·726 (0·682-0·770)	0.0003	..	..	0·771 (0·697-0·845)	0·070
Clinical model	0·726 (0·683-0·769)	0.0006	..	..	0·768 (0·694-0·842)	0.056
Clinical model (without HPV) + MTV	0·690 (0·645-0·735)	<0·0001	0·664 (0·595-0·733)	0·0087	..	..
Clinical model (wo HPV)	0·684 (0·640-0·729)	<0·0001	0.664 (0.596-0.731)	0.011	..	..
DeepPET-OPSCC score	0·707 (0·658-0·757)	0.0023	0·688 (0·621-0·756)	0.186	0·714 (0·607-0·822)	0.034
Radiomics signature	0·621 (0·570-0·672)	<0·0001	0·607 (0·537-0·676)	0·0004	0·694 (0·588-0·800)	0·022
SUVmax	0·559 (0·504-0·614)	<0·0001	0·495 (0·421-0·569)	<0·0001	0·563 (0·445-0·681)	0·0003
MTV	0·641 (0·592-0·690)	<0·0001	0·629 (0·560-0·698)	0·0027	0·677 (0·562-0·791)	0·015

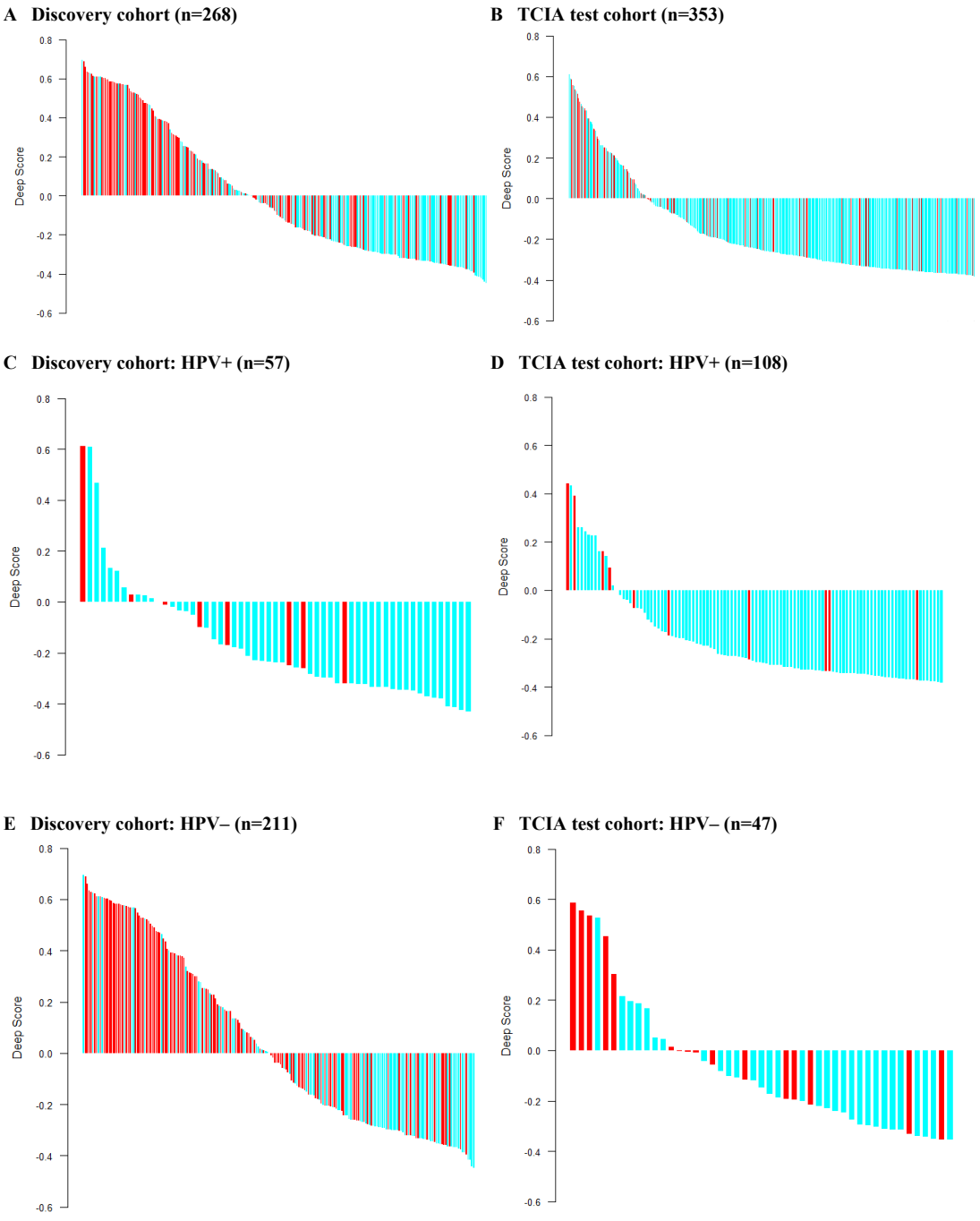
Integrated nomogram: combining DeepPET-OPSCC score, age, sex, HPV status, cT, cN, and cTNM stage.
Clinical model: combining age, sex, HPV status, cT, cN, and cTNM stage.
p-value is measured by the dependent Student *t* test. TCIA=The Cancer Imaging Archive. HPV=human papillomavirus. SUVmax=maximum standard uptake value. MTV=metabolic tumor volume.

Table S19: 2-year area under the curves (AUCs) of integrated nomograms, DeepPET-OPSCC score, and clinical factors in the discovery and TCIA test cohorts

Variables	Discovery cohort (n=268)		TCIA test cohort (n=348)		TCIA test cohort with known HPV status (n=151)	
	AUC (95% CI)	p value	AUC (95% CI)	p value	AUC (95% CI)	p value
Integrated nomogram	0.804 (0.758-0.848)	1 (ref)	..	..	0.867 (0.797-0.931)	1 (ref)
Integrated nomogram without HPV	0.772 (0.718-0.823)	0.022	0.754 (0.659-0.841)	1 (ref)	..	..
Clinical model + MTV	0.775 (0.724-0.820)	0.034	0.724 (0.634-0.809)	0.15	0.849 (0.778-0.916)	0.18
Clinical model	0.778 (0.727-0.823)	0.059	0.720 (0.632-0.803)	0.14	0.846 (0.775-0.912)	0.15
Clinical model (wo HPV) + MTV	0.738 (0.683-0.793)	0.0037				
Clinical model (wo HPV)	0.734 (0.678-0.788)	0.0036				
DeepPET-OPSCC score	0.741 (0.686-0.795)	0.0057	0.724 (0.625-0.815)	0.23	0.762 (0.626-0.887)	0.033
HPV	0.621 (0.587-0.654)	<0.0001	..	..	0.751 (0.647-0.849)	0.0063
cT	0.700 (0.643-0.758)	0.0003	0.701 (0.612-0.786)	0.14	0.704 (0.551-0.843)	0.022
cN	0.615 (0.565-0.663)	<0.0001	0.557 (0.483-0.629)	0.0008	0.618 (0.542-0.696)	<0.0001
cTNM	0.618 (0.570-0.666)	<0.0001	0.568 (0.492-0.639)	0.0053	0.596 (0.519-0.676)	<0.0001
SUVmax	0.596 (0.531-0.656)	<0.0001	0.529 (0.432-0.629)	<0.0001	0.558 (0.412-0.705)	0.0004
MTV	0.671 (0.610-0.731)	<0.0001	0.700 (0.612-0.790)	0.099	0.744 (0.588-0.881)	0.045

Integrated nomogram: combining DeepPET-OPSCC score, age, sex, HPV status, cT, cN, and cTNM stage.
Clinical model: combining age, sex, HPV status, cT, cN, and cTNM stage.
p-value is measured by the *z* test using 1000 bootstrap replicates. TCIA=The Cancer Imaging Archive. HPV=human papillomavirus.

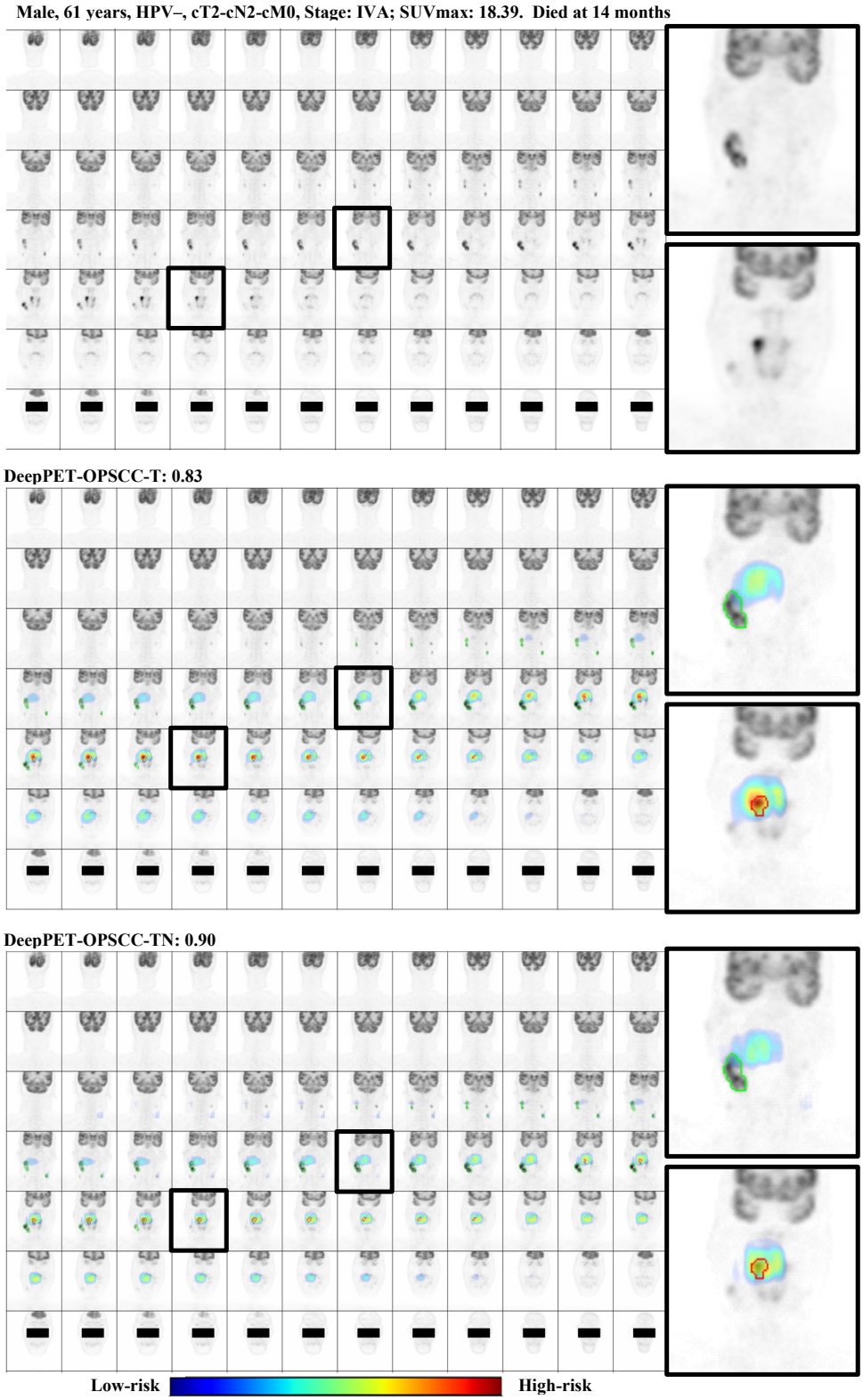
Figure S1: Risk distribution of DeepPET-OPSCC score and patient overall survival status. Results shown on discovery and TCIA test cohorts



Patients (horizontal axis) are sorted by the predicted risk scores (vertical axis). Red bars indicate events (i.e., death), Cyan bars indicate censored. TCIA=The Cancer Imaging Archive.

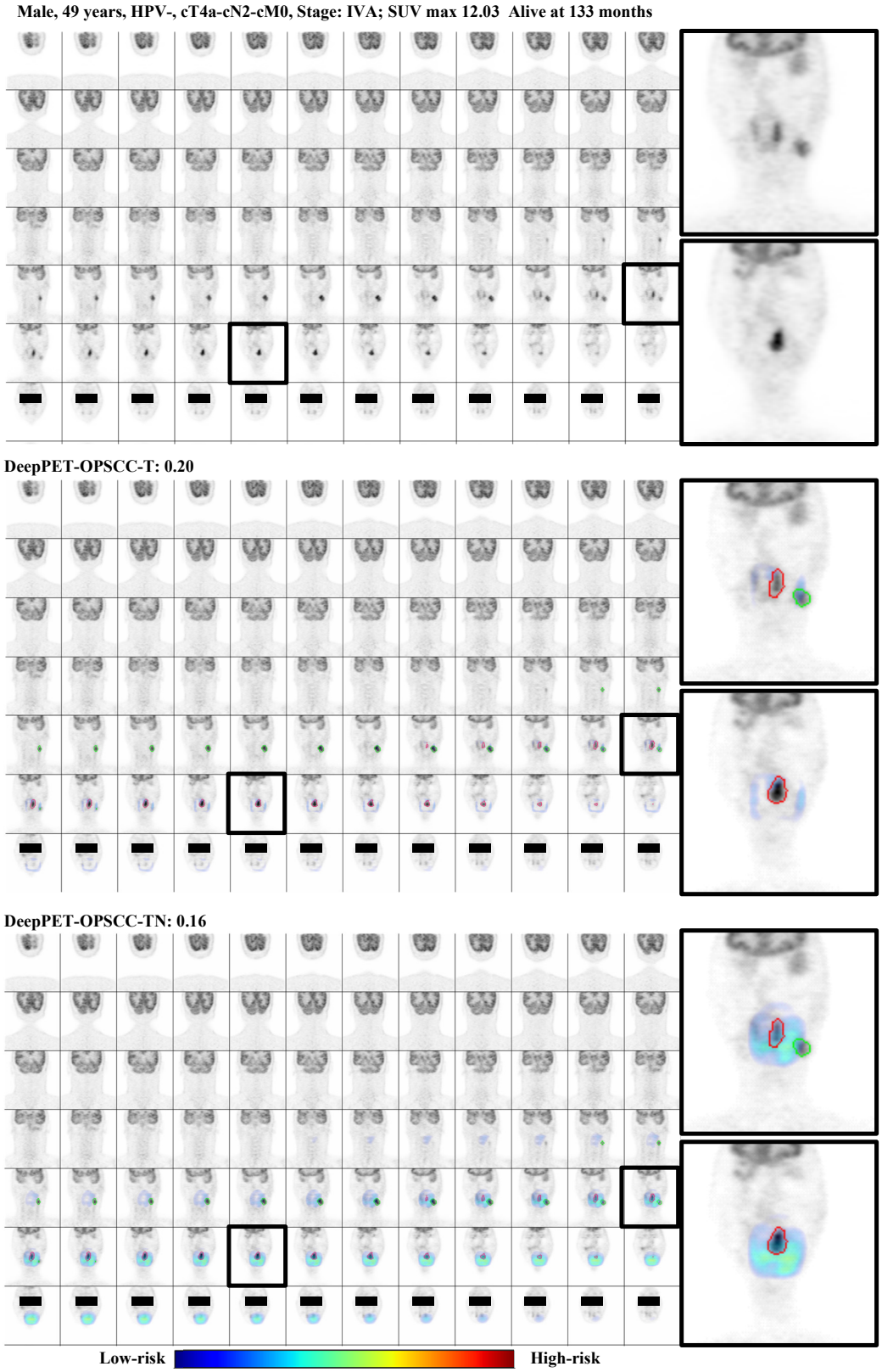
Figure S2: Examples of 3D PET images and corresponding heatmaps

Figure S2A: Examples of 3D PET images (consecutive image slices), corresponding activation maps (heatmaps), and two enlarged images with heatmaps for better visual observation



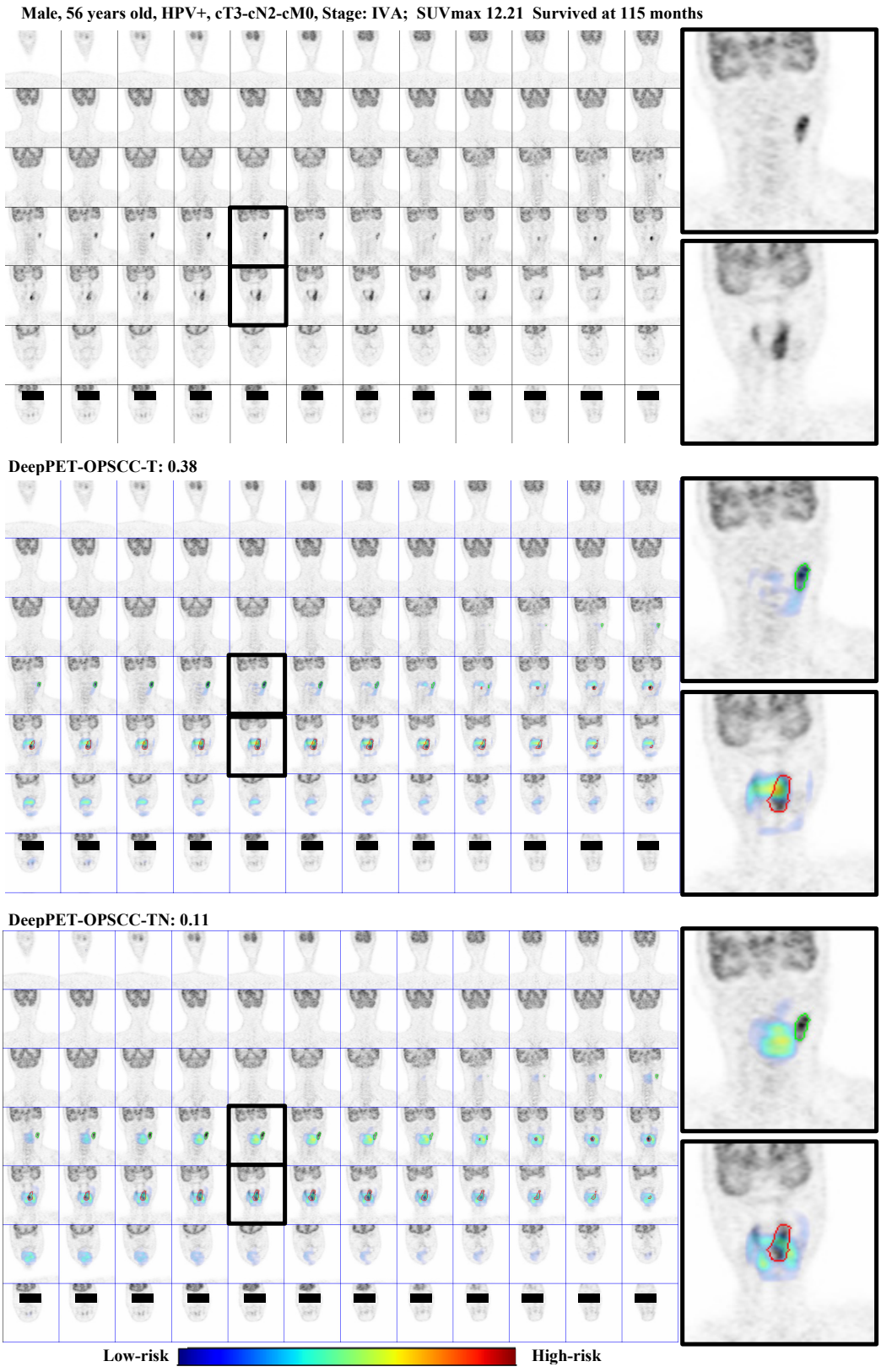
Red curves: auto-segmented tumor boundaries; Green curves: auto-segmented lymph node boundaries. The PET images are anonymized by blocking the eye region with black boxes.

Figure S2B: Additional examples of 3D PET images and corresponding heatmaps



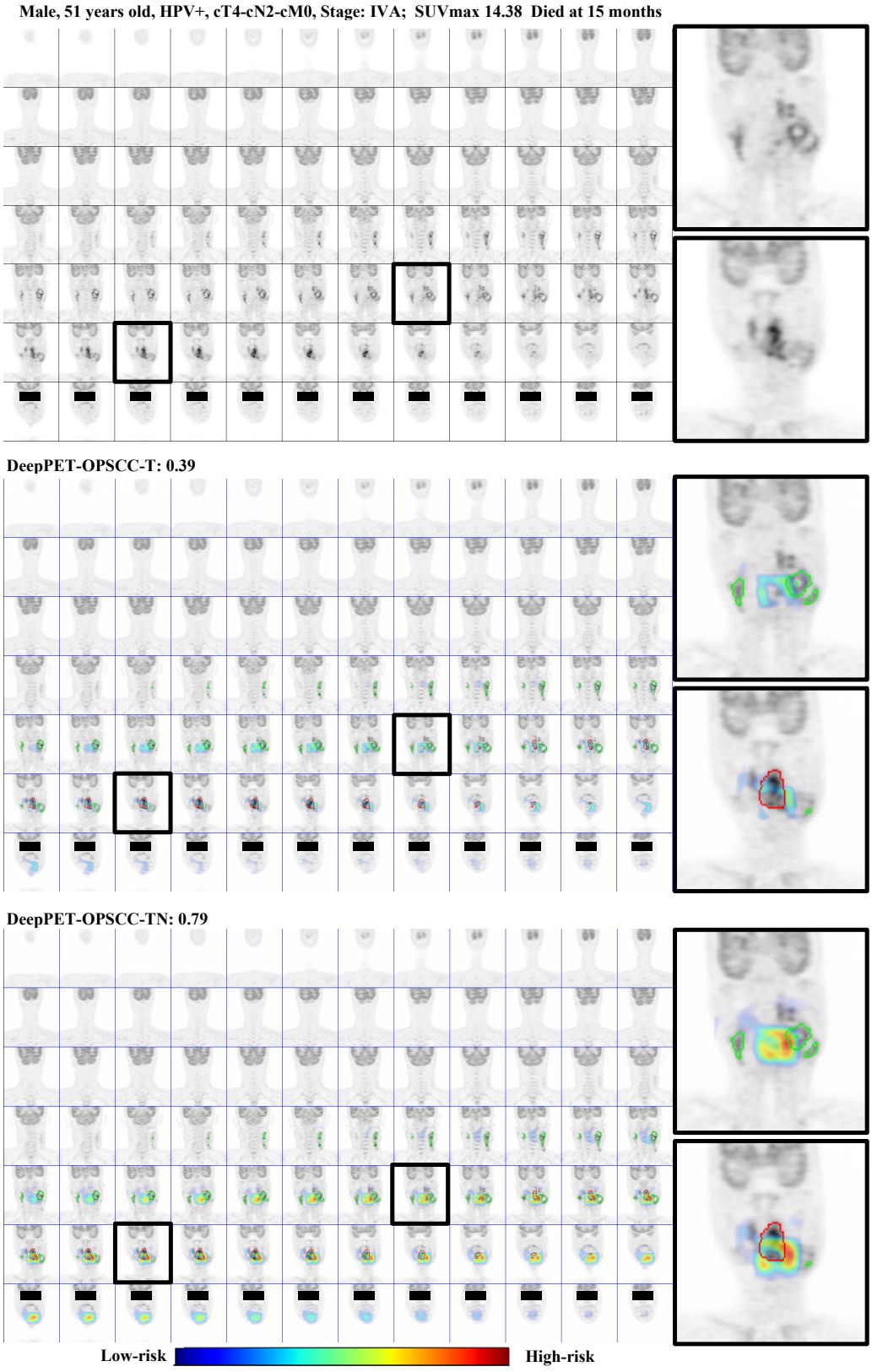
Red curves: auto-segmented tumor boundaries; Green curves: auto-segmented lymph node boundaries. The PET images are anonymized by blocking the eye region with black boxes.

Figure S2C: Additional examples of 3D PET images and corresponding heatmaps



Red curves: auto-segmented tumor boundaries; Green curves: auto-segmented lymph node boundaries. The PET images are anonymized by blocking the eye region with black boxes.

Figure S2D: Additional examples of 3D PET images and corresponding heatmaps



Red curves: auto-segmented tumor boundaries; Green curves: auto-segmented lymph node boundaries. The PET images are anonymized by blocking the eye region with black boxes.

Figure S3: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in the entire cohort with known HPV status

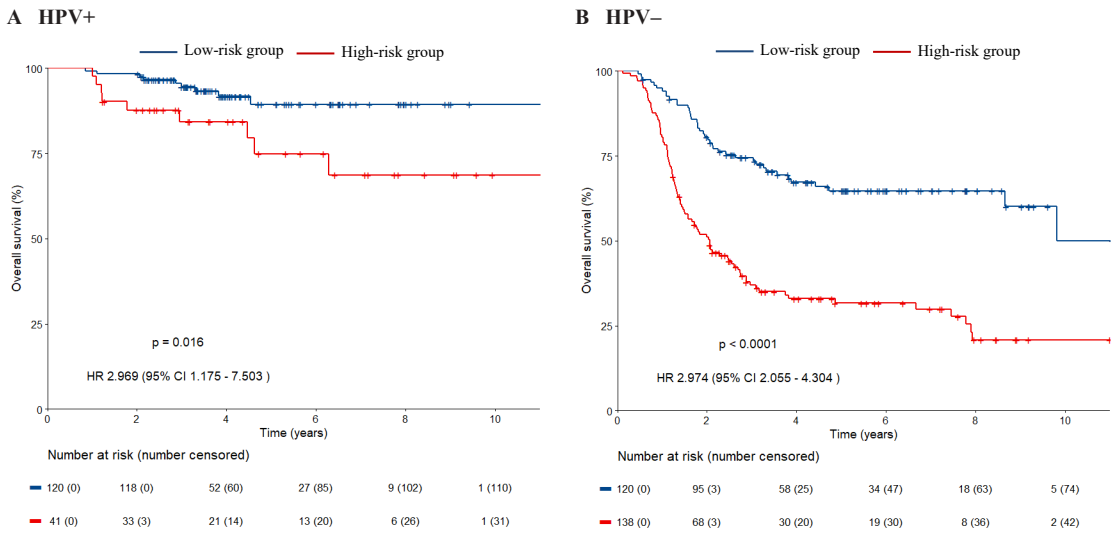
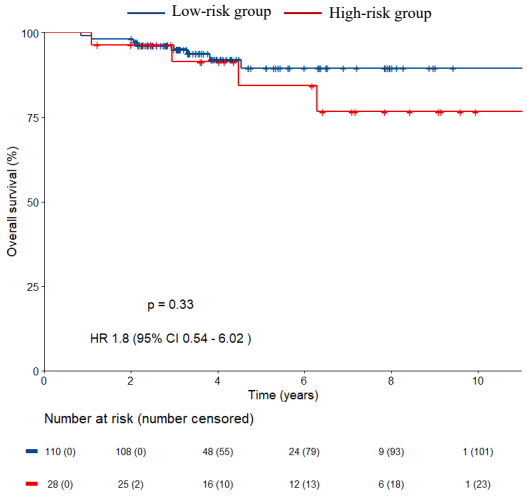


Figure S4: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in cT stages of the entire HPV+ cohort

A HPV+ and cT1-3



B HPV+ and cT4a

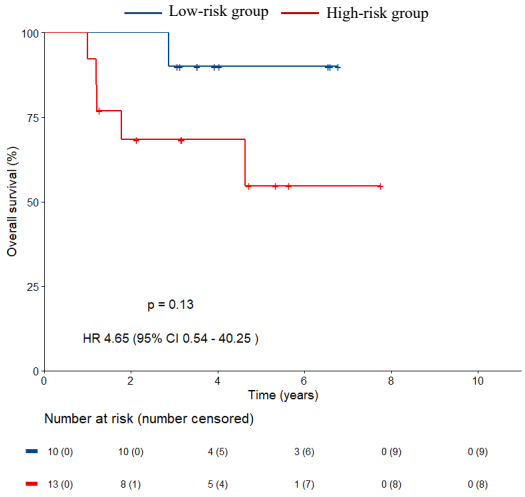


Figure S5: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in cN stages of the entire HPV+ cohort

A HPV+ and cN2

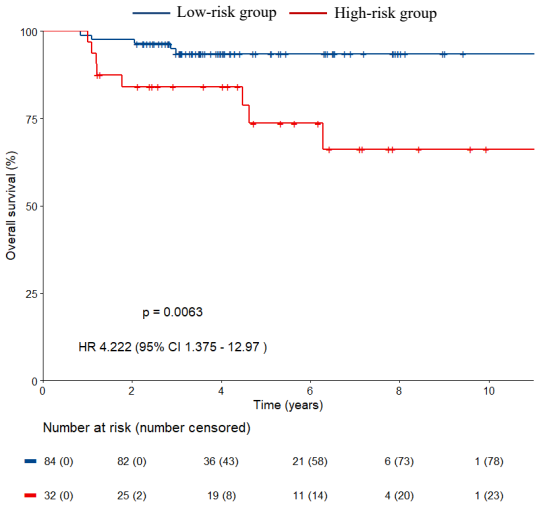


Figure S6: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in cTNM stages of the entire HPV+ cohort

A HPV+ and cTNM I-III

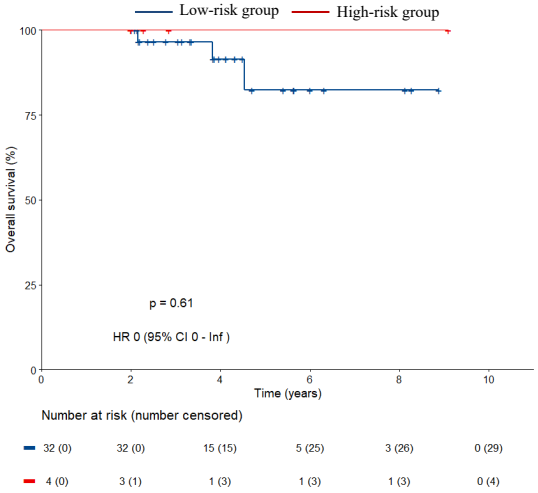
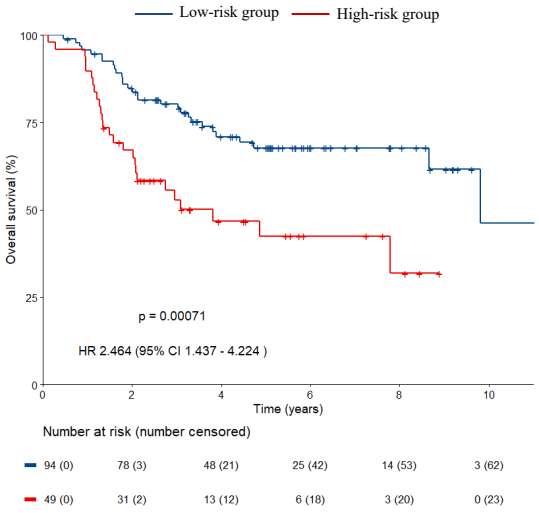
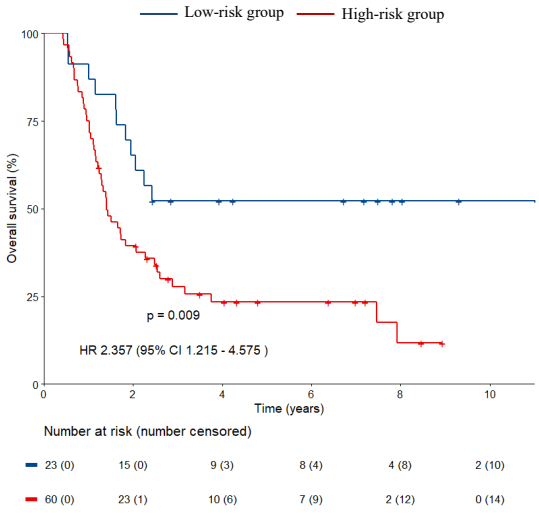


Figure S7: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in cT stages of the entire HPV– cohort

A HPV– and cT1-3



B HPV– and cT4a



C HPV– and cT4b

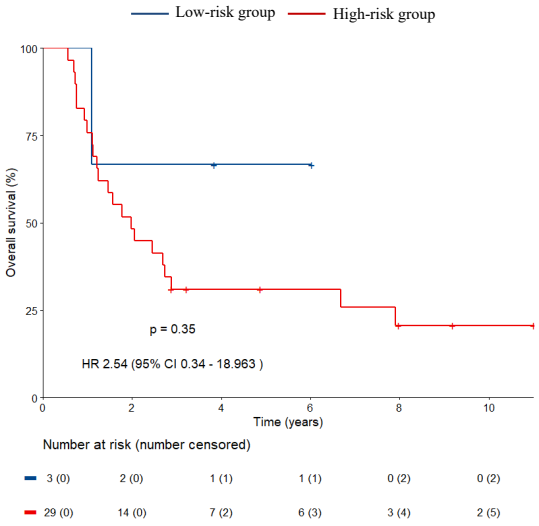


Figure S8: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in cN stages of the entire HPV– cohort

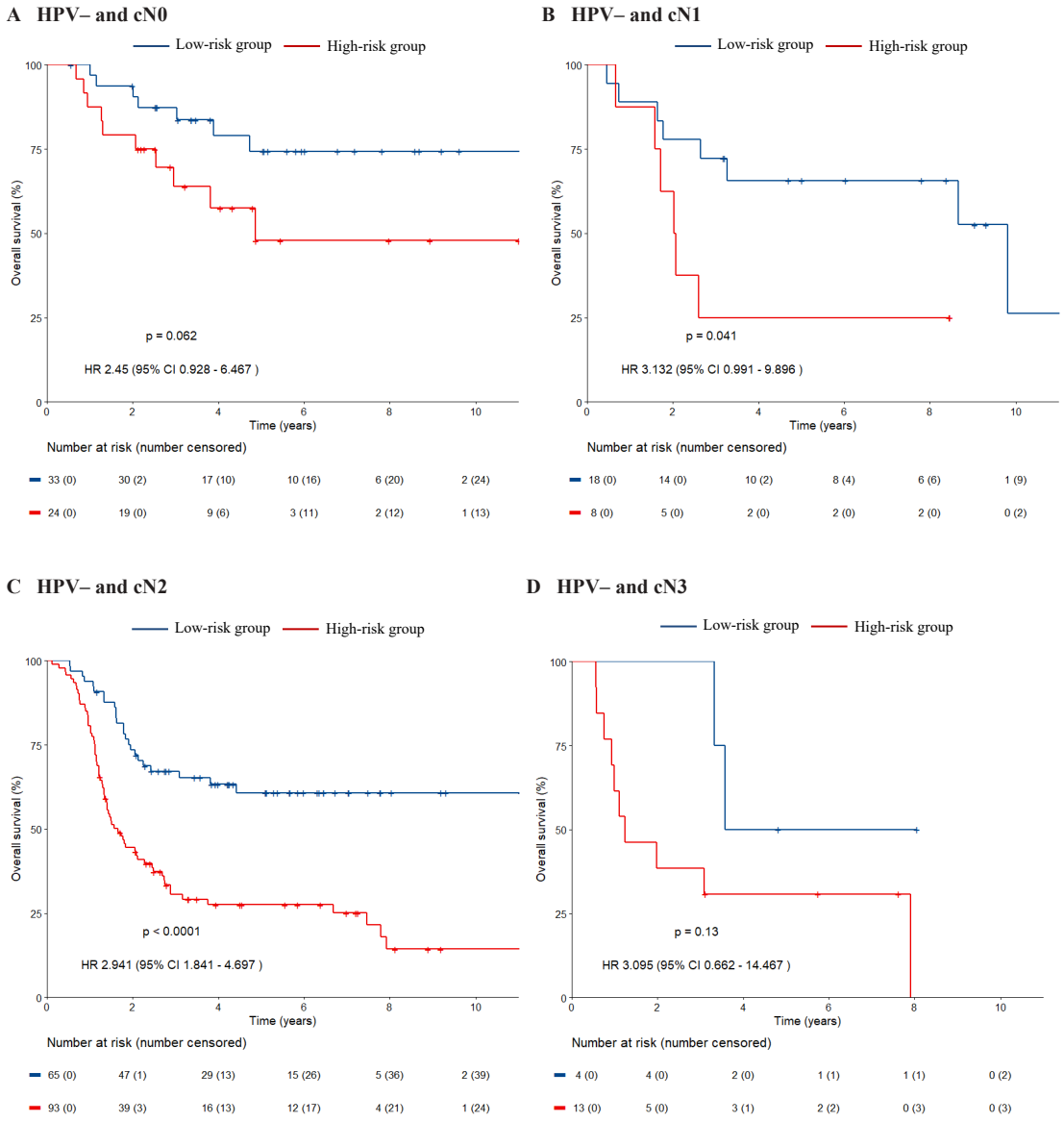


Figure S9: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in cTNM stages of the entire HPV– cohort

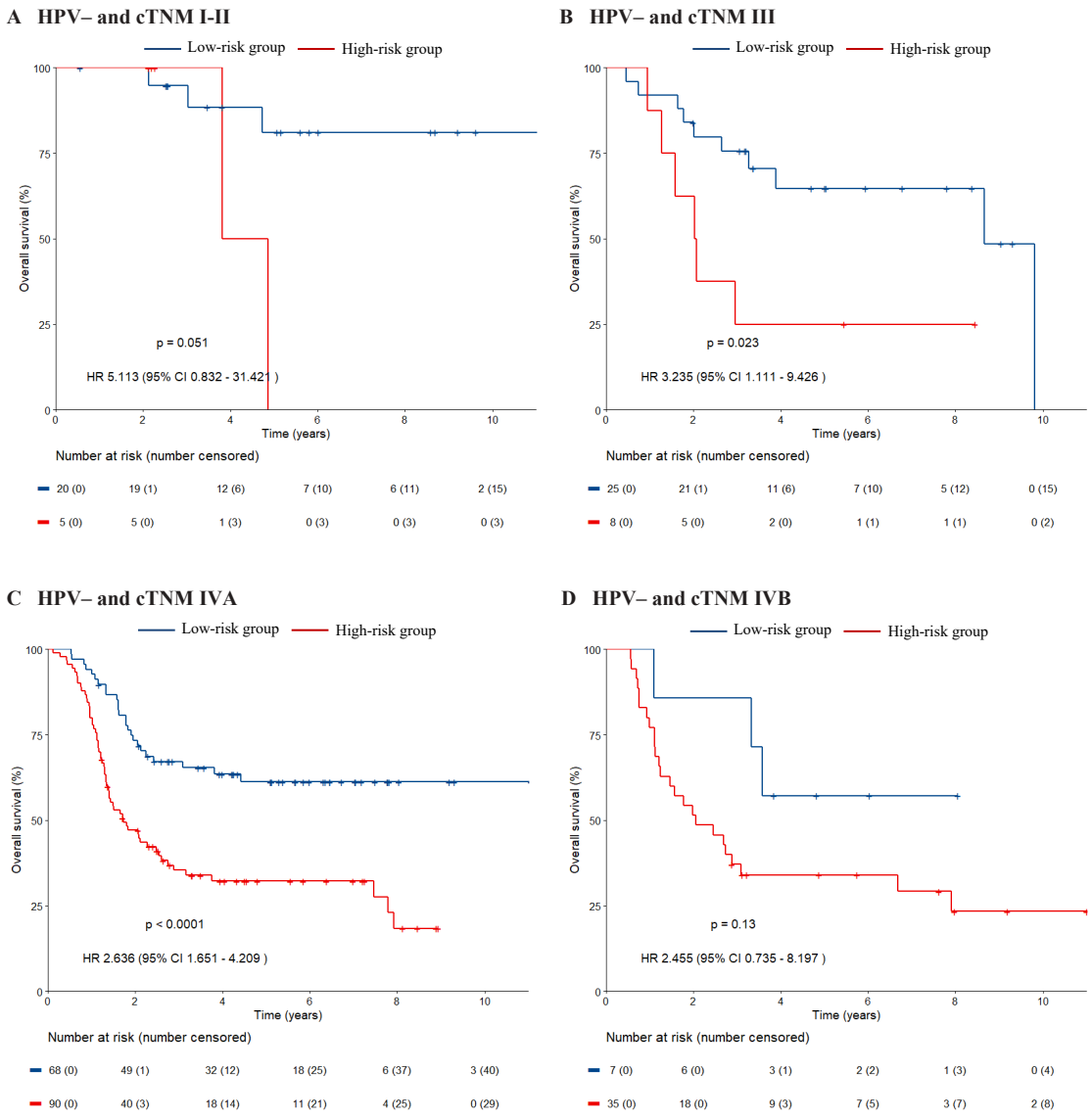
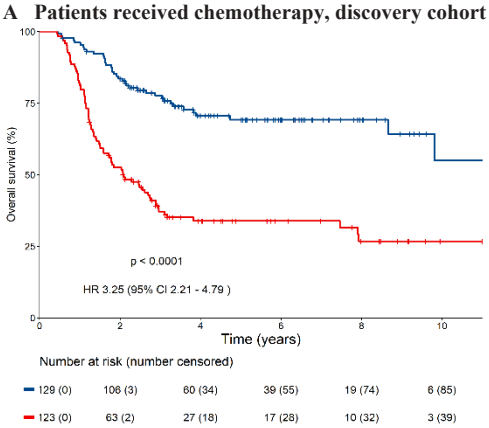
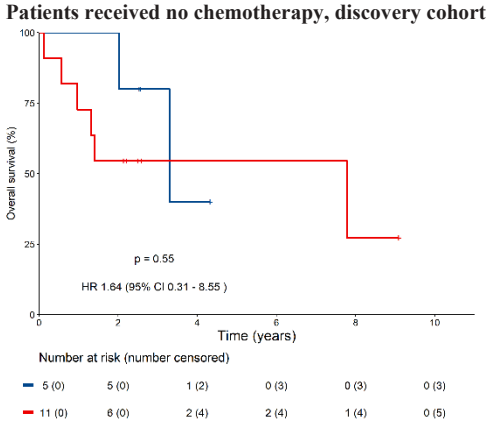


Figure S10: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category, evaluated in receiving chemotherapy or no chemotherapy subgroups in the discover, TCIA test, and the entire cohort with known HPV status

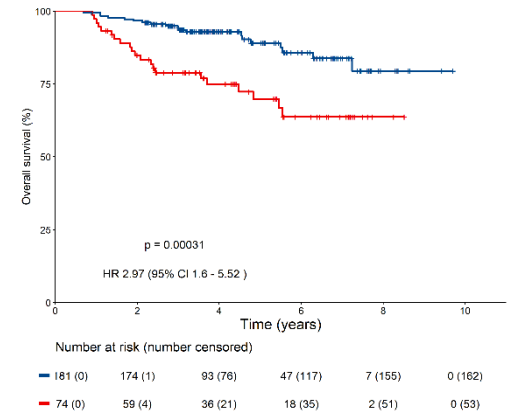
A Patients received chemotherapy, discovery cohort



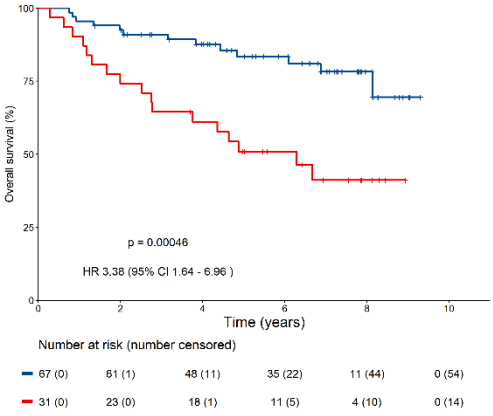
B Patients received no chemotherapy, discovery cohort



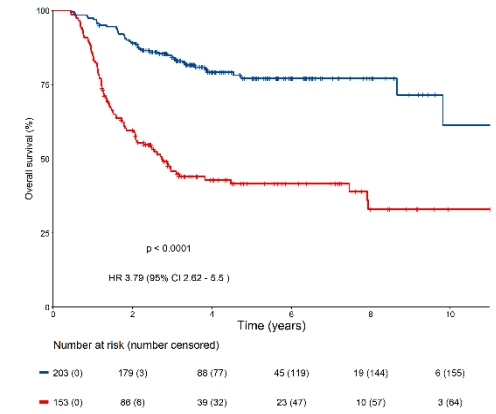
C Patients received chemotherapy, TCIA test cohort



D Patients received no chemotherapy, TCIA test cohort



E Patients received chemotherapy, entire cohort



F Patients received no chemotherapy, entire cohort

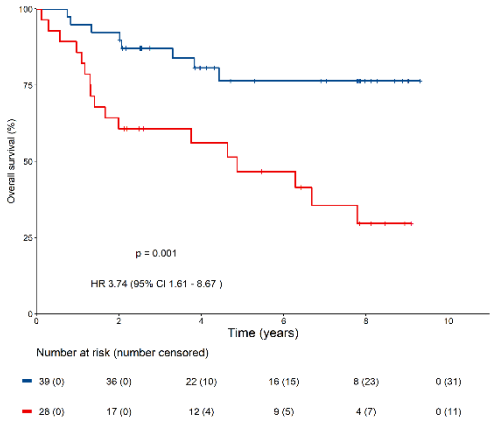


Figure S11: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category with three groups defined by tertiles of the risk scores in the discovery cohort

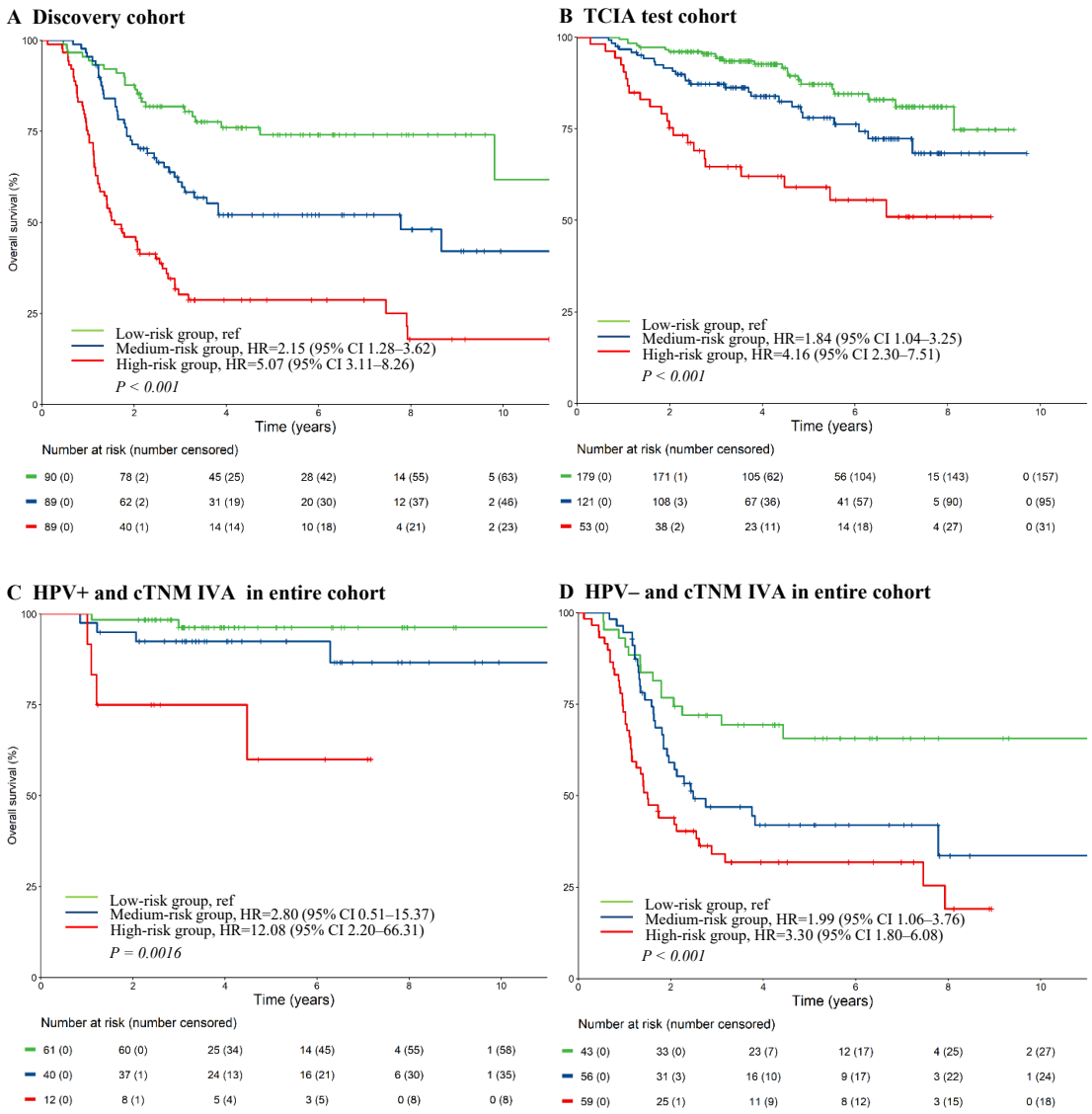


Figure S12: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category with four groups defined by quartiles of the risk scores in the discovery cohort

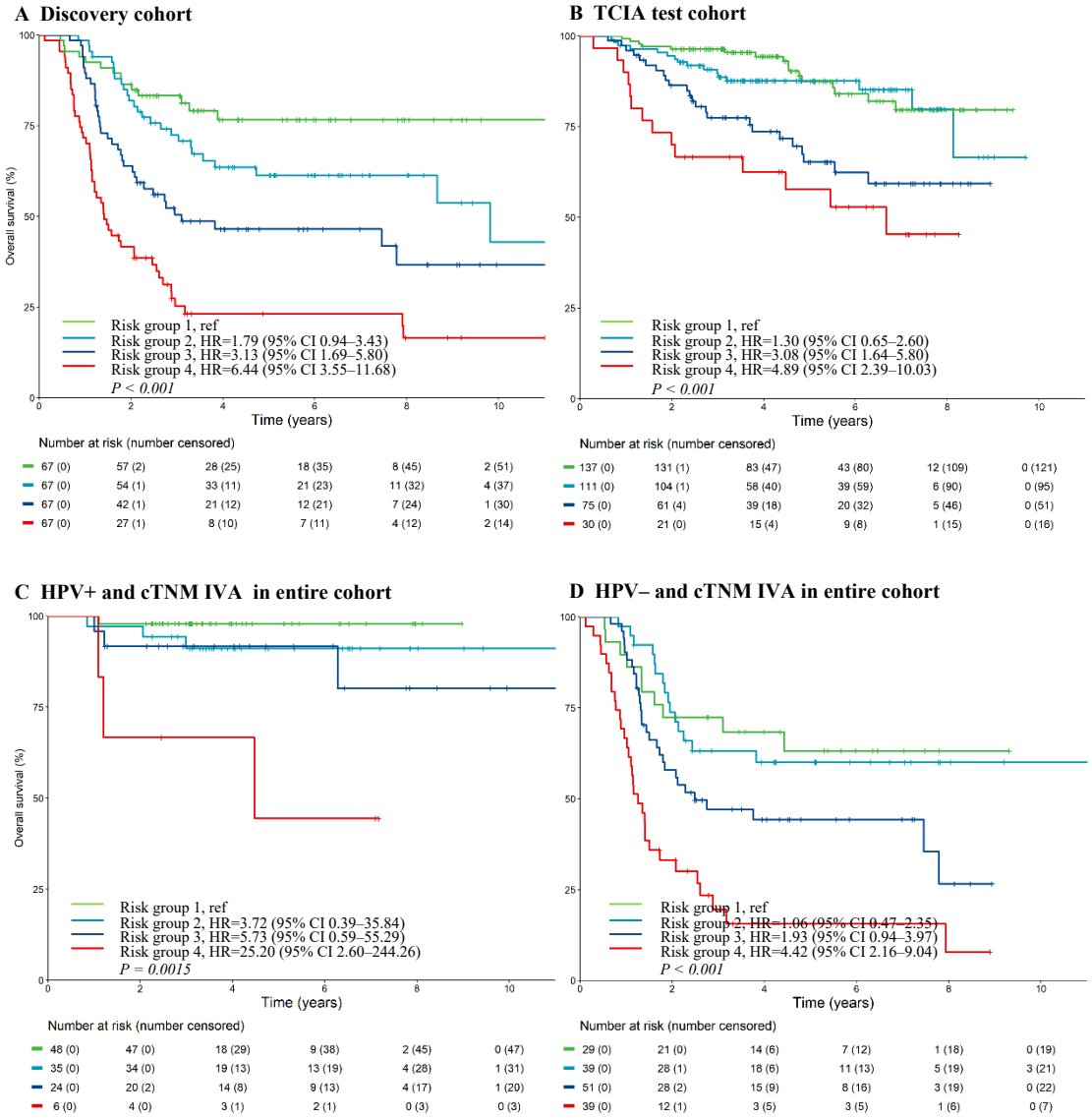


Figure S13: Kaplan-Meier estimates of overall survival by the DeepPET-OPSCC risk category with five groups defined by quintiles of the risk scores in the discovery cohort

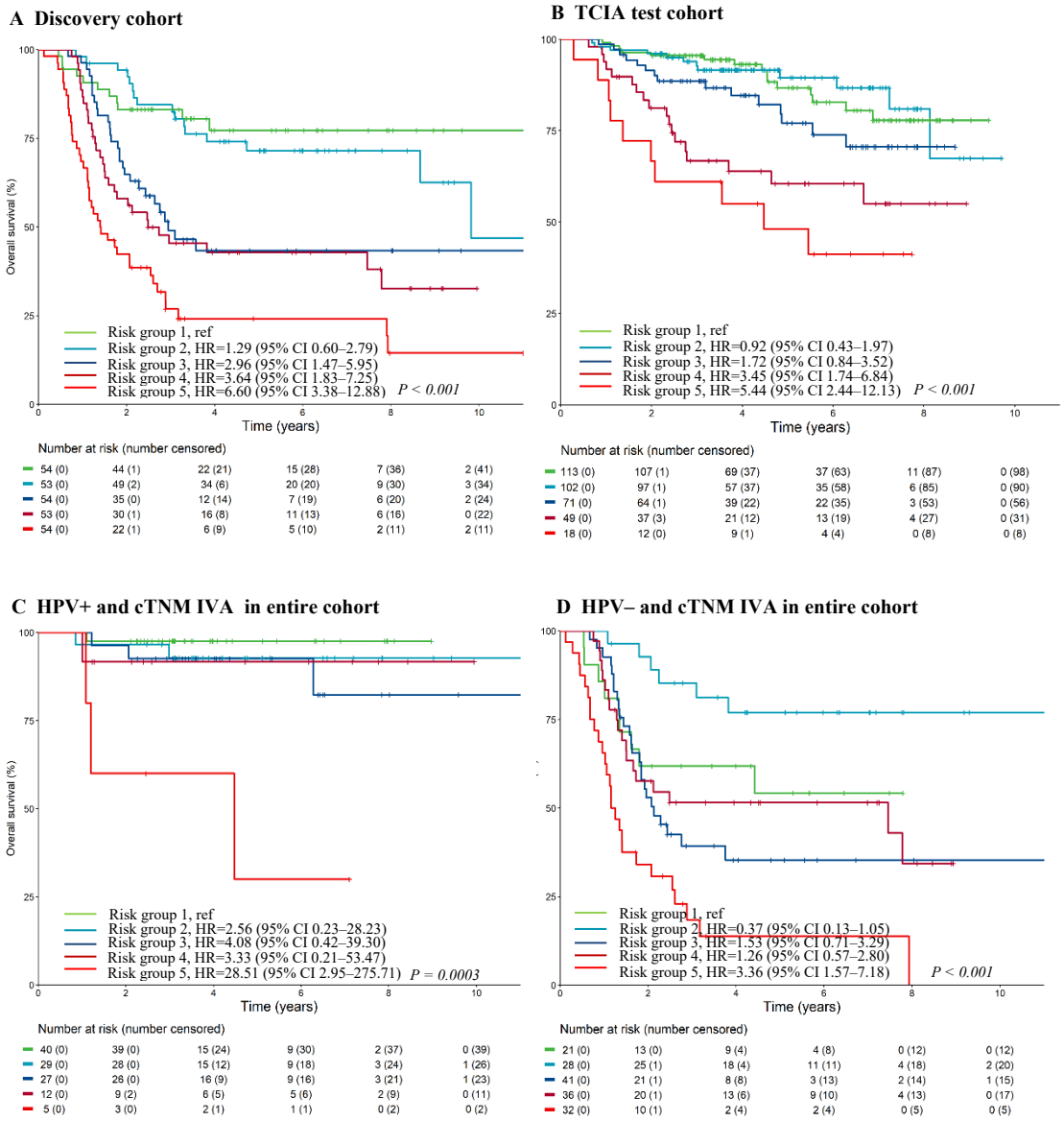
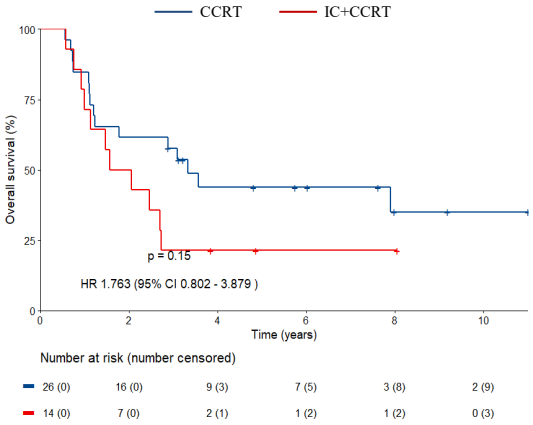


Figure S14: Relationship between the DeepPET-OPSCC risk category and the usage of induction chemotherapy (IC) before chemoradiotherapy (CCRT) in patients with HPV– and stage IVB oropharyngeal cancer in the entire cohort with known HPV status

A Patients with HPV– and stage IVB disease



B Patients with HPV– and stage IVB disease and DeepPET-OPSCC high risk group

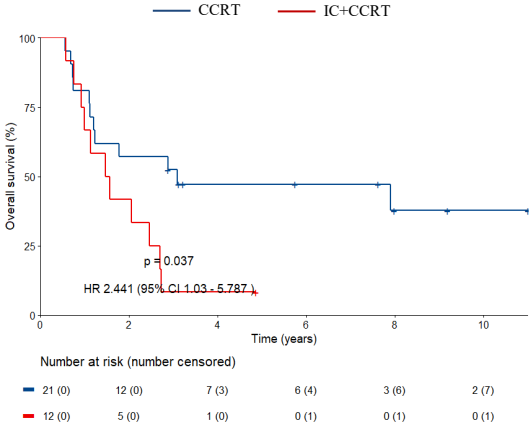
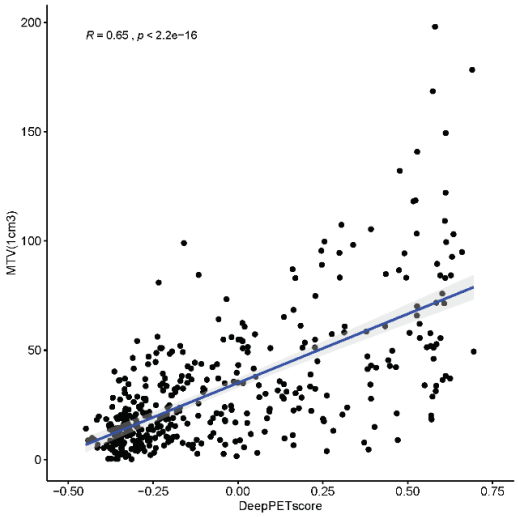


Figure S15: Scatter plots of the relationship between SUVmax/MTV and DeepPET-OPSCC scores in the entire cohort with known HPV status (n=419).

A Scatter plots of MTV and DeepPET-OPSCC scores



B Scatter plots of SUV_{max} and DeepPET-OPSCC scores

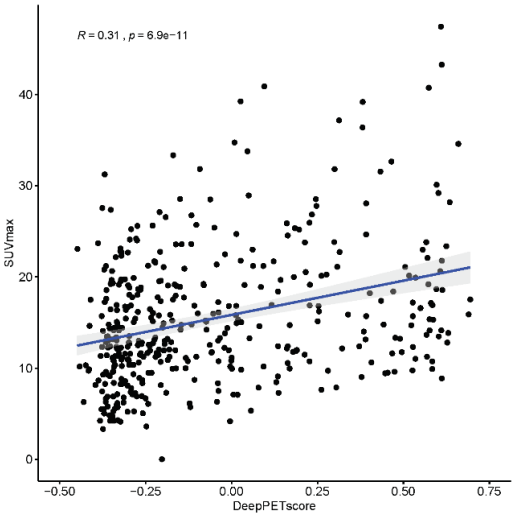


Figure S16: Time-dependent receiver operating characteristic (ROC) curves and AUCs of the DeepPET-OPSCC score, clinical factors, and an integrated nomogram that combines the DeepPET-OPSCC score and clinical factors (age, gender, with or without HPV status, cT, cN, and cTNM stages), to predict overall survival at 2 years evaluated on the discovery (n=268) and TCIA test cohorts (n=151 and n=348 with HPV and without HPV status, respectively)

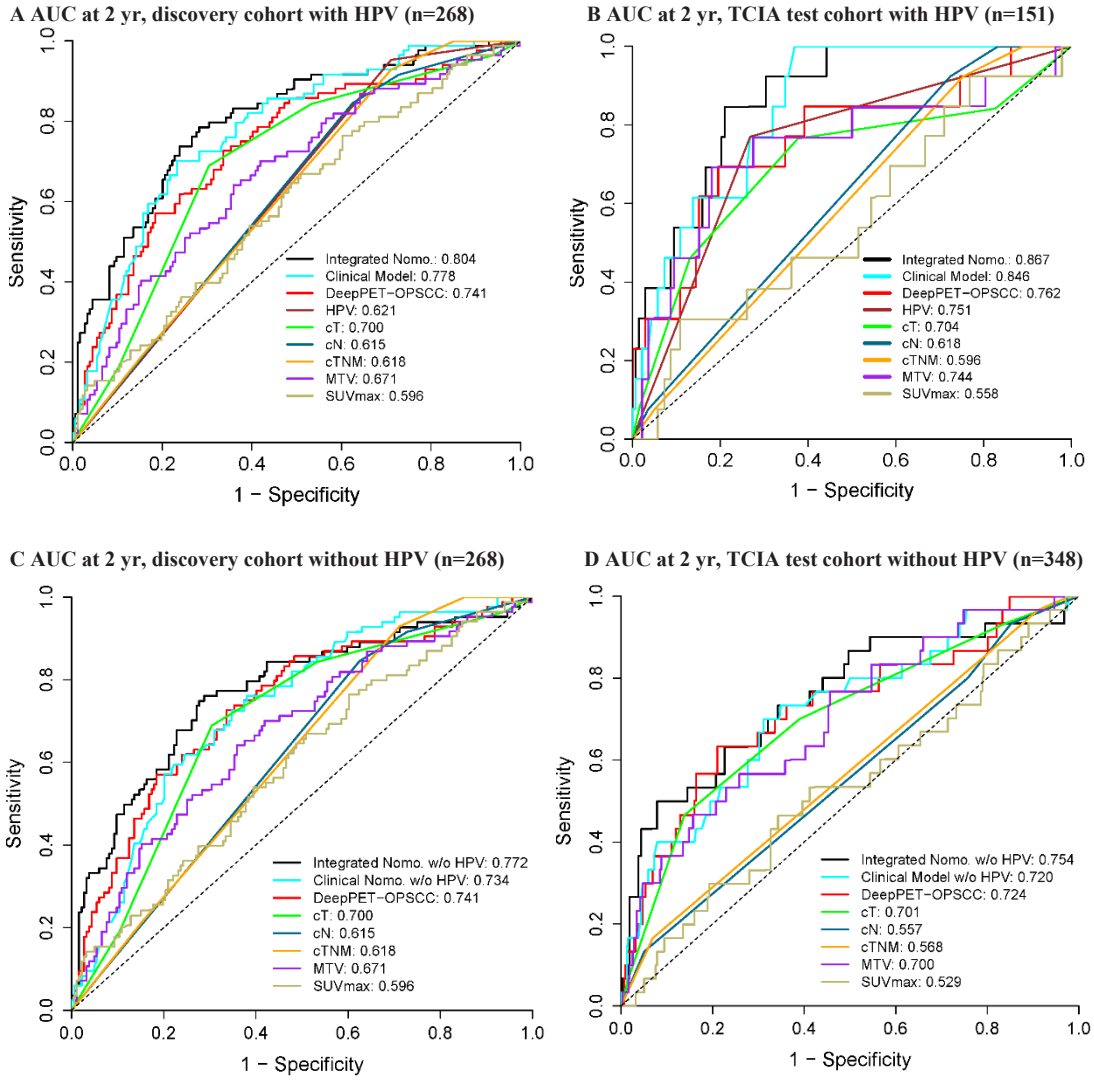


Figure S17: Time-dependent receiver operating characteristic (ROC) curves and AUCs of the DeepPET-OPSCC score, clinical factors, and an integrated nomogram that combines the DeepPET-OPSCC score and clinical factors (age, gender, without HPV status, cT, cN, and cTNM stages), to predict overall survival at 5 years evaluated on the discovery cohort (n=268)

A AUC at 5 years, discovery cohort without HPV

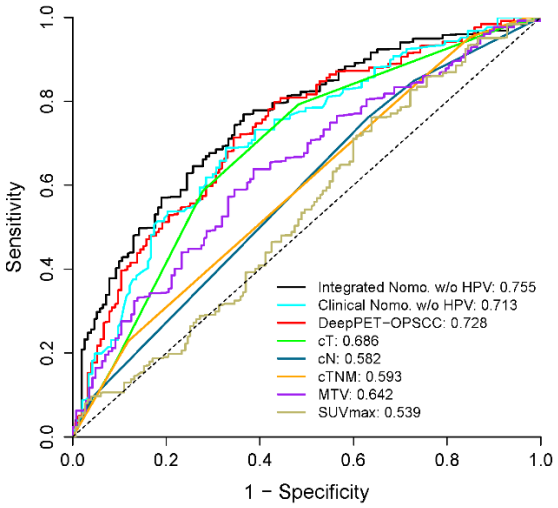


Figure S18: Kaplan-Meier analysis of the DeepPET-OPSCC risk category for the patients in the clinical deployment test cohort (n=31)

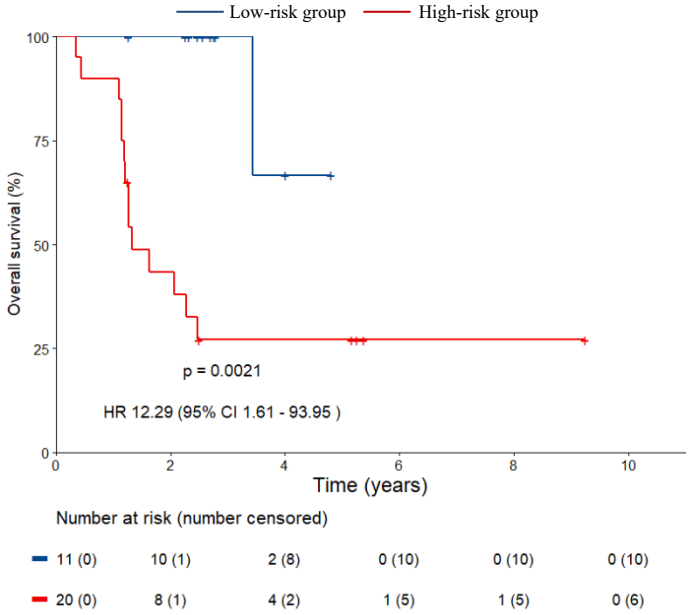
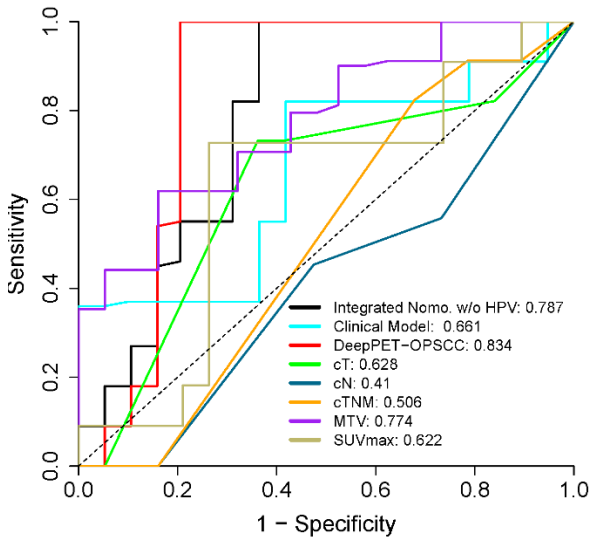


Figure S19: Time-dependent receiver operating characteristic (ROC) curves and AUCs of the DeepPET-OPSCC score, clinical Model, and an integrated nomogram that combines the DeepPET-OPSCC score and clinical factors (age, gender, cT, cN, and cTNM stages), to predict overall survival at 2 years evaluated on the clinical deployment test cohort



The follow-up times of this cohort is 2.3 (1.3–2.8) years and thus the endpoint of 5-year survival was not suitable for this cohort.