Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

Deep Learning for Fully Automated Prediction of Overall Survival in Patients with Oropharyngeal Cancer Using FDG-PET Imaging

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[14]

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 (FDG)-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% confidence interval (CI), 1.31–3.28 and HR = 2.39; 95% CI, 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 (95% CI, 0.621–0.757) in the TCIA test cohort, and 0.787 (95% CI, 0.675–0.899) in the clinical deployment test cohort; the average time taken was 2 minutes 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) vs. 0.749 (95% CI, 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 papillomavirus (HPV) infection (1). However, there are significant differences in 5-year overall survival (OS) rates between HPV-related (HPV+) and tobacco- and alcohol-related (HPV–) cases (75%–80% vs. 45%–50%, respectively; ref. 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).

[1]Department of Nuclear Medicine, Chang Gung Memorial Hospital, Keelung, and Chang Gung University, Taoyuan City, Taiwan, ROC. [2]PAII Inc., Bethesda, Maryland. [3]Department of Radiotherapy, The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, China. [4]Departments of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee. [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]Department of Biostatistics, Harvard University T.H. Chan School of Public Health, Boston, Massachusetts. [8]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee. [9]Department of Otorhinolaryngology, Head and Neck Surgery, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC. [10]Department of Medical Oncology, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC. [11]Department of Radiation Oncology, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC. [12]Department of Pathology, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC. [13]Ping An Technology Co., Ltd., Shenzhen, China. [14]Department of Medicine and Molecular Imaging Center, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan City, Taiwan, ROC.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

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

Corresponding Authors: Ling Zhang, PAII Inc., 6720b Rockledge Dr, Bethesda, MD 20817. Phone: 319-512-6453; E-mail: zhangling300@paii-labs.com; and Tzu-Chen Yen, Department of Nuclear Medicine, Chang Gung Memorial Hospital at Linkou, 5 Fu-Shin Street, Kueishan, Taoyuan 333, Taiwan, ROC. Phone: 886-3-328-1200 x 2673; E-mail: yentc1110@gmail.com

Clin Cancer Res 2021;XX:XX–XX

doi: 10.1158/1078-0432.CCR-20-4935

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

Translational Relevance

Although 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 [^{18}F]fluorodeoxyglucose (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.

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, noninvasive, independent of sampling bias, and not limited to the portion of tested tissue (17). Although 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). Although 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 methodologic barriers, the reliance on manually selected two-dimensional (2D) 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: (i) nonmetastatic (M0) OPSCC and absence of other concomitant malignancies, (ii) availability of baseline pretreatment PET images covering the head and neck region, (iii) treatment with curative intent, and (iv) 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 test cohorts. The need for informed consent was waived. This study was conducted in accordance with the Declaration of Helsinki.

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). 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 [interquartile range (IQR) 3–14] days from the pathologic diagnosis. HPV status was ascertained using p16 IHC. 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 IHC. 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 IHC), 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

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

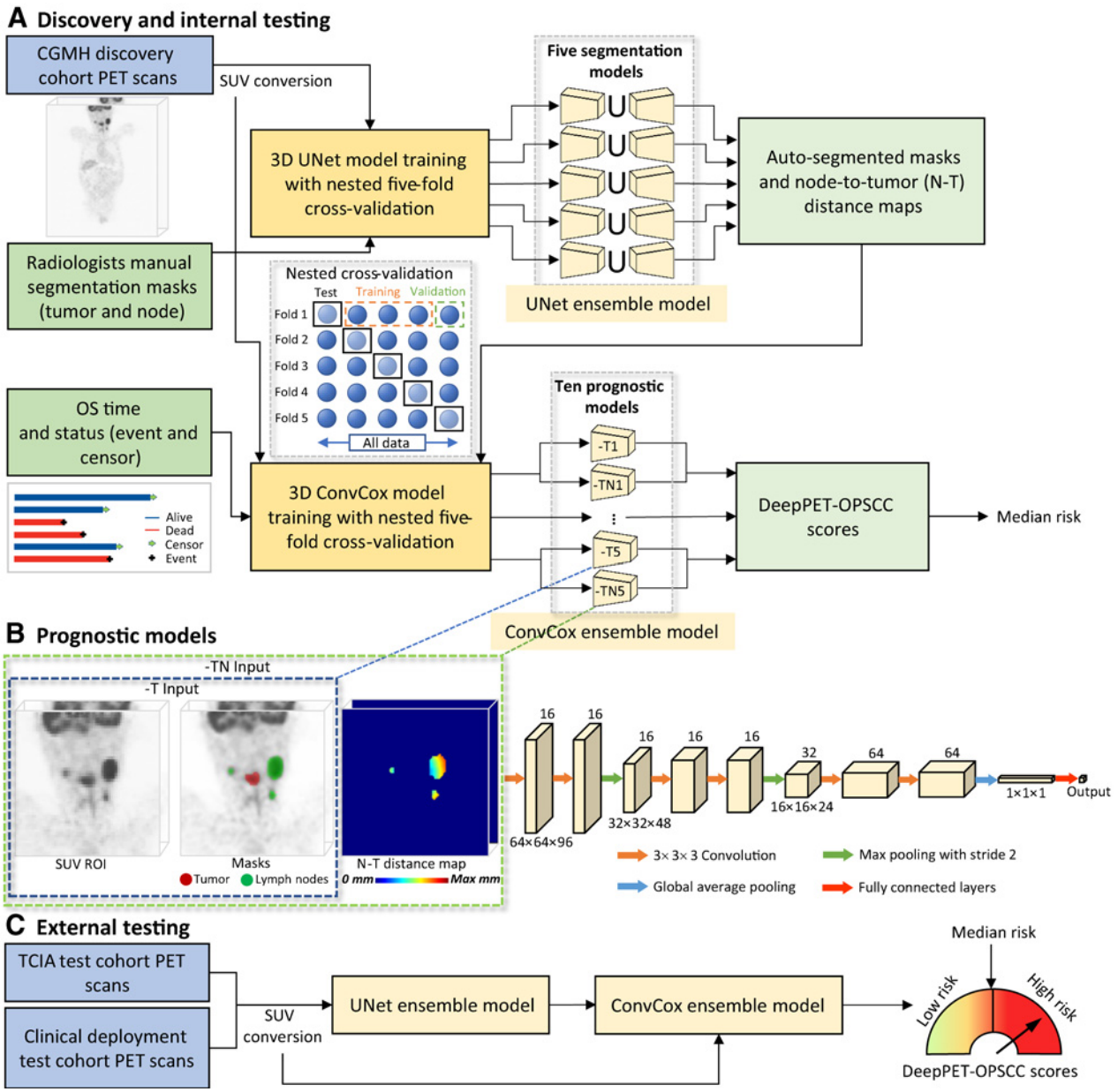

Figure 1.

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, 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 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: (i) a DeepPET-OPSCC-T model with two input channels (SUV and tumor mask), and (ii) 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,** Architecture, input, and output of the 3D ConvCox network in the DeepPET-OPSCC-T/-TN prognostic models. **C,** 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. -T1, DeepPET-OPSCC-T model 1; -TN1, DeepPET-OPSCC-TN model 1.

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 to avoid the overoptimistic issue inherent to conventional cross-validation, as individual DeepPET-OPSCC scores in

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

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 semiautomatically by an experienced nuclear radiologist (N.-M. Cheng), 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; **Fig. 1B**] using OS time and patient status (alive vs. 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; ref. 32). Nonlinear associations between 3D images and time-dependent censored OS were directly modeled. The architecture (**Fig. 1B**) 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: (i) DeepPET-OPSCC-T with two input channels (concatenating SUV and tumor mask), and (ii) DeepPET-OPSCC-TN with three input channels (concatenating SUV, tumor mask, and N-T distance map). Given that there is more variability in the appearance image (N-T distance map) as compared with that in the binary image (tumor mask), the deep learning model may not capture adequate information in the tumor mask. Therefore, 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; ref. 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.

To determine the DeepPET-OPSCC risk category (i.e., dichotomized into high vs. 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 (**Fig. 1C**), 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/deepmed/DeepPET-OPSCC-Example). All of the data in the TCIA test cohort can be accessed at TCIA (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 (Supplementary Tables S1 and S2; ref. 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 AUCs for OS at 2 and 5 years. The overall c-index and AUC in the discovery cohort were calculated by concatenating all normalized scores from the five test sets. To assess the improvements in the c-indexes between the compared models, the Student t test for dependent samples was used (36). A similar approach has been implemented in previous studies (37, 38). The 95% confidence intervals (CI) for AUC were constructed from 1,000 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

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

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 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%)

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%)
OS 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 (IQR ranges) or counts (percentages) unless otherwise specified.

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

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 and cT, cN, and cTNM stages are summarized in Supplementary 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 Supplementary Fig. S1. The median DeepPET-OPSCC score (–0.12) in the test sets of the discovery cohort was used as the cutoff to obtain the DeepPET-OPSCC risk category (dichotomized into high risk vs. low risk), which was a strong predictor of OS in all three study cohorts (**Figs. 2A** and **2B**; Supplementary 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**; Supplementary 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 (Supplementary Tables S6 and S7). The continuous DeepPET-OPSCC score was also retained as a strong predictor in the multivariable model (Supplementary Table S8). Validation with additional clinical variables (e.g., smoking) as well as pathologic (e.g., tumor grade) and IHC-based (e.g., Cyclin D1) markers in the discovery cohort is provided in Supplementary 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 (**Table 3**, Supplementary Tables S4–S7, S12–S13). For example, DeepPET-OPSCC yielded a significantly higher ($P = 0.012$) c-index than the -TN model in the discovery cohort, with a borderline significantly higher ($P = 0.10$) c-index than the -T model in the TCIA test cohort (Supplementary Table S13). 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 ($P = 0.21$; Supplementary Table S13; Supplementary Protocol, Section 3.5).

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 status and cT, cN, and cTNM stages, or the use of chemotherapy (Supplementary Figs. S2–S9). 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 vs. low risk, HR = 4.20; 95% CI, 1.18–14.92; $P = 0.016$; **Fig. 2C**; HR = 2.64; 95% CI, 1.65–4.21; $P < 0.001$; **Fig. 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; $P = 0.037$; Supplementary Table S14; Supplementary Fig. S13).

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 (Supplementary Table S15) and TCIA test cohorts (Supplementary Table S16). The DeepPET-OPSCC score also showed significant correlations with SUV_{max} ($R = 0.31$) and MTV ($R = 0.65$; Supplementary Fig. S14). 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 test 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 5-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% CI, 0.649–0.842), clinical model plus MTV: 0.754 (95% CI, 0.659–0.843), HPV: 0.624 (95% CI, 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; $P < 0.05$; **Fig. 2E**; Supplementary Table S17]. A full description—including c-indices and 2-year AUCs—is provided in Tables S18 and S19 and Fig. S15 and S16 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 minutes 6 seconds per PET exam on an NVIDIA Titan RTX-6000 GPU. The fully-automated tool significantly predicted OS ($P = 0.002$; Supplementary Fig. S17) 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 (Supplementary Fig. S18).

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

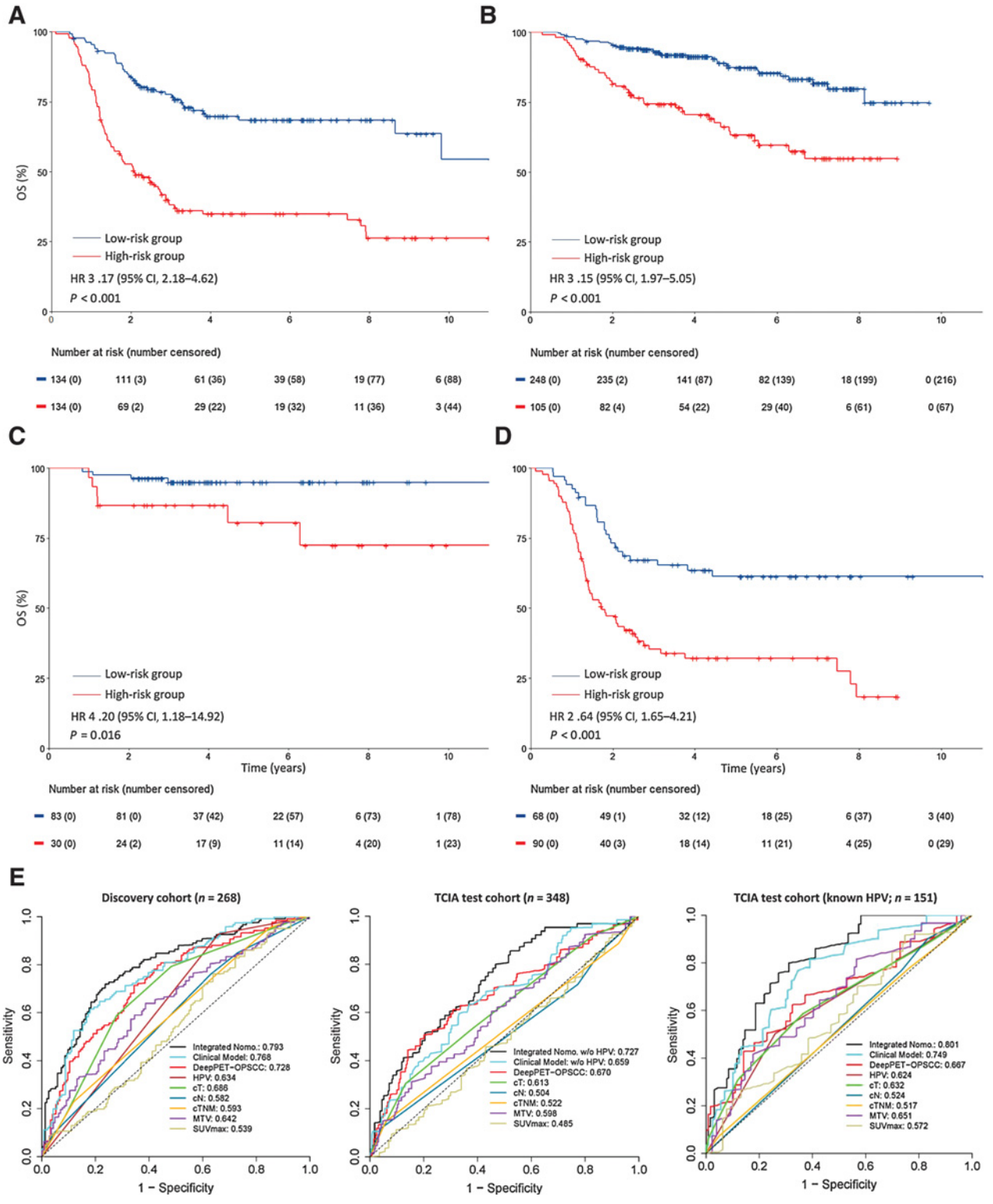

Figure 2.

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 S10-S12 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 **A** to **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).

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

Table 2. Multivariable Cox regression analysis of OS 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

[a]Cutoff threshold was the median value in the discovery cohort.

Visualization

Our tool allowed obtaining a renormalized heatmap that can depict risk at both voxel and patient levels through a hot-cold color code (**Fig. 3**; Supplementary Figs. S19A–S19D). We found that the DeepPET-OPSCC-T model focused mostly on the tumor's interior, whereas the DeepPET-OPSCC-TN model tended to fixate on the interface between the tumor and lymph nodes. This illustrative example also shows that the -T and -TN models can complement each other.

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 minutes) 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, and 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. Although 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 3D tumor characteristics from 2D slices. Moreover, fully automated prediction

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

Table 3. c-index, HR, and AUC at 5 years, all with 95% CIs, of different deep learning and radiomics approaches evaluated on the discovery and TCIA test cohorts.

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

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. Although 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 2 and 5 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 5-year OS). Moreover, DeepPET-OPSCC followed the path of the eighth AJCC staging system, which downstaged stage IV to stages I to 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.

When DeepPET-OPSCC was included in multivariable analyses, we unexpectedly found that high SUV_{max} (HR = 0.60; $P = 0.010$) in the discovery cohort and high MTV (HR = 0.40; $P = 0.014$) in the TCIA test cohort had protective effects. A potential explanation may be related to the presence of necrosis or abscess tumors, which was known to portend poor outcomes in head and neck malignancies while being associated with low SUV_{max} and MTV values (43). Alternatively, this result may stem from the presence of collinearity in multivariable analysis. In this regard, DeepPET-OPSCC was significantly correlated ($P < 0.001$) with both SUV_{max} and MTV. Nevertheless, this effect was not observed in the entire cohort with known HPV status. Similar counterintuitive results can be found in multivariable analyses of published clinical studies (38, 41), in which, for example, cT3 or cT4 versus cT1 yielded an HR of 0.4 (38).

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. Although we hypothesize that personalized radiation plans with higher tumoricidal doses could potentially target the identified high-risk regions (12, 44), this requires further investigation.

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

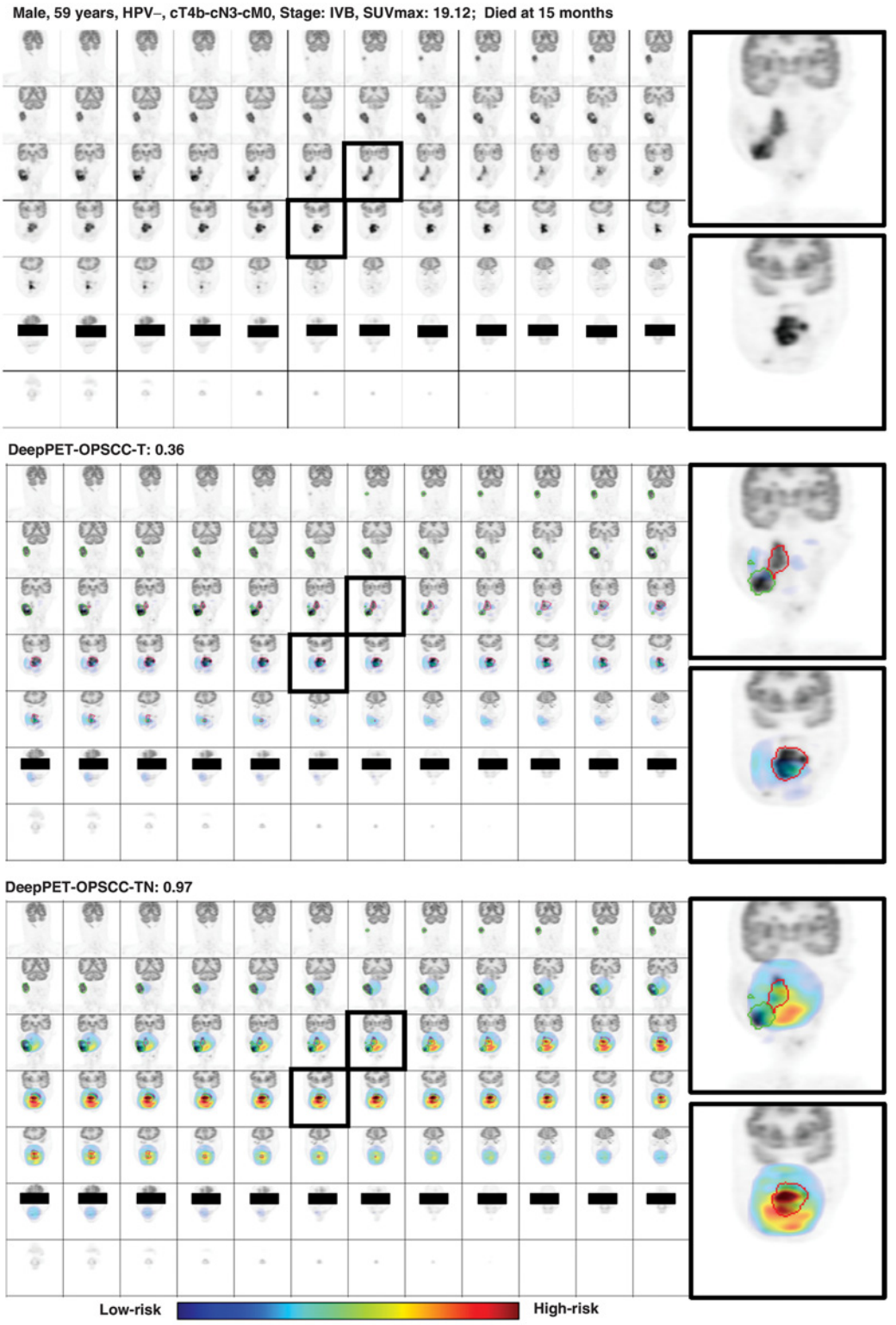

Figure 3.
Examples of 3D PET images (consecutive image slices), corresponding activation maps (heatmaps), and two enlarged images with heatmaps for better visual observation. In this illustrative example, auto-segmented tumors and lymph node boundaries are indicated by red and green curves, respectively. The PET images are anonymized by blocking the eye region with black boxes.

Published OnlineFirst May 4, 2021; DOI: 10.1158/1078-0432.CCR-20-4935

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 (**Fig. 2C**; Supplementary Figs. S4 and S10–S12) 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; ref. 45)]. Conversely, certain subgroups of HPV– patients with cT1–3, T4a, N1, N2, AJCC III, or IVA stages (**Fig. 2D**; Supplementary Figs. S6–S8 and S10–S12) 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; ref. 10)]. Interestingly, CCRT was associated with a better OS compared with 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.

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

N.-M. Cheng: Conceptualization, resources, data curation, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **J. Yao:** Conceptualization, resources, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **J. Cai:** Software, investigation, visualization, methodology. **X. Ye:** Resources, data curation, software, validation. **S. Zhao:** Formal analysis, investigation, methodology. **K. Zhao:** Resources, data curation. **W. Zhou:** Resources, data curation. **I. Nogues:** Writing–review and editing. **Y. Huo:** Investigation. **C.-T. Liao:** Resources, data curation. **H.-M. Wang:** Resources, data curation, formal analysis, writing–review and editing. **C.-Y. Lin:** Resources, data curation. **L.-Y. Lee:** Resources, data curation. **J. Xiao:** Formal analysis, supervision, funding acquisition. **L. Lu:** Resources, formal analysis, supervision, funding acquisition, project administration, writing–review and editing. **L. Zhang:** Conceptualization, resources, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **T.-C. Yen:** Conceptualization, resources, formal analysis, supervision, funding acquisition, investigation, methodology, project administration, writing–review and editing.

Acknowledgments

The authors are indebted to TCIA for data availability. This study was financially supported by grants from the Ministry of Science and Technology of ROC (MOST 106–2314-B-182A-025-MY3) and the Chang Gung Memorial Hospital Research Fund (CORPG3J0342).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 22, 2020; revised March 9, 2021; accepted April 30, 2021; published first May 4, 2021.

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;363:24–35.
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;17:440–51.
3. Chow LQM. Head and neck cancer. N Engl J Med 2020;382:60–72.
4. Cramer JD, Burtness B, Le QT, Ferris RL. The changing therapeutic landscape of head and neck cancer. Nat Rev Clin Oncol 2019;16:669–83.
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 2019;393:40–50.
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 2019;393:51–60.
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. Lancet Oncol 2017;18:803–11.
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. J Clin Oncol 2017;35:490–7.
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. J Natl Cancer Inst 2021 Jan 12 [Epub ahead of print].
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. Lancet Oncol 2020;21:1173–87.
11. Budach V, Tinhofer I. Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer: a systematic review. Lancet Oncol 2019;20:e313–e26.

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. Lancet Oncol 2017;18:e266–e73.
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. J Natl Cancer Inst 2020;112:855–8.
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;123:1778–90.
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;26:22–30.
16. Rosenberg AJ, Vokes EE. Optimizing treatment de-escalation in head and neck cancer: current and future perspectives. Oncologist 2021;26:40–8.
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;69:127–57.
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;12:1778.
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 Oncol 2015;54:1423–9.
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;61:327–36.
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. J Clin Oncol 2020;38:1304–11.
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. Lancet Oncol 2019;20:728–40.
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. Lancet Oncol 2019;20:1645–54.
25. Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep 2019;9:2764.
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:e1002711.
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 Jan 6 [Epub ahead of print].
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;25:4271–9.
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. Nat Methods 2021;18:203–11.
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;122:534–45.
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;247:2543–6.
34. Sauerbrei W, Taube SE, McShane LM, Cavenagh MM, Altman DG. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): an abridged explanation and elaboration. J Natl Cancer Inst 2018;110:803–11.
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;66:370–4.
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 2008;24:2200–8.
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. J Natl Cancer Inst 2021;113:606–15.
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;296:216–24.
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;12:77.
40. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008;26:1364–70.
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 2020;395:350–60.
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;48:441–6.
43. Yamazaki H, Ogita M, Himei K, Nakamura S, Suzuki G, Kotsuma T, et al. Effect of intratumoral abscess/necrosis on the outcome for head and neck cancer patients treated by hypofractionated stereotactic re-irradiation using CyberKnife®. Mol Clin Oncol 2017;7:336–40.
44. Horsman MR, Mortensen LS, Petersen JB, Busk M, Overgaard J. Imaging hypoxia to improve radiotherapy outcome. Nat Rev Clin Oncol 2012;9:674–87.
45. 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. J Clin Oncol 2019;37:2661–9.

AACR American Association for Cancer Research

Clinical Cancer Research

Deep Learning for Fully Automated Prediction of Overall Survival in Patients with Oropharyngeal Cancer Using FDG-PET Imaging

Nai-Ming Cheng, Jiawen Yao, Jinzheng Cai, et al.

Clin Cancer Res Published OnlineFirst May 4, 2021.