Aus der

Klinik und Poliklinik für Hals-Nasen-Ohrenheilkunde

der Ludwig-Maximilians-Universität München

PET/CT Radiomics and Machine Learning for Non-Invasive Molecular and Prognostic Characterization of Oropharyngeal Squamous Cell Carcinomas

Dissertation

zum Erwerb des Doktorgrades der Medizin

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität München

vorgelegt von

Stefan Philipp Markus Haider

aus Wien

2023

Mit Genehmigung der Medizinischen Fakultät

der Ludwig-Maximilians-Universität zu München

Erster Gutachter:	Prof. Dr. med. Philipp Baumeister
Zweiter Gutachter:	Prof. Dr. Seyedmehdi Payabvash, MD
Dritter Gutachter:	Priv. Doz. Dr. Christian Thieke
ggf. weitere Gutachter:	Prof. Dr. Michael Ingrisch
Mitbetreuung durch den promovierten Mitarbeiter:	Dr. med. Kariem Sharaf
Dekan:	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 17.04.2023

1. Contents

2. Abbreviations and Acronyms

AI	Artificial intelligence
AJCC	American Joint Committee on Cancer
CRAN	The Comprehensive R Archive Network
CT	Computed tomography
DNA	Deoxyribonucleic acid
ENT	Ear, nose, and throat
FDG	^{18}F-2-Fluor-2-deoxy-D-glucose
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papillomavirus
LRP	Locoregional progression
ML	Machine learning
OPSCC	Oropharyngeal squamous cell carcinoma
PCR	Polymerase chain reaction
PET	Positron emission tomography
RNA	Ribonucleic acid
TCIA	The Cancer Imaging Archive
TNM	Tumor-node-metastasis
UICC	Union for International Cancer Control
VOI	Volume of interest

3. List of Publications

3.1 Original Articles

1. Haider SP, Mahajan A, Zeevi T, Baumeister P, Reichel C, Sharaf K, et al. **PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma.** Eur J Nucl Med Mol Imaging. 2020;47(13):2978-91.
2. 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. 2020;12(7).
3. Haider SP, Sharaf K, Zeevi T, Baumeister P, Reichel C, Forghani R, et al. **Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics.** Transl Oncol. 2020;14(1):100906.
4. Haider SP, Qureshi AI, Jain A, Tharmaseelan H, Berson ER, Zeevi T, et al. **Admission computed tomography radiomic signatures outperform hematoma volume in predicting baseline clinical severity and functional outcome in the ATACH-2 trial intracerebral hemorrhage population.** European Journal of Neurology. 2021;28(9):2989-3000.
5. Haider SP, Qureshi AI, Jain A, Tharmaseelan H, Berson ER, Majidi S, et al. **The coronal plane maximum diameter of deep intracerebral hemorrhage predicts functional outcome more accurately than hematoma volume.** International Journal of Stroke. 2021:17474930211050749.
6. Sharaf K, Eggersmann TK, Haider SP, Schwenk-Zieger S, Zhou J, Gires O, et al. **Human Adipose-Derived Stem/Stromal Cells Promote Proliferation and Migration in Head and Neck Cancer Cells.** Cancers. 2021;13(11):2751.
7. Sharaf K, Lechner A, Haider SP, Wiebringhaus R, Walz C, Kranz G, et al. **Discrimination of Cancer Stem Cell Markers ALDH1A1, BCL11B, BMI-1, and CD44 in Different Tissues of HNSCC Patients.** Current Oncology. 2021;28(4):2763-74.
8. Liu X, Maleki F, Muthukrishnan N, Ovens K, Huang SH, Pérez-Lara A, Romero-Sanchez G, Bhatnagar SH, Chatterjee A, Pusztaszeri MP, Spatz A, Batist G, Payabvash S, Haider SP, et al. **Site-Specific Variation in Radiomic Features of Head and Neck Squamous Cell Carcinoma and Its Impact on Machine Learning Models.** Cancers. 2021;13(15):3723.
9. Gross M, Spektor M, Jaffe A, Kücükkaya AS, Iseke S, Haider SP, et al. **Improved performance and consistency of deep learning 3D liver segmentation with heterogenous cancer stages in magnetic resonance imaging.** PLOS ONE. 2021;16(12):e0260630.
10. Petukhova-Greenstein A, Zeevi T, Yang J, Chai N, DiDomenico P, Deng Y, Ciarleglio M, Haider SP, et al. **Biomarkers for Prediction of Outcome After Radiofrequency Ablation of Hepatocellular Carcinoma: Qualitative and Quantitative Assessment of LI-RADS and Radiomic Features.** Journal of Vascular and Interventional Radiology. 2022.

11. Avery EW, Behland J, Mak A, Haider SP, Zeevi T, Sanelli PC, et al. **CT angiographic radiomics signature for risk stratification in anterior large vessel occlusion stroke.** NeuroImage: Clinical. 2022:103034.
12. Lechner A, Haider SP, Paul B, Escrihuela Branz PFF, Felicio-Briegel A, Widmann M, et al. **Misjudgment of skills in clinical examination increases in medical students due to a shift to exclusively online studies during the COVID-19 pandemic.** Accepted for publication in Journal of Personalized Medicine. 2022.

3.2 Review Articles

1. Haider SP, Burtness B, Yarbrough WG, Payabvash S. **Applications of radiomics in precision diagnosis, prognostication and treatment planning of head and neck squamous cell carcinomas.** Cancers of the Head & Neck. 2020;5(1):6.
2. Haider SP, Sharaf K, Baumeister P, Reichel CA. **Künstliche Intelligenz in der Hals-Nasen-Ohrenheilkunde.** HNO. 2021:1-7.

3.3 Conference Abstracts

1. Haider SP, Mahajan A, Zeevi T, Baumeister P, Reichel C, Sharaf K, et al. **Radiomics signature of Human Papillomavirus (HPV) status in oropharyngeal squamous cell carcinoma (OPSCC).** In: ECR 2020 Book of Abstracts. Insights into Imaging. 2020;11(1):34. Presented at the European Congress of Radiology 2020 (Oral Presentation).
2. Haider SP, Mahajan A, Zeevi T, Forghani R, Kann B, Judson BL, et al. **Devising novel imaging biomarkers for Human Papillomavirus (HPV) status in oropharyngeal squamous cell carcinoma (OPSCC): applying radiomics and machine learning algorithms.** Laryngorhinootologie. 2020;99(S 02). Presented at Jahresversammlung der Deutschen Gesellschaft für Hals-Nasen-Ohren-Heilkunde 2020/2021 (Oral Presentation).
3. Haider SP, Mahajan A, Zeevi T, Baumeister P, Reichel C, Sharaf K, et al. **PET/CT radiomics improves survival prognostication in patients with human papillomavirus (HPV)-associated oropharyngeal squamous cell carcinoma (OPSCC) beyond AJCC 8th edition staging.** Presented at the Annual Meeting of the Radiological Society of North America 2020 (Poster).
4. Haider SP, Qureshi AI, Jain A, Tharmaseelan H, Berson ER, Zeevi T, et al. **Admission CT radiomic signatures are stronger predictors of long-term outcome and baseline clinical severity than hematoma volume in patients with intracerebral hemorrhage (ICH).** Presented at the Annual Meeting of the Radiological Society of North America 2021 (Oral Presentation).
5. Haider SP, Sharaf K, Zeevi T, Mahajan A, Forghani R, Judson BL, et al. **PET/CT radiomics potentially improves progression-free survival (PFS) and overall survival (OS) prognostication beyond UICC TNM staging in oropharyngeal squamous cell carcinoma (OPSCC) patients.** Jahresversammlung der Deutschen Gesellschaft für Hals-Nasen-Ohren-Heilkunde 2022 (Accepted for Oral Presentation).

6. Payabvash S, Haider SP, Kann B, Forghani R, Baumeister P, Reichel C, et al. **Prediction of HPV status in oropharyngeal squamous cell carcinoma: applying machine-learning classifiers and radiomics features from primary tumor lesions and metastatic lymph nodes on pretreatment PET/CT.** Presented at the 53rd annual ASHNR meeting 2019.
7. Payabvash S, Haider SP, Kann B, Forghani R, Baumeister P, Reichel C, et al. **Pre-treatment PET/CT radiomics of oropharyngeal squamous cell carcinoma can improve survival prognostication beyond the AJCC 8th staging classification.** Presented at the 53rd annual ASHNR meeting 2019.
8. Kücükkaya AS, Zeevi T, Raju R, Chai N, Haider S, Elbanan M, et al. **Proof-of-concept use of machine learning to predict tumor recurrence of early-stage hepatocellular carcinoma before therapy using baseline magnetic resonance imaging.** Journal of Hepatology. 2020;73:S130-S1.
9. Batra R, Kuecuekkaya A, Zeevi T, Raju R, Chai N, Haider S, et al. **PROOF-OF-CONCEPT USE OF MACHINE LEARNING TO PREDICT TUMOR RECURRENCE OF EARLY-STAGE HEPATOCELLULAR CARCINOMA BEFORE THERAPY USING BASELINE MAGNETIC RESONANCE IMAGING.** Transplantation. 2020;104(S3):S43-S4.
10. Payabvash S, Acosta J, Haider S, Noche R, Kirsch E, Matouk C, et al. **Abstract WMP101: Prediction of Clinical Outcome in Supratentorial Intracerebral Hemorrhage: Application of Baseline Ct Scan Radiomics Feature Extraction and Machine Learning Classifiers.** Stroke. 2020;51(Suppl_1): AWMP101-AWMP.
11. Avery EW, Behland J, Haider SP, Zeevi T, Nguyen CK, Peshwe K, et al. **Abstract P510: Ct Angiographic Radiomics Signature Predicts Functional Outcome Following Endovascular Thrombectomy in Large Vessel Occlusion Stroke.** Stroke. 2021;52(Suppl_1):AP510-AP.
12. Avery EW, Behland J, Mak A, Haider SP, Zeevi T, Sanelli PC, et al. **CT angiographic radiomics signature for risk-stratification in anterior large vessel occlusion (LVO) stroke.** Presented at the Annual Meeting of the Radiological Society of North America 2021.
13. Weber C, Haider SP, Lake E, Scheinost D, Bamford NS, Constable T, Payabvash S. **Age-related Changes Of White Matter (WM) Microstructure In Autism Spectrum Disorder (ASD).** Presented at the Annual Meeting of the Radiological Society of North America 2021.
14. Weber C, Lake E, Scheinost D, Haider SP, Bamford N, Constable T, Payabvash S. **White Matter (WM) Microstructural Correlates of Autism Spectrum Disorder (ASD) across Different Age Spectrums.** Presented at the Annual Meeting of the American Society of Functional Neuroradiology 2021.
15. Acosta J, Haider SP, Rivier C, Leasure AC, Sheth KN, Falcone GJ, Payabvash S. **Pervasive White Matter Microstructure Dysintegrity Related To High Blood Pressure Among Asymptomatic Population.** Stroke. 2022;53(Suppl_1):A9-A.
16. Weber CF, Haider SP, Mukherjee P, Lake E, Scheinost D, Bamford NS, et al. **Edge Density Imaging in Autism Spectrum Disorder: Microstructural Connectome Alterations from Infancy to Adulthood.** Presented at the Annual Meeting of the American Society of Pediatric Neuroradiology 2022.
17. Weber CF, Haider SP, Mukherjee P, Lake E, Scheinost D, Bamford NS, et al. **Age-associated White Matter Microstructure and Connectome Abnormalities in Autism Spectrum Disorder.** Presented at the Annual Meeting of the International Society for Magnetic Resonance in Medicine 2022.

18. Sharaf K, Haider S, Gires O, Canis M, Lechner A, Zhou J, et al. **Humane adipogene Stammzellen fördern die Proliferation und Migration von Kopf-Hals-Karzinomzellen.** Jahresversammlung der Deutschen Gesellschaft für Hals-Nasen-Ohren-Heilkunde 2022 (Accepted for presentation).
19. Lechner A, Haider S, Paul B, Escrihuela Branz P, Huber J, Canis M, et al. **ToSkORL-2: Selbst- / Fremdeinschätzung studentischer HNO-Untersuchungskompetenz nach Online-Kursen während der COVID-Pandemie.** Jahresversammlung der Deutschen Gesellschaft für Hals-Nasen-Ohren-Heilkunde 2022 (Accepted for presentation).

4. Introduction

4.1 Oropharyngeal Squamous Cell Carcinoma

Oropharyngeal squamous cell carcinomas (OPSCC) elicited considerable interest across various research fields in recent years, given the dramatic increase in incidence of human papillomavirus (HPV)-attributable OPSCC in numerous regions around the world, including Europe and North America [1, 2]. Notably, this marked trend, which some termed a *cancer epidemic* [1, 3], contributed to the adoption of primary preventive measures: Prophylactic HPV vaccination is now recommended for boys aged 9-14 years in addition to the already established vaccination for girls in Germany and elsewhere [4].

OPSCC arise in the squamous epithelial lining of pharyngeal subsites collectively termed the *oropharynx*, including the inferior surface of the soft palate and uvula, the tonsillar pillars, the palatine tonsils, the glossotonsillar sulci, the base of the tongue, the valleculae, and the portion of the lateral and posterior pharyngeal walls extending between the level of the free border of the soft palate and the level of the superior surface of the hyoid bone [5].

In 2018, oral and pharyngeal cancers accounted for 1,9 % and 3,7 % of all new cancer diagnoses in German females and males, respectively [6, 7]. Approximately 38 % of all oral and pharyngeal cancers were localized in the oropharynx [7]. Squamous cell carcinoma is the most frequent histological type, representing 84 % of oral and pharyngeal cancers [6].

4.1.1 Etiopathogenesis

In carcinogenesis of HPV-negative head and neck squamous cell carcinomas (HNSCC), chronic exposure to mutagens such as tobacco and alcohol typically leads to progressive accumulation of numerous genetic and epigenetic alterations, which may present clinically as precursor lesions such as leukoplakia, and ultimately cancer [8, 9]. Frequently mutated genes thought to act as cancer driver genes include CDKN2A (tumor suppressor gene, encoding the $p16^{INK4A}$ protein) and TP53 (tumor suppressor, encoding the p53 protein), which are key actors in cell cycle control; FAT1 and NOTCH1 (tumor suppressors, encoding eponymous proteins), whose inactivation may disinhibit WNT signaling; and KMT2D (tumor suppressor, encoding Histone-lysine N-methyltransferase), which is involved in epigenetic regulation [8, 10].

In addition to tobacco and alcohol, sustained infection of epithelial cells with high-risk HPV strains has been identified as a major exogenic risk factor for OPSCC – especially for cancer in the palatine and lingual tonsils [1, 3, 8]. Sexual transmission is the principal

cause for oral and pharyngeal HPV infection, with the lifetime number of oral intercourse partners considered the decisive factor [1, 11-13]. On a molecular level, carcinogenesis of HPV-associated OPSCC is remarkably different from the HPV-negative form. Expression of viral oncoproteins E6 and E7 initiates malignant transformation of predominantly tonsillar crypt epithelial cells: E6 binds and induces proteolysis of the p53 protein, and E7 inactivates retinoblastoma pocket proteins (pRB, p107 and p13, tumor suppressor proteins), leading to dysfunctional cell cycle control, inhibited apoptosis and cell immortalization, and therefore setting the stage for accumulation of mutations and carcinogenesis [3, 8, 14-16].

These findings, among others, support the notion that two distinct OPSCC entities prevail in the current epidemiological landscape: HPV-associated and HPV-negative cancers. This is a common theme in the work presented. We sought to investigate the utility of advanced imaging features of OPSCC in prediction of HPV “status” and in prognostication of outcome of HPV-associated cancers.

4.1.2 Clinical Presentation and Diagnostic Workup

Signs and symptoms of OPSCC may include palpable or visible neck masses, sore throat, dysphagia, globus sensation, odynophagia, otalgia, bleeding and hemoptysis, voice changes, and B-symptoms [9, 17]. Patients with HPV-associated OPSCC are more likely to note a neck mass as the initial symptom, while HPV-negative patients tend to complain of a sore throat, dysphagia or odynophagia first [9, 17]. HPV-positive patients tend to present with smaller primary tumors, and more advanced lymphadenopathy, compared to their HPV-negative counterparts [3, 9, 16, 18]. HPV-driven OPSCC tend to affect younger, healthier, patients with higher socioeconomic status who are more frequently male and more often lack a significant smoking history [1, 9, 16, 18, 19].

Initial diagnostic OPSCC workup typically involves clinical examinations including an ear-nose-throat (ENT)-examination, imaging studies such as B-mode sonography, computed tomography (CT), magnetic resonance imaging or positron emission tomography combined with CT (PET/CT, see section 4.1.2.1), and a diagnostic panendoscopy under general anesthesia with tissue sampling for histopathological examination and HPV testing (see sections 4.1.2.2 and 4.1.3.1) [20-24]. The above examinations are essential for determining the cancer entity, anatomic location, and extent of the primary tumor and any regional or distant metastatic spread, for cancer staging (see section 4.1.3.2), and for detecting any synchronous malignancies or dysplastic precursor lesions [20-23]. The obtained findings, in addition to the patient’s performance status and personal preferences, will guide the therapeutic approach.

In this work, we focused on pretreatment PET/CT imagery and applied advanced image analysis to facilitate molecular and prognostic characterization of OPSCC.

4.1.2.1 PET/CT Imaging

Both CT and PET are cross-sectional diagnostic imaging techniques, which, if a sufficient number of cross sections is obtained, can generate digital volumetric representations of the scanned object. PET/CT combines a CT and a PET scanner in a single gantry and can therefore obtain co-registered CT and PET images, which may be digitally fused for anatomic localization of PET findings and attenuation correction of PET images [25].

In CT scanners, an x-ray tube and diametrically opposite detectors synchronously rotate around the object [26, 27]. The attenuation of x-ray beams is measured as they pass through the patient at various angles of rotation, and the measurements are repeated regularly along the patient's z-axis as they translate through the scanner [26, 27]. Computer algorithms reconstruct tomographic images from the raw data, wherein voxel values represent physical density expressed in Hounsfield units [26, 28].

Intravenous injection of a radiotracer precedes PET imaging, allowing the tracer to distribute in the body and to accumulate in certain tissues, depending on its pharmaceutical characteristics. Radiolabeled glucose, namely ^{18}F-2-Fluor-2-deoxy-D-glukose (FDG), is the mainstay among radiotracers currently utilized in clinical oncologic PET [29]. Following cellular uptake, the enzyme hexokinase phosphorylates FDG, leading to metabolic "trapping" of the radiotracer in the cell [29]. Given the altered glucose-dependent metabolism of cancer cells, comparatively greater amounts of FDG accumulate therein, providing the image contrast for clinically viable PET [29]. As the radioisotope (^{18}F in FDG) decays, positrons are emitted (beta plus decay) [30]. The positron travels a short distance, typically few millimeters, before interacting with an electron. This triggers a matter-antimatter-reaction known as annihilation, in which the particles' masses are converted into a pair of 511 kilo electron volt photons [30]. The two photons are emitted in diametrically opposite directions (i.e., at almost exactly 180 degrees to each other), and may hit detector elements arranged in detector rings, in the center of which the patient resides [30]. The scanner can now narrow down the location of the annihilation event to an approximately straight line connecting two opposing detector elements activated by the photon pair. After recording a great number of annihilations over a period of typically several minutes, a series of data processing steps is applied, including attenuation correction [30]. Computer algorithms reconstruct tomographic images, in which voxel values usually represent radioactivity concentration expressed in Becquerel per milliliter.

In summary, CT and PET gather complementary information: CT measures physical density at high spatial resolution, thus reflecting structural/morphological tissue properties. FDG-PET indirectly measures glucose metabolism at lower resolution, thus reflecting functional properties. One objective of this work was to investigate whether combined analysis of tissue density features form CT and functional metabolism features

from PET would leverage greater value for molecular and prognostic characterization of OPSCC than CT or PET features alone.

While essential for detection and staging of OPSCC, cross-sectional imaging provides few visual clues that would allow the reader to discern HPV-associated from HPV-negative OPSCC [31-34]. We therefore aimed to investigate if advanced quantitative imaging features of OPSCC would enable precise differentiation.

The role of PET/CT in initial workup and staging of non-recurrent OPSCC with non-occult primaries varies across different centers and geographic regions and is yet to be definitively established: However, a US National Comprehensive Cancer Network guideline [20] states that PET/CT is the preferred imaging modality to screen for distant metastases in patients with locoregionally advanced HNSCC and acknowledges a limited role in workup of nodal metastases; and a joint guideline by the European Head and Neck-, Medical Oncology- and Radiation Oncology-Societies [21] strongly recommends PET/CT to assess for distant metastases, with chest CT as a possible alternative. In Germany's statutory health insurance system, especially reimbursement restrictions presently limit the use of PET/CT to guiding treatment decisions regarding neck dissections in patients with advanced cancer [35, 36].

4.1.2.2 HPV Testing

Testing for HPV association is recommended and routinely performed on all newly diagnosed OPSCC using tissue samples obtained, for example, during panendoscopy [20, 24]. Different assays may be employed, including high-risk HPV ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) *in situ* hybridization, viral DNA amplification by polymerase chain reaction (PCR), detection of viral E6 and E7 messenger RNA by reverse transcription PCR, and immunohistochemical staining for $p16^{INK4A}$ protein overexpression [16, 24, 37, 38]. p16 overexpression is induced by the HPV E7 protein – therefore, p16 immunohistochemistry is a surrogate marker for HPV [8, 24, 37]. Per the latest edition of cancer staging manuals, a different tumor-node-metastasis (TNM) and stage group classification applies for p16 positive OPSCC, making p16-testing a prerequisite for staging (see section 4.1.3.2) [5, 39].

In this work, we aimed to devise exploratory PET/CT imaging biomarkers for HPV association by extracting advanced imaging features of OPSCC primary tumors and metastatic cervical lymphadenopathy. If sufficient diagnostic accuracy is attained, such imaging markers may offer non-invasive, inexpensive, and readily available alternatives to tumor specimen-based HPV testing, given that tomographic imaging is routinely performed as part of the diagnostic workup.

4.1.3 Determinants of Prognosis

A wide range of variables affect the prognosis of OPSCC patients, including HPV "status", TNM stages, patient age, performance status, presence of comorbid conditions, and tobacco exposure [3, 5, 18, 19, 39-45]. The variables introduced in detail below are most relevant to this work.

4.1.3.1 HPV "Status"

Compared to the HPV-negative form, HPV-associated OPSCC carries a more favorable prognosis, with longer overall survival and lower rates of locoregional progression [18, 41, 46-50]. This is in part attributed to differences in demographic and risk profiles, with HPV-driven cancer afflicting a younger and healthier patient population with little tobacco exposure [1, 3, 18]. However, studies demonstrate HPV positivity remains significantly associated with favorable outcome after adjustment for such confounders, suggesting it is an independent prognostic marker [18, 41, 47-50]. To date, the underlying biologic mechanisms remain somewhat elusive, with much of the pertinent research focusing on increased sensitivity to treatment which is evident in HPV-positive patients [3, 16, 18, 19, 51]. In addition to HPV classification, we investigated the possible added prognostic value of advanced imaging features of OPSCC in prognostically favorable HPV-associated cancers.

4.1.3.2 TNM Staging of HPV-associated Oropharyngeal Cancer

Although performed as part of diagnostic workup (see section 4.1.2), cancer staging is discussed here since this work is centered around its relevance for prognosis. The previous seventh edition of the Union for International Cancer Control (UICC) and American Joint Committee on Cancer (AJCC) TNM staging system failed to reflect the pronounced outcome discrepancies between HPV-driven and HPV-negative OPSCC at equal disease stages, and was found inadequate for prognostic stratification of HPV-mediated cancer in general, warranting reconsideration of the staging approach for OPSCC [42, 52]. Therefore, the eighth and latest edition of the UICC/AJCC TNM staging system introduced separate anatomic stages and prognostic groups for $p16^{INK4A}$ positive OPSCC [5, 39, 42, 53]. Effective since 2017, it is a relatively new prognostic tool, and may evolve in the future. In this work, we aimed to devise prognostic models based on advanced imaging features of OPSCC, to compare their performance with TNM models, and to explore the potential added value of combining TNM staging and imaging features.

4.1.3.3 Locoregional Disease Progression

Locoregional disease progression (LRP) is the most frequent type of treatment failure after definitive therapy of OPSCC in curative intent, occurring in approximately 10% of HPV positive patients [41, 54, 55]. LRP worsens the prognosis of those affected and correlates strongly with reduced overall survival [54, 56, 57]. Salvage therapy after LRP is often morbid and impairs patients' functional outcome and quality of life [58, 59]. LRP is thus an endpoint of major clinical interest. We aimed to devise advanced imaging biomarkers for LRP risk prediction in HPV-associated OPSCC. Such biomarkers may improve patient selection for therapeutic strategies before treatment, and may allow early identification of patients at elevated risk of LRP post treatment, who may benefit from closer surveillance and swift initiation of diagnostic and therapeutic interventions.

4.2 Radiomics

4.2.1 General Concepts of Head-and-Neck-Cancer Radiomics

In current clinical practice, interpretation of tomographic imaging findings primarily relies upon qualitative visual evaluation: Taking an HNSCC primary as an example, the anatomic tumor location and extent, presence and pattern of contrast uptake, signs of necrosis, degree of margin irregularity and the displacement or infiltration of surrounding structures are *qualitative* imaging features of interest, to name a few [60, 61]. In addition, simple length measurements are commonly performed, mostly of the in-plane tumor size.

Enabled by advancements in computational capabilities and the rise of artificial intelligence (AI) technology, radiomic analysis emerged as an alternate, exploratory means for medical image evaluation in recent years [60-65]. In radiomics, advanced computer algorithms automatically extract large numbers of so-called *features*, usually from contoured and annotated volumes-of-interest corresponding to disease sites in standard-of-care tomographic images [61-65]. Radiomic features are *quantitative* descriptors of shape and size, pixel/voxel grey scale intensity and intensity distribution as well as texture characteristics [60-63]. Many studies report the use of image filters or decompositions prior to feature extraction to enhance the granularity of analyses [61-63]. Figure 1 defines and summarizes important traits of radiomics.

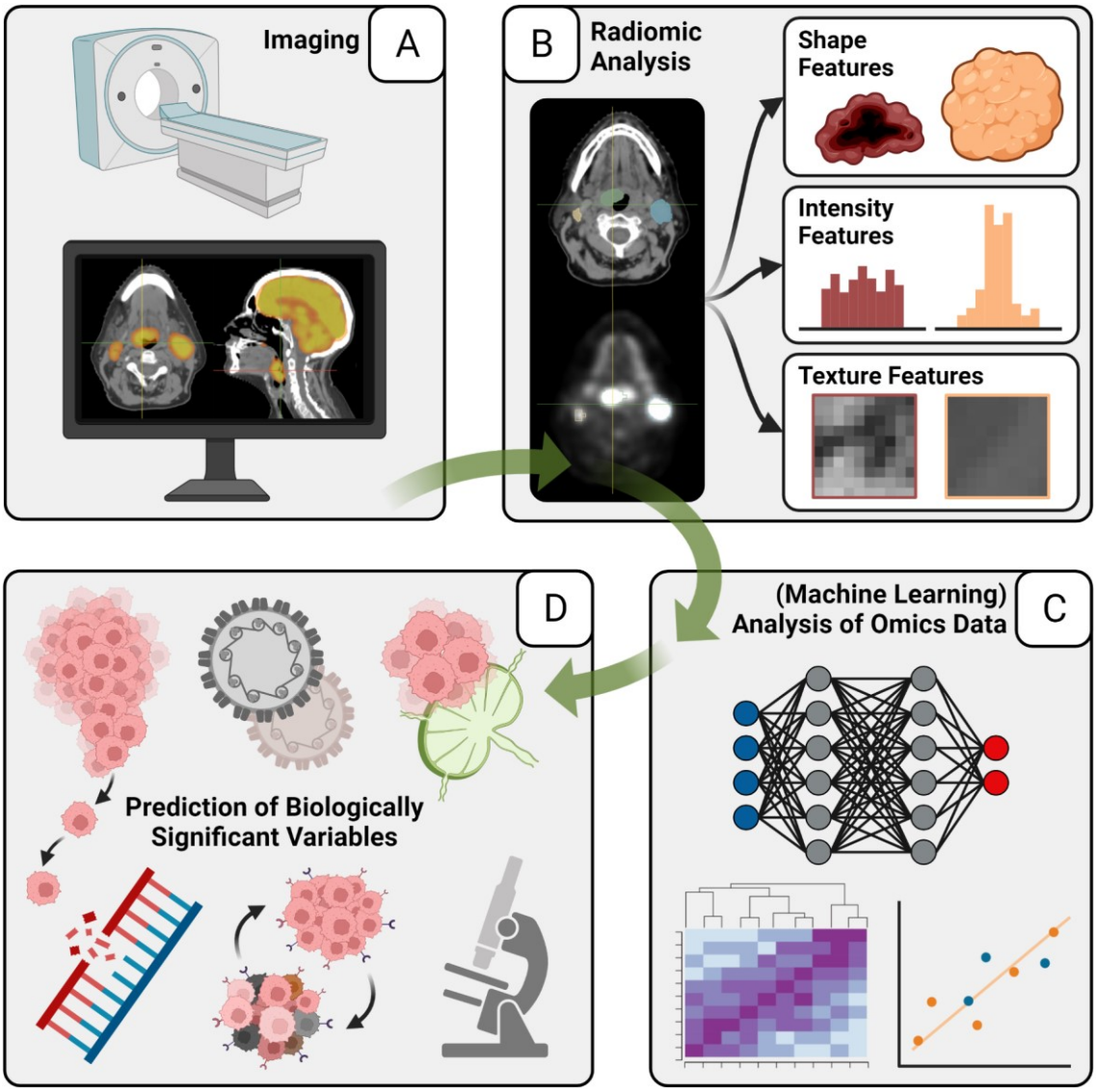


Figure 1. Radiomics entails the transfer and application of the "omics" concept (defined as comprehensive high-throughput characterization of biological matter and known from disciplines like genomics, proteomics, and transcriptomics) to medical imaging [66]. It encompasses extraction of high-dimensional sets of quantitative radiomic features from tomographic medical imagery (A, B) as well as mapping of imaging phenotypes to medically significant variables to facilitate automation of detection and diagnosis, predictive and prognostic modelling, and imaging biomarker development (C, D) [61-64].*

Radiomic features, combinations thereof – termed radiomic signatures – and predictions from more complex radiomics-based models were shown to correlate with variables of interest across numerous diseases and may serve as the basis for imaging biomarker development [60-65]. In HNSCC, studies explored associations with histological grade, nodal status and extra-nodal extension, treatment response and oncologic outcome,

* Created with BioRender.com PET/CT image in (A) and (B) sourced from „Head-Neck-PET-CT" collection [67].

tumor biology including HPV “status”, TP53 mutation status, and PD-L1 expression, as well as chemoradiotherapy side effects, among others [68-73].

The main hypothesis underlying radiomics is that features may decode medically significant imaging patterns which are not visually perceivable, therefore postulating that standard-of-care imaging data is underutilized in current clinical practice, and that radiomics may provide added clinical value [60-62]. For example, certain features may reflect spatial intratumor heterogeneity of biological traits, which poses a diagnostic challenge on the one hand as it is not reliably assessed by single-site biopsies, and on the other hand may carry important implications for tumor behavior [65, 74-77]. Therefore, radiomics may be applied to devise imaging biomarkers of tumor heterogeneity [65, 75-77].

This work sought to characterize OPSCC using PET/CT radiomics.

4.2.2 Machine Learning-Analysis of Radiomic Data

Machine learning (ML) refers to the study and application of computer algorithms which learn from experience [78-80]. Put another way, ML algorithms derive intelligent functionality from data, without requiring explicit programming with rules [78-80]. Therefore, ML research typically relies on training and test samples for model development and performance validation, respectively. ML is considered a subfield of AI [79, 80].

Radiomics produces large feature sets, often supplemented by clinical or other “omics” data. The discovery of medically meaningful patterns and relationships in such extensive datasets is challenging. Rather than relying on hypothesis-driven evaluation of individual features, radiomics research is largely data driven, whereby a priori no assumptions are made about the biological significance of individual features [61, 65, 81]. ML has proven to be very useful in this setting, with a considerable fraction of radiomics studies relying on ML for model and biomarker generation [81].

In this work, we applied a series of ML algorithms to derive predictive models from radiomic data.

4.3 Radiomics-based Machine Learning-Models for HPV Classification and Locoregional Progression Prognostication

This work sought to explore the potential of radiomics in OPSCC patients. The primary aim was to investigate the utility of pretreatment FDG-PET and non-contrast CT radiomic features for molecular and prognostic characterization of therapy-naïve OPSCC.

Recognizing the etiopathogenetic, molecular, demographic, clinicopathologic and prognostic differences between HPV-driven and HPV-negative cancers as well as the profound impact of HPV on OPSCC epidemiology (see section 4.1), we selected HPV association as the *molecular* trait of interest. Applying ML classification algorithms, we aimed to devise radiomic biomarkers indicating HPV association.

LRP of OPSCC carries major clinical significance, given its strong association with decreased overall survival, and the added morbidity salvage therapy often entails (see section 4.1.3.3). Therefore, LRP was selected as the endpoint for *prognostic* analyses in HPV-attributable cancer in this work. ML was used to explore the prognostic significance of radiomics compared to clinical variables including TNM staging, as well as its added prognostic value.

An additional overarching goal was investigation of the complementary value of primary tumor features and features of cervical metastatic lymphadenopathy, as well as of metabolic PET features and structural CT features. Given the magnitude of radiomic feature sets, we applied various combinations of feature dimensionality reduction techniques and ML algorithms to devise optimized biomarkers and models. To enhance the generalizability of findings, we sought to acquire multicentric data, and applied cross validation techniques throughout this work, as well as independent and external validation of select HPV biomarkers.

5. Summary

Every day, extensive amounts of tomographic imagery including CT-, PET- and magnetic resonance-images are acquired in hospitals. Recent advances in radiomics suggest computer algorithms quantifying size, shape, intensity, and texture of imaging findings may provide added clinical value beyond physicians' qualitative visual evaluation. Radiomics may be a readily available, cost-efficient, and non-invasive means to expand the scope of standard-of-care imaging to include provision of quantifiable, objective biomarkers and prognostic models. The data-driven analysis of radiomic features is often accomplished by ML – a type of AI which is especially capable of exploiting the vast datasets generated in "omics" research.

High-risk HPV strains are causally linked to a marked increase in OPSCC incidence around the world in recent decades. HPV-driven cancer affects younger, healthier patients, carries more favorable prognosis, and has recently been assigned a separate TNM classification – compared to frequently tobacco- and alcohol-related HPV-negative OPSCC. HPV testing is routinely performed on tissue samples of OPSCC.

To devise non-invasive radiomic HPV biomarkers, we gathered a multicentric, multinational cohort and extracted radiomic features of 435 OPSCC primary tumors and 741 metastatic cervical lymph nodes on pretreatment FDG-PET and non-contrast CT. Combining different sources of radiomics input, feature dimensionality reduction techniques, and ML algorithms, we trained, optimized, and compared 360 candidate HPV classification models, reaching moderate to high performance in cross validation. Twelve select top performing models did satisfactorily generalize to an independent validation dataset, and the best PET-based models were additionally validated in an external set. A model combining radiomic PET and CT features of primary tumors as input and applying ridge regression feature selection and an extreme gradient boosting ML classifier achieved the highest overall performance.

Locoregional treatment failure occurs in approximately 10% of HPV-associated cancers and entails worse outcome and morbid salvage therapies. Using a subset of 190 patients with HPV-attributable OPSCC and sufficient follow-up intervals, we pursued a similar analysis approach to develop ML models for prognostication of LRP after definitive therapy in curative intent. We generated and compared models relying on radiomic features or clinical variables including eighth edition UICC/AJCC TNM staging or a combination of both. A random survival forest ML model with radiomic PET and CT features of primary tumors as input was superior, and addition of clinical variables did not improve performance. A random forest classifier using the same radiomics input achieved significant stratification into high- and low-LRP-risk groups. All models were cross validated.

Across HPV and LRP analyses, combining radiomic PET and CT features tended to yield higher performance than single-modality models, likely reflecting the complementarity of information gathered by metabolic PET and morphological CT imaging. Conversely, supplementing radiomic features of primary tumors with metastatic lymph node radiomics did not reliably improve model performance. Models relying on lymph node radiomics alone fared worse than primary tumor models in predicting HPV “status”.

The radiomic HPV biomarkers and LRP prognostication models devised in this work are promising. In the future, non-invasive PET/CT biomarkers may supplement or substitute tissue-based HPV testing, and LRP prediction models may guide therapy planning and alert physicians to increased risk of progression after treatment in certain patients who may especially benefit from tight surveillance. Radiomics and ML could become key enablers of personalized “precision medicine”, which may help achieve the next major leap forward in cancer care. Before routine clinical application may be considered, higher model accuracy must be attained, and additional validation in large, prospective studies performed.

6. Zusammenfassung (German Summary)

Tagtäglich werden umfangreiche Schnittbilddaten bestehend aus etwa CT-, PET- und Magnetresonanztomografie-Bildern in Kliniken aufgenommen. Jüngste Erkenntnisse auf dem Gebiet der Radiomics legen nahe, dass Computeralgorithmen, die Größe, Form, Intensität und Textur von Bildbefunden quantifizieren, klinischen Mehrwert gegenüber qualitativer visueller Befundung durch ÄrztInnen erbringen könnten. Radiomics könnten niederschwellig, kosteneffizient und nicht-invasiv das Leistungsspektrum klinischer Standardbilddiagnostik um die Bereitstellung quantifizierbarer, objektiver Biomarker und prognostischer Modelle erweitern. Zur Analyse von Radiomics Features wird häufig ML herangezogen – es handelt sich um AI, welche besonders effektiv die gewaltigen Datenmengen der „Omics"-Forschung verwerten kann.

Hochrisiko-HPV sind ursächlich für einen deutlichen Anstieg an OPSCC-Fällen weltweit in den letzten Jahrzehnten. HPV-assoziierter Krebs betrifft jüngere, gesündere PatientInnen, geht mit besserer Prognose einher und wird anhand neuer separater TNM-Stadien eingeteilt – im Vergleich zu HPV-negativen OPSCC, die häufig Tabak- und Alkohol-assoziiert sind. OPSCC-Gewebeproben werden routinemäßig auf HPV getestet.

Zur Entwicklung von Radiomics-Biomarkern für HPV akquirierten wir eine multizentrische, multinationale OPSCC-Kohorte und extrahierten Radiomics Features aus 435 Primärtumoren und 741 metastatischen zervikalen Lymphknoten in prätherapeutischen FDG-PET- und nativ-CT-Bildern. Durch Kombination verschiedener Radiomics-Inputs, Feature-Dimensionsreduktionstechniken und ML-Algorithmen trainierten, optimierten und verglichen wir 360 Kandidatenmodelle, welche moderate bis hohe HPV-Klassifikationsleistung im Kreuzvalidierungsversuch erzielten. Zwölf ausgewählte leistungsstarke Modelle wiesen eine gute Generalisierbarkeit in einem unabhängigen Validierungsdatensatz auf; die besten PET-basierten Modelle wurden zusätzlich in einem externen Datensatz validiert. Die höchste Gesamtleistung erzielte ein Modell, das PET- und CT-Radiomics Features von Primärtumoren kombinierte und mittels Ridge Regression-Featureauswahl und Extreme Gradient Boosting-ML-Klassifikationsalgorithmus erstellt wurde.

Lokoregionäres Therapieversagen tritt bei etwa 10% der HPV-assoziierten Karzinome auf und zieht regelhaft schlechteres onkologisches Outcome und nebenwirkungsreiche Zweitlinientherapien nach sich. Mittels einer Teilkohorte bestehend aus 190 HPV-assoziierten OPSCC mit ausreichenden Nachsorgeintervallen verfolgten wir einen ähnlichen Analyseansatz, um ML-Modelle zur Vorhersage von LRP nach kurativ intendierter Therapie zu entwickeln. Wir erstellten und verglichen Modelle, welche Radiomics Features oder klinische Variablen inklusive der UICC/AJCC TNM-Klassifikation in der achten Edition oder eine Kombination beider nutzten. Ein Random Survival Forest-ML-Modell basierend auf PET- und CT-Radiomics Features von Primärtumoren erzielte die präzisesten Vorhersagen, und die Integration klinischer

Variablen verbesserte das Model nicht. Ein auf selbigem Radiomics-Input beruhender Random Forest-Klassifikationsalgorithmus erreichte eine signifikante LRP-Risikostratifizierung. Alle Modelle wurden kreuzvalidiert.

In sowohl HPV- als auch LRP-Analysen führte die Kombination von PET- und CT-Features zu tendenziell höheren Modellleistungen als unimodale Analysen, was die Komplementarität metabolischer PET- und struktureller CT-Features widerspiegeln könnte. Umgekehrt verbesserte die Addition von Lymphknoten- zu Primärtumor-Radiomics Features die Modellleistungen nicht zuverlässig. Radiomics-HPV-Biomarker, die allein auf Lymphknotenmodellen beruhten, waren zudem weniger präzise als Primärtumormodelle.

Die im Rahmen dieser Arbeit entwickelten radiomischen HPV-Biomarker und LRP-Prognosemodelle sind vielversprechend. Zukünftig könnten nicht-invasive PET/CT-Biomarker gewebebasierte HPV-Tests ergänzen oder ersetzen. Gleichermaßen könnten LRP-Prognosemodelle Therapieplanungen präzisieren und BehandlerInnen posttherapeutisch auf ein erhöhtes Progressrisiko bei PatientInnen hinweisen, die folglich von engmaschiger Surveillance besonders profitieren würden. Radiomics und ML könnten elementare Bestandteile der personalisierten „Präzisionsmedizin“ der Zukunft darstellen und Krebstherapien und -diagnostik entscheidend verbessern. Bevor eine routinemäßige klinische Anwendung in Betracht kommt, müssen höhere Modellpräzision erreicht und prospektive Validierungsstudien durchgeführt werden.

7. Contribution to Publications

7.1 Publication 1

7.1.1 Publication Reference

Haider SP, Mahajan A, Zeevi T, Baumeister P, Reichel C, Sharaf K, Forghani R, Kucukkaya AS, Kann BH, Judson BL, Prasad ML, Burtness B, Payabvash S. **PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma.** Eur J Nucl Med Mol Imaging. 2020;47(13):2978-91.

7.1.2 Contribution

Hypothesis and Plan: My contribution included optimizing the research hypotheses and methodological strategy, including researching literature, leading to adoption of radiomic analysis of metastatic cervical lymph nodes, “virtual” consensus volumes of interest (VOI), multiple delineation-analysis and -feature selection, a CT artifact handling strategy, and fully automated hyperparameter tuning using Bayesian optimization.

Data Curation: I assumed responsibility for data acquisition from institutional archives and from a public repository (*The Cancer Imaging Archive*, TCIA), including acquisition of all PET/CT imaging data and of all clinical data. My contribution included selecting TCIA collections and individual subjects from TCIA and institutional repositories based on exclusion and inclusion criteria, downloading, archiving, integrating, and processing all imaging and clinical data.

Methodology and Formal Analysis: To facilitate review and segmentation of PET/CT imagery in *3D-Slicer* software, I developed standard operating procedures defining loading, segmenting, and saving PET/CT images and corresponding segmentations using *3D-Slicer* modules and functions. I performed the bulk of image review and segmentation tasks, comprising exclusion of subjects or individual axial CT slices affected by CT artifacts, initial segmentation of all VOIs on PET and CT images, including one of two repeat segmentation rounds to enable multiple delineation-analysis, as well as preliminary TNM staging. Configuring the *Pyradiomics* pipeline required literature research and preliminary analyses to determine image preprocessing and radiomic feature extraction settings, which I conducted. I implemented all conventional statistical and ML analyses in *R*, which included radiomic and clinical data preprocessing, multiple delineation-based feature selection, data allocation to training and validation sets, feature standardization, generation of “virtual” consensus VOIs, feature dimensionality reduction, ML classifiers, a cross validation framework, Bayesian hyperparameter

optimization, independent and external validation, performance quantification and statistical comparison of the performance of different models, feature importance score quantification, visualization and archiving of pipeline outputs. These analyses relied upon custom written code, externally sourced packages and functions from *The Comprehensive R Archive Network* (CRAN) as well as *R* base functions. I designed, wrote, and validated all custom code, and integrated all externally sourced code in our custom pipeline, which included researching literature to facilitate selection, integration, and configuration of externally sourced CRAN packages and functions. In close collaboration with the principal investigator and a ML specialist, I made substantial contributions to the configuration and streamlining of the *R* analysis pipeline as a whole, which encompassed defining the data allocation strategy to training and validation sets, defining methodological approaches such as dimensionality reduction techniques and ML algorithms, setting of model hyperparameters or hyperparameter bounds, and researching literature to support the aforementioned. Finally, the statistical analysis of patients' characteristics was my responsibility.

Presentation of Results and Manuscript Preparation: My contribution encompassed preparation of the initial manuscript and supplement drafts in their entirety, including visualization of methodology and results in figures, diagrams, and tables, drafting all manuscript and supplemental text, and formatting the manuscript and supplement. I prepared the submission, formally submitted all material, and prepared the initial draft of all revisions based on co-authors' and peer reviewers' feedback. In addition, I presented the results at scientific conferences as oral abstracts (European Congress of Radiology, Annual Meeting of the German Society of Otorhinolaryngology, see section 3.3).

7.2 Publication 2

7.2.1 Publication Reference

Haider SP, Sharaf K, Zeevi T, Baumeister P, Reichel C, Forghani R, Kann BH, Petukhova A, Judson BL, Prasad ML, Liu C, Burtness B, Mahajan A, Payabvash S. **Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics.** Transl Oncol. 2021;14(1):100906.

7.2.2 Contribution

Hypothesis and Plan: Adding to contributions for publication 1 which were adopted herein (see section 7.1.2), I defined several additional methodological aspects of the research plan and reviewed the pertinent literature. These included ascertainment of patient

inclusion- and exclusion criteria, implementation of random survival forest models, and statistical approaches for quantification of random forest performance in the setting of right-censored survival data.

Data Curation: I gathered all clinical data including oncologic treatment and outcome information required for analysis of LRP and for inclusion of appropriate patients. These data originated from the same institutional and TCIA archives; I assumed responsibility for downloading, archiving, integrating, and processing all data.

Methodology and Formal Analysis: I implemented all conventional statistical and ML analyses in *R*, which included radiomic and clinical data preprocessing, multiple delineation-based feature selection, feature standardization, generation of “virtual” consensus VOIs, feature dimensionality reduction, random forest ML models, integration of clinical variables in ML models, a cross validation framework, model performance quantification in the setting of right-censored survival data, and visualization and archiving of pipeline outputs. These analyses relied upon custom written code, externally sourced packages and functions from CRAN as well as *R* base functions; some implementations were adopted from publication 1 (see section 7.1.2). I designed, wrote, and validated all custom code, and integrated all externally sourced code in our custom pipeline, which included researching literature to facilitate selection, integration, and configuration of externally sourced CRAN packages and functions. Under the supervision of the principal investigator and with inputs from a ML specialist, I designed, implemented, and finetuned the *R* analysis pipeline as a whole, which encompassed defining methodological approaches such as dimensionality reduction techniques and the use of random survival forest and random classification forest ML algorithms, setting of model hyperparameters, and researching literature to facilitate the aforementioned. Finally, the statistical analysis of patients’ characteristics was my responsibility.

Presentation of Results and Manuscript Preparation: My contribution encompassed preparation of most of the initial manuscript draft and of the supplement draft in its entirety, including visualization of methodology and results in figures, diagrams, and tables, drafting most of the manuscript and all supplemental text, and formatting the manuscript and supplement. I prepared the submission, formally submitted all material, and prepared the initial draft of all revisions based on co-authors’ and peer reviewers’ feedback.

8. References

1. Gillison ML, Chaturvedi AK, Anderson WF, Fakhry C. Epidemiology of Human Papillomavirus–Positive Head and Neck Squamous Cell Carcinoma. Journal of Clinical Oncology. 2015;33(29):3235-42.

2. Wittekindt C, Wagner S, Bushnak A, Prigge ES, von Knebel Doeberitz M, Würdemann N, et al. Increasing Incidence rates of Oropharyngeal Squamous Cell Carcinoma in Germany and Significance of Disease Burden Attributed to Human Papillomavirus. Cancer Prev Res (Phila). 2019;12(6):375-82.

3. Marur S, D'Souza G, Westra WH, Forastiere AA. HPV-associated head and neck cancer: a virus-related cancer epidemic. Lancet Oncol. 2010;11(8):781-9.

4. AG HPV der Ständigen Impfkommission (STIKO). Wissenschaftliche Begründung für die Empfehlung der HPV-Impfung für Jungen im Alter von 9 bis 14 Jahren. Epidemiologisches Bulletin.2018(26):233–50.

5. Brierley JD, Gospodarowicz MK, Wittekind C. TNM Classification of Malignant Tumours. 8 ed. Chicester: John Wiley & Sons, Incorporated; 2017.

6. Erdmann F, Spix C, Katalinic A, Christ M, Folkerts J, Hansmann J, et al. Krebs in Deutschland für 2017/2018. 13 ed. Berlin: Robert Koch-Institut and Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V.; 2021.

7. Zentrum für Krebsregisterdaten des Robert Koch-Institut. Datenbankabfrage: Robert Koch-Institut; 2022 [cited 2022 February 2nd]. Available from: https://www.krebsdaten.de/Krebs/DE/Datenbankabfrage/datenbankabfrage_stufe1_node.html.

8. Leemans CR, Snijders PJF, Brakenhoff RH. The molecular landscape of head and neck cancer. Nat Rev Cancer. 2018;18(5):269-82.

9. Chi AC, Day TA, Neville BW. Oral cavity and oropharyngeal squamous cell carcinoma—an update. CA: A Cancer Journal for Clinicians. 2015;65(5):401-21.

10. Lawrence MS, Sougnez C, Lichtenstein L, Cibulskis K, Lander E, Gabriel SB, et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576-82.

11. Drake VE, Fakhry C, Windon MJ, Stewart CM, Akst L, Hillel A, et al. Timing, number, and type of sexual partners associated with risk of oropharyngeal cancer. Cancer. 2021;127(7):1029-38.

12. Heck JE, Berthiller J, Vaccarella S, Winn DM, Smith EM, Shan'gina O, et al. Sexual behaviours and the risk of head and neck cancers: a pooled analysis in the International Head and Neck Cancer Epidemiology (INHANCE) consortium. International Journal of Epidemiology. 2010;39(1):166-81.

13. Gillison ML, Broutian T, Pickard RK, Tong ZY, Xiao W, Kahle L, et al. Prevalence of oral HPV infection in the United States, 2009-2010. Jama. 2012;307(7):693-703.

14. Rampias T, Sasaki C, Weinberger P, Psyrri A. E6 and e7 gene silencing and transformed phenotype of human papillomavirus 16-positive oropharyngeal cancer cells. J Natl Cancer Inst. 2009;101(6):412-23.

15. Steenbergen RD, Snijders PJ, Heideman DA, Meijer CJ. Clinical implications of (epi)genetic changes in HPV-induced cervical precancerous lesions. Nat Rev Cancer. 2014;14(6):395-405.

16. Taberna M, Mena M, Pavon MA, Alemany L, Gillison ML, Mesia R. Human papillomavirus-related oropharyngeal cancer. Ann Oncol. 2017;28(10):2386-98.

17. McIlwain WR, Sood AJ, Nguyen SA, Day TA. Initial Symptoms in Patients With HPV-Positive and HPV-Negative Oropharyngeal Cancer. JAMA Otolaryngology–Head & Neck Surgery. 2014;140(5):441-7.

18. Benson E, Li R, Eisele D, Fakhry C. The clinical impact of HPV tumor status upon head and neck squamous cell carcinomas. Oral Oncol. 2014;50(6):565-74.

19. Wang MB, Liu IY, Gornbein JA, Nguyen CT. HPV-Positive Oropharyngeal Carcinoma:A Systematic Review of Treatment and Prognosis. Otolaryngology–Head and Neck Surgery. 2015;153(5):758-69.

20. Pfister DG, Spencer S, Adkins D, Birkeland AC, Brizel DM, Busse PM, et al. Head and Neck Cancers, Version 1.2022: National Comprehensive Cancer Network; 2022 [cited 2022 February 15th]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/head-and-neck.pdf.

21. Machiels JP, René Leemans C, Golusinski W, Grau C, Licitra L, Gregoire V. Squamous cell carcinoma of the oral cavity, larynx, oropharynx and hypopharynx: EHNS-ESMO-ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2020;31(11):1462-75.

22. Mehanna H, Evans M, Beasley M, Chatterjee S, Dilkes M, Homer J, et al. Oropharyngeal cancer: United Kingdom National Multidisciplinary Guidelines. J Laryngol Otol. 2016;130(S2):S90-s6.

23. PDQ® Adult Treatment Editorial Board. Oropharyngeal Cancer Treatment (Adult) (PDQ®): Health Professional Version. PDQ Cancer Information Summaries. Bethesda (MD): National Cancer Institute; 2002-2022.

24. Lewis JS, Jr., Beadle B, Bishop JA, Chernock RD, Colasacco C, Lacchetti C, et al. Human Papillomavirus Testing in Head and Neck Carcinomas: Guideline From the College of American Pathologists. Arch Pathol Lab Med. 2018;142(5):559-97.

25. Mohnike W, Lampe M. Einführung. In: Mohnike W, Hör G, Hertel A, Schelbert H, editors. PET/CT-Atlas: Interdisziplinäre PET/CT- und PET/MR-Diagnostik und Therapie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 1-4.

26. Flohr T. Messprinzip, Bildrekonstruktion, Gerätetypen und Aufnahmetechniken. Wie funktioniert CT? Eine Einführung in Physik, Funktionsweise und klinische Anwendungen der Computertomographie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 3-13.

27. Flohr T. Detektoren, Röntgenstrahler und Gantry. Wie funktioniert CT? Eine Einführung in Physik, Funktionsweise und klinische Anwendungen der Computertomographie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 15-22.

28. Fornaro J. Postprocessing. Wie funktioniert CT? Eine Einführung in Physik, Funktionsweise und klinische Anwendungen der Computertomographie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 41-6.

29. Rösch F, Piel M. Radiochemie/Tracer. In: Mohnike W, Hör G, Hertel A, Schelbert H, editors. PET/CT-Atlas: Interdisziplinäre PET/CT- und PET/MR-Diagnostik und Therapie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 35-73.

30. Rausch I, Beyer T. Physikalisch-technische Grundlagen. In: Mohnike W, Hör G, Hertel A, Schelbert H, editors. PET/CT-Atlas: Interdisziplinäre PET/CT- und PET/MR-Diagnostik und Therapie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 5-34.

31. Cantrell SC, Peck BW, Li G, Wei Q, Sturgis EM, Ginsberg LE. Differences in imaging characteristics of HPV-positive and HPV-Negative oropharyngeal cancers: a blinded matched-pair analysis. AJNR Am J Neuroradiol. 2013;34(10):2005-9.

32. Fujita A, Buch K, Truong MT, Qureshi MM, Mercier G, Jalisi S, et al. Imaging characteristics of metastatic nodes and outcomes by HPV status in head and neck cancers. Laryngoscope. 2016;126(2):392-8.

33. Subramaniam RM, Alluri KC, Tahari AK, Aygun N, Quon H. PET/CT imaging and human papilloma virus-positive oropharyngeal squamous cell cancer: evolving clinical imaging paradigm. J Nucl Med. 2014;55(3):431-8.

34. Corey AS, Hudgins PA. Radiographic imaging of human papillomavirus related carcinomas of the oropharynx. Head and neck pathology. 2012;6 Suppl 1(Suppl 1):S25-S40.

35. Bundesministerium für Gesundheit und Gemeinsamer Bundesausschuss. Bekanntmachung eines Beschlusses des Gemeinsamen Bundesausschussesüber eine Änderung der Richtlinie Methoden vertragsärztliche Versorgung: Positronenemissionstomographie (PET)/Computertomographie (CT) bei Kopf-Hals-Tumoren. Bundesanzeiger. 2017;2017(BAnz AT 06.06.2017).

36. Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie. Positronenemissionstomographie (PET) in der Onkologie - Einsatz zur Steuerung von Diagnostik und Therapie. 2021 [cited 2022 February 15th]. Available from: https://www.dgho.de/publikationen/stellungnahmen/g-ba/pet-pet-ct/positronenemissionstomographie-in-der-onkologie-2021.pdf/view.

37. Prigge ES, Arbyn M, von Knebel Doeberitz M, Reuschenbach M. Diagnostic accuracy of p16(INK4a) immunohistochemistry in oropharyngeal squamous cell carcinomas: A systematic review and meta-analysis. Int J Cancer. 2017;140(5):1186-98.

38. Lewis JS, Jr. Human Papillomavirus Testing in Head and Neck Squamous Cell Carcinoma in 2020: Where Are We Now and Where Are We Going? Head Neck Pathol. 2020;14(2):321-9.

39. Amin M, Edge S, Greene F, Byrd D, Brookland R, Washington M, et al. AJCC Cancer Staging Manual. 8 ed: Springer International Publishing; 2017.

40. 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.

41. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010;363(1):24-35.

42. 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.

43. Gillison ML, Zhang Q, Jordan R, Xiao W, Westra WH, Trotti A, et al. Tobacco smoking and increased risk of death and progression for patients with p16-positive and p16-negative oropharyngeal cancer. J Clin Oncol. 2012;30(17):2102-11.

44. Woolgar JA. Histopathological prognosticators in oral and oropharyngeal squamous cell carcinoma. Oral Oncology. 2006;42(3):229-39.

45. Woolgar JA, Triantafyllou A. Pitfalls and procedures in the histopathological diagnosis of oral and oropharyngeal squamous cell carcinoma and a review of the role of pathology in prognosis. Oral Oncology. 2009;45(4):361-85.

46. Ragin CC, Taioli E. Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: review and meta-analysis. Int J Cancer. 2007;121(8):1813-20.

47. Rischin D, Young RJ, Fisher R, Fox SB, Le QT, Peters LJ, et al. Prognostic significance of p16INK4A and human papillomavirus in patients with oropharyngeal cancer treated on TROG 02.02 phase III trial. J Clin Oncol. 2010;28(27):4142-8.

48. Posner MR, Lorch JH, Goloubeva O, Tan M, Schumaker LM, Sarlis NJ, et al. Survival and human papillomavirus in oropharynx cancer in TAX 324: a subset analysis from an international phase III trial. Ann Oncol. 2011;22(5):1071-7.

49. Lassen P, Eriksen JG, Krogdahl A, Therkildsen MH, Ulhøi BP, Overgaard M, et al. The influence of HPV-associated p16-expression on accelerated fractionated radiotherapy in head and neck cancer: evaluation of the randomised DAHANCA 6&7 trial. Radiother Oncol. 2011;100(1):49-55.

50. O'Rorke MA, Ellison MV, Murray LJ, Moran M, James J, Anderson LA. Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis. Oral Oncol. 2012;48(12):1191-201.

51. Cleary C, Leeman JE, Higginson DS, Katabi N, Sherman E, Morris L, et al. Biological Features of Human Papillomavirus-related Head and Neck Cancers Contributing to Improved Response. Clin Oncol (R Coll Radiol). 2016;28(7):467-74.

52. Dahlstrom KR, Calzada G, Hanby JD, Garden AS, Glisson BS, Li G, et al. An evolution in demographics, treatment, and outcomes of oropharyngeal cancer at a major cancer center: a staging system in need of repair. Cancer. 2013;119(1):81-9.

53. Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67(2):122-37.

54. Gronhoj C, Jakobsen KK, Jensen DH, Rasmussen J, Andersen E, Friborg J, et al. Pattern of and survival following loco-regional and distant recurrence in patients with HPV+ and HPV- oropharyngeal squamous cell carcinoma: A population-based study. Oral Oncol. 2018;83:127-33.

55. Fakhry C, Zhang Q, Nguyen-Tan PF, Rosenthal D, El-Naggar A, Garden AS, et al. Human Papillomavirus and Overall Survival After Progression of Oropharyngeal Squamous Cell Carcinoma. Journal of Clinical Oncology. 2014;32(30):3365-73.

56. Caula A, Boukhris M, Guerlain J, Tao Y, Breuskin I, Mirghani H, et al. Correlation between the duration of locoregional control and survival in T1–T2 oropharyngeal cancer patients. European Archives of Oto-Rhino-Laryngology. 2019;276(4):1161-6.

57. Joseph AW, Guo T, Hur K, Xie Y, Yin L, Califano JA, et al. Disease-free survival after salvage therapy for recurrent oropharyngeal squamous cell carcinoma. Head Neck. 2016;38 Suppl 1:E1501-9.

58. Gazda P, Gauche C, Chaltiel L, Chabrillac E, Vairel B, De Bonnecaze G, et al. Functional and oncological outcomes of salvage transoral robotic surgery: a comparative study. Eur Arch Otorhinolaryngol. 2022;279(1):457-66.

59. Smirk R, Kyzas P. Outcome of salvage procedures for recurrent oral and oropharyngeal cancer. Br J Oral Maxillofac Surg. 2018;56(9):847-53.

60. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: A Cancer Journal for Clinicians. 2019;69(2):127-57.

61. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-77.

62. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology. 2017;14(12):749-62.

63. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to Radiomics. J Nucl Med. 2020;61(4):488-95.

64. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019;46(13):2656-72.

65. Tomaszewski MR, Gillies RJ. The Biological Meaning of Radiomic Features. Radiology. 2021;298(3):505-16.

66. Sepiashvili L, Bruce JP, Huang SH, O'Sullivan B, Liu FF, Kislinger T. Novel insights into head and neck cancer using next-generation "omic" technologies. Cancer Res. 2015;75(3):480-6.

67. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Khaouam N, et al. Data from Head-Neck-PET-CT. The Cancer Imaging Archive2017.

68. Bruixola G, Remacha E, Jiménez-Pastor A, Dualde D, Viala A, Montón JV, et al. Radiomics and radiogenomics in head and neck squamous cell carcinoma: Potential contribution to patient management and challenges. Cancer Treatment Reviews. 2021;99:102263.

69. Guha A, Connor S, Anjari M, Naik H, Siddiqui M, Cook G, et al. Radiomic analysis for response assessment in advanced head and neck cancers, a distant dream or an inevitable reality? A systematic review of the current level of evidence. Br J Radiol. 2019:20190496.

70. Tanadini-Lang S, Balermpas P, Guckenberger M, Pavic M, Riesterer O, Vuong D, et al. Radiomic biomarkers for head and neck squamous cell carcinoma. Strahlentherapie und Onkologie. 2020;196(10):868-78.

71. Haider SP, Burtness B, Yarbrough WG, Payabvash S. Applications of radiomics in precision diagnosis, prognostication and treatment planning of head and neck squamous cell carcinomas. Cancers of the Head & Neck. 2020;5(1):6.

72. Zhu Y, Mohamed ASR, Lai SY, Yang S, Kanwar A, Wei L, et al. Imaging-Genomic Study of Head and Neck Squamous Cell Carcinoma: Associations Between Radiomic Phenotypes and Genomic Mechanisms via Integration of The Cancer Genome Atlas and The Cancer Imaging Archive. JCO Clin Cancer Inform. 2019;3:1-9.

73. Chen RY, Lin YC, Shen WC, Hsieh TC, Yen KY, Chen SW, et al. Associations of Tumor PD-1 Ligands, Immunohistochemical Studies, and Textural Features in (18)F-FDG PET in Squamous Cell Carcinoma of the Head and Neck. Sci Rep. 2018;8(1):105.

74. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883-92.

75. Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol. 2016;61(13):R150-66.

76. Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton JD, Snyder A, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. Clin Radiol. 2017;72(1):3-10.

77. Choi ER, Lee HY, Jeong JY, Choi YL, Kim J, Bae J, et al. Quantitative image variables reflect the intratumoral pathologic heterogeneity of lung adenocarcinoma. Oncotarget. 2016;7(41):67302-13.

78. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-58.

79. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology. 2018;288(2):318-28.

80. Uribe CF, Mathotaarachchi S, Gaudet V, Smith KC, Rosa-Neto P, Bénard F, et al. Machine Learning in Nuclear Medicine: Part 1—Introduction. Journal of Nuclear Medicine. 2019;60(4):451-8.

81. Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. Med Phys. 2020;47(5):e185-e202.

9. Publication 1

Haider SP, Mahajan A, Zeevi T, Baumeister P, Reichel C, Sharaf K, Forghani R, Kucukkaya AS, Kann BH, Judson BL, Prasad ML, Burtness B, Payabvash S. **PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma.** Eur J Nucl Med Mol Imaging. 2020;47(13):2978-91.*

* Reprinted by permission from: Springer Nature: European Journal of Nuclear Medicine and Molecular Imaging.

European Journal of Nuclear Medicine and Molecular Imaging (2020) 47:2978–2991
https://doi.org/10.1007/s00259-020-04839-2

ORIGINAL ARTICLE

PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma

Stefan P. Haider[1,2] · Amit Mahajan[1] · Tal Zeevi[3] · Philipp Baumeister[2] · Christoph Reichel[2] · Kariem Sharaf[2] · Reza Forghani[4] · Ahmet S. Kucukkaya[1] · Benjamin H. Kann[5] · Benjamin L. Judson[6] · Manju L. Prasad[7] · Barbara Burtness[8] · Seyedmehdi Payabvash[1]

Received: 19 January 2020 / Accepted: 24 April 2020 / Published online: 12 May 2020

Abstract

Purpose To devise, validate, and externally test PET/CT radiomics signatures for human papillomavirus (HPV) association in primary tumors and metastatic cervical lymph nodes of oropharyngeal squamous cell carcinoma (OPSCC).

Methods We analyzed 435 primary tumors (326 for training, 109 for validation) and 741 metastatic cervical lymph nodes (518 for training, 223 for validation) using FDG-PET and non-contrast CT from a multi-institutional and multi-national cohort. Utilizing 1037 radiomics features per imaging modality and per lesion, we trained, optimized, and independently validated machine-learning classifiers for prediction of HPV association in primary tumors, lymph nodes, and combined "virtual" volumes of interest (VOI). PET-based models were additionally validated in an external cohort.

Results Single-modality PET and CT final models yielded similar classification performance without significant difference in independent validation; however, models combining PET and CT features outperformed single-modality PET- or CT-based models, with receiver operating characteristic area under the curve (AUC) of 0.78, and 0.77 for prediction of HPV association using primary tumor lesion features, in cross-validation and independent validation, respectively. In the external PET-only validation dataset, final models achieved an AUC of 0.83 for a virtual VOI combining primary tumor and lymph nodes, and an AUC of 0.73 for a virtual VOI combining all lymph nodes.

Conclusion We found that PET-based radiomics signatures yielded similar classification performance to CT-based models, with potential added value from combining PET- and CT-based radiomics for prediction of HPV status. While our results are promising, radiomics signatures may not yet substitute tissue sampling for clinical decision-making.

Keywords Radiomics · Oropharyngeal squamous cell carcinoma · PET/CT · Quantitative imaging · HPV · Imaging biomarker

Stefan P. Haider and Amit Mahajan contributed equally to this work.

This article is part of the Topical Collection on *Advanced Image Analyses (Radiomics and Artificial Intelligence)*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00259-020-04839-2) contains supplementary material, which is available to authorized users.

✉ Seyedmehdi Payabvash
sam.payabvash@yale.edu

[1] Section of Neuroradiology, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 789 Howard Ave, PO Box 208042, New Haven, CT 06519, USA

[2] Department of Otorhinolaryngology, University Hospital of Ludwig Maximilians Universität München, Munich, Germany

[3] Center for Translational Imaging Analysis and Machine Learning, Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA

[4] Department of Diagnostic Radiology and Augmented Intelligence & Precision Health Laboratory, McGill University Health Centre & Research Institute, Montreal, Quebec, Canada

[5] Department of Radiation Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

[6] Division of Otolaryngology, Department of Surgery, Yale School of Medicine, New Haven, CT, USA

[7] Department of Pathology, Yale School of Medicine, New Haven, CT, USA

[8] Section of Medical Oncology, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

Introduction

The incidence of oropharyngeal squamous cell carcinomas (OPSCC) associated with high-risk human papillomavirus (HPV) infection has dramatically increased over the past few decades [1]. HPV-associated OPSCC has distinct biologic and clinical characteristics compared to the HPV-negative form—with longer overall survival, more favorable outcome, and improved treatment response [2, 3]. Both the Union for International Cancer Control (UICC) and American Joint Committee on Cancer (AJCC) adjusted their staging manuals accordingly, and classified HPV-associated OPSCC as a distinct entity with separate staging rules, different from HPV-negative carcinomas [4, 5]. In clinical practice, HPV association is ascertained after tissue sampling by means of p16 immunohistochemical staining of histology slides or HPV-specific tests, such as high-risk HPV DNA in situ hybridization (ISH), or polymerase chain reaction (PCR) [6].

Radiomics—the latest addition to the "-omics" concept—refers to high-throughput extraction of quantitative "features" from medical images for biomarker design with a wide range of applications [7]. Recent studies have demonstrated the feasibility of quantitative bioimaging features (i.e., radiomics) for prediction of HPV status in OPSCC using contrast-enhanced computed tomography (CT) scans [8–11]. [18F]fluorodeoxyglucose positron emission tomography (PET) and simultaneous non-contrast CT scans are routinely used for diagnosis, staging, treatment planning, and surveillance of head and neck squamous cell carcinomas (HNSCC), providing imaging biodata amenable for quantitative data mining. However, there are few reports suggesting that quantitative measures from PET scans are associated with HPV status [12, 13].

In this study, we aimed to devise, optimize, and independently validate radiomics signatures for prediction of HPV status in OPSCC based on radiomics features extracted from primary tumor lesions and metastatic cervical lymph nodes on PET and non-contrast CT scans from multi-institutional and multi-national cohorts. While combining PET and CT radiomics into holistic machine-learning models may provide comprehensive information regarding tissue heterogeneity, lesion morphology, and metabolic characteristics, the application of such models for prediction of OPSCC HPV status has barely been studied before. We also used radiomics signatures to predict HPV association in OPSCC metastases in cervical lymph nodes. Finally, we expanded the use of virtual volumes of interest (VOI) combining several lesions in one given subject into a uniform predictive radiomic feature set [9].

Material and methods

Acquisition of imaging and clinical data

Imaging and clinical data were retrospectively acquired from (I) the patients' registry of Yale's Smilow Cancer Hospital supplemented by a review of the electronic medical records and imaging archive from 2009 through 2019 ("Yale" cohort) and (II) four independent collections in The Cancer Imaging Archive (TCIA) including [14] (II-a) the "Head-Neck-PET-CT" collection from four Canadian institutions ("Canadian" cohort) [15, 16]; (II-b) the "Data from Head and Neck Cancer CT Atlas" collection from MD Anderson Cancer Center ("MD Anderson" cohort) [17, 18]; (II-c) the "TCGA-HNSC" collection from various institutions across America ("TCGA" cohort) [19]; and (II-d) the "Head-Neck-Radiomics-HN1" collection from the Netherlands ("MAASTRO" cohort) [20, 21]. Institutional review board approval was obtained for Yale data collection with informed consent waived. TCIA collections host de-identified data and the providing institutions are responsible for consents and approvals.

Patients were included if they (1) had pretreatment PET and non-contrast CT scans of the neck available, (2) biopsy-proven OPSCC, and (3) HPV status ascertained by HPV-specific testing [6] and/or p16 immunohistochemistry. Patients were excluded if they had (1) recurrent OPSCC, (2) unknown primary site, or (3) > 50% of the primary tumor volume affected by artifacts on visual assessment of neck CTs [22]. Biopsies or fine needle aspiration procedures of the primary tumor and lymph nodes before PET/CT imaging were permitted. Imaging and image reconstruction were performed according to clinical protocols specific to the source institution. HPV status ascertainment in the "Yale" dataset was based on the Guideline from the College of American Pathologists [6]. Institutional protocols were used for MAASTRO and Canadian cohorts and an overall HPV annotation is provided in TCIA. High-risk ISH results were available for the "MD Anderson" and p16 and/or high-risk ISH for the "TCGA" cohort.

Lesion segmentation

Figure 1 summarizes the segmentation, preprocessing, and feature extraction workflow. First, the co-registered PET/CT scans were retrieved and reviewed and the gross demonstrable extent and location of primary tumor and metastatic cervical lymph nodes was assessed—i.e., gross tumor volume (GTV) as defined by the "ICRU 83" report [23]. Next, hypermetabolic areas of the primary tumor and every node GTV were separately delineated on PET axial slices using the "Paint" and "Erase" tools of the "Segment Editor" module to define the PET VOIs (i.e., manual "slice by slice" delineation); segmentations were then overlaid onto co-registered CT scans and

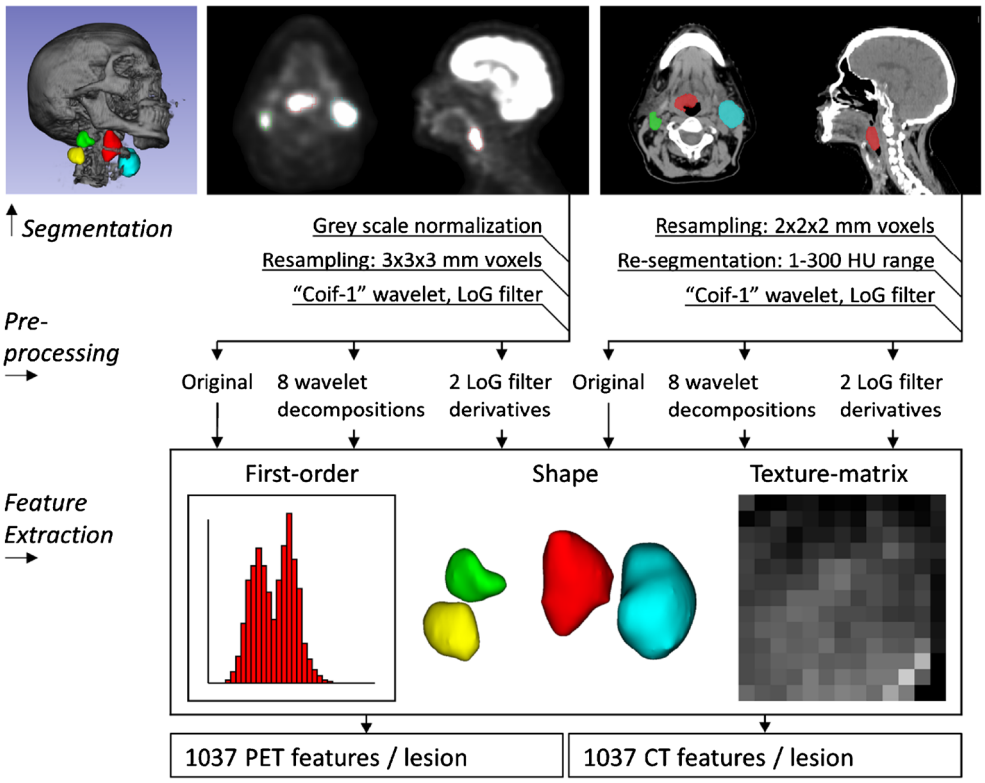


Fig. 1 Radiomics pipeline depicting segmentation, image preprocessing, and feature extraction steps

manually adapted to the GTV extent on CT to create CT VOIs, accounting for bone, air, and surrounding fat planes, using the "Paint" and "Erase" tools. Axial CT slices with substantial CT streak artifact were excluded from analysis, and individual lymph nodes with > 50% of volume affected by artifact were entirely excluded [22]. All segmentations were verified and adjusted by a neuroradiologist (SP) with greater than 8 years of experience in head and neck radiology. All segmentations were created using 3D-Slicer version 4.10.1 [24].

Image preprocessing and feature extraction

PET voxel intensities were normalized by dividing each voxel's raw value by the maximum intensity of the left lentiform nucleus, which was shown to improve PET discriminative ability as an internal, image-derived standardization method [25]. To mitigate the effects of different slice thicknesses and voxel sizes on radiomics features [26], and guarantee rotational invariance of texture features [27], we applied trilinear interpolation to generate isotropic $3 \times 3 \times 3$ and $2 \times 2 \times 2$ mm voxels in PET and CT scans, respectively. CT segmentation masks were restricted to a 1–300 Hounsfield unit (HU) threshold range, limiting the analysis to soft tissue densities. Derivative images were generated to enhance specific imaging characteristics: A "coif-1" wavelet transform yielded eight decompositions per original image by applying high-pass and low-pass filtering in each spatial direction [21, 28]. Edge-enhancement Laplacian of Gaussian (LoG) filtering with 3 and 6 mm for PET scans and 2 and 4 mm for CT scans "sigma" settings generated two derivates per each imaging modality [28, 29]. Computation of texture and certain first-order features requires image grayscale discretization into "bins" [27]. We applied a fixed bin width of 2 units [28], which yielded homogenous numbers of bins in our cohort, and may be more appropriate under certain circumstances than a fixed bin number for comparative analysis [30]. Eighteen first-order, fourteen volumetric shape, and 75 texture-matrix radiomic features were extracted from each set of original images. First-order and texture features were additionally extracted from eight wavelet- and two LoG-filtering derivatives per imaging modality, yielding a total of 1037 PET and 1037 CT radiomics features per each primary tumor or node. A customized Pyradiomics version 2.1.2 pipeline was applied for preprocessing, generation of derivative images, and feature extraction [28]. A comprehensive feature list is included in Supplementary Table 1.

Assessment of radiomics feature stability

Prior studies demonstrated limited robustness of certain radiomic features with manual segmentation inconsistencies due to intra- and inter-observer variabilities [21, 26, 31, 32]. Applying multiple delineation analysis to select stable radiomics features, we selected a random sample of 50 patients stratified by dataset from pooled TCIA cohorts—excluding the MAASTRO cohort, which was reserved for

external validation. Observer #1 re-segmented all primary tumors and two randomly selected metastatic nodes per patient > 2 months after creating the original segmentations (validation by a neuroradiologist (SP)), and an additional, independent observer #2 created a third set of segmentations validated by another independent neuroradiologist (AM) with greater than 12 years of experience in head and neck radiology. A two-way random effects, absolute agreement, single rater/measurement intra-class correlation coefficient (ICC) was calculated for each feature to assess inter-rater agreement, and a two-way mixed effects, absolute agreement, single rater ICC to assess intra-rater agreement [33]. Features with an ICC 95% confidence interval (CI) lower bound < 0.8 in intra-observer or inter-observer assessment were excluded from analysis. Our ICC cutoff was set at 0.8 based on prior publications using this cutoff in analysis of feature stability and reproducibility [31, 32]. The ICC were separately calculated for primary tumors, lymph nodes, and the combined set of tumors and nodes. Segmentation set #1 from observer #1 and segmentations from observer #2 were used to assess inter-rater agreement, and segmentation sets #1 and #2 from observer #1 were utilized for intra-rater assessment. All analyses in this study were performed using various statistical packages in R version 3.6.0 in conjunction with custom-written code [34]. The R "psych" package [35] was used for ICC calculation.

Data allocation

Figure 2 depicts the data allocation, feature selection, and machine-learning classification pipeline. To avoid overfitting, and improve model generalizability, we allocated separate datasets for model development, independent, and external validation. The MAASTRO dataset ($n = 29$) contains PET and contrast-enhanced CT scans, and thus, only the PET scans were used for external validation of final models. The remaining cohorts were pooled, and a random sample of 80 patients (approximately 20%)—stratified by HPV status—was selected for independent validation. This set and the MAASTRO cohort were kept fully isolated from model development. The remaining subjects ($n = 326$) were used for model development, hyperparameter optimization, cross-validation, and final model training. The 50 multiple delineation cases were assigned to the training set.

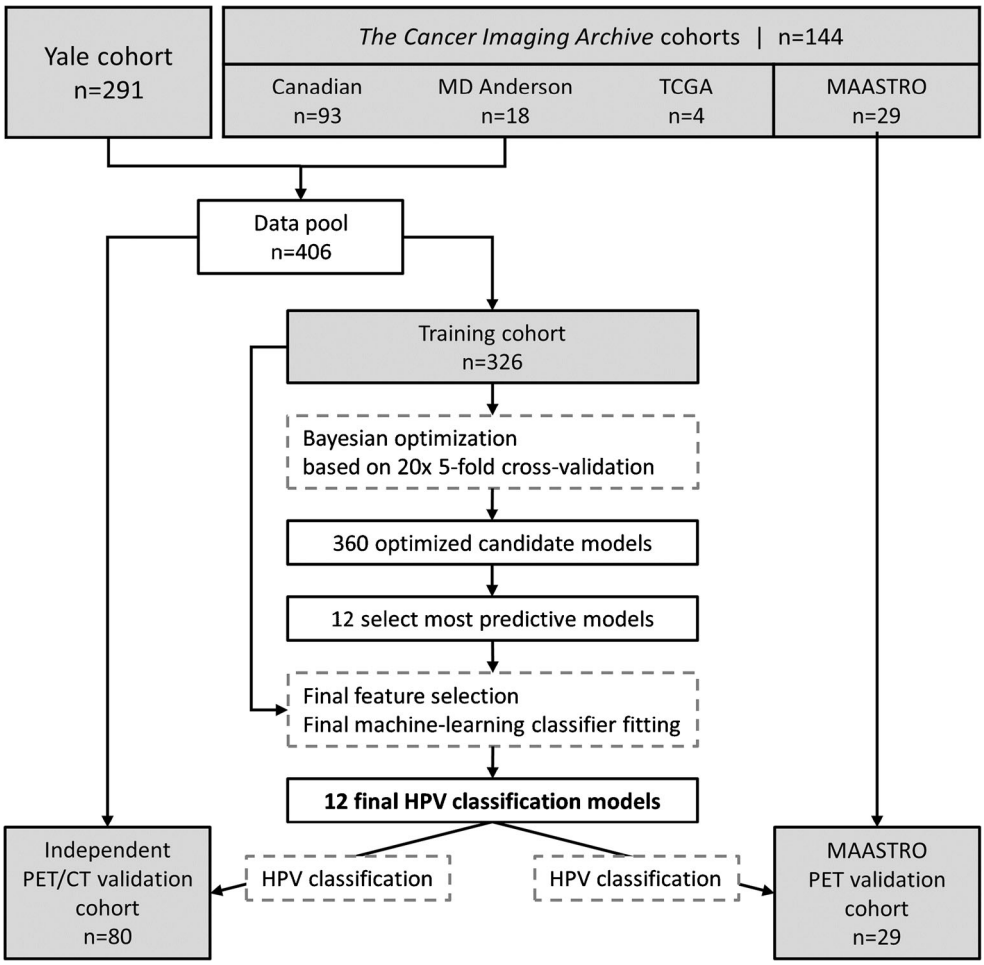


Fig. 2 Data allocation and analysis pipeline—summarizing the allocation strategy of Yale / TCIA data into training, independent validation, and external validation cohorts as well as the model building process from Bayesian optimization to final model generation and validation

Candidate HPV classification models

To develop optimized classification models, we compared all combinations of six feature selection, and five machine-learning classification algorithms (30 pairs). Feature selection and machine-learning classification algorithms are listed in Table 1, and a detailed description is provided in the supplementary methods section. We defined four VOI sources of radiomics input for classification models: (1) primary tumor lesion, (2) each individual lymph node, (3) consensus of primary tumor and all metastatic nodes, and (4) consensus of all metastatic nodes. The algorithmic approach for generation of "virtual" consensus VOI combining primary tumors and nodal lesions was described by Yu et al. [9]. Radiomics features originated from three imaging modalities (i.e., PET, CT, PET and CT). Subsequently, 360 candidate HPV classification models were devised (6 feature selection methods × 5 machine-learning classification algorithms × 4 VOI sources of radiomics input × 3 imaging modalities).

Cross-validation and Bayesian hyperparameter optimization

A framework applying 20 repeats of stratified 5-fold cross-validation (with HPV status as strata) was designed for hyperparameter optimization and internal performance validation of candidate models (Supplementary Fig. 1) [36–39]. Classification performance was measured in the validation folds and averaged across all 100 permutations to generate stable outcome results. In each cross-validation round, feature standardization and selection were performed on the training folds, followed by machine-learning classifier training. This approach avoids information leakage from training to validation folds, reduces overfitting, and therefore provides accurate estimates of final model performance in independent validation sets. For models using radiomics of individual nodes as the VOI source, all lymph nodes corresponding to each patient were allocated to the same fold to avoid training and validation on the same scan.

Bayesian optimization [40] was implemented to feed sets of hyperparameters to the cross-validation framework, evaluate their performance, and iteratively optimize and feed new hyperparameter sets. The machine-learning classifiers' upper and lower hyperparameter bounds and tuning repetition counts are provided in Supplementary Table 2. The "rBayesianOptimization" package for R was used [41].

After tuning, we applied the cross-validation framework one last time per each candidate model with optimized hyperparameters to measure the internally validated cross-validation performance and to establish a prediction probability cutoff for binary HPV classification: In each validation fold, the cutoff closest to achieving balanced sensitivity and specificity for detection of HPV positivity was determined, and all validation fold cutoffs were averaged to determine a model's global cutoff.

Final model training and validation

For each combination of source VOI and imaging modality, the highest performing candidate model ($4 \times 3 = 12$ models) was selected. Using the full training dataset, feature standardization and feature selection were performed followed by fitting of machine-learning classifiers with optimized hyperparameters. Subsequently, final models were applied in the independent validation set, and PET-only models were additionally applied in the MAASTRO external validation set.

Performance metrics

The receiver operating characteristic (ROC) area under the curve (AUC) was used to evaluate cross-validation and final model performance. Mean and standard deviation (SD) of the AUC

Table 1 List of feature selection methods and machine-learning classifiers

Feature selection method	Abbreviation
Hierarchical clustering	HClust
Minimum redundancy maximum relevance filter	MRMR
No feature selection applied	noFS
PCA-based feature selection	PCA
Pearson correlation-based redundancy reduction combined with a mutual information maximization filter	pMIM
Logistic regression with RIDGE regularization adapted for feature selection	RIDGE
Machine-learning classifiers	
Logistic regression with elastic net regularization	ElNet
Naive Bayes classifier	NBayes
Random forest	RF
Support vector machine with radial kernel	SVM
XGBoost	XGB

distribution over validation folds from repeated cross-validation are reported. DeLong's test was used to compare paired AUCs, and to calculate p values and AUC 95% confidence intervals (CI) [42]. We used the "pROC" package in R to compute, compare, and analyze AUC metrics [43]. In the independent validation set, confusion matrix parameters (including recall, precision, F1-score, and geometric mean of precision and recall) were additionally calculated based on each model's global cutoff. Finally, we calculated "importance scores" for all selected features or feature clusters to assess their impact in final model predictions as detailed in the supplementary methods.

Results

Characteristics of patients, primary tumor lesions, and metastatic nodes

A total of 435 patients with OPSCC met our inclusion criteria—291 (66.9%) from Yale and 144 (33.1%) from TCIA. Of these, 315 (72.4%) were high-risk HPV-associated, and 120 (27.6%) were HPV-negative. Table 2 depicts demographics and basic imaging characteristics of different datasets. Detailed characteristics of Yale and TCIA datasets are included in Supplementary Table 3, and prior publications [15, 18, 19, 21], respectively. A total of 2248 segmentations were generated, corresponding to PET and CT representations of 435 primary tumors and 741 metastatic cervical lymph nodes.

Radiomics feature stability

Following image preprocessing and feature extraction, multiple delineation analysis was performed on 50 primary tumors, 65 randomly selected lymph nodes, and the combined set of all lesions to assess feature robustness and identify stable features. The ICC-based feature stability assessment, including numbers of retained features, is summarized in Table 3. PET features exhibited greater stability in primary tumors, whereas CT features were more robust in lymph nodes. All feature sets showed similar reproducibility in inter- and intra-rater testing.

Cross-validation and model optimization

Figure 3 depicts a heatmap summary of cross-validation performance for all candidate models. Of 360 candidate models, the top 12 models with the greatest averaged AUC in cross-validation were selected as final models and are listed in Table 4—one model for each combination of VOI source and imaging modality. Across all candidates, those models using XGBoost classifiers were most frequently selected for final validation (9/12 models). Final models most frequently utilized hierarchical clustering and logistic regression with RIDGE regularization for feature selection (Table 4).

Independent and external validation of optimized models

Table 4, Fig. 4, and Supplementary Fig. 2 depict final models' performance and ROC curves. Overall, models combining PET and CT radiomics features achieved higher classification performance compared to single-modality PET- or CT-based models. Top performing optimized models generalized well to independent and external validation cohorts, with AUC performances well within 2 SD of their respective averaged AUC in cross-validation.

Overall, primary tumor radiomics signatures exhibited the highest discriminative ability in both cross-validation and independent validation, with the XGBoost classifier using RIDGE feature selection from combined PET/CT radiomics predicting HPV status with an averaged AUC of 0.78 (SD = 0.06) in cross-validation, and AUC of 0.77 (95% CI = 0.65–0.89) in independent validation.

Models using radiomics features from individual nodes or consensus of all nodes per each patient yielded lower performance than primary tumor models (Table 4). Nevertheless, combining features from all metastatic nodes in each patient appears to be advantageous, given the performance superiority of models using consensus VOI over individual node features. In the external PET-only MAASTRO validation cohort, the XGBoost classifier with hierarchical feature clustering from consensus sets of all lymph nodes yielded an AUC of 0.73 (95% CI = 0.47–0.94).

The consensus VOI model derived from primary tumor lesions and metastatic lymph nodes yielded the highest classification performance in the external PET-only MAASTRO cohort, with the XGBoost classifier achieving an AUC of 0.83 (95% CI = 0.68–0.98).

Of note, the AUC classification performance of models utilizing different imaging modalities within one VOI source, or utilizing different VOI in paired radiomics datasets (primary tumor and consensus of tumor and all nodes), was not significantly different in independent validation ($p > 0.05$, DeLong's test).

Supplementary Tables 4.1 to 4.4 list feature importance scores for 12 final models. Radiomics features representing histogram metrics of CT attenuation and heterogeneity in CT and PET intensities were ranked first among different models.

Table 2 Patients' characteristics

	Training/cross-validation cohort	Independent validation cohort	External validation cohort (MAASTRO—PET only)	*p* value training vs. independent	*p* value training vs. external
Number of patients	326	80	29		
Number of lymph nodes	518	148	75		
Sex, n (%)				0.76	1.00
M	272 (83.4%)	65 (81.3%)	24 (82.8%)		
F	54 (16.6%)	15 (18.8%)	5 (17.2%)		
Age [years], mean (SD)	60.59 (9.32)	60.53 (9.06)	60.86 (7.06)	0.78	0.79
HPV status, n (%)				1.00	< 0.001[b]
Positive	244 (74.8%)	60 (75.0%)	11 (37.9%)		
Negative	82 (25.2%)	20 (25.0%)	18 (62.1%)		
T stage[a], n (%)				0.54	0.99
T1	46 (14.1%)	6 (7.5%)	5 (17.2%)		
T2	120 (36.8%)	35 (43.8%)	10 (34.5%)		
T3	106 (32.5%)	24 (30.0%)	8 (27.6%)		
T4	54 (16.6%)	15 (18.8%)	6 (20.7%)		
N stage[a], n (%)				0.64	0.23
N0	64 (19.6%)	14 (17.5%)	5 (17.2%)		
N1	143 (43.9%)	41 (51.3%)	9 (31.0%)		
N2	111 (34.0%)	24 (30.0%)	15 (51.7%)		
N3	8 (2.5%)	1 (1.3%)	0 (0.0%)		
M stage[a], n (%)				0.68	0.52
M0	312 (95.7%)	78 (97.5%)	29 (100.0%)		
M1	14 (4.3%)	2 (2.5%)	0 (0.0%)		
Overall stage[a], n (%)				0.48	0.009[b]
I	110 (33.7%)	31 (38.8%)	6 (20.7%)		
II	100 (30.7%)	23 (28.8%)	6 (20.7%)		
III	53 (16.3%)	11 (13.8%)	4 (13.8%)		
IV	63 (19.3%)	15 (18.8%)	13 (44.8%)		
Included lymph nodes/patient, range	0–8	0–7	0–9		
PET[c], mean (SD)					
Slice thickness [mm]	3.36 (0.38)	3.35 (0.37)	3[d]		
In-plane pixel spacing [mm]	4.33 (0.92)	4.34 (0.93)	2.67[d]		
In-plane image matrix [n x n]	150.56 (60.58) x idem	150.78 (61.63) x idem	256 × 256[d]		
CT[c], mean (SD)			n/a		
Slice thickness [mm]	3.15 (0.53)	3.20 (0.46)			
In-plane pixel spacing [mm]	1.11 (0.19)	1.12 (0.19)			
In-plane image matrix [n x n]	512 × 512	512 × 512			

[a] UICC/AJCC 8th edition staging manuals TNM/overall stage [4, 5]

[b] The external validation cohort (MAASTRO with PET scans only) had a larger proportion of HPV-negative and thus stage IV patients compared to training/cross-validation cohort

[c] Values are from original pretreatment images before preprocessing

[d] Identical imaging characteristics in entire dataset

Discussion

In many institutions, whole body PET/CT is the mainstay of initial staging in patients with HNSCC, delineating the extent of primary tumors, identifying metastatic cervical nodes, and screening for distant metastasis. Hidden bioimaging patterns related to tumor shape, texture, and hypermetabolism can provide additional information regarding the tumor molecular subtype, such as HPV status in OPSCC. In this study, we devised, validated, and comparatively analyzed PET and CT radiomics signatures for prediction of HPV status in OPSCC primary tumors and metastatic cervical lymph nodes. We

Table 3 Inter- and intra-class correlation coefficient (ICC) for feature stability assessment

	Multiple delineation lesions (n)	Averaged Inter-rater ICC (SD)	Averaged Intra-rater ICC (SD)	n (%) retained features (intra- and inter-rater ICC lower CI bound ≥ 0.8)
Primary tumors	50	PET: 0.92 (0.12)	PET: 0.91 (0.11)	PET: 751 (72.4%)
		CT: 0.86 (0.16)	CT: 0.89 (0.13)	CT: 586 (54.7%)
Lymph nodes	65	PET: 0.86 (0.17)	PET: 0.85 (0.18)	PET: 582 (56.1%)
		CT: 0.90 (0.17)	CT: 0.90 (0.17)	CT: 770 (74.3%)
All lesions	115	PET: 0.88 (0.15)	PET: 0.87 (0.16)	PET: 651 (62.8%)
		CT: 0.91 (0.13)	CT: 0.93 (0.11)	CT: 854 (82.4%)

Mean (SD) ICC of PET and CT radiomics feature sets in primary tumor, lymph node, and combined sets in inter- and intra-rater analysis, and number of retained features in each feature set

showed that PET-based radiomics signatures yielded similar classification performance to CT-based models, with a trend suggesting improved predictive performance when combining PET and CT radiomics features. However, paired models´ performance was not significantly different in independent validation. Overall, radiomics signatures based on primary tumor lesions had higher predictive performance compared to those based on metastatic cervical lymph nodes; however,

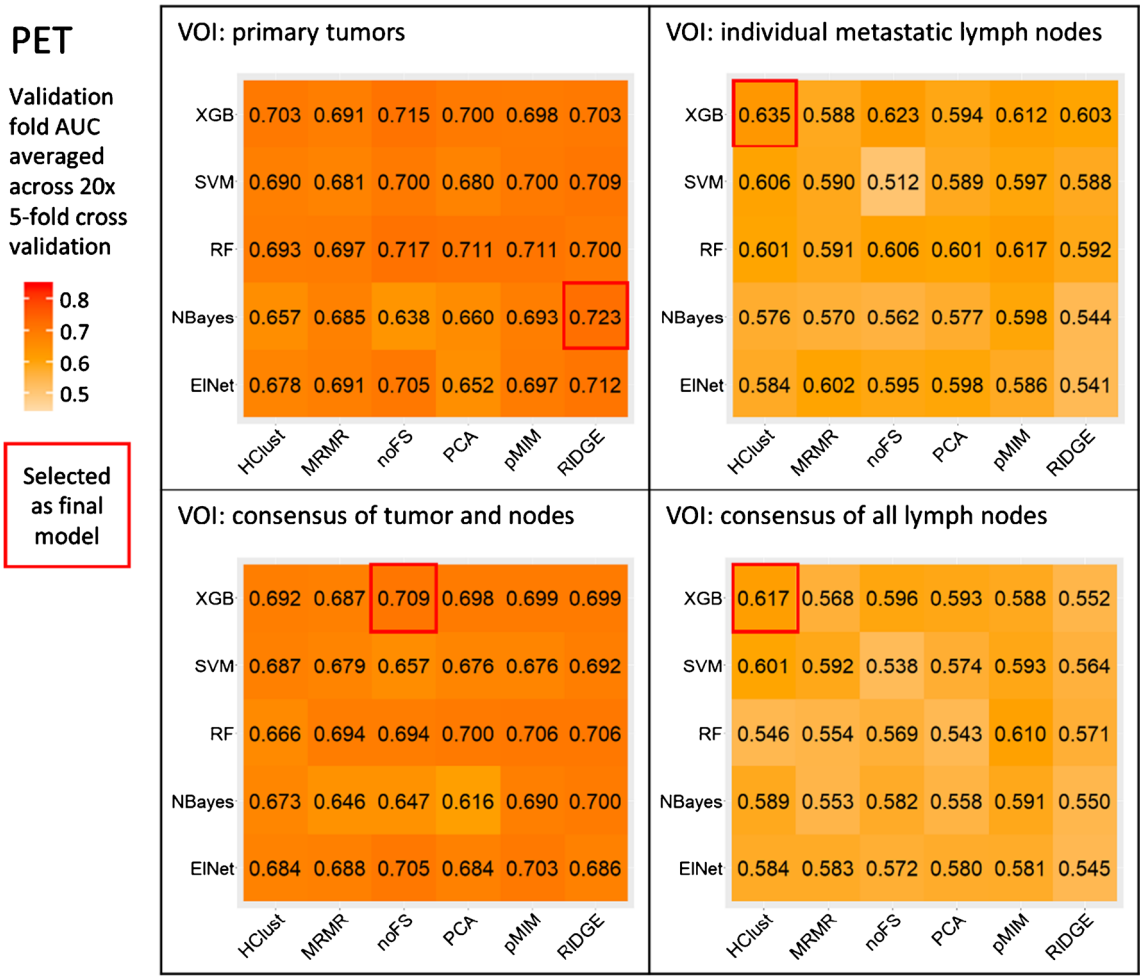


Fig. 3 Candidate model performance heatmap. The values represent the averaged AUC across 100 permutations from 20 repeats of (HPV status stratified) 5-fold cross-validation. A total of 360 candidate classification models were devised: 6 feature selection methods × 5 machine-learning classification algorithms × 4 VOI sources of radiomics input × 3 imaging modalities. The 12 highest performing candidate models (one per 4 VOI sources of radiomics input × 3 imaging modalities) were used for independent and external validation (Table 4)

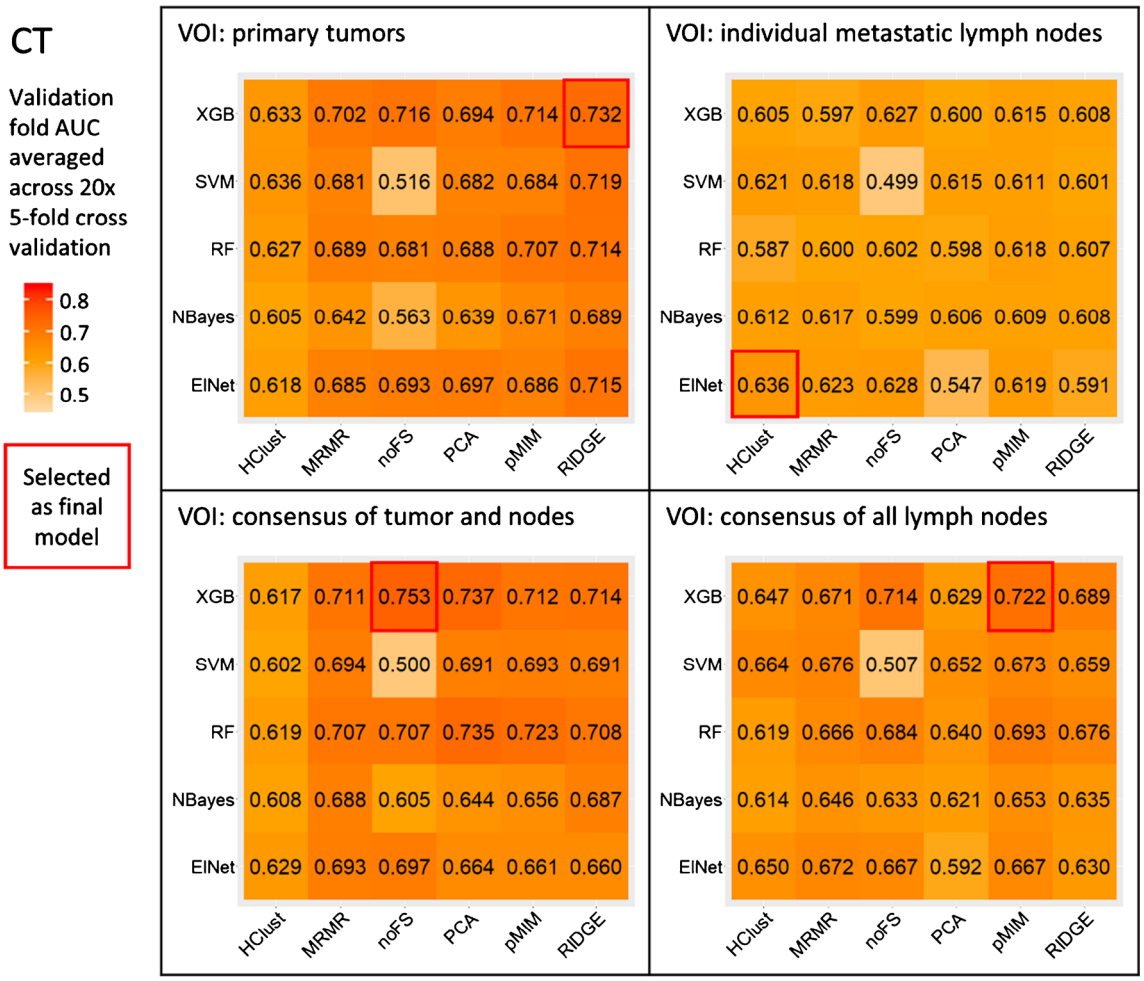


Fig. 3 (continued)

models using consensus VOI of all nodes had better classification performance than individual nodes. Nevertheless, in the external validation cohort (MAASTRO), the PET-based radiomics signature from the consensus VOI of primary tumor and metastatic nodes achieved the highest AUC of 0.83 (95% CI 0.68–0.98) in prediction of HPV status, suggesting a potential benefit from combining tumoral and nodal radiomics features.

Assessment of HPV status in OPSCC plays a crucial role in cancer staging and treatment planning [44]. The Guideline from the College of American Pathologists recommends HPV-specific tests—such as DNA ISH or PCR—in certain p16 positive cervical nodes and multisite primary tumors [6]. Radiomics-based HPV biomarkers could substitute such confirmatory second-line tests. In addition, PET/CT biomarkers suggestive of HPV-associated cervical lymphadenopathy in patients presenting with cancer of unknown primary origin can initiate endoscopic examinations and rigorous tissue sampling for identification of potential OPSCC origin. In the future, our approach should be refined and validated using larger cohorts, with strict standardization of image acquisition techniques. Quantitative imaging biomarkers must conform to the same requirements as the conventional biomarkers, which they are designed to substitute or supplement before translation to clinical application can be considered.

In this study, we applied PET-guided segmentation of primary tumors and metastatic nodes, wherein all lesions were first segmented on PET images, and segmentations were then overlaid and adapted onto the co-registered non-contrast CT scans. Identification and delineation of HNSCC lesions on CT scans—especially in the absence of contrast administration—can be challenging. Indeed, prior studies demonstrated benefits of PET-guided radiotherapy planning over CT-guidance [45, 46]. On the other hand, non-contrast CT scans provide standardized tissue density values for evaluation of tumor texture heterogeneity, devoid of variability inherent to contrast-enhanced CT scan acquisition with variable contrast accumulation which can affect radiomics feature measurement [47]. PET-guided segmentation may enable precise non-contrast

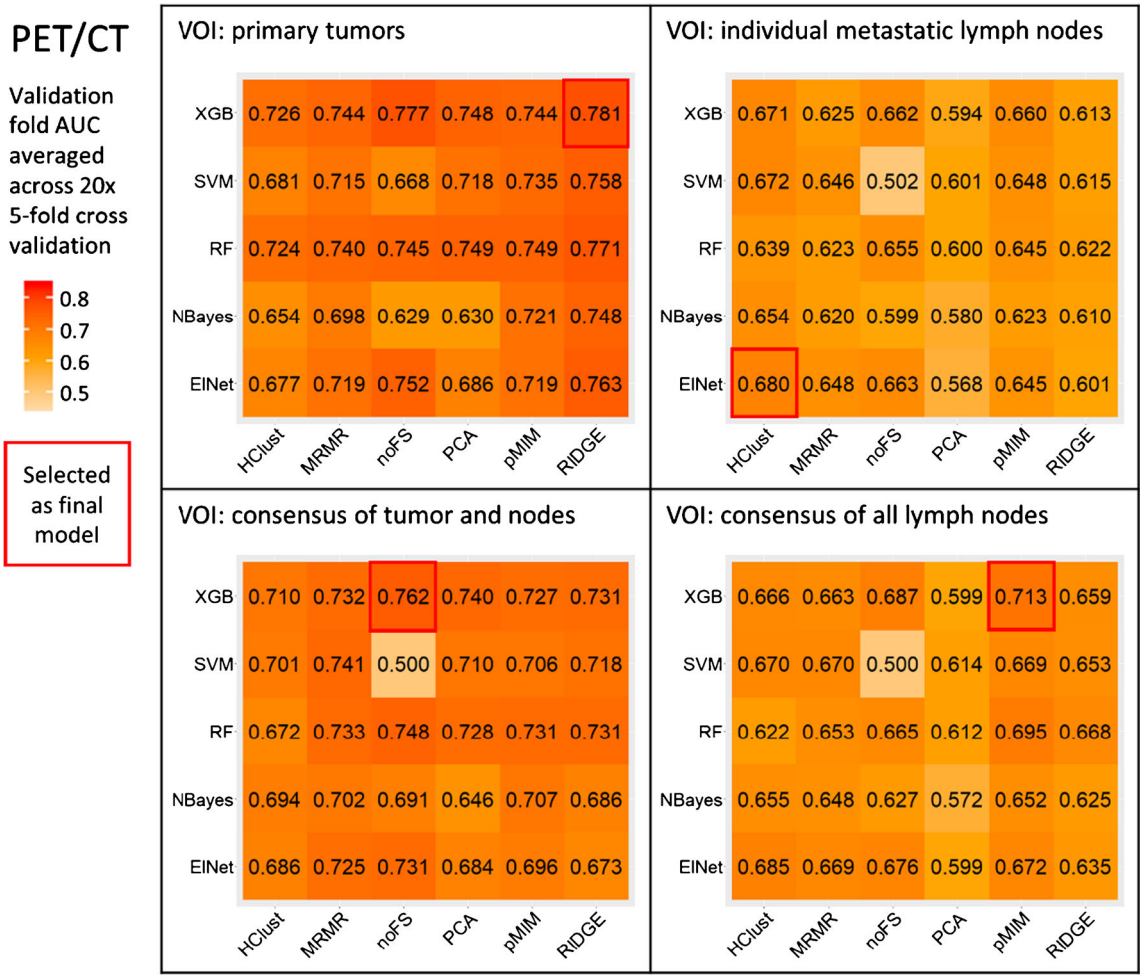


Fig. 3 (continued)

CT radiomics analysis, overcoming inherent segmentation difficulties in the absence of contrast administration.

Our methodology may also guide future research in devising PET/CT radiomics biomarkers for HNSCC. Testing an array of feature selection methods and machine-learning classifiers using cross-validation is an effective and easily interpretable approach to determine the most promising models for subsequent validation in independent and external datasets. Particularly, this is the first radiomics-based HPV classification study to apply state-of-the-art machine-learning classifiers in a comparative analysis, where 9 out of 12 selected models utilized the Extreme Gradient Boosting ("XGBoost") algorithm [48], outperforming similar algorithms in recent machine-learning competitions [49]. In addition, we applied automated hyperparameter optimization of machine-learning classifiers to avoid bias introduced by manual tuning, and augment model comparability across imaging modalities, VOIs, and algorithmic approaches. Tuning of machine-learning classifiers resembles optimization of "black box" functions and may be approached with purpose-built algorithms like Bayesian Optimization [40]. Future studies may similarly automate tuning when seeking to compare models in an unbiased fashion. In addition, similar averaged performance in cross-validation (20 repetitions of 5-fold cross-validation—i.e., 100 test folds), independent, and external validation cohorts confirms the stability of our results.

A shortcoming of previous radiomics studies, including HPV classification studies, was relying on homogeneous single center imaging data and/or absence of external validation [8–11, 50]. Such study design limits generalizability and applicability [50]. Our results based on a multi-institutional cohort confirmed the generalizability of our radiomics signatures and their potential as universal non-invasive biomarkers for molecular phenotyping. We acquired a multi-institutional and multi-national cohort from four different TCIA collections and our own center, and applied a three-dataset approach for model building and validation. The bulk of data was pooled and divided into training and independent validation datasets, and the PET-only MAASTRO cohort was held out for additional external

Table 4 Final models' performance in cross-validation, independent, and external validation

Source VOI	Imaging modality	Machine-learning, feature selection	Cross-validation AUC (SD)[a]	Independent validation					External validation
				AUC (95% CI)[b]	Recall[c]	Precision[c]	F1-score[c]	Geometric mean[c,d]	AUC (95% CI)
Primary tumors	PET/CT	XGB, RIDGE	0.78 (0.06)	0.77 (0.65–0.89)	0.73	0.88	0.80	0.80	
	PET	NBayes, RIDGE	0.72 (0.06)	0.70 (0.55–0.85)	0.77	0.84	0.80	0.80	0.75 (0.55–0.94)
	CT	XGB, RIDGE	0.73 (0.06)	0.76 (0.63–0.88)	0.70	0.91	0.80	0.80	
Individual metastatic lymph nodes	PET/CT	ElNet, HClust	0.68 (0.07)	0.61 (0.49–0.73)	0.58	0.81	0.68	0.69	
	PET	XGB, HClust	0.63 (0.07)	0.60 (0.50–0.71)	0.52	0.86	0.64	0.67	0.58 (0.45–0.71)
	CT	ElNet, HClust	0.64 (0.08)	0.59 (0.48–0.70)	0.52	0.80	0.63	0.64	
Consensus of tumor and nodes	PET/CT	XGB, noFS	0.76 (0.06)	0.75 (0.62–0.87)	0.63	0.88	0.74	0.75	
	PET	XGB, noFS	0.71 (0.07)	0.71 (0.58–0.85)	0.67	0.88	0.76	0.77	0.83 (0.68–0.98)
	CT	XGB, noFS	0.75 (0.07)	0.67 (0.54–0.80)	0.58	0.81	0.68	0.69	
Consensus of all lymph nodes	PET/CT	XGB, pMIM	0.71 (0.07)	0.60 (0.42–0.78)	0.65	0.89	0.76	0.76	
	PET	XGB, HClust	0.62 (0.08)	0.65 (0.47–0.84)	0.62	0.91	0.74	0.75	0.73 (0.47–0.94)
	CT	XGB, pMIM	0.72 (0.08)	0.61 (0.44–0.77)	0.60	0.89	0.71	0.73	

Among 360 candidate models, the 12 highest performing models (one per 4 VOI sources of radiomics input × 3 imaging modalities) were used for independent and external validation (Fig. 3)

[a] The average (SD) of the AUC values in validation folds across 100 cross-validation permutations is reported

[b] DeLong's test for paired model AUC comparison found no significant performance difference among paired models in independent validation [42]

[c] The HPV-associated class was defined as the positive instance

[d] Geometric mean of precision and recall

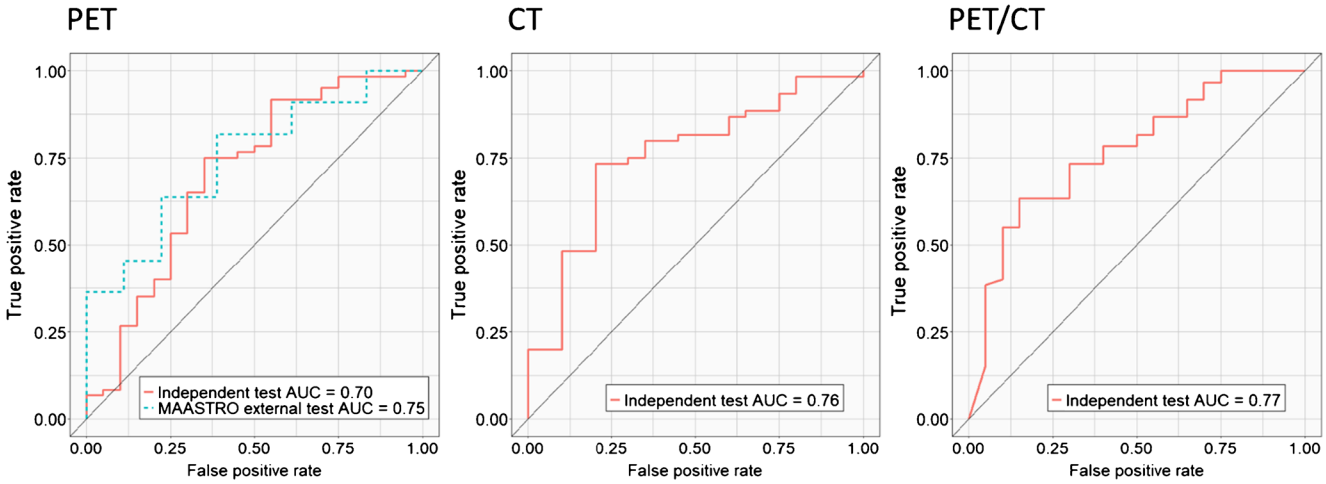


Fig. 4 ROC curve analysis depicting independent validation of HPV classifiers using radiomics features from primary tumor lesions (Table 4—primary tumor source VOI)

validation. All PET and CT models yielded similar performance in cross-validation, independent, and external validation, confirming model generalizability and suggesting robustness across institutions, heterogenous imaging protocols, and scanning equipment. Our three-dataset-approach may serve as a paradigm for future radiomics studies seeking to expand their scope beyond exploratory investigation. Finally, this approach may pave the way for application of PET radiomics in treatment response prediction and molecular pattern detection beyond HPV.

The feature importance scores in Supplementary Tables 4.1 to 4.4 provide insight about PET/CT radiomics markers with the highest impact on model predictions. Notably, features representative of heterogeneity in tumor soft tissue on CT and metabolic activity on PET, as well as high histogram percentile CT density—which may represent tissue hypercellularity—had the highest impact on model decisions. These may be reflective of histological differences between the (commonly) keratinizing HPV-negative OPSCC, versus the HPV-associated subtype, which is usually associated with poorly differentiated histologic morphology, minimal interstitial space, and basaloid cells with scant cytoplasm [44, 51].

To address the class imbalance in our cohort—given the 4:1 ratio of HPV-associated to HPV-negative OPSCC subjects—we used the AUC of ROC as our primary performance metric, which in contrary to classification accuracy is not inherently susceptible to class imbalance.

One limitation of our study is the lack of uniform testing for HPV association, which was ascertained using institutional protocols in TCIA cohorts. Ideally, a full set of p16 and select HPV-specific tests would be available for comparative analysis. Also, the external validation cohort was small ($n = 29$ patients and $n = 75$ lymph nodes) and lacked CT scans; thus, future validation in larger external PET/CT cohorts is crucial. Metastatic involvement of cervical lymph nodes was determined by expert read of PET/CT studies, but without pathologic correlation in all involved nodes. Finally, inclusion of demographic information can improve the discriminative accuracy of radiomics models for prediction of HPV status in OPSCC, as Ravanelli et al. [52] showed that addition of smoking status to MRI diffusion metrics increased model performance. However, some demographic information may not be reliably available in all patients, and some are prone to cultural and geographical variations, which may impose limitations to the generalizability and applicability of models; thus, we limited the scope of our study to develop an objective biomarker solely using quantitative radiomics features extracted from clinical PET/CT scans.

Conclusion

Using independent and external datasets, we devised and validated non-invasive HPV biomarkers for molecular OPSCC subtyping. We demonstrate discriminative ability of both PET and CT classifiers, with the highest predictive performance achieved when PET and CT radiomic features were combined. Furthermore, in our external validation, we found potential added predictive performance when PET radiomic markers derived from the primary tumor and metastatic cervical nodes were combined. Applying a rigorous test to identify stable radiomics features, comparative analysis of different feature selection methods, and machine-learning classifiers, and benefitting from a multi-institutional cohort, our models demonstrate reliable generalizability in scans obtained at different institutions. Of note, the performance of radiomics models for distinction of HPV association remains too low to substitute tissue sampling; however, pending validation in larger cohorts

and attainment of sufficient diagnostic accuracy in the future, such non-invasive HPV biomarkers may potentially supplement tissue sampling in equivocal instances, or provide preliminary assessment before the biopsy.

Code availability Our code is publicly available from our "OPSCC-Radiomics" GitHub-repository (https://github.com/nafets200/OPSCC-Radiomics).

Compliance with ethical standards

Conflict of interest SPH has nothing to disclose. AM has nothing to disclose. TZ has nothing to disclose. PB has nothing to disclose. CR has nothing to disclose. KS has nothing to disclose. RF has acted as speaker and consultant for GE Healthcare and has a research agreement (beta tester) and support from GE Healthcare. RF is also a founder and stockholder of 4intelligent Inc., and a clinical research scholar (chercheur-boursier clinician) supported by the Fonds de recherche en santé du Québec (FRQS). ASK has nothing to disclose. BHK has nothing to disclose. BLJ has nothing to disclose. MLP has nothing to disclose. BB has nothing to disclose. SP has nothing to disclose.

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the corresponding institutional research committees and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all participants in prospective trials. The Yale University ethics committee waived consent for retrospective analysis of patients' information under IRB protocol #2000024295 ("Imaging biomarkers for tumor classifications in brain and head/neck tumors using radiomics and machine-learning algorithms").

Abbreviations AJCC, American Joint Committee on Cancer; AUC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; HNSCC, head-and-neck squamous cell carcinoma; HPV, human papillomavirus; HU, Hounsfield unit; ICC, inter-/intra-class correlation coefficient; ISH, in situ hybridization; LoG, Laplacian of Gaussian; OPSCC, oropharyngeal squamous cell carcinoma; PCR, polymerase chain reaction; PET, [18F]fluorodeoxyglucose positron emission tomography; GTV, gross tumor volume; ROC, receiver operating characteristic; SD, standard deviation; TCIA, The Cancer Imaging Archive; UICC, Union for International Cancer Control; VOI, volume of interest

References

1. Gillison ML, Chaturvedi AK, Anderson WF, Fakhry C. Epidemiology of human papillomavirus-positive head and neck squamous cell carcinoma. J Clin Oncol. 2015;33:3235–42. https://doi.org/10.1200/JCO.2015.61.6995.
2. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010;363:24–35. https://doi.org/10.1056/NEJMoa0912217.
3. Benson E, Li R, Eisele D, Fakhry C. The clinical impact of HPV tumor status upon head and neck squamous cell carcinomas. Oral Oncol. 2014;50:565–74. https://doi.org/10.1016/j.oraloncology.2013.09.008.
4. AJCC Cancer staging manual (8th edition): Springer International Publishing; 2017.
5. TNM Classification of Malignant Tumours, 8th Edition. Wiley-Blackwell: Union for International Cancer Control; 2016.
6. Lewis JS Jr, Beadle B, Bishop JA, Chernock RD, Colasacco C, Lacchetti C, et al. Human papillomavirus testing in head and neck carcinomas: guideline from the College of American Pathologists. Arch Pathol Lab Med. 2018;142:559–97. https://doi.org/10.5858/arpa.2017-0286-CP.
7. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278:563–77. https://doi.org/10.1148/radiol.2015151169.
8. Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. Computed tomography radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. Int J Radiat Oncol Biol Phys. 2017;99:921–8. https://doi.org/10.1016/j.ijrobp.2017.06.002.
9. Yu K, Zhang Y, Yu Y, Huang C, Liu R, Li T, et al. Radiomic analysis in prediction of human papilloma virus status. Clin Transl Radiat Oncol. 2017;7:49–54. https://doi.org/10.1016/j.ctro.2017.10.001.
10. Mungai F, Verrone GB, Pietragalla M, Berti V, Addeo G, Desideri I, et al. CT assessment of tumor heterogeneity and the potential for the prediction of human papillomavirus status in oropharyngeal squamous cell carcinoma. Radiol Med. 2019;124:804–11. https://doi.org/10.1007/s11547-019-01028-6.
11. Zhu Y, Mohamed ASR, Lai SY, Yang S, Kanwar A, Wei L, et al. Imaging-genomic study of head and neck squamous cell carcinoma: associations between radiomic phenotypes and genomic mechanisms via integration of the Cancer Genome Atlas and the Cancer Imaging Archive. JCO Clin Cancer Inform. 2019;3:1–9. https://doi.org/10.1200/CCI.18.00073.
12. Elhalawani H, Mackin D, Ger RB, Lin T, Mohamed AS, Rock C, et al. FDG-PET imaging-derived radiomics correlates of human papillomavirus status: connecting the dots in the oropharyngeal cancer biology, metabolism, and imaging interplay. International Journal of Radiation Oncology Biology Physics. 2018;102:e262. doi:https://doi.org/10.1016/j.ijrobp.2018.07.856.
13. Cheng NM, Chang JT, Huang CG, Tsan DL, Ng SH, Wang HM, et al. Prognostic value of pretreatment (1)(8)F-FDG PET/CT and human papillomavirus type 16 testing in locally advanced oropharyngeal squamous cell carcinoma. Eur J Nucl Med Mol Imaging. 2012;39:1673–84. https://doi.org/10.1007/s00259-012-2186-9.
14. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57. https://doi.org/10.1007/s10278-013-9622-7.
15. Vallieres M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts H, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7:10117. https://doi.org/10.1038/s41598-017-10371-5.
16. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Khaouam N, et al. Data from head-neck-PET-CT. The Cancer Imaging Archive; 2017.
17. Grossberg A, Mohamed A, Elhalawani H, Bennett W, Smith K, Nolan T, et al. Data from Head and Neck Cancer CT Atlas. The Cancer Imaging Archive; 2017.
18. Grossberg AJ, Mohamed ASR, Elhalawani H, Bennett WC, Smith KE, Nolan TS, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci Data. 2018;5:180173. https://doi.org/10.1038/sdata.2018.173.
19. Zuley ML, Jarosz R, Kirk S, Lee Y, Colen R, Garcia K, et al. Radiology data from the Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma [TCGA-HNSC] collection. The Cancer Imaging Archive; 2016.

20. Wee L, Dekker A. Data from head-neck-Radiomics-HN1 [data set]. The Cancer Imaging Archive; 2019.
21. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006. https://doi.org/10.1038/ncomms5006.
22. Ger RB, Craft DF, Mackin DS, Zhou S, Layman RR, Jones AK, et al. Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis. Comput Med Imaging Graph. 2018;69:134–9. https://doi.org/10.1016/j.compmedimag.2018.09.002.
23. 4. Definition of volumes. Journal of the ICRU. 2010;10:41–53. https://doi.org/10.1093/jicru_ndq009.
24. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging. 2012;30:1323–41. https://doi.org/10.1016/j.mri.2012.05.001.
25. Britz-Cunningham SH, Millstine JW, Gerbaudo VH. Improved discrimination of benign and malignant lesions on FDG PET/CT, using comparative activity ratios to brain, basal ganglia, or cerebellum. Clin Nucl Med. 2008;33:681–7. https://doi.org/10.1097/RLU.0b013e318184b435.
26. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys. 2018;102:1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.
27. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv e-prints; 2016.
28. Pyradiomics-community. Pyradiomics Documentation Release 2.1.2. 2018.
29. Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging. 2012;3:573–89. https://doi.org/10.1007/s13244-012-0196-6.
30. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015;5:11075. https://doi.org/10.1038/srep11075.
31. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of radiomic features in [(11)C]choline and [(18)F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. Mol Imaging Biol. 2016;18:935–45. https://doi.org/10.1007/s11307-016-0973-6.
32. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52:1391–7. https://doi.org/10.3109/0284186X.2013.812798.
33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
34. R Development Core team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
35. Revelle W. psych: procedures for psychological, psychometric, and personality research. Version 1.8.12 ed. Northwestern University, Evanston, Illinois, USA; 2018.
36. Lu CF, Hsu FT, Hsieh KL, Kao YJ, Cheng SJ, Hsu JB, et al. Machine learning-based radiomics for molecular subtyping of gliomas. Clin Cancer Res. 2018;24:4429–36. https://doi.org/10.1158/1078-0432.CCR-17-3445.
37. Qiu W, Kuang H, Nair J, Assis Z, Najm M, McDougall C, et al. Radiomics-based intracranial thrombus features on CT and CTA predict recanalization with intravenous Alteplase in patients with acute ischemic stroke. AJNR Am J Neuroradiol. 2019;40:39–44. https://doi.org/10.3174/ajnr.A5918.
38. Zhou L, Zhang Z, Chen YC, Zhao ZY, Yin XD, Jiang HB. A deep learning-based radiomics model for differentiating benign and malignant renal tumors. Transl Oncol. 2019;12:292–300. https://doi.org/10.1016/j.tranon.2018.10.012.
39. Wei L, Rosen B, Vallières M, Chotchutipan T, Mierzwa M, Eisbruch A, et al. Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling. Phys Imaging Radiat Oncol. 2019;10:49–54. https://doi.org/10.1016/j.phro.2019.05.001.
40. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Proces Syst. 2012;25:2960–8.
41. Yan Y. rBayesian optimization: Bayesian optimization of hyperparameters. Version 1.1.0 ed; 2016.
42. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.
44. Payabvash S, Chan A, Jabehdar Maralani P, Malhotra A. Quantitative diffusion magnetic resonance imaging for prediction of human papillomavirus status in head and neck squamous-cell carcinoma: a systematic review and meta-analysis. Neuroradiol J. 2019;32:232–40. https://doi.org/10.1177/1971400919849808.
45. Liu F, Xi XP, Wang H, Han YQ, Xiao F, Hu Y, et al. PET/CT-guided dose-painting versus CT-based intensity modulated radiation therapy in locoregional advanced nasopharyngeal carcinoma. Radiat Oncol. 2017;12:15. https://doi.org/10.1186/s13014-016-0739-y.
46. Delouya G, Igidbashian L, Houle A, Belair M, Boucher L, Cohade C, et al. (1)(8)F-FDG-PET imaging in radiotherapy tumor volume delineation in treatment of head and neck cancer. Radiother Oncol. 2011;101:362–8. https://doi.org/10.1016/j.radonc.2011.07.025.
47. Forghani R, Savadjiev P, Chatterjee A, Muthukrishnan N, Reinhold C, Forghani B. Radiomics and artificial intelligence for biomarker and prediction model development in oncology. Comput Struct Biotechnol J. 2019;17:995–1008. https://doi.org/10.1016/j.csbj.2019.07.001.
48. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv e-prints; 2016.
49. Jin Y. Tree boosting with XGBoost—why does XGBoost win "every" machine learning competition? ; 2017.
50. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20:405–10. https://doi.org/10.3348/kjr.2019.0025.
51. Payabvash S. Quantitative diffusion magnetic resonance imaging in head and neck tumors. Quant Imaging Med Surg. 2018;8:1052–65. https://doi.org/10.21037/qims.2018.10.14.
52. Ravanelli M, Grammatica A, Tononcelli E, Morello R, Leali M, Battocchio S, et al. Correlation between human papillomavirus status and quantitative MR imaging parameters including diffusion-weighted imaging and texture features in oropharyngeal carcinoma. AJNR Am J Neuroradiol. 2018;39:1878–83. https://doi.org/10.3174/ajnr.A5792.

European Journal of Nuclear Medicine and Molecular Imaging

Supplement

for the original article

PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma

Stefan P. Haider; Amit Mahajan, MD; Tal Zeevi, MSc; Philipp Baumeister, MD; Christoph Reichel, MD; Kariem Sharaf, MD, DDS; Reza Forghani, MD, PhD; Ahmet S. Kucukkaya; Benjamin H. Kann, MD; Benjamin L. Judson, MD; Manju L. Prasad, MD; Barbara Burtness, MD; Seyedmehdi Payabvash, MD

Correspondence:

Seyedmehdi Payabavsh, MD

Department of Radiology and Biomedical Imaging,
Yale School of Medicine
789 Howard Ave,
PO Box 208042
Fitkin CB30-E
New Haven, CT 06519
Phone: +1 (203) 214 4650
Fax: +1 (203) 737 1241
Email sam.payabvash@yale.edu

Table of Contents

1. Supplementary methods

1.1 Dimensionality reduction techniques

1.1.1 HClust – Hierarchical clustering

First, a "euclidean" feature distance matrix of all radiomics features was computed using the R "stats" package (version 3.6.0) [1] "dist" function, followed by hierarchical clustering by means of the "stats" "hclust" function using Ward clustering with Ward's clustering criterion implemented (i.e. "ward.D2" package option) [2]. The dendrogram was cut until 20 clusters remained. We extracted "meta-features" by averaging all features in any remaining cluster. Hierarchical clustering was performed with (cross-validation) training data, and only the averaging operations were applied to all cases. The 20 resulting meta features were used for training and classification.

1.1.2 MRMR – Maximum relevance minimum redundancy filter

We used the "mRMRe" package (version 2.0.9) [3] for R to perform "classical" MRMR feature selection, and the n most predictive feature were selected. n was treated as a model hyperparameter and tuned in Bayesian optimization.

1.1.3 noFS – No feature selection

Feature selection was not performed, and the classifiers were fitted on the entire feature set.

1.1.4 PCA – Principal component analysis

Singular value decomposition based principal component analysis was performed on the (cross-validation) training data using the R "stats" package (version 3.6.0) [1] "prcomp" function. We used "scheme 1" of a method proposed by Song et al. [4] to adapt principal component analysis for feature selection. 30 eigenvectors were used to perform step 2 and features were sorted according to their contribution to the feature extraction result in step 4 of the Song et al. method. Based on the ranking created in step 4, the n features contributing the most to the feature extraction result were used for classifier fitting, and n was treated as a model hyperparameter and tuned in Bayesian optimization.

1.1.5 pMIM – Pearson correlation-based redundancy reduction with mutual information filter

We applied the R "stats" package (version 3.6.0) [1] "cor" function to compute Pearson's correlation coefficient (r) for all possible radiomics feature pairs. To exclude highly correlated, "redundant" features, we at random excluded one feature from every pair exhibiting $r > 0.9$ or $r < -0.9$.
Thereafter, we applied a mutual information maximization filter to the non-redundant feature set using the "MIM" function of the "praznik" package (version 6.0.0) for R [5]. The n most predictive features were selected, and n was treated as a model hyperparameter and tuned in Bayesian optimization.

1.1.6 RIDGE – RIDGE regularized logistic regression for feature selection

Using the R "glmnet" package (version 2.0-18) [6] "cv.glmnet" function, we fit a ridge regularized logistic regression model on the (cross-validation) training data. The function's internal 10-fold cross-validation mode was used to determine the "lambda" parameter. Each feature's regression coefficient was queried from the fitted "glmnet" object at the "lambda" value maximizing the mean cross-validated area under the curve. All features were subsequently sorted in descending order of their absolute regression coefficient value, and the n highest ranked features were selected. n was treated as a model hyperparameter and tuned in Bayesian optimization.

1.2 Machine-learning classifiers

1.2.1 ElNet – Elastic net regularized logistic regression

We used the R "glmnet" package (version 2.0-18) [6] "cv.glmnet" function to fit elastic net regularized logistic regression models. The function's internal 10-fold cross-validation mode was used to determine the "lambda" parameter. New predictions were made at the "lambda" value maximizing the mean cross-validated area under the curve. The elastic net mixing parameter "alpha" was tuned in Bayesian optimization.

1.2.2 NBayes – Naïve Bayes classifier

The "naive_bayes" function of the "naivebayes" package (version 0.9.6) [7] for R was used to create the models. Laplace smoothing was not applied, the kernel was not used, and the Gaussian distribution was applied for all features. Naïve Bayes candidate models had no tunable hyperparameters, and Bayesian optimization was only used to tune the feature count, if applicable.

1.2.3 RF – Random forest classifier

We used the "randomForest" package (version 4.6-14) [8] in R to build random forest models. The "randomForest" function was configured to grow 1000 trees and perform sampling of cases with replacement. The "mtry" (number of features randomly sampled as candidates at each split) and "maxnodes" (maximum number of terminal nodes in a tree) parameters were tuned in Bayesian optimization. All other function parameters were kept in default.

1.2.4 SVM – Support vector machine classifier

The "gamma" and "cost" parameters were optimized in Bayesian optimization and a radial kernel was applied. We used the "e1071" (version 1.7-2) package [9] to implement support vector machines in R and class weights were specified to be inversely proportional to the class distribution in the training data; all other parameters were set to the default options.

1.2.5 XGB – XGBoost

We implemented extreme gradient boosting based on the "xgboost" package [10, 11] for R in tree-booster mode ("gbtree" option) and tuned the following hyperparameters using Bayesian optimization: "eta", "gamma", "max_depth", "min_child_weight", "subsample", "colsample_bytree" and "lambda". One parameter ("nrounds") was fine-tuned: While fitting the classifier on cross-validation training data, cross-validation test set class probabilities were iteratively predicted after completing each boosting round, and a test set classification area under the curve measurement was computed and stored to represent each boosting round count. Following completion of the 20x repeated 5-fold cross-validation framework, we averaged all area under the curve measurements pertaining to one boosting round count across all 100 cross-validation test set permutations. The "nrounds" setting yielding the highest mean area under the curve performance was saved as the final parameter setting. The rest of parameters were set by the default package recommendation.

1.3 Calculation of radiomics feature importance

Feature importance scores were calculated for all selected features or feature clusters utilized for final model classifier fitting on the training cohort. Each feature´s importance rank and score are reported in supplementary table 4. Importance score quantification relies on methodology specific to the classification algorithm applied; hence, importance scores are comparable only among models utilizing identical feature selection and classification methodology.

Feature importance scores for models utilizing the XGBoost classifier (XGB) were obtained through a package [10, 11] function: The “xgb.importance” function returns a score quantifying the total gain of each given feature´s splits (“Gain”), expressed as a fraction. I.e. higher fractions indicate higher feature importance.

Models using elastic net regularized logistic regression (ElNet) by default produce a feature importance score during model fitting: Each feature’s regression coefficient was queried from the fitted “glmnet” [6] object at the “lambda” value maximizing the mean cross-validated area under the curve. All features were subsequently sorted in descending order of their absolute regression coefficient value to generate the feature importance rank; and each feature´s regression coefficient is used as the importance score. I.e. higher absolute values indicate higher feature importance.

One final model utilized the Naïve Bayes classifier (NBayes), combined with RIDGE feature selection. In this case, we report feature regression coefficients derived from RIDGE feature selection as importance scores, and sorted features in descending order of their absolute regression coefficient to generate the importance rank. Please refer to the dimensionality reduction techniques section of the supplementary methods for details.

Several final models relied on hierarchical clustering (HClust) to reduce feature dimensionality. One meta feature was calculated to represent each cluster (see dimensionality reduction techniques section of the supplementary methods). Supplementary table 4 reports importance rank and score for each meta feature (denoted as “Cluster_#”). For those models which did not apply feature selection (“noFS”) we reported the ranks and scores of the 20 highest ranked radiomics features.

2. Supplementary tables

Supplementary table 1 List of extracted radiomics features

Feature Family		Feature name
First-order	1	10th percentile
	2	90th percentile
	3	Energy
	4	Entropy
	5	Interquartile Range
	6	Kurtosis
	7	Maximum
	8	Mean
	9	Mean Absolute Deviation
	10	Median
	11	Minimum
	12	Range
	13	Robust Mean Absolute Deviation
	14	Root Mean Squared
	15	Skewness
	16	Total Energy
	17	Uniformity
	18	Variance
Shape	1	Elongation
	2	Flatness
	3	Least Axis Length
	4	Major Axis Length
	5	Maximum 2D Diameter (Column)
	6	Maximum 2D Diameter (Row)
	7	Maximum 2D Diameter (Slice)
	8	Maximum 3D Diameter
	9	Mesh Volume
	10	Minor Axis Length
	11	Sphericity
	12	Surface Area
	13	Surface Area to Volume Ratio
	14	Voxel Volume
Texture - Gray Level Cooccurrence Matrix Features (glcm)	1	Autocorrelation
	2	Cluster Prominence
	3	Cluster Shade
	4	Cluster Tendency
	5	Contrast
	6	Correlation
	7	Difference Average
	8	Difference Entropy
	9	Difference Variance

	10	Informational Measure of Correlation 1
	11	Informational Measure of Correlation 2
	12	Inverse Difference
	13	Inverse Difference Moment
	14	Inverse Difference Moment Normalized
	15	Inverse Difference Normalized
	16	Inverse Variance
	17	Joint Average
	18	Joint Energy
	19	Joint Entropy
	20	Maximal Correlation Coefficient
	21	Maximum Probability
	22	Sum Average
	23	Sum Entropy
	24	Sum of Squares
Texture - Gray Level Size Zone Matrix Features (glszm)	1	Gray Level Non-Uniformity
	2	Gray Level Non-Uniformity Normalized
	3	Gray Level Variance
	4	High Gray Level Zone Emphasis
	5	Large Area Emphasis
	6	Large Area High Gray Level Emphasis
	7	Large Area Low Gray Level Emphasis
	8	Low Gray Level Zone Emphasis
	9	Size Zone Non-Uniformity
	10	Size Zone Non-Uniformity Normalized
	11	Small Area Emphasis
	12	Small Area High Gray Level Emphasis
	13	Small Area Low Gray Level Emphasis
	14	Zone Entropy
	15	Zone Percentage
	16	Zone Variance
Texture - Gray Level Run Length Matrix Features (glrlm)	1	Gray Level Non-Uniformity
	2	Gray Level Non-Uniformity Normalized
	3	Gray Level Variance
	4	High Gray Level Run Emphasis
	5	Long Run Emphasis
	6	Long Run High Gray Level Emphasis
	7	Long Run Low Gray Level Emphasis
	8	Low Gray Level Run Emphasis
	9	Run Entropy
	10	Run Length Non-Uniformity
	11	Run Length Non-Uniformity Normalized
	12	Run Percentage
	13	Run Variance

	14	Short Run Emphasis
	15	Short Run High Gray Level Emphasis
	16	Short Run Low Gray Level Emphasis
Texture - Neighboring Gray Tone Difference Matrix Features (ngtdm)	1	Busyness
	2	Coarseness
	3	Complexity
	4	Contrast
	5	Strength
Texture - Gray Level Dependence Matrix Features (gldm)	1	Dependence Entropy
	2	Dependence Non-Uniformity
	3	Dependence Non-Uniformity Normalized
	4	Dependence Variance
	5	Gray Level Non-Uniformity
	6	Gray Level Variance
	7	High Gray Level Emphasis
	8	Large Dependence Emphasis
	9	Large Dependence High Gray Level Emphasis
	10	Large Dependence Low Gray Level Emphasis
	11	Low Gray Level Emphasis
	12	Small Dependence Emphasis
	13	Small Dependence High Gray Level Emphasis
	14	Small Dependence Low Gray Level Emphasis

Complete list of Pyradiomics [12] features used in this study. Exact feature definitions are provided in ref. [13].

Supplementary table 2 Bayesian optimization parameter bounds

Machine-learning classifier	Rounds (n) [a]	Hyperparameter [b]	Lower bound	Upper bound
ElNet [6]	100	Features (n) [c]	2	30
		alpha	0	1
NBayes [7]	50	Features (n) [c]	2	30
RF [8]	150	Features (n) [c]	2	30
		mtry	2	30 [d]
		maxnodes	2	32768 [e]
SVM [9]	150	Features (K) [c]	2	30
		gamma	0	0.5
		cost	0.5	10
XGB [10, 11]	200	Features (n) [c]	2	30
		eta	0	1
		gamma	0	10
		max_depth	3	15
		min_child_weight	0	20
		subsample	0.4	1
		colsample_bytree	0.4	1
		lambda	0.5	1

[a] Per each tuned hyperparameter, 50 rounds of tuning were performed, or a maximum of 200 rounds per model.

[b] Refer to refs. in column 1 for parameter definitions

[c] Feature count was not tuned for every feature selection method. Refer to the supplementary methods for details.

[d] This bound was increased to 40 for the RF model combined with noFS (no feature selection performed) to adjust for higher data dimensionality.

[e] As a general rule, the maximum possible number of terminal nodes in a decision tree is 2 to the power of the depth. The depth was limited to 15, resulting in a maximum of 32768 terminal nodes.

The "rBayesianOptimization" package [14] for R was used to tune machine-learning classifier hyperparameters in candidate HPV classification models. The table summarizes the number of performed optimization rounds as well as the lower and upper preset parameter bounds for each hyperparameter.

Supplementary table 3 Yale cohort patients' characteristics

Variables	Yale cohort
Patients – n	291
Sex – n (%)	
M	245 (84.2 %)
F	46 (15.8 %)
Age [years] – mean (SD)	60.58 (9.55)
Race / ethnicity – n (%)	
white	246 (84.5 %)
black or African American	24 (8.2 %)
Hispanic or Latino	16 (5.5 %)
other / unknown	5 (1.7 %)
HPV status – n (%)	
positive	218 (74.9 %)
negative	73 (25.1 %)
T stage [a] – n (%)	
T1	40 (13.7 %)
T2	111 (38.1 %)
T3	91 (31.3 %)
T4	49 (16.8 %)
N stage [a] – n (%)	
N0	56 (19.2 %)
N1	128 (44.0 %)
N2	100 (34.4 %)
N3	7 (2.4 %)
M stage [a] – n (%)	
M0	275 (94.5 %)
M1	16 (5.5 %)
Overall stage [a] – n (%)	
I	99 (34.0 %)
II	91 (31.3 %)
III	44 (15.1 %)
IV	57 (19.6 %)
Included lymph nodes – n	486
Included lymph nodes / patient – range	0 – 8
Smoking – n (%)	
never-smoker	96 (33.0 %)
smoker	193 (66.3 %)
pack-years – median (IQR)	20 (10-40)
pack-years unknown – (n)	20
unknown	2 (0.7 %)
PET [b] – mean (SD)	
slice thickness [mm]	3.28 (0.36)
in-plane pixel spacing [mm]	4.46 (0.97)
in-plane image matrix [n x n]	155.71 (70.04) x 155.71 (70.04)
administered [18F]FDG activity [MBq]	449.72 (100.42)
time from [18F]FDG administration to scan [min]	64.20 (10.21)
CT [b] – mean (SD)	
slice thickness [mm]	3.27 (0.41)
in-plane pixel spacing [mm]	1.14 (0.20)
in-plane image matrix [n x n]	512 x 512

CT tube current [mA]	124.97 (64.80)
CT tube peak voltage [kV]	122.16 (11.13)
Scanner manufacturer – n (%)	
GE Medical Systems	183 (62.9 %)
Siemens	108 (37.1 %)

[a] UICC/AJCC 8th edition staging manuals TNM / overall stage [15, 16]

[b] Values are from original images before pre-processing

[18F]FDG = [18F]fluorodeoxyglucose; MBq = Megabecquerel; kV = Kilovolt; mA = Milliampere

Supplementary table 4 Feature importance in final models

Supplementary table 4.1 Source VOI: Primary tumors

Imaging modality					**PET/CT**	
Machine-learning classifier					**XGB**	
Feature selection method					**RIDGE**	
Feature identifier [a]					**Feature Importance** [b]	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
CT	LoG	4mm	firstorder	90Percentile	1	0.156
PET	wavelet	HLL	ngtdm	Complexity	2	0.152
PET	wavelet	HLL	firstorder	Minimum	3	0.098
PET	wavelet	HLL	glszm	HighGrayLevelZoneEmphasis	4	0.088
CT	wavelet	LLL	firstorder	Skewness	5	0.073
CT	wavelet	LHL	glcm	Correlation	6	0.061
CT	wavelet	LLH	firstorder	Mean	7	0.044
CT	LoG	2mm	firstorder	90Percentile	8	0.041
PET	wavelet	HHH	gldm	LowGrayLevelEmphasis	9	0.037
CT	wavelet	LLH	glcm	DifferenceAverage	10	0.034
CT	wavelet	LLH	firstorder	MeanAbsoluteDeviation	11	0.034
CT	LoG	4mm	ngtdm	Strength	12	0.033
CT	wavelet	LLH	firstorder	10Percentile	13	0.028
PET	wavelet	HLL	glcm	Contrast	14	0.027
PET	wavelet	LHL	gldm	LargeDependenceLowGrayLevel Emphasis	15	0.020
CT	wavelet	HLL	firstorder	90Percentile	16	0.018
CT	LoG	2mm	firstorder	Maximum	17	0.016
CT	wavelet	HLL	firstorder	Mean	18	0.015
PET	wavelet	HLH	glrlm	LongRunLowGrayLevelEmphasis	19	0.014
CT	wavelet	HHL	firstorder	Maximum	20	0.012
CT	wavelet	LLH	firstorder	Median	21	0.000
PET	wavelet	HLH	gldm	LargeDependenceLowGrayLevel Emphasis	22	0.000
CT	wavelet	LLH	gldm	DependenceNonUniformity Normalized	23	0.000
PET	wavelet	HLL	glrlm	LongRunHighGrayLevelEmphasis	24	0.000
CT	wavelet	LLH	ngtdm	Contrast	25	0.000
PET	wavelet	HLL	glrlm	HighGrayLevelRunEmphasis	26	0.000
PET	wavelet	HLL	gldm	HighGrayLevelEmphasis	27	0.000
PET	wavelet	HLL	glrlm	ShortRunHighGrayLevelEmphasis	28	0.000
CT	LoG	4mm	glcm	ClusterShade	29	0.000
PET	wavelet	HLL	glcm	Autocorrelation	30	0.000

Imaging modality					**PET**	
Machine-learning classifier					**NBayes**	
Feature selection method					**RIDGE**	
Feature identifier [a]					**Feature Importance** [b]	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
PET	wavelet	HLH	gldm	LargeDependenceLowGrayLevel Emphasis	1	0.023
PET	wavelet	LHL	gldm	LargeDependenceLowGrayLevel Emphasis	2	-0.022
PET	wavelet	HLH	glrlm	LongRunLowGrayLevelEmphasis	3	0.021
PET	wavelet	HLL	glcm	Contrast	4	-0.019
PET	wavelet	HLL	glrlm	LongRunHighGrayLevelEmphasis	5	-0.018
PET	wavelet	HHH	gldm	LowGrayLevelEmphasis	6	-0.018
PET	wavelet	HLL	gldm	HighGrayLevelEmphasis	7	-0.017
PET	wavelet	HLL	glrlm	HighGrayLevelRunEmphasis	8	-0.017
PET	wavelet	HHH	gldm	LargeDependenceLowGrayLevel Emphasis	9	-0.017
PET	wavelet	HLL	glrlm	ShortRunHighGrayLevelEmphasis	10	-0.017
PET	wavelet	HLL	glszm	HighGrayLevelZoneEmphasis	11	-0.017
PET	wavelet	HLL	glcm	Autocorrelation	12	-0.017
PET	wavelet	HHH	glrlm	LongRunLowGrayLevelEmphasis	13	-0.016
PET	wavelet	HLL	firstorder	Minimum	14	0.016
PET	wavelet	HLL	glszm	GrayLevelVariance	15	-0.016
PET	wavelet	HLL	ngtdm	Complexity	16	-0.016
PET	wavelet	HLL	glcm	JointAverage	17	-0.016
PET	wavelet	HLL	glcm	SumAverage	18	-0.016
PET	wavelet	HHH	glrlm	LowGrayLevelRunEmphasis	19	-0.016
PET	wavelet	LHL	firstorder	Kurtosis	20	-0.016
PET	wavelet	HLL	glcm	DifferenceAverage	21	-0.015

Imaging modality					**CT**	
Machine-learning classifier					**XGB**	
Feature selection method					**RIDGE**	
Feature identifier [a]					**Feature Importance** [b]	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
CT	LoG	4mm	firstorder	90Percentile	1	0.261
CT	wavelet	LHL	glcm	Correlation	2	0.140
CT	LoG	2mm	firstorder	90Percentile	3	0.135
CT	wavelet	LLH	firstorder	10Percentile	4	0.132
CT	wavelet	LLL	firstorder	Skewness	5	0.125
CT	wavelet	LLH	firstorder	Median	6	0.066
CT	LoG	2mm	firstorder	Maximum	7	0.063
CT	wavelet	HLL	firstorder	90Percentile	8	0.046
CT	wavelet	HHL	firstorder	Maximum	9	0.031

Supplementary table 4.2 Source VOI: Individual metastatic lymph nodes

Imaging modality	**PET/CT**	
Machine-learning classifier	**ElNet**	
Feature selection method	**HClust**	
	Feature Importance [b]	
Feature identifier [a]	**Rank**	**Score**
Cluster_10	1	-0.612
Cluster_19	2	0.489
Cluster_8	3	0.369
Cluster_13	4	-0.350
Cluster_11	5	0.301
Cluster_18	6	-0.295
Cluster_20	7	0.291
Cluster_14	8	-0.271
Cluster_6	9	-0.258
Cluster_7	10	-0.166
Cluster_16	11	-0.159
Cluster_4	12	0.143
Cluster_1	13	0.121
Cluster_2	14	0.102
Cluster_9	15	-0.079
Cluster_12	16	-0.068
Cluster_5	17	-0.062
Cluster_3	18	0.046
Cluster_17	19	0.038
Cluster_15	20	-0.010

Imaging modality	**PET**	
Machine-learning classifier	**XGB**	
Feature selection method	**HClust**	
	Feature Importance [b]	
Feature identifier [a]	**Rank**	**Score**
Cluster_8	1	0.180
Cluster_3	2	0.164
Cluster_13	3	0.133
Cluster_19	4	0.122
Cluster_6	5	0.114
Cluster_20	6	0.097
Cluster_15	7	0.082
Cluster_17	8	0.070
Cluster_10	9	0.022
Cluster_14	10	0.016
Cluster_1	11	0.000
Cluster_2	12	0.000
Cluster_4	13	0.000
Cluster_5	14	0.000

Cluster_7	15	0.000
Cluster_9	16	0.000
Cluster_11	17	0.000
Cluster_12	18	0.000
Cluster_16	19	0.000
Cluster_18	20	0.000

Imaging modality	**CT**	
Machine-learning classifier	**ElNet**	
Feature selection method	**HClust**	
	Feature Importance [b]	
Feature identifier [a]	**Rank**	**Score**
Cluster_15	1	0.569
Cluster_17	2	0.516
Cluster_7	3	0.494
Cluster_10	4	-0.472
Cluster_2	5	-0.347
Cluster_14	6	-0.296
Cluster_8	7	0.234
Cluster_9	8	-0.227
Cluster_16	9	-0.214
Cluster_19	10	-0.197
Cluster_20	11	0.153
Cluster_4	12	-0.149
Cluster_11	13	-0.146
Cluster_13	14	0.119
Cluster_5	15	-0.099
Cluster_18	16	0.067
Cluster_12	17	0.052
Cluster_1	18	-0.043
Cluster_3	19	-0.033
Cluster_6	20	-0.033

Supplementary table 4.3 Source VOI: Consensus of tumor and nodes

Imaging modality					**PET/CT**	
Machine-learning classifier					**XGB**	
Feature selection method					**noFS [c]**	
Feature identifier [a]					**Feature Importance [b]**	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
CT	LoG	2mm	firstorder	90Percentile	1	0.091
PET	wavelet	HLL	firstorder	Minimum	2	0.085
CT	LoG	4mm	firstorder	90Percentile	3	0.083
CT	wavelet	LLL	glcm	Idn	4	0.052
PET	wavelet	HLL	ngtdm	Complexity	5	0.047
PET	wavelet	HLL	glszm	SmallAreaHighGrayLevelEmphasis	6	0.047
CT	wavelet	LLL	firstorder	Mean	7	0.041

CT	LoG	4mm	glrlm	RunPercentage	8	0.040
PET	wavelet	HHL	glrlm	LongRunHighGrayLevelEmphasis	9	0.035
CT	wavelet	LLH	glrlm	RunEntropy	10	0.032
PET	wavelet	LLL	glszm	GrayLevelNonUniformity	11	0.029
PET	wavelet	LHH	glszm	ZoneEntropy	12	0.028
CT	LoG	4mm	glrlm	ShortRunEmphasis	13	0.028
CT	wavelet	HLH	glszm	ZoneEntropy	14	0.025
CT	LoG	4mm	gldm	LargeDependenceHighGrayLevel Emphasis	15	0.021
CT	wavelet	HHH	glszm	LargeAreaHighGrayLevelEmphasis	16	0.020
CT	wavelet	LLH	firstorder	Uniformity	17	0.019
CT	wavelet	LLH	glszm	ZoneEntropy	18	0.018
PET	LoG	3mm	glcm	SumAverage	19	0.017
PET	wavelet	LLH	glcm	Contrast	20	0.017

Imaging modality					**PET**	
Machine-learning classifier					**XGB**	
Feature selection method					**noFS** [c]	
Feature identifier [a]					**Feature Importance** [b]	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
PET	wavelet	HHL	glrlm	LongRunHighGrayLevelEmphasis	1	0.078
PET	LoG	3mm	glcm	SumAverage	2	0.050
PET	wavelet	HLL	glszm	HighGrayLevelZoneEmphasis	3	0.048
PET	wavelet	HLL	glszm	SmallAreaHighGrayLevelEmphasis	4	0.045
PET	wavelet	LLH	glcm	SumSquares	5	0.041
PET	original		glrlm	GrayLevelNonUniformity	6	0.040
PET	wavelet	HLL	ngtdm	Complexity	7	0.038
PET	wavelet	HLL	glszm	SizeZoneNonUniformity	8	0.035
PET	wavelet	HHH	glrlm	LowGrayLevelRunEmphasis	9	0.031
PET	wavelet	LHH	glcm	ClusterShade	10	0.027
PET	LoG	3mm	glrlm	ShortRunHighGrayLevelEmphasis	11	0.026
PET	wavelet	HLL	glcm	ClusterShade	12	0.025
PET	wavelet	HLH	glrlm	GrayLevelVariance	13	0.024
PET	wavelet	LHL	glszm	HighGrayLevelZoneEmphasis	14	0.023
PET	wavelet	HHH	glrlm	HighGrayLevelRunEmphasis	15	0.022
PET	wavelet	LHH	firstorder	Kurtosis	16	0.022
PET	wavelet	LHL	glcm	Autocorrelation	17	0.018
PET	wavelet	HHH	gldm	HighGrayLevelEmphasis	18	0.018
PET	wavelet	HLL	glrlm	HighGrayLevelRunEmphasis	19	0.017
PET	wavelet	HLH	glcm	ClusterTendency	20	0.016

Imaging modality					**CT**	
Machine-learning classifier					**XGB**	
Feature selection method					**noFS** [c]	
Feature identifier [a]					**Feature Importance** [b]	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
CT	LoG	4mm	firstorder	90Percentile	1	0.104
CT	LoG	2mm	firstorder	90Percentile	2	0.101
CT	LoG	4mm	glrlm	ShortRunEmphasis	3	0.082
CT	wavelet	LLH	glszm	ZoneEntropy	4	0.053
CT	wavelet	HHH	glszm	LargeAreaHighGrayLevelEmphasis	5	0.037
CT	wavelet	LLL	firstorder	Skewness	6	0.036
CT	wavelet	HLL	glcm	Idn	7	0.034
CT	wavelet	LLH	firstorder	Range	8	0.033
CT	original		glszm	SmallAreaLowGrayLevelEmphasis	9	0.032
CT	wavelet	LLH	glrlm	RunEntropy	10	0.032
CT	LoG	4mm	glszm	GrayLevelVariance	11	0.025
CT	LoG	4mm	glrlm	RunPercentage	12	0.022
CT	wavelet	LHL	glcm	Correlation	13	0.019
CT	wavelet	HLH	gldm	DependenceEntropy	14	0.018
CT	wavelet	LLL	glcm	Idn	15	0.018
CT	LoG	4mm	glrlm	RunLengthNonUniformity Normalized	16	0.017
CT	LoG	4mm	glcm	Correlation	17	0.017
CT	wavelet	HHL	glszm	ZoneEntropy	18	0.017
CT	wavelet	HLL	firstorder	10Percentile	19	0.016
CT	wavelet	LLH	glcm	SumEntropy	20	0.013

Supplementary table 4.4 Source VOI: Consensus of all lymph nodes

Imaging modality					**PET/CT**	
Machine-learning classifier					**XGB**	
Feature selection method					**pMIM**	
Feature identifier [a]					**Feature Importance** [b]	
Modality	**Pre-processing**		**Family**	**Feature name**	**Rank**	**Score**
CT	wavelet	LLH	glszm	ZoneEntropy	1	0.329
CT	wavelet	HHL	glcm	JointEntropy	2	0.276
CT	wavelet	HHL	gldm	LargeDependenceHighGrayLevel Emphasis	3	0.169
CT	wavelet	HHL	glcm	Idn	4	0.116
CT	wavelet	HHH	gldm	DependenceEntropy	5	0.027
CT	LoG	2mm	glcm	DifferenceAverage	6	0.025
CT	wavelet	HHL	gldm	DependenceEntropy	7	0.022
CT	LoG	4mm	glszm	ZoneEntropy	8	0.020
CT	wavelet	LLL	glszm	ZoneEntropy	9	0.007
CT	wavelet	LLH	glcm	JointEntropy	10	0.004
CT	wavelet	LHH	gldm	DependenceEntropy	11	0.004
CT	wavelet	LHL	glcm	JointEntropy	12	0.000

Imaging modality	PET	
Machine-learning classifier	XGB	
Feature selection method	HClust	
	Feature Importance [b]	
Feature identifier [a]	Rank	Score
Cluster_15	1	0.123
Cluster_1	2	0.109
Cluster_16	3	0.102
Cluster_3	4	0.087
Cluster_13	5	0.067
Cluster_5	6	0.066
Cluster_20	7	0.062
Cluster_9	8	0.062
Cluster_4	9	0.062
Cluster_18	10	0.059
Cluster_19	11	0.046
Cluster_2	12	0.040
Cluster_11	13	0.037
Cluster_10	14	0.035
Cluster_7	15	0.022
Cluster_12	16	0.022
Cluster_6	17	0.000
Cluster_8	18	0.000
Cluster_14	19	0.000
Cluster_17	20	0.000

Imaging modality			CT		
Machine-learning classifier			XGB		
Feature selection method			pMIM		
Feature identifier [a]				Feature Importance [b]	
Modality	Pre-processing		Family	Feature name	Rank	Score
CT	wavelet	HHL	glcm	JointEntropy	1	0.223
CT	wavelet	LLH	glszm	ZoneEntropy	2	0.219
CT	wavelet	HHL	gldm	LargeDependenceHighGrayLevel Emphasis	3	0.159
CT	original		glszm	SmallAreaLowGrayLevelEmphasis	4	0.078
CT	wavelet	HLL	glcm	JointEntropy	5	0.076
CT	wavelet	HHL	glcm	Idn	6	0.069
CT	wavelet	LHL	glcm	Idn	7	0.046
CT	wavelet	LHH	ngtdm	Contrast	8	0.044
CT	LoG	2mm	glcm	DifferenceAverage	9	0.033
CT	LoG	4mm	glszm	ZoneEntropy	10	0.023
CT	wavelet	HHH	gldm	DependenceEntropy	11	0.018
CT	LoG	2mm	firstorder	RootMeanSquared	12	0.012
CT	wavelet	LLH	glcm	JointEntropy	13	0.000
CT	wavelet	LHH	gldm	DependenceEntropy	14	0.000
CT	wavelet	LHL	glcm	SumEntropy	15	0.000

CT	wavelet	LLL	glcm	JointEntropy	16	0.000
CT	original		shape	LeastAxisLength	17	0.000

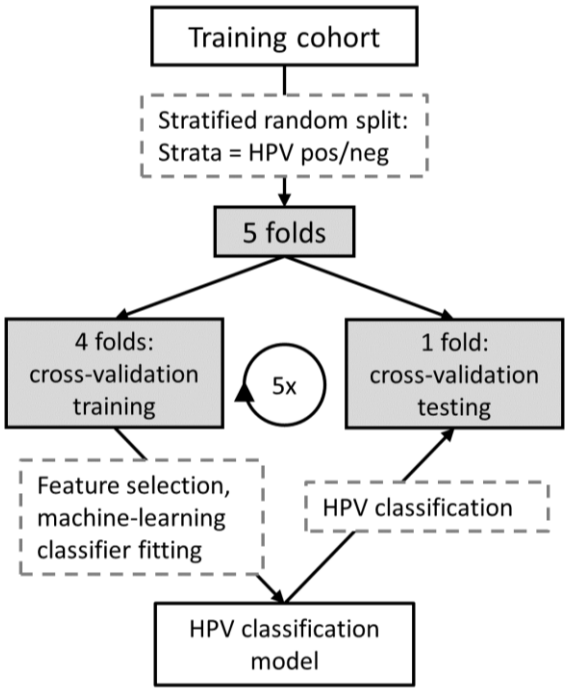
[a] Feature identifiers are composed of an imaging modality abbreviation (PET or CT), a pre-processing specification (left column: type of pre-processing, i.e. wavelet- or LoG-filtering or original; right column: 3-letter directional specification of wavelet [13, 17], or LoG sigma setting), and the feature family and feature name (supplementary table 1)

[b] Refer to the supplementary methods (section 1.3) for details regarding importance rank/score calculations.

[c] Feature selection was omitted – the listed features reflect the 20 highest-ranked features in term of feature importance.

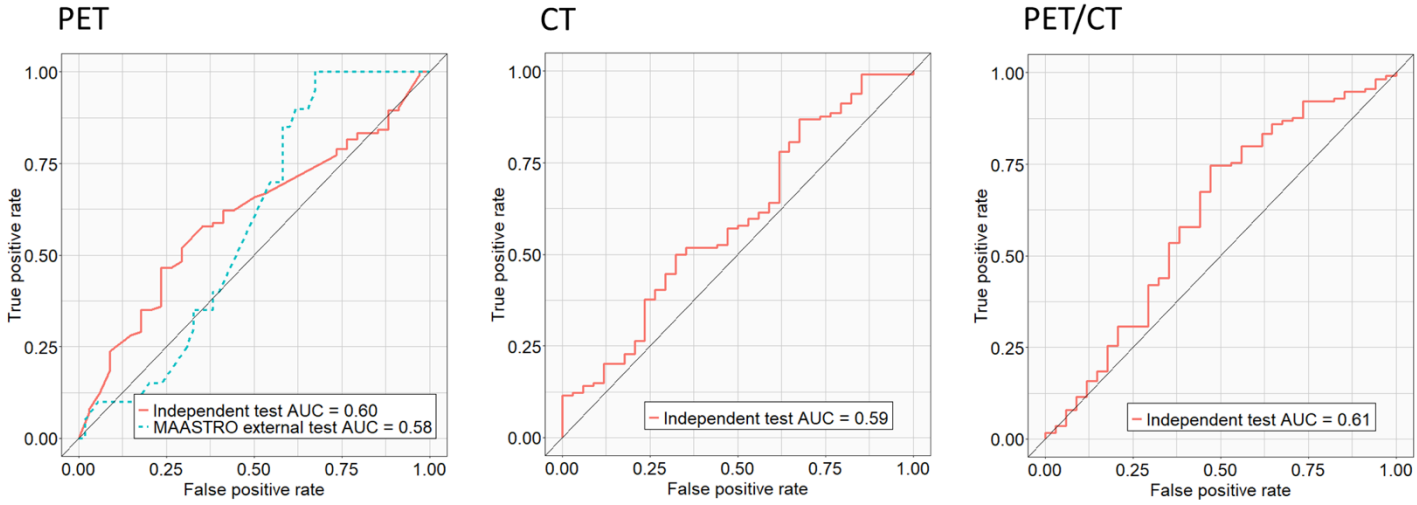
Importance ranks and scores reflecting each selected feature´s or feature cluster´s contribution to the classification result were calculated for all features utilized in each final model.

3. Supplementary figures

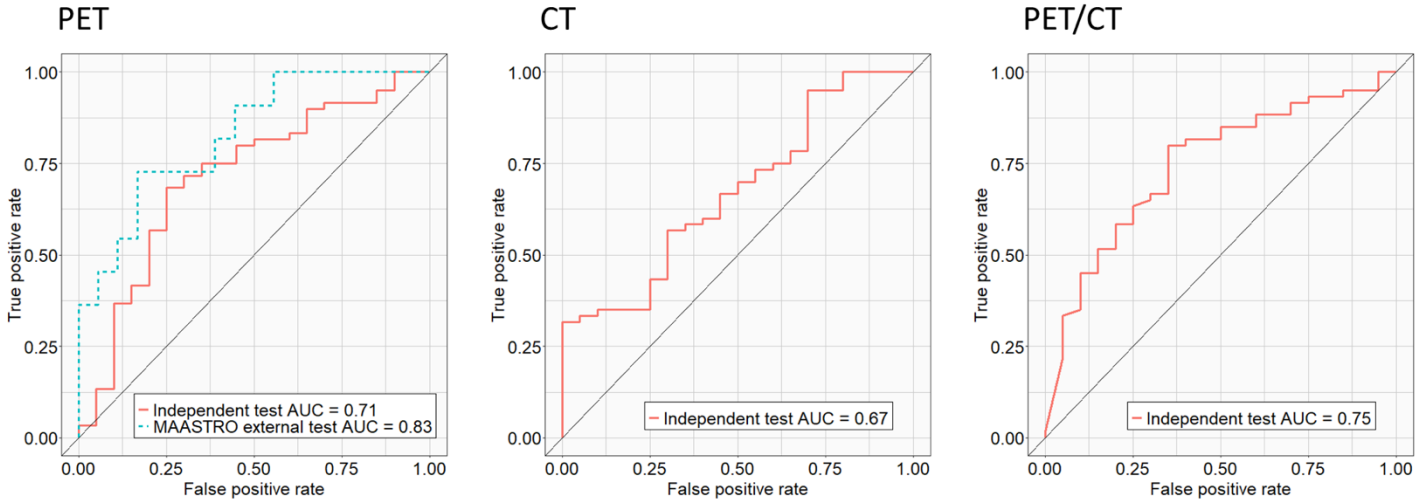


Supplementary fig. 1 Cross-validation framework

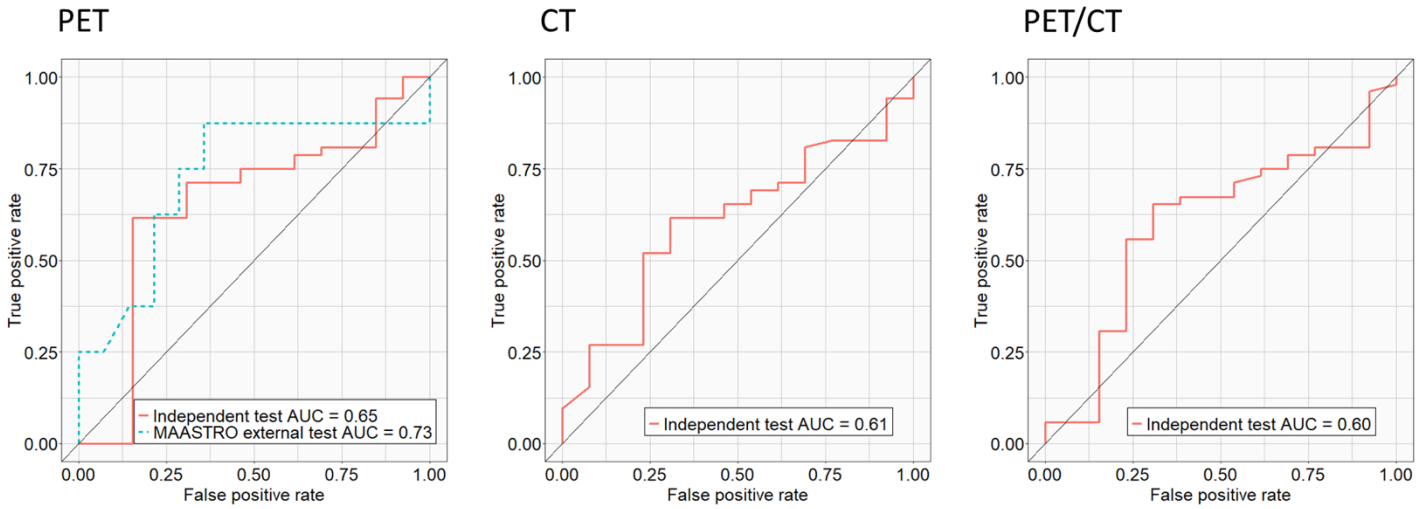
VOI: individual metastatic lymph nodes



VOI: consensus of tumor and nodes



VOI: consensus of all lymph nodes



Supplementary fig. 2 Final model validation ROC curves

4. References

1. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
2. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification. 2014;31:274-95. doi:10.1007/s00357-014-9161-z.
3. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics. 2013;29:2365-8. doi:10.1093/bioinformatics/btt383.
4. Song F, Guo Z, Mei D. Feature Selection Using Principal Component Analysis. 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization; 2010. p. 27-30.
5. Kursa MB. praznik: Collection of Information-Based Feature Selection Filters. Version 6.0.0 ed; 2019.
6. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. 2010. 2010;33:22. doi:10.18637/jss.v033.i01.
7. Majka M. naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R. Version 0.9.6 ed; 2019.
8. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2:18-22.
9. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Version 1.7-2 ed; 2019.
10. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. xgboost: Extreme Gradient Boosting. Version 0.90.0.2 ed; 2019.
11. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv e-prints; 2016.
12. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77:e104-e7. doi:10.1158/0008-5472.CAN-17-0339.
13. Pyradiomics-community. Pyradiomics Documentation Release 2.1.2. 2018.
14. Yan Y. rBayesianOptimization: Bayesian Optimization of Hyperparameters. Version 1.1.0 ed; 2016.
15. AJCC Cancer Staging Manual (8th edition): Springer International Publishing; 2017.
16. TNM Classification of Malignant Tumours, 8th Edition. Wiley-Blackwell: Union for International Cancer Control; 2016.
17. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006. doi:10.1038/ncomms5006.

10. Publication 2

Haider SP, Sharaf K, Zeevi T, Baumeister P, Reichel C, Forghani R, Kann BH, Petukhova A, Judson BL, Prasad ML, Liu C, Burtness B, Mahajan A, Payabvash S. **Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics.** Transl Oncol. 2021;14(1):100906.*

Translational Oncology 14 (2021) 100906

Contents lists available at ScienceDirect

Translational Oncology

journal homepage: www.elsevier.com/locate/tranon

Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics

Stefan P. Haider[a,b,1], Kariem Sharaf[b,1], Tal Zeevi[c], Philipp Baumeister[b], Christoph Reichel[b], Reza Forghani[d], Benjamin H. Kann[e], Alexandra Petukhova[a], Benjamin L. Judson[f], Manju L. Prasad[g], Chi Liu[h], Barbara Burtness[i], Amit Mahajan[a], Seyedmehdi Payabvash[a,*]

[a] Section of Neuroradiology, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 789 Howard Ave, PO Box 208042, New Haven, CT 06519, United States
[b] Department of Otorhinolaryngology, University Hospital of Ludwig Maximilians Universität München, Marchioninistrasse 15, 81377 Munich, Germany
[c] Center for Translational Imaging Analysis and Machine Learning, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, United States
[d] Department of Diagnostic Radiology and Augmented Intelligence & Precision Health Laboratory (AIPHL), McGill University Health Centre & Research Institute, 1650 Cedar Avenue, Montreal, Quebec QC H3G 1A4, Canada
[e] Department of Radiation Oncology, Dana-Farber Cancer Institute, Harvard Medical School, 450 Brookline Avenue, Boston, MA 02215, United States
[f] Division of Otolaryngology, Department of Surgery, Yale School of Medicine, 330 Cedar Street, New Haven, CT 06520, United States
[g] Department of Pathology, Yale School of Medicine, 310 Cedar Street, New Haven, CT 06520, United States
[h] Division of Bioimaging Sciences, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, United States
[i] Section of Medical Oncology, Department of Internal Medicine, Yale School of Medicine and Yale Cancer Center, 25 York Street, New Haven, CT 06520, United States

ARTICLE INFO

Keywords:
Radiomics
Oropharyngeal squamous cell carcinoma
PET/CT
HPV
Imaging biomarker
Risk stratification

ABSTRACT

Locoregional failure remains a therapeutic challenge in oropharyngeal squamous cell carcinoma (OPSCC). We aimed to devise novel objective imaging biomarkers for prediction of locoregional progression in HPV-associated OPSCC. Following manual lesion delineation, 1037 PET and 1037 CT radiomic features were extracted from each primary tumor and metastatic cervical lymph node on baseline PET/CT scans. Applying random forest machine-learning algorithms, we generated radiomic models for censoring-aware locoregional progression prognostication (evaluated by Harrell's C-index) and risk stratification (evaluated in Kaplan-Meier analysis). A total of 190 patients were included; an optimized model yielded a median (interquartile range) C-index of 0.76 (0.66-0.81; $p = 0.01$) in prognostication of locoregional progression, using combined PET/CT radiomic features from primary tumors. Radiomics-based risk stratification reliably identified patients at risk for locoregional progression within 2-, 3-, 4-, and 5-year follow-up intervals, with log-rank p-values of $p = 0.003$, $p = 0.001$, $p = 0.02$, $p = 0.006$ in Kaplan-Meier analysis, respectively. Our results suggest PET/CT radiomic biomarkers can predict post-radiotherapy locoregional progression in HPV-associated OPSCC. Pending validation in large, independent cohorts, such objective biomarkers may improve patient selection for treatment de-intensification trials in this prognostically favorable OPSCC entity, and eventually facilitate personalized therapy.

Introduction

Head and neck squamous cell carcinoma (HNSCC) is among the most morbid cancers [1]. Sustained high-risk human papillomavirus (HPV)-infection in the oropharynx is the cause of a large and increasing proportion of oropharyngeal squamous cell carcinomas (OPSCC) [2], characterized by distinct demographic, biologic and – most notably – prognostic attributes compared to HPV-negative OPSCC [3]. Consequently, the

Abbreviations: AJCC, American Joint Committee on Cancer; AUC, area under the receiver operating characteristic curve; CT, computed tomography; GTV, gross tumor volume; Gy, Gray; HNSCC, head and neck squamous cell carcinoma; HPV, human papillomavirus; IQR, interquartile range; LRP, locoregional progression; OPSCC, oropharyngeal squamous cell carcinoma; PET, [18F]fluorodeoxyglucose positron emission tomography; RCF, random classification forest; RSF, random survival forest; TCIA, The Cancer Imaging Archive; VOI, volume of interest.

* Corresponding author.
E-mail address: sam.payabvash@yale.edu (S. Payabvash).

[1] Both the authors contributed equally to this work.

https://doi.org/10.1016/j.tranon.2020.100906
Received 27 July 2020; Received in revised form 27 September 2020; Accepted 28 September 2020

American Joint Committee on Cancer (AJCC) adopted separate staging schemes for HPV-associated and HPV-negative OPSCC in the 8th edition staging manual [4,5].

Improved responsiveness to treatment and more indolent natural history of HPV-associated OPSCC may potentially render this prognostically favorable subtype amenable to treatment de-intensification with reduced treatment-related toxicity [6,7]. Nevertheless, treatment failure with locoregional disease progression (LRP) is a negative prognostic factor in HPV-associated OPSCC, often entailing salvage resection or irradiation which are commonly associated with increased morbidity and impaired functionality, and ultimately resulting in reduced overall survival [8–10]. Thus, there is a pressing need for novel biomarkers to identify patients amenable for safe treatment de-escalation and ultimately personalized clinical decision-making.

The notion that quantitative characterization of increasingly larger sets of biomedical data may pave the way for precision diagnosis, prognostication and treatment decision-making has shaped the "-omics" concept – e.g. genomics, metabolomics, proteomics. Radiomics analysis has expanded the scope of "-omics" to quantitative characterization of medical images by extracting high-dimensional sets of "features" from volumes of interest (VOI) such as primary tumor lesions, which capture lesion shape, image intensity and texture patterns. The resulting imaging biomarkers may be correlated with treatment outcome, tumor microenvironment, tissue heterogeneity and pathophysiology; and may enable development of prognostic tools substituting or supplementing traditional outcome predictors such as cancer staging [11–15]. Depending on the imaging modality used, radiomic features can represent a variety of tumor characteristics; [^{18}F]fluorodeoxyglucose positron emission tomography (PET) radiomics may provide wholistic quantification of tumor metabolic activity and activity distribution; whereas computed tomography (CT) radiomics can describe structural properties and tissue density. In many centers, PET/CT imaging is an integral part of cancer staging and work-up.

Prior studies have demonstrated the predictive value of radiomic biomarkers for LRP in HNSCC, but HPV status was rarely available in all studied OPSCC patients, and subgroup analysis of HPV-associated OPSCC was not reported [16–19]. Radiomics analysis can predict HPV status, and thus the results of prior studies may in part reflect the differences between HPV-associated and HPV-negative subgroups [20,21]. In this study, we aim to apply machine-learning algorithms using combined PET and non-contrast CT radiomic features extracted from baseline clinical scans for prediction and risk stratification of post-radiotherapy LRP in an HPV-associated OPSCC cohort. We acquired a multi-institutional cohort, and devised prognostic biomarkers using radiomic features from the primary tumor as well as metastatic cervical lymph nodes in addition to clinical variables.

Material and methods

Imaging and clinical data

Imaging data and corresponding clinical information were retrospectively acquired from (1) Yale's Smilow Hospital cancer registry from 2009 to 2019; and (2) public collections in The Cancer Imaging Archive (TCIA) [22]: (2a) the "Head-Neck-PET-CT" collection provides data from four institutions in Canada ("Canadian" cohort) [23]; and (2b) the "Data from Head and Neck Cancer CT Atlas" collection holds an MD Anderson Cancer Center dataset ("MD Anderson" cohort) [24]. Our institutional review board approved this study under IRB protocol #2,000,024,295 and waived informed consent; TCIA provides de-identified data with consents obtained and ethical compliance ensured by source institutions.

We included cases of histopathologically confirmed OPSCC with (1) confirmed HPV-association, (2) pre-treatment PET and non-contrast CT scans of the neck, (3) LRP events or ≥18 months of adequate follow-up documentation, and (4) patients who received radiotherapy as part of definitive or adjuvant treatment after surgery, with or without concurrent platinum-based chemotherapy or targeted therapy with cetuximab.

We excluded (1) HPV-negative subjects, (2) patients receiving palliation only and/or denying treatment, (3) patients with recurrent OPSCC, (4) with M1 disease at initial staging, (5) with >50% of the primary gross tumor volume affected by artifacts on visual evaluation of CT scans [25], and (6) with <60 Gray (Gy) in the adjuvant, and <66 Gy in the definitive radiotherapy setting delivered to the gross tumor volume [26].

Post-treatment cancer surveillance at our institution included regular physical examinations, endoscopy and imaging, with additional tissue sampling performed at specialists' discretion. Locoregional disease progression was ascertained by tissue sampling or unequivocal imaging evidence; the latter was confirmed in retrospective data review by documented response to therapy or additional histopathological examination. Study endpoints in TCIA cohorts were based on annotations provided in the datasets.

Lesion segmentation and staging

The segmentation, radiomic feature extraction and disease progression modelling pipeline employed in our study is illustrated in Fig. 1. Separate PET and CT VOI corresponding to the primary tumor lesion and each individual metastatic cervical lymph node were generated as a first step in our radiomics pipeline. Regional metastatic spread was determined based on tissue sampling or unequivocal PET scan findings. We utilized 3D-Slicer version 4.10.1 [27] for image review and segmentation.

The co-registered pre-treatment PET/CT scans were retrieved and reviewed in 3D-Slicer, and the gross tumor volume (GTV) as defined by the "ICRU 83" report [28] was assessed. Using the "Paint" and "Erase" tools in the 3D-Slicer "Segment Editor" module, hypermetabolic areas of the primary tumor and every metastatic lymph node were manually delineated (i.e. slice-by-slice segmentation on axial PET reconstructions). PET segmentations were then copied onto the co-registered CT and manually adjusted to the GTV outline on CT using the "Paint" and "Erase" tools to generate the CT VOI, excluding air, adjacent uninvolved bone, and preserved fat planes. Axial CT slices with streak artifacts involving the lesion upon visual assessment were excluded from analysis; and metastatic lymph nodes with >50% of the GTV involved were entirely excluded [25].

A trained research associate (SPH) initially segmented all lesions; followed by VOI verification and adaption by a neuroradiologist (SP) with greater than 8 years of experience in head and neck cancer imaging. SP and AM (neuroradiologist with greater than 12 years of experience) performed OPSCC staging in accordance with the AJCC 8th edition staging manual [5].

Pre-processing, feature extraction and stability-based feature pre-selection

PET/CT imaging and image reconstruction were performed at the scan source institutions utilizing standard clinical protocols. PET and CT pre-processing was applied before radiomic feature extraction to homogenize the imaging data: PET grey scale normalization, PET and CT voxel size interpolation to isotropic dimensions, CT re-segmentation, generation of ten image derivates per original PET or CT enhancing certain image characteristics, and grey scale discretization were consecutively performed – a detailed description of our automated pre-processing pipeline is included in the supplementary methods [20].

A set of 1037 PET and 1037 CT radiomic features was subsequently extracted from each primary tumor and metastatic node VOI; comprised of volumetric shape features ($n = 14$ features) extracted from only the original image ($n = 1$ image); and first-order ($n = 18$ features) and texture-matrix features ($n = 75$ features) extracted from

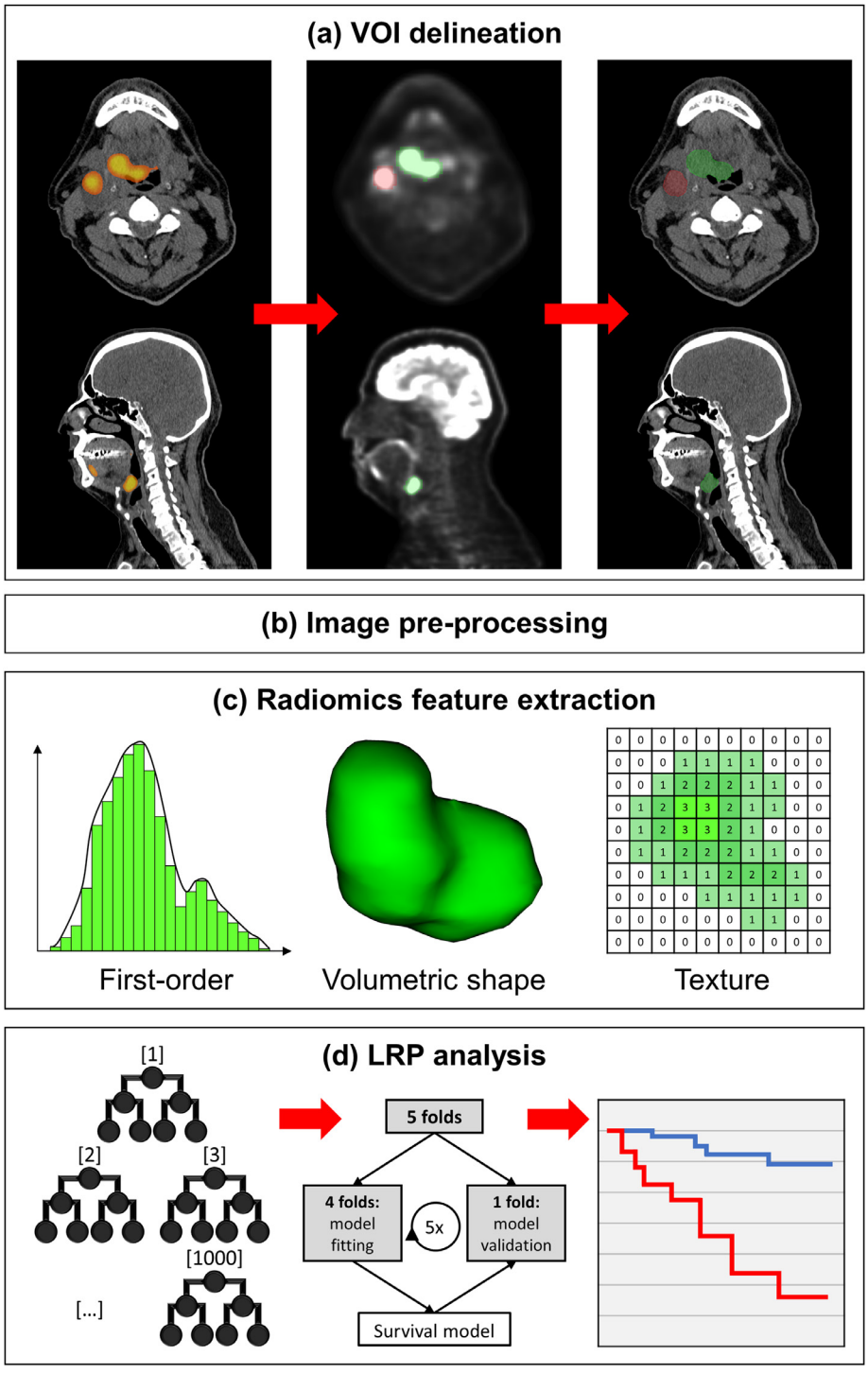


Fig. 1. Radiomics pipeline: (a) VOI delineation – after reviewing the co-registered scans, all lesions were manually delineated on PET axial images, and segmentations were transferred and adapted to the corresponding CT; (b) image pre-processing – details are included in the supplementary methods; (c) radiomics features extraction – 1037 PET and 1037 CT features corresponding to three categories (first-order, volumetric shape, texture) were extracted from each lesion, a comprehensive feature list is included in the supplement; (d) LRP analysis – prognostication and risk stratification was based on random forest machine-learning models with 1000 decision trees internally validated in 20-repeat 5-fold cross-validation, wherein models were iteratively trained on 4 folds, and evaluated in the 5th fold.

both the originals ($n=1$ image) and image derivates ($n=10$ images) generated in pre-processing. This approach yielded a total of ($14\times1+18\times11+75\times11=$) 1037 PET and 1037 CT features per lesion. A complete list of radiomic features utilized in this study is included in supplementary Table 1. A Pyradiomics version 2.1.2 pipeline was customized and applied for radiomics analysis [20, 29].

We investigated radiomic feature stability in an inter-rater and intra-rater setting to pre-select features prior to disease progression modelling, given the volatile robustness of individual features to delineation variability reported in previous studies [30–32]. Unstable features were excluded; the methodology and results are reported in the supplementary methods and supplementary Table 2, respectively [20].

Disease progression modelling and prognostication

We defined locoregional progression (LRP) as the event of interest, with time-to-LRP defined as the time interval from OPSCC initial diagnosis to progression. Right-censoring was applied for loss to follow-up, death, or diagnosis of distant metastases. Subsequently, patients without an LRP event and <18 months of follow-up from diagnosis to censoring were excluded.

We devised and compared three types of LRP models [15]: (1) "Radiomics" models used radiomic signatures, (2) the "clinical" model incorporated AJCC staging (T-, N- and overall-stage), patient age at initial diagnosis, and the treatment modality, and (3) "combined" mod-

els utilized the combined set of radiomics and above-mentioned clinical predictors for LRP prognostication. AJCC T-, N- and overall-stage were included as ordinal variables with four (T1-T4), four (N0-N3) and three (overall stages I-III) levels, respectively. No overall-stage IV cancers were present, since subjects with distant metastases were excluded. Radiomic features, feature clusters and patient age were numeric variables, and the treatment modality was included as a categorical variable.

An array of different approaches was implemented and compared to generate and select features for prognostically optimized radiomic signatures in combined and radiomic models: (1) Radiomic features from three imaging modalities (PET, CT, PET and CT) were used for LRP modelling; (2) we derived radiomic features from two VOI sources: the primary tumor and a "virtual" VOI combining the primary tumor and all metastatic nodes in a given subject as described by Yu et al. [21]; and (3) the prognostic ability of unreduced feature sets was compared to three dimensionality reduction techniques (abbreviations in Fig. 2, details in the supplementary methods). We applied and compared all methodological approaches (3 imaging modalities x 2 VOI sources x 4 dimensionality reduction techniques) for LRP modelling.

The generic model types ("radiomics", "clinical", "combined") introduced above were implemented applying random survival forest (RSF) [33] machine-learning algorithms for prognostication, which were configured to grow 1000 decision trees using a C-index split rule [34] with the remaining parameters in default. Statistical analysis was performed in R version 3.6.0 [35] using extension packages, R base functions and custom-written code. We used the "ranger" package (version 0.12.1) [33] for RSF modelling.

To limit overfitting and enhance generalizability, all models were internally validated in a framework applying 20 repeats of stratified 5-fold cross-validation (i.e. 100 permutations) using the event/non-event groups and follow-up duration as strata. Consensus VOI generation (if applicable), radiomic feature standardization, dimensionality reduction (if applicable), and RSF fitting were consecutively performed on the training folds, and RSF performance was quantified in the validation fold of each cross-validation iteration. This approach avoids "information leakage" from training to validation data and generates realistic estimates of RSF performance in new datasets.

We quantified models' prognostic abilities in each validation fold with a right-censoring adjusted concordance index (Harrell's C-index [34]), and the median score was calculated across 20 cross-validation iterations to represent models' overall performance.

We further investigated the performance of three select models: the clinical model, and the best – in terms of C-index score – radiomic and combined model, respectively. Models' validation fold C-index distribution across 20-repeat 5-fold cross validation was compared against random predictions (i.e. C-index calculated with the same model predictions but randomly resampled validation fold LRP outcome) using a corrected paired *t*-test ("corrected repeated k-fold cv test" [36]). P-values <0.05 ascertained significance.

We generated time-dependent performance curves to track and compare model performance throughout follow-up by calculating Uno's estimator of cumulative/dynamic area under the curve (AUC) for right-censored survival data [37] in each validation fold ("survAUC" package [38] for R), and averaging AUC scores across 20×5-fold cross validation. The resulting performance curves were plotted for the first five years of follow-up.

Risk stratification and Kaplan-Meier analysis

The potential role of radiomics for LRP risk stratification was investigated by generating radiomics risk groups (high-risk vs. low-risk) in binary classification analysis [15]. We subsequently conducted Kaplan-Meier analysis with radiomics risk groups. For comparison, AJCC-staging (T-, N- overall-stage), patient age (age ≥ cohort median vs. < cohort median), and treatment modality variables served as Kaplan-Meier risk groups. A log-rank test generated p-values with $p<0.05$ considered significant.

To generate radiomics risk groups, our framework applying 20-repeat stratified 5-fold cross-validation was adapted for binary classification, using event/non-event groups as strata, and a random classification forest (RCF) algorithm ("ranger" package version 0.12.1) [33] configured to grow 1000 decision trees for risk score computation (i.e. probability of experiencing an event). RCF case weights in a given outcome class (event or non-event) were specified to be inversely proportional to the class distribution in the training data to account for imbalance, with the remaining RCF parameters in default.

Patients' RCF risk scores were averaged across validation folds, and a risk cutoff was calculated by maximizing Youden's statistic in receiver operating characteristic-analysis. Patients with averaged risk scores greater than the cutoff were allocated to the radiomics high-risk group. RCF models were trained on the radiomics-only dataset of the radiomic LRP model selected for further evaluation (previous subsection) without feature selection applied.

Patients were labelled for Kaplan-Meier analysis using 2-, 3-, 4- and 5-year follow-up cutoffs; subjects diagnosed with LRP before a given cutoff were labelled positive, subjects lost to follow-up before a cutoff were excluded, and the remainder was labelled negative and censored at the cutoff. Separate RCF models were generated for each cutoff, and the resulting radiomics risk groups and were investigated in separate Kaplan-Meier plots. Equivalently, Kaplan-Meier analysis with clinical variables was conducted separately for each follow-up cutoff. This strategy avoids censoring before a cutoff (i.e. "dense" survival data) and thus enables RCF performance maximization, while allowing both exploring the majority of the documented follow-up period as well as comparing radiomics risk stratification with clinical variables in an easily interpretable fashion.

Results

Cohort characteristics

A total of 190 patients with HPV-associated OPSCC met inclusion criteria; thereof, 15 (~8%) had LRP events at a median (interquartile range, IQR) of 14.5 (11.0–21.6) months after initial diagnosis. Patients were followed-up for a median (IQR) of 40.7 (30.7–53.5) months after initial diagnosis. Table 1 summarizes demographics, treatment, imaging and staging characteristics of our study cohort.

In addition to 190 OPSCC primary tumors, 266 metastatic lymph nodes were segmented. Thereof, 422 (19.2%) out of 2193 primary tumor lesion axial slices, and 155 (5.6%) out of 2778 lymph node lesion axial slices were affected by streak artifact on CT, and were excluded (details in supplementary Table 3).

Prognostication of locoregional disease progression

The best radiomics LRP model in our study yielded a median (IQR) C-index of 0.76 (0.66-0.81; $p = 0.01$) using the full set of PET/CT primary tumor radiomic features (Fig. 2). The model using clinical variables did not exhibit prognostic value in cross validation, yielding a C-index (IQR) of 0.49 (0.39-0.58; $p = 0.46$), and combined models achieved median scores similar to those of corresponding radiomic models (Fig. 2). Combined PET/CT radiomic models achieved higher prognostic performance than single imaging modality models in the majority of permutations. Models combining radiomic features from primary tumors and metastatic cervical lymph nodes ("virtual" consensus VOI) improved PET-based LRP prognostication, with the best combined model yielding a median C-index (IQR) of 0.65 (0.52–0.76) using random forest-based feature selection ("pRF"); whereas the corresponding PET primary tumor model yielded a median C-index of 0.64 (0.50–0.71).

Select models subjected to further evaluation are highlighted in Fig. 2. Performance curve plotting (Fig. 3) again revealed similar

Table 1
Cohort characteristics.

Number of OPSCC patients – n	190
Included metastatic lymph nodes – n	266
LRP events – n (%)	15 (7.9%)
Follow-up [months] – median (IQR)	40.7 (30.7–53.5)
Time-to-event [months] – median (IQR)	14.5 (11.0–21.6)
Data source – n (%)	
Yale	112 (58.9%)
TCIA	78 (41.1%)
Sex – n (%)	
male	154 (81.1%)
female	36 (18.9%)
Age [years] – mean (SD)	59.83 (8.51)
HPV status – n (%)	
positive	190 (100%)
Smoking – n (%)	
never-smoker	48 (25.3%)
smoker	77 (40.5%)
pack-years – median (IQR)	15 (7.75–30)
pack-years unknown – n	15
unknown	65 (34.2%)
T stage [a] – n (%)	
T1	26 (13.7%)
T2	77 (40.5%)
T3	64 (33.7%)
T4	23 (12.1%)
N stage [a] – n (%)	
N0	35 (18.4%)
N1	108 (56.8%)
N2	43 (22.6%)
N3	4 (2.1%)
Overall stage [a] – n (%)	
I	85 (44.7%)
II	78 (41.1%)
III	27 (14.2%)
Included lymph nodes / patient – range	0 – 6
Primary treatment – n (%)	
CCRT or CBRT	135 (71.1%)
Surgery with adjuvant RT, CCRT or CBRT	34 (17.9%)
RT alone	21 (11.1%)
PET [b] – mean (SD)	
slice thickness [mm]	3.44 (0.40)
in-plane pixel spacing [mm]	4.28 (0.90)
in-plane image matrix [n x n]	148.25 (60.17) x idem
CT [b] – mean (SD)	
slice thickness [mm]	3.06 (0.60)
in-plane pixel spacing [mm]	1.12 (0.18)
in-plane image matrix [n x n]	512 × 512

[a] AJCC 8th edition staging manual T/N/overall stage [5].
[b] Values from image originals before preprocessing.
CBRT = concurrent bioradiotherapy with cetuximab; CCRT = concurrent platinum-based chemoradiotherapy; RT = radiotherapy alone; SD = standard deviation.

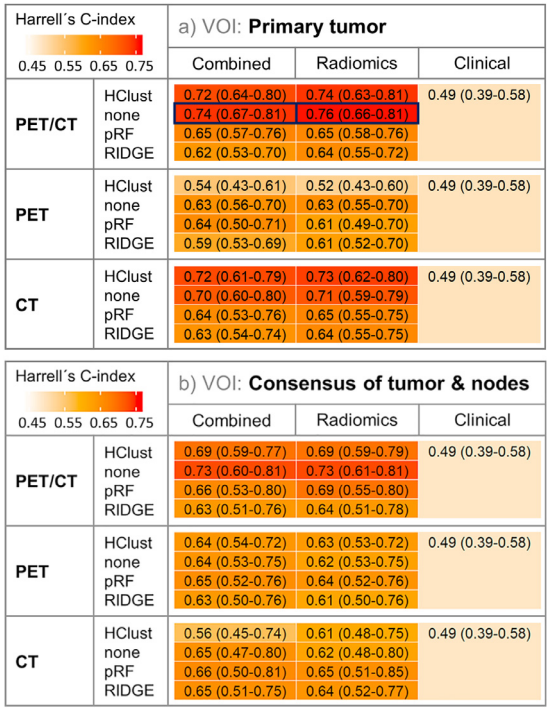


Fig. 2. Heatmap summary of LRP model performance quantified by the median (IQR) validation fold Harrell´s C-index across 20-repeat 5-fold cross-validation. The radiomics and combined models selected for further evaluation are highlighted (blue frame). All methodological combinations to generate radiomics signatures for radiomics and combined models were applied (3 imaging modalities x 2 VOI sources x 4 dimensionality reduction techniques (HClust, none, pRF, RIDGE)).
Clinical = clinical model; Combined = combined model; HClust = hierarchical clustering; none = no dimensionality reduction applied; pRF = Pearson correlation-based redundancy reduction with random survival forest variable importance; Radiomics = radiomics model; RIDGE = Cox regression with RIDGE regularization adapted for feature selection.

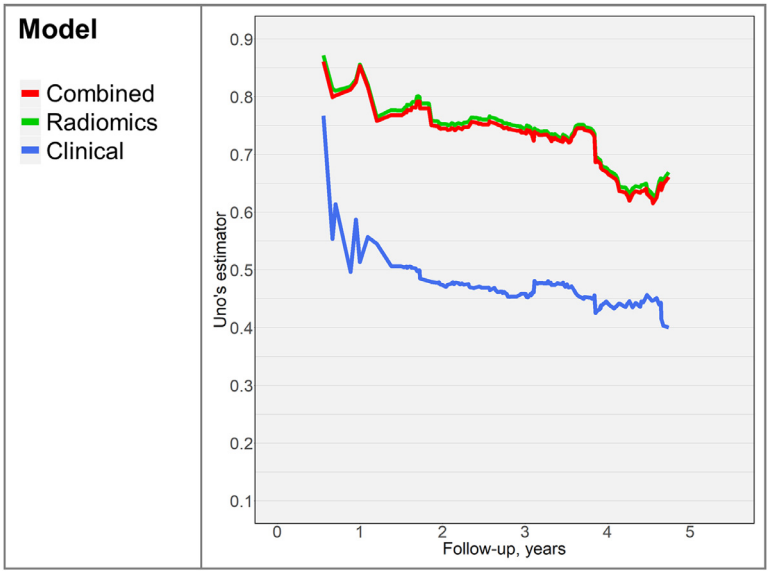


Fig. 3. Time-dependent performance curves depict selected models' (highlighted in Fig. 2) prognostic performance throughout 5-years of follow-up. The corresponding clinical model is presented for comparison.

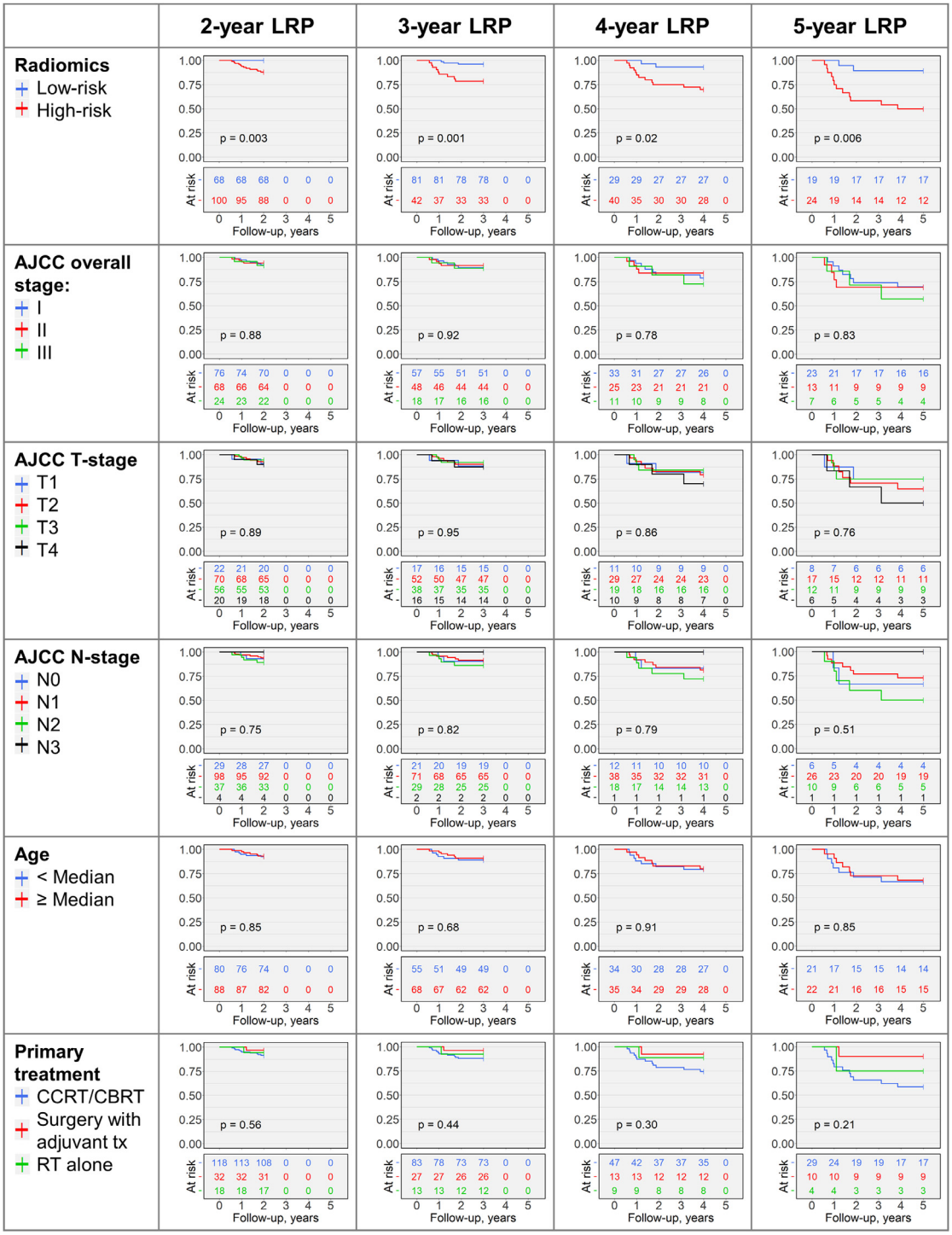


Fig. 4. Kaplan-Meier plots and log-rank test p-values depicting risk stratification based on radiomics analysis and clinical variables.

superiority of the radiomic model and the combined model over the clinical model. Notably, while radiomic modelling exhibited high prognostic performance in follow-up years 1 through 4, model performance was moderate in the fifth year.

Risk stratification of locoregional disease progression

In Kaplan-Meier analysis, radiomics high-risk groups exhibited significantly higher rates of LRP than corresponding low-risk groups in analysis of all follow-up cutoffs, achieving log-rank p-values of $p = 0.003$, $p = 0.001$, $p = 0.02$, $p = 0.006$ for the 2-, 3-, 4-, and 5-year follow-up intervals, respectively (Fig. 4). Risk groups derived from clinical variables (AJCC staging, age, treatment) did not differ significantly ($p > 0.05$, Fig. 4).

Discussion

Improved responsiveness to treatment and more indolent natural history of HPV-associated OPSCC – as compared to HPV-negative cancers – may render this prognostically favorable subtype amenable to de-intensified therapy with reduced treatment-related toxicity and morbidity [6,7]. Accurate prognostication and risk stratification are, how-

ever, the first steps in personalized treatment decision-making. Using a multi-institutional cohort, we investigated the prognostic value of baseline PET/CT radiomics in prediction of LRP. Applying machine-learning analysis, we devised prognostic models utilizing PET/CT features from primary tumor lesions (with and without metastatic cervical nodes). Pending validation in larger cohorts, these novel objective biomarkers can provide decision assistance tools for precision treatment planning in patients with HPV-associated OPSCC.

LRP represents treatment failure of definite therapy in curative intent, with few remaining satisfactory options: Salvage surgery and irradiation are commonly associated with increased morbidity and impaired functionality; and while new immune checkpoint inhibitors alone or in addition to conventional systemic treatment improved outcome in patients not amenable to localized therapy, long-term control of relapsed HNSCC often remains fairly poor [39–41]. Locoregional relapse is also strongly tied to poor overall survival in HPV-associated OPSCC [8, 10], substantiating the importance of this endpoint for therapeutic decision making in OPSCC.

While pre-treatment PET/CT imaging is a mainstay of disease work-up and cancer staging, human visual interpretation cannot seize the full prognostic utility encoded in metabolic and structural bioimaging patterns [11–14]. By capturing such bioimaging features, radiomic biomarkers may help identify patients who are at increased risk for LRP, and may potentially improve patient selection in future trials of treatment de-intensification for HPV-associated OPSCC, and guide personalized clinical treatment planning.

In this study, we showed the merits of radiomic features quantifying tumor intensity, volumetric shape and texture for LRP prognostication and risk stratification. Employing a pre-processing pipeline designed to mitigate heterogeneity in imaging data, a multiple-delineation feature pre-selection approach retaining only stable radiomic features, and a rigorous cross-validation scheme avoiding data leakage from training to test sets, our results depict the prognostic potentials of machine-learning-generated radiomic biomarkers for LRP in a realistic fashion. A model utilizing the full set of combined PET/CT primary tumor radiomic features was reliably prognostic of LRP in cross validation, yielding a median (IQR) C-index of 0.76 (0.66-0.81; $p = 0.01$); whereas a clinical model combining AJCC overall-, T- and N-stage as well as treatment modality and patient age did not exhibit prognostic abilities. This result may be linked to the low event rate of ~8% and the relatively small cohort size. Models combining clinical variables and radiomic signatures performed similarly to the corresponding radiomic models. Additionally, radiomics-based risk-stratification biomarkers identified patients at increased risk of LRP in different follow-up cutoffs (2-, 3-, 4- and 5-year follow-up with $p<0.05$); whereas clinical variables could not significantly stratify LRP risk ($p>0.05$). These findings suggest radiomics analysis may be a more powerful means for LRP risk stratification and prognostication than the tested set of potential clinical predictors.

Notably, models integrating PET and CT radiomics outperformed single modality models in most permutations in both primary tumor and combined tumor/lymph node analysis, suggesting complementary prognostic value of “metabolic” and “structural” features derived from PET and CT imaging, respectively. Additionally, consensus VOI combining PET radiomics information from primary tumors and metastatic cervical lymph nodes yielded performance improvements over most corresponding PET primary tumor models. This finding may suggest added prognostic value from PET lymph node features.

Performance curves (Fig. 3) are a valuable tool to investigate model prognostic accuracy throughout a relevant follow-up period, providing a more granular understanding of model performance than summary measures such as Harrell´s C index. We plotted performance curves for select radiomic and combined models as well as the clinical model, again revealing superiority of radiomics-based prognostication. While radiomic models achieved high prognostic performance in follow-up years 1 through 4, model performance was moderate in the fifth year, which could be related to data sparsity in model training secondary to right censoring.

Accurate contouring of head-and-neck cancer lesions on CT is challenging – especially on pre-contrast images. In this study, we applied manual PET-guided segmentation, allowing full utilization of both accurate PET-guided lesion contouring and standardized CT tissue densities devoid of contrast-induced variability. Notably, analysis of contrast-enhanced CT scans may be limited due to variabilities in contrast accumulation, affecting radiomic feature extraction and reproducibility [14]. Additionally, combined PET and non-contrast CT radiomics analysis extracts both “metabolic” and “structural” tissue density features, allowing comprehensive assessment of primary tumor and metastatic nodes.

Methodologically, our modelling approach relied on random forest machine-learning algorithms: we applied RSF algorithms designed to handle right-censored survival data as well as “classical” RCF models for binary classification [33]. Machine-learning has proven effective in handling the high variable dimensionality commonly associated with radiomics analysis, with random forest models in particular often outperforming other approaches due to superior robustness [11,14,19].

We acquired a multi-national and multi-institutional cohort incorporating data from our institution and several additional centers in Canada and in the United States to increase the cohort size. Additionally, using multi-center data may help augment model robustness to variations among imaging protocols, scanner hardware and image reconstruction and ultimately lead to more generalizable models and model performance estimates.

LRP in HPV-associated OPSCC treated with radiotherapy is rare – in our cohort of 190 subjects, ~8% experienced events – making allocation of independent validation sets in our study challenging. Thus, we applied a rigorous cross validation framework, with particular attention given to avoiding data leakage from training to validation folds; i.e. consensus VOI generation, feature standardization, dimensionality reduction, and RSF fitting were performed on the training folds, and model performance was quantified in the corresponding validation folds. This approach is expected to yield realistic quantification of model performance in new datasets. Nevertheless, future prospective studies with larger study cohorts and higher absolute event counts are required to confirm the prognostic value of quantitative imaging models for LRP prediction. Additionally, our models require independent validation in external cohorts before translation to clinical application may be considered.

Our cohort of 190 patients with HPV-associated OPSCC was acquired from Yale’s Smilow Hospital (2009 to 2019) and two public collections in The Cancer Imaging Archive. PET/CT acquisition and image reconstruction protocols varied over the years and between different cancer centers. This limitation was addressed by adopting a comprehensive image pre-processing pipeline designed to reduce heterogeneity, denoise our dataset, and homogenize PET/CT scans. Nonetheless, standardization of both PET and CT image acquisition across centers and scanner manufacturers may harbor potential for improved radiomics capabilities in OPSCC outcome prognostication and should be pursued as a long-term goal in the field of quantitative imaging.

Manual lesion segmentation is inherently prone to inter- and intra-rater variability as well as limited reproducibility. Despite our efforts to pre-select a subset of robust radiomics features in multiple delineation analysis, fully or partially automating the lesion delineation process may help reduce the aforementioned limitations and ultimately contribute to improved LRP prognostic performance.

Future studies should also incorporate further established LRP predictors into clinical and combined models – e.g. smoking status was unavailable in a considerable portion of our dataset. Finally, metastatic involvement of cervical lymph nodes was determined by expert radiologist assessment, but without histopathological examination of all nodes.

Conclusion

Radiomics analysis decoding metabolic and structural bioimaging patterns of the primary tumor lesion and metastatic nodes in pre-treatment PET/CT scans can provide novel quantitative imaging biomarkers for risk stratification and prediction of post-radiotherapy LRP in HPV-associated OPSCC. Pending independent validation in large external cohorts, such biomarkers may supplement patient selection for trials of treatment de-intensification for prognostically favorable HPV-associated OPSCC, and ultimately guide personalized treatment decision-making.

Declaration of Competing Interest

SPH: None declared. KS: None declared. TZ: None declared. PB: None declared. CR: None declared. RF has acted as speaker and consultant for GE Healthcare and has a research agreement (beta tester) and support from GE Healthcare. RF is also a founder and stockholder of 4intelligent Inc.; and a clinical research scholar (chercheur-boursier clinician) supported by the Fonds de recherche en santé du Québec (FRQS). BHK: None declared. AP: None declared. BLJ: None declared. MLP: None declared. CL has research agreements with Siemens Medical Solutions and GE Healthcare. BB: None declared. AM: None declared. SP: None declared.

CRediT authorship contribution statement

Stefan P. Haider: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Kariem Sharaf:** Conceptualization, Data curation, Investigation, Writing - original draft, Writing - review & editing. **Tal Zeevi:** Formal analysis, Investigation, Methodology, Writing - review & editing. **Philipp Baumeister:** Investigation, Writing - review & editing. **Christoph Reichel:** Investigation, Writing - review & editing. **Reza Forghani:** Investigation, Writing - review & editing. **Benjamin H. Kann:** Data curation, Investigation, Writing - review & editing. **Alexandra Petukhova:** Investigation, Writing - review & editing. **Benjamin L. Judson:** Investigation, Writing - review & editing. **Manju L. Prasad:** Data curation, Investigation, Writing - review & editing. **Chi Liu:** Investigation, Writing - review & editing. **Barbara Burtness:** Investigation, Writing - review & editing. **Amit Mahajan:** Data curation, Formal analysis, Investigation, Validation, Writing - review & editing. **Seyedmehdi Payabvash:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing - review & editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tranon.2020.100906.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA Cancer J. Clin. 69 (1) (2019) 7–34.
[2] M.L. Gillison, et al., Epidemiology of Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma, J. Clin. Oncol. 33 (29) (2015) 3235–3242.
[3] E. Benson, et al., The clinical impact of HPV tumor status upon head and neck squamous cell carcinomas, Oral Oncol. 50 (6) (2014) 565–574.
[4] W.M. Lydiatt, et al., Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual, CA Cancer J. Clin. 67 (2) (2017) 122–137.
[5] AJCCCancer Staging Manual, 8th edition, Springer International Publishing, 2017.
[6] D.J. Adelstein, et al., Role of treatment deintensification in the management of p16+ oropharyngeal cancer: ASCO provisional clinical opinion, J. Clin. Oncol. 37 (18) (2019) 1578–1589.
[7] S. Marur, 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. 35 (5) (2017) 490–497.
[8] C. Gronhoj, et al., Pattern of and survival following loco-regional and distant recurrence in patients with HPV+ and HPV- oropharyngeal squamous cell carcinoma: a population-based study, Oral Oncol. 83 (2018) 127–133.
[9] A.W. Joseph, et al., Disease-free survival after salvage therapy for recurrent oropharyngeal squamous cell carcinoma, Head Neck 38 (Suppl 1) (2016) E1501–E1509.
[10] A. Caula, et al., Correlation between the duration of locoregional control and survival in T1-T2 oropharyngeal cancer patients, Eur. Arch. Otorhinolaryngol. 276 (4) (2019) 1161–1166.
[11] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, Radiology 278 (2) (2016) 563–577.
[12] S.S. Yip, H.J. Aerts, Applications and limitations of radiomics, Phys. Med. Biol. 61 (13) (2016) R150–R166.
[13] S.P. Haider, et al., Applications of radiomics in precision diagnosis, prognostication and treatment planning of head and neck squamous cell carcinomas, Cancers Head Neck 5 (1) (2020) 6.
[14] R. Forghani, et al., Radiomics and artificial intelligence for biomarker and prediction model development in oncology, Comput. Struct. Biotechnol. J. 17 (2019) 995–1008.
[15] S.P. Haider, et al., Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma, Cancers (Basel) 12 (7) (2020).
[16] M.R. Folkert, et al., Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal cancer based on FDG-PET image characteristics, Phys. Med. Biol. 62 (13) (2017) 5327–5343.
[17] M. Bogowicz, et al., Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer, Sci. Rep. 9 (1) (2019) 15198.
[18] G. Feliciani, et al., Radiomic profiling of head and neck cancer: (18)F-FDG PET texture analysis as predictor of patient survival, Contrast Media Mol Imaging 2018 (2018) 3574310.
[19] S. Leger, et al., A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling, Sci. Rep. 7 (1) (2017) 13206.
[20] S.P. Haider, et al., PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma, Eur. J. Nucl. Med. Mol. Imaging (2020).
[21] K. Yu, et al., Radiomic analysis in prediction of Human Papilloma Virus status, Clin. Transl. Radiat. Oncol. 7 (2017) 49–54.
[22] K. Clark, et al., The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (6) (2013) 1045–1057.
[23] M. Vallières, et al., Data from Head-Neck-PET-CT, The Cancer Imaging Archive, 2017.
[24] A. Grossberg, et al., Data from Head and Neck Cancer CT Atlas, The Cancer Imaging Archive, 2017.
[25] R.B. Ger, et al., Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis, Comput. Med. Imaging Graph. 69 (2018) 134–139.
[26] D.J. Sher, et al., Radiation therapy for oropharyngeal squamous cell carcinoma: executive summary of an ASTRO evidence-based clinical practice guideline, Pract Radiat Oncol 7 (4) (2017) 246–253.
[27] R. Kikinis, S.D. Pieper, K.G. Vosburgh, 3.D. Slicer, Springer New York, New York, NY, 2014, pp. 277–289. Editor.
[28] Definition of Volumes, J. ICRU 10 (1) (2010) 41–53.
[29] Pyradiomics-communityPyradiomics Documentation Release 2.1.2, 2018 [cited 2019 December 15th, 2019]; Available from:https://readthedocs.org/projects/pyradiomics/downloads/pdf/2.1.2/.
[30] A. Traverso, et al., Repeatability and reproducibility of radiomic features: a systematic review, Int. J. Radiat. Oncol. Biol. Phys. 102 (4) (2018) 1143–1158.
[31] L. Lu, et al., Robustness of radiomic features in [(11)C]Choline and [(18)F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization, Mol. Imaging Biol. 18 (6) (2016) 935–945.
[32] R.T. Leijenaar, et al., Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability, Acta Oncol. 52 (7) (2013) 1391–1397.
[33] M.N. Wright, A. Ziegler, Ranger: a fast implementation of random forests for high dimensional data in C plus plus and R, J. Stat. Softw. 77 (1) (2017) 1–17.
[34] M. Schmid, M. Wright, A. Ziegler, On the use of Harrell's C for clinical risk prediction via random survival forests, arXiv e-prints (2015).
[35] R Development Core TeamR: A Language and Environment For Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019.
[36] R.R. Bouckaert, E. Frank, in: Evaluating the Replicability of Significance Tests For Comparing Learning Algorithms, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 3–12.
[37] H. Uno, et al., Evaluating prediction rules for t-year survivors with censored regression models, J. Am. Stat. Assoc. 102 (478) (2007) 527–537.
[38] Potapov, S., W. Adler, and M. Schmid, survAUC: Estimators of Prediction Accuracy for Time-to-Event Data. 2012.

[39] P. Pipkorn, et al., Outcomes of relapsed human papillomavirus-related oropharyngeal squamous cell carcinoma treated with curative intent, Head Neck 41 (5) (2019) 1312–1319.
[40] J. Zenga, et al., Salvage of recurrence after surgery and adjuvant therapy: a multi-institutional study, Otolaryngol. Head Neck Surg. 161 (1) (2019) 74–81.
[41] B. Burtness, et al., Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): a randomised, open-label, phase 3 study, Lancet 394 (10212) (2019) 1915–1928.

Supplement

for the original article

Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics

Table of Contents

1. Supplementary methods

1.1 Image pre-processing pipeline

For normalization of PET scan voxel values, we divided each voxel's intensity by the left lentiform nucleus' maximum intensity to improve the inter-scanner and inter-institutional generalizability of PET-based quantitative metrics [1]. To ensure texture feature rotational invariance [2] and even out voxel size and slice thickness dissimilarities [3-6], we generated isotropic 3x3x3 and 2x2x2 mm PET and CT voxels, respectively, using trilinear image interpolation [7]. A re-segmentation process of CT VOIs only retaining voxels within a 1-300 Hounsfield unit (HU) range was applied to restrict radiomics analysis to soft tissue densities. We generated ten image derivates per original PET or CT scan to refine radiomics analysis of specific characteristics. High and low frequency analysis was enhanced using a "coif-1" wavelet transform to generate eight decompositions per original [7, 8]. Laplacian of Gaussian (LoG) filtering for edge-enhancement with "sigma" settings of 3 and 6 mm for PET, and 2 and 4 mm for CT images yielded two additional derivates per original [7, 9]. To enable extraction of texture and first-order features [2], voxel intensities were discretized using a fixed-bin-width method [7, 10] with a 2 unit width for PET and CT scans. We customized a Pyradiomics version 2.1.2 pipeline to facilitate image pre-processing [7, 11].

1.2 Ancillary study to determine feature robustness

Given the variable robustness of individual radiomic features to segmentation inconsistencies, we conducted a multiple delineation-based feature pre-selection study to exclude features with low inter- and intra-observer stability from the feature set utilized in this study. Imaging data was acquired from three collections provided by a public imaging repository ("The Cancer Imaging Archive", TCIA) [12] – including (1) the "Head-Neck-PET-CT" collection from four Canadian centers [13, 14]; (2) the "Head and Neck Cancer CT Atlas" collection from MD Anderson Cancer Center dataset [15, 16]; and (3) the "TCGA-HNSC" collection from various institutions across America [17].

Subjects with (1) pre-treatment PET and non-contrast CT scans of the neck, (2) biopsy-confirmed OPSCC, and (3) known p16 or high-risk HPV status were included. Patients with (1) recurrent OPSCC, or (2) >50% of the primary tumor VOI affected by CT artifacts [18] were excluded.

A randomly sampled cohort of 50 patients from the pooled TCIA cohorts (stratified by dataset) was selected. Observer 1 segmented all primary tumors and two randomly selected metastatic nodes in each patient; and re-segmented the same set of lesions >2 months after initial review and segmentation. A second observer created a third set of segmentations. After feature extraction, two intraclass correlation coefficient (ICC) statistics were calculated for each radiomic feature: To assess inter-rater agreement, a two-way random effects, absolute agreement, single rater/measurement ICC was applied; and the two-way mixed effects, absolute agreement, single rater/measurement ICC was used to quantify intra-rater agreement [19, 20]. Features with a lower 95% confidence interval bound ≥0.8 in both inter-and intra-rater assessments were considered stable and retained for further analysis. ICC metrics were separately calculated for primary tumors and the combined set of tumors and nodes. The R "psych" package [21] "ICC" function was used for ICC calculations.

Supplementary table 2 summarizes the results – feature sets exhibited similar inter- and intra-rater ICC scores. PET feature reproducibility was superior to CT in primary tumors, but inferior in the combined set of all lesions. Prior studies reported similar PET and CT primary tumor feature reproducibility scores and ratios [3, 22, 23]. For example, Leijenaar et al. [23] reported 71% of all assessed PET features were highly reproducible (defined as ICC ≥ 0.8). The number and ratio of stable features retained for further analysis in each subset are reported in supplementary table 2.

1.3 Dimensionality reduction techniques

1.3.1 **HClust** – Hierarchical clustering

The R "stats" package (version 3.6.0) [24] "dist" function was used to generate a "euclidean" radiomic feature distance matrix. Next, the "stats" "hclust" function performed hierarchical clustering, applying Ward clustering with Ward's clustering criterion implemented (i.e. "ward.D2" package option) [25]. We cut the dendrogram and retained 30 clusters ("stats" "cutree" function). One "meta-feature" was extracted from each cluster by averaging all radiomic features. Clustering was performed with cross-validation training data only, and meta-feature computation was subsequently applied in all subjects.

1.3.2 **none** – No feature selection

Feature dimensionality reduction was omitted, and the random survival forest models were fit on the unreduced feature set.

1.3.3 **pRF** – Pearson correlation-based redundancy reduction with random forest variable importance

The R "stats" package (version 3.6.0) [24] "cor" function was configured to compute a radiomic feature correlation matrix utilizing Pearson's correlation coefficient (r) based on the cross-validation training set. To reduce pair-wise feature correlation, we excluded the feature with higher mean absolute correlation from any given feature pair with $r > 0.9$ or $r < -0.9$ ("findCorrelation" function of "caret" package [26]).
Thereafter, a random survival forest model was fit on the dimensionality-reduced cross-validation training data ("ranger" package version 0.12.1 [27]). A C-index based split rule [28] was applied to grow 1000 decision trees with the remaining function arguments kept in default. Radiomic feature variable importance scores were queried from the random forest object, and features were ranked in descending order of their respective importance score. The 30 highest-ranked features were selected for survival modelling.

1.3.4 **RIDGE** – RIDGE regularized Cox regression for feature selection

Ridge-regularized Cox survival regression models were trained using the cross-validation training folds ("glmnet" package version 2.0-18 [29] "cv.glmnet" function). The "lambda" parameter was automatically determined in 10-fold cross validation within the "cv.glmnet" fitting process, and each feature's regression coefficient was derived from the fit model at the "lambda" value minimizing the mean cross-validated error. Features were ranked in descending order of their respective absolute regression coefficient value, and the 30 highest-ranked features were selected for survival modelling.

2. Supplementary tables

Supplementary table 1 List of extracted radiomic features

Feature Family		Feature name
First-order	1	10th percentile
	2	90th percentile
	3	Energy
	4	Entropy
	5	Interquartile Range
	6	Kurtosis
	7	Maximum
	8	Mean
	9	Mean Absolute Deviation
	10	Median
	11	Minimum
	12	Range
	13	Robust Mean Absolute Deviation
	14	Root Mean Squared
	15	Skewness
	16	Total Energy
	17	Uniformity
	18	Variance
Shape	1	Elongation
	2	Flatness
	3	Least Axis Length
	4	Major Axis Length
	5	Maximum 2D Diameter (Column)
	6	Maximum 2D Diameter (Row)
	7	Maximum 2D Diameter (Slice)
	8	Maximum 3D Diameter
	9	Mesh Volume
	10	Minor Axis Length
	11	Sphericity
	12	Surface Area
	13	Surface Area to Volume Ratio
	14	Voxel Volume
Texture - Gray Level Cooccurrence Matrix Features	1	Autocorrelation
	2	Cluster Prominence
	3	Cluster Shade
	4	Cluster Tendency
	5	Contrast
	6	Correlation
	7	Difference Average
	8	Difference Entropy
	9	Difference Variance
	10	Informational Measure of Correlation 1

Feature Family		Feature name
	11	Informational Measure of Correlation 2
	12	Inverse Difference
	13	Inverse Difference Moment
	14	Inverse Difference Moment Normalized
	15	Inverse Difference Normalized
	16	Inverse Variance
	17	Joint Average
	18	Joint Energy
	19	Joint Entropy
	20	Maximal Correlation Coefficient
	21	Maximum Probability
	22	Sum Average
	23	Sum Entropy
	24	Sum of Squares
Texture - Gray Level Size Zone Matrix Features	1	Gray Level Non-Uniformity
	2	Gray Level Non-Uniformity Normalized
	3	Gray Level Variance
	4	High Gray Level Zone Emphasis
	5	Large Area Emphasis
	6	Large Area High Gray Level Emphasis
	7	Large Area Low Gray Level Emphasis
	8	Low Gray Level Zone Emphasis
	9	Size Zone Non-Uniformity
	10	Size Zone Non-Uniformity Normalized
	11	Small Area Emphasis
	12	Small Area High Gray Level Emphasis
	13	Small Area Low Gray Level Emphasis
	14	Zone Entropy
	15	Zone Percentage
	16	Zone Variance
Texture - Gray Level Run Length Matrix Features	1	Gray Level Non-Uniformity
	2	Gray Level Non-Uniformity Normalized
	3	Gray Level Variance
	4	High Gray Level Run Emphasis
	5	Long Run Emphasis
	6	Long Run High Gray Level Emphasis
	7	Long Run Low Gray Level Emphasis
	8	Low Gray Level Run Emphasis
	9	Run Entropy
	10	Run Length Non-Uniformity
	11	Run Length Non-Uniformity Normalized
	12	Run Percentage
	13	Run Variance
	14	Short Run Emphasis
	15	Short Run High Gray Level Emphasis
	16	Short Run Low Gray Level Emphasis
	1	Busyness

Feature Family		Feature name
Texture - Neighboring Gray Tone Difference Matrix Features	2	Coarseness
	3	Complexity
	4	Contrast
	5	Strength
Texture - Gray Level Dependence Matrix Features	1	Dependence Entropy
	2	Dependence Non-Uniformity
	3	Dependence Non-Uniformity Normalized
	4	Dependence Variance
	5	Gray Level Non-Uniformity
	6	Gray Level Variance
	7	High Gray Level Emphasis
	8	Large Dependence Emphasis
	9	Large Dependence High Gray Level Emphasis
	10	Large Dependence Low Gray Level Emphasis
	11	Low Gray Level Emphasis
	12	Small Dependence Emphasis
	13	Small Dependence High Gray Level Emphasis
	14	Small Dependence Low Gray Level Emphasis

Complete list of Pyradiomics [11] features used in this study. Exact feature definitions are provided in ref. [7].

Supplementary table 2 Multiple delineation-based feature stability assessment

Radiomics source VOI	Number of lesions (n)	Mean inter-rater ICC (SD)	Mean intra-rater ICC (SD)	Number (percentage) of stable features
Primary tumors	50	**PET**: 0.92 (0.12) **CT**: 0.86 (0.16)	**PET**: 0.91 (0.11) **CT**: 0.89 (0.13)	**PET**: 751 (72.4 %) **CT**: 586 (54.7 %)
Primary tumors and lymph nodes	50 (tumor lesions) 65 (lymph nodes)	**PET**: 0.88 (0.15) **CT**: 0.91 (0.13)	**PET**: 0.87 (0.16) **CT**: 0.93 (0.11)	**PET**: 651 (62.8 %) **CT**: 854 (82.4 %)

Based on three VOI sets created by two observers, inter- and intra-rater ICC were calculated for each feature for primary tumor lesions and a combined set of tumor and lymph node VOI. The mean (standard deviation, SD) ICC in PET and CT feature subsets is reported as well as the numbers and percentages of stable features retained for analysis (lower 95% confidence interval bound of inter- and intra-rater ICC ≥0.8). Unstable features were excluded from any further analysis.

Supplementary table 3 Streak artifact related exclusion of axial CT slices

Radiomics source VOI	Number of lesions (n)	Number (percentage) of lesions with artifacts [a]	Total number of slices (n) [b]	Total number (percentage) of slices with artifacts [c]	Average number (percentage) of slices with artifacts per artifact-positive lesion
Primary tumors	190	91 (47.9 %)	2193	422 (19.2 %)	4.64 (37.1 %)
Lymph nodes	266	38 (14.3 %)	2778	155 (5.6 %)	4.08 (31.0 %)

[a] Number and percentage of lesions with at least one axial CT slice excluded from analysis due to streak artifacts
[b] Total number of axial CT slices across study cohort with visible tumor tissue – including slices with artifacts
[c] Total number and percentage of axial CT slices across study cohort which were excluded from analysis due to streak artifacts

---

Axial CT slices with streak artifacts involving the OPSCC lesion upon visual assessment of scans were excluded from analysis; and metastatic lymph nodes with >50% of the lesion volume involved were entirely excluded. Patients with >50% of the primary gross tumor volume affected by artifacts were not included in this study.

3. References

1. Britz-Cunningham, S.H., J.W. Millstine, and V.H. Gerbaudo, *Improved discrimination of benign and malignant lesions on FDG PET/CT, using comparative activity ratios to brain, basal ganglia, or cerebellum.* Clin Nucl Med, 2008. **33**(10): p. 681-7.
2. Zwanenburg, A., et al. *Image biomarker standardisation initiative*. arXiv e-prints, 2016.
3. Traverso, A., et al., *Repeatability and Reproducibility of Radiomic Features: A Systematic Review.* Int J Radiat Oncol Biol Phys, 2018. **102**(4): p. 1143-1158.
4. Shafiq-Ul-Hassan, M., et al., *Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels.* Med Phys, 2017. **44**(3): p. 1050-1062.
5. Zhao, B., et al., *Reproducibility of radiomics for deciphering tumor phenotype with imaging.* Sci Rep, 2016. **6**: p. 23428.
6. Larue, R.T., et al., *Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures.* Br J Radiol, 2017. **90**(1070): p. 20160665.
7. Pyradiomics-community. *Pyradiomics Documentation Release 2.1.2*. 2018 [cited 2019 December 15th, 2019]; Available from: https://readthedocs.org/projects/pyradiomics/downloads/pdf/2.1.2/.
8. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nat Commun, 2014. **5**: p. 4006.
9. Davnall, F., et al., *Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?* Insights Imaging, 2012. **3**(6): p. 573-89.
10. Leijenaar, R.T., et al., *The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis.* Sci Rep, 2015. **5**: p. 11075.
11. van Griethuysen, J.J.M., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.
12. Clark, K., et al., *The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository.* J Digit Imaging, 2013. **26**(6): p. 1045-57.
13. Vallières, M., et al., *Data from Head-Neck-PET-CT*. 2017: The Cancer Imaging Archive.
14. Vallieres, M., et al., *Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer.* Sci Rep, 2017. **7**(1): p. 10117.
15. Grossberg, A., et al., *Data from Head and Neck Cancer CT Atlas*. 2017: The Cancer Imaging Archive.
16. Grossberg, A.J., et al., *Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy.* Sci Data, 2018. **5**: p. 180173.
17. Zuley, M.L., et al., *Radiology Data from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma [TCGA-HNSC] collection*. 2016: The Cancer Imaging Archive.
18. Ger, R.B., et al., *Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis.* Comput Med Imaging Graph, 2018. **69**: p. 134-139.
19. McGraw, K.O. and S.P. Wong, *Forming inferences about some intraclass correlation coefficients.* Psychological Methods, 1996. **1**(1): p. 30-46.
20. Koo, T.K. and M.Y. Li, *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.* J Chiropr Med, 2016. **15**(2): p. 155-63.
21. Revelle, W., *psych: Procedures for Psychological, Psychometric, and Personality Research*. 2018: Northwestern University, Evanston, Illinois, USA.
22. Lu, L., et al., *Robustness of Radiomic Features in [(11)C]Choline and [(18)F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization.* Mol Imaging Biol, 2016. **18**(6): p. 935-945.
23. Leijenaar, R.T., et al., *Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability.* Acta Oncol, 2013. **52**(7): p. 1391-7.
24. R Development Core Team, *R: A language and environment for statistical computing*. 2019, R Foundation for Statistical Computing: Vienna, Austria.

25. Murtagh, F. and P. Legendre, *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?* Journal of Classification, 2014. **31**(3): p. 274-295.
26. Kuhn, M., et al., *caret: Classification and Regression Training*. 2019.
27. Wright, M.N. and A. Ziegler, *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C plus plus and R.* Journal of Statistical Software, 2017. **77**(1): p. 1-17.
28. Schmid, M., M. Wright, and A. Ziegler *On the use of Harrell's C for clinical risk prediction via random survival forests*. arXiv e-prints, 2015.
29. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* 2010, 2010. **33**(1): p. 22.

11. Acknowledgements

I would like to extend my deep gratitude to the principal investigator of the publications included herein, Ass. Prof. Seyedmehdi Payabvash, MD (Yale Neuroradiology), who envisioned the studies, welcomed me into his laboratory, entrusted me with the implementation of the project, provided continuous mentorship over the course of three years, committed the bulk of resources, and never hesitated to utilize his professional network to help this project succeed.

Likewise, I am very grateful to Prof. Dr. Philipp Baumeister and Prof. Dr. Christoph Reichel (both LMU Otorhinolaryngology), who provided guidance and feedback over the course of the project, and most importantly contributed helpful insights from the perspective of ENT specialists, which – given the interdisciplinary nature of our research – proved to be invaluable in this project.

I would especially like to thank Dr. Kariem Sharaf (LMU Otorhinolaryngology), who not only served as a scientific mentor from the very beginning of our research endeavor, but also co-envisioned the LMU-Yale cooperation which constitutes the foundation for my dissertation project. He played a major support role in many organizational matters and relied on his personal network to enable and advance our research.

I would also like to extend my gratitude to Tal Zeevi, MSc (Yale Radiology) for providing unwavering engineering support regarding machine learning analyses.

Furthermore, I am grateful to Prof. Barbara Burtness, MD (Yale Medical Oncology, Chair of ECOG-ACRIN Head and Neck Therapeutics Committee), Prof. Manju L. Prasad, MD (Yale Pathology), Prof. Benjamin L. Judson, MD (Yale Otolaryngology), Ass. Prof. Benjamin H. Kann, MD (Harvard Radiation Oncology, Dana-Farber Cancer Institute), Assoc. Prof. Reza Forghani, MD, PhD (McGill University Radiology) and Assoc. Prof. Amit Mahajan, MD (Yale Neuroradiology), who provided scientific guidance in matters related to their respective fields and/or supported conceptualization of our studies and/or facilitated data acquisition.

Finally, I would like to acknowledge monetary support from the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes, merit-based scholarship, 2018 – 2022), the Biomedical Education Program (scholarship supporting my research at Yale, 2019 – 2020, Univ.-Prof. Dr. Torsten Doenst, Univ.-Prof. Dr. Dr. Michael Marschollek, Univ.-Prof. Dr. Hans-Christoph Pape, et al.), the Klee Family Foundation (student research grant, 2019, Dr. Christiane Fechner, et al.), and all those entities who provided travel support for conference attendance (European Society of Radiology; German Society of Oto-Rhino-Laryngology, Head and Neck Surgery; Rolf W. Günther Foundation for Radiological Sciences).

12. Affidavit

Stefan Philipp Markus Haider

I hereby declare, that the submitted thesis entitled:

“PET/CT Radiomics and Machine Learning for Non-Invasive Molecular and Prognostic Characterization of Oropharyngeal Squamous Cell Carcinomas”

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been submitted to any other university for the purpose of obtaining an academic degree.

Munich, 19.04.2023 Stefan Philipp Markus Haider