Largest diameter delineations can substitute 3D tumor volume delineations for radiomics prediction of human papillomavirus status on MRI's of oropharyngeal cancer

Physica Medica 101 (2022) 36–43

Contents lists available at ScienceDirect

Physica Medica

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

Original paper

Largest diameter delineations can substitute 3D tumor volume delineations for radiomics prediction of human papillomavirus status on MRI's of oropharyngeal cancer

Paula Bos [a,b,c,*], Michiel W.M. van den Brekel [b,d], Marjaneh Taghavi [a], Zeno A.R. Gouw [e], Abrahim Al-Mamgani [e], Selam Waktola [a], Hugo J.W.L. Aerts [a,f,g], Regina G.H. Beets-Tan [a,c,h], Jonas A. Castelijns [a], Bas Jasperse [a,i]

[a] Department of Radiology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[b] Department of Head and Neck Oncology and Surgery, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[c] GROW School for Oncology and Developmental Biology, University of Maastricht, Maastricht, the Netherlands
[d] Department of Oral and Maxillofacial Surgery, Amsterdam University Medical Center (AUMC), Amsterdam, the Netherlands
[e] Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[f] Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, United States
[g] Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands
[h] Department of Regional Health Research, University of Southern Denmark, Denmark
[i] Department of Radiology, Amsterdam University Medical Center, Amsterdam the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Human papillomavirus
Radiomics
Segmentation

ABSTRACT

Purpose: Laborious and time-consuming tumor segmentations are one of the factors that impede adoption of radiomics in the clinical routine. This study investigates model performance using alternative tumor delineation strategies in models predictive of human papillomavirus (HPV) in oropharyngeal squamous cell carcinoma (OPSCC).

Methods: Of 153 OPSCC patients, HPV status was determined using p16/p53 immunohistochemistry. MR-based radiomic features were extracted within 3D delineations by an inexperienced observer, experienced radiologist or radiation oncologist, and within a 2D delineation of the largest axial tumor diameter and 3D spheres within the tumor. First, logistic regression prediction models were constructed and tested separately for each of these six delineation strategies. Secondly, the model trained on experienced delineations was tested using these delineation strategies. The latter methodology was repeated with the omission of shape features. Model performance was evaluated using area under the curve (AUC), sensitivity and specificity.

Results: Models constructed and tested using single-slice delineations (AUC/Sensitivity/Specificity: 0.84/0.75/0.84) perform better compared to 3D experienced observer delineations (AUC/Sensitivity/Specificity: 0.76/0.76/0.71), where models based on 4 mm sphere delineations (AUC/Sensitivity/Specificity: 0.77/0.59/0.71) show similar performance. Similar performance was found when experienced and largest diameter delineations (AUC/Sens/Spec: 0.76/0.75/0.65 vs 0.76/0.69/0.69) was used to test the model constructed using experienced delineations without shape features.

Conclusion: Alternative delineations can substitute labor and time intensive full tumor delineations in a model that predicts HPV status in OPSCC. These faster delineations may improve adoption of radiomics in the clinical setting. Future research should evaluate whether these alternative delineations are valid in other radiomics models.

Abbreviations: AUC, Area under the curve; 95% CI, 95% Confidence interval; CRT, Chemoradiation therapy; DSC, Dice Similarity Coefficient; GTV, Gross Tumor Volume; HD, Hausdorff Distance; HPV, Human Papilloma Virus; ICC, Intraclass Correlation Coefficient; LoG, Laplacian of Gaussian; OPSCC, Oropharyngeal squamous cell carcinoma.

* Corresponding author.
E-mail address: paulabos01@gmail.com (P. Bos).

https://doi.org/10.1016/j.ejmp.2022.07.004
Received 13 October 2021; Received in revised form 11 July 2022; Accepted 13 July 2022
Available online 23 July 2022

Introduction

Background

Radiomics is a promising tool for the non-invasive detection of clinically relevant tumor characteristics. These characteristics can be used to predict treatment response[1,2], classify tumor types[3,4] or discriminate tumor properties[5,6]. Radiomics analysis requires various steps that include image acquisition, image pre-processing, tumor delineation, feature extraction, feature selection and model construction. These steps can be controlled easily within research settings, but poses challenges with regard to reproducibility and repeatability in daily clinical practice[7–9]. Even if these challenges and other requirements for clinical implementation[10,11] are overcome, time consuming expert tumor delineations, taking valuable hours to complete, hampers further adoption of radiomics in daily clinical practice[12].

Time reduction with regard to tumor delineation can be achieved by either automated delineation strategies or manual delineation strategies which are easier to implement. Previous studies have shown that variability of tumor delineations can impact model performance. However, these studies[9,13] mainly focused on the consequences of (semi-) automatic alteration of available manual full tumor delineations on model performance. The methods used in these studies cannot be translated to adequate delineation strategies that would reduce time and labor consumption of manual tumor delineations needed for the implementation of radiomics in a clinical setting. A study comparing models based on rough and precise tumor delineations found that radiomic features extracted from precise delineations were more informative for prediction of overall survival in non-small cell lung cancer patients[14]. These interesting findings show that the choice of delineation strategy can lead to substantial variations in radiomic results[14]. Consensus of the most suitable delineation strategy is therefore highly recommended to standardize the radiomic workflow and increase clinical implementation.

In this study we investigate whether the performance of a previously published[5] radiomics model predictive of human papillomavirus (HPV) status of oropharyngeal squamous cell carcinomas (OPSCC) is similar when fast (“simple”, “rough”) or readily available tumor delineations are used compared to the time consuming standard expert tumor delineations. The following fast or readily available tumor delineations will be considered: tumor volumes delineated by a non-experienced observer, the readily available gross tumor volumes (GTV) delineated by radiation oncologists, tumor delineations extracted on the axial slice with the largest diameter and a simple strategy where a sphere was drawn within the tumor volume. Radiomic features (i.e. radiomics signature) are selected during model construction and may depend on the delineation strategy used. To ensure that the same radiomic features are detected when the model is applied to a new case, one can assume that the same delineation strategy should be used when implementing the model. Under this assumption, separate models will be constructed for each delineation strategy. On the other hand, alternative delineations may be able to adequately quantify relevant features that were selected in a model trained using the optimal expert 3D tumor delineations. Under this assumption, the performance of the model constructed using optimal delineations will be applied using the alternative delineations. The latter approach will be repeated while omitting shape and size features, as some of the alternative delineations are not able to quantify these features.

Materials and methods

The study was approved by the local institutional review board (IRBd18047). Due to the retrospective nature of the study, informed consent was waived.

Study population

A cohort of 240 patients with histologically proven primary OPSCC, treated with chemoradiation (CRT) between January 2010 and December 2015 at our Institute was considered. All patients had no history of previous head and neck malignancies. The main exclusion criteria were (a) no determined HPV status of the tumor, (b) no available pretreatment MRI, (c) poor image quality, (d) undetectable tumors, and, (e) a second head and neck primary tumor. In total, 153 patients were eligible for this study. HPV status of the tumor was determined on biopsy material using p16 and p53 immunohistochemistry using the methodology described in Henneman et al.[15].

Image acquisition

Pretreatment MR and CT images were acquired as part of the clinical routine. T1-weighted postcontrast (postcontrast T1W) MRI was used for analysis, with a slice thickness ranging between 0.8 and 1.0 mm, TR/TE: 4300–10000/1.7–4.6 ms, echo train length of 60–90 and 10° flip angle.

CT images for GTV delineation were acquired during treatment planning from two scanners. All CT images had a slice thickness of 3 mm, a tube current of 120 kV, and an exposure ranging from 19 to 509 mAs.

Tumor delineations

Primary tumors were delineated using six delineation strategies (see below and Fig. 1), including three delineations covering the whole tumor volume and three delineations including only a part of the tumor (“simple delineations”, e.g. spherical volumes). Whole tumor volumes represent the full 3D tumor volume, where “simple delineations” evaluates tumor delineation strategies which might easily implementable in the clinic. Tumors were delineated on postcontrast T1W MRI, except for the GTV delineation. Observers were allowed to review other available imaging modalities to improve tumor delineation and were blinded to HPV status. Delineations were performed using the 3D slicer software (version 4.8.0, https://www.slicer.org). The annotation time for each delineation time was recorded.

1. *3D Non-experienced observer*: One observer in training (PB, 1 year of experience in head and neck diagnosis) delineated the 3D tumor volume.
2. *3D Experienced observer*: An experienced radiologist (BJ, >7 years of expertise in head and neck diagnosis) reviewed and corrected the *Non-experienced* tumor delineation.
3. *3D GTV*: GTV was delineated on contrast-enhanced planning CT-scan for radiotherapy treatment purposes by a radiotherapist, with the allowance to review planning MRI when available. Planning CT and its GTV contouring were registered to post contrast T1W using B-spline registration (SimpleElastix[16], see Appendix A).
4. *2D Largest Diameter*: The slice with the largest axial tumor diameter was automatically selected from the 3D *Experienced* manual tumor delineation using Python scripting (version 3.4, https://www.python.org).
5. *3D Spherical ⌀4mm*: A sphere of 4 mm was placed in the most solid part of the tumor by the *Non-experienced* observer. A size of 4 mm was selected since this was the minimum maximal tumor diameter included in the cohort.
6. *3D Spherical ⌀BestFit*: A sphere with the largest possible diameter (best fit) was placed in the most solid tumor area by the *Non-experienced* observer.

The spherical tumor delineations were delineated one year after initial delineation of the *Non-experienced* observer, blinded to the initial delineation to prevent memory bias.

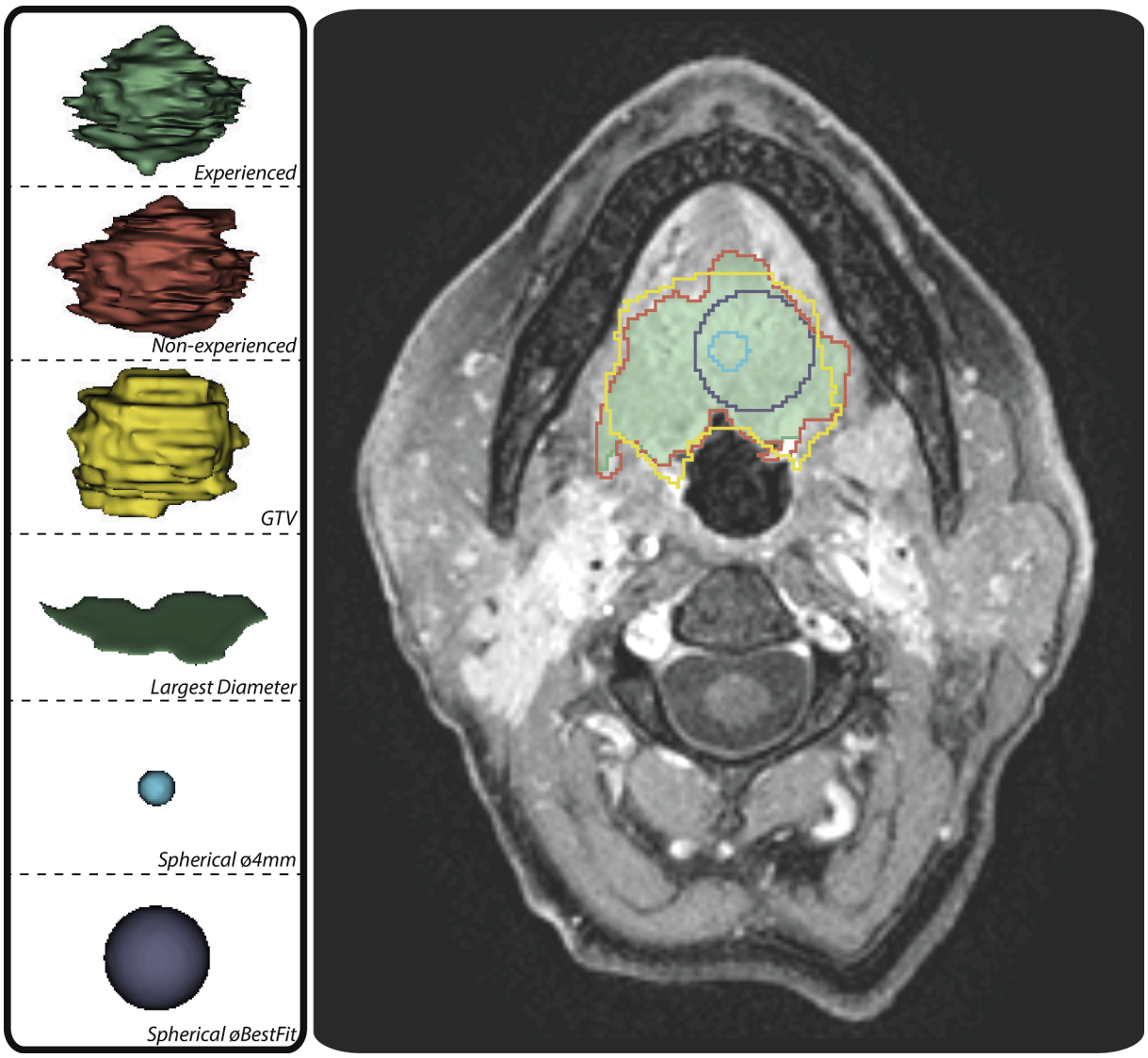

Fig. 1. An illustration of the six manual delineations. The six individual delineations are visualized in the left box. The right box illustrates these delineations on postcontrast T1w MRI on the slide with the largest axial diameter.

Image pre-processing

Prior to analysis, MR images were normalized, resampled and discretized. Image normalization was applied with zero mean and unit standard deviation to avoid inhomogeneity between MRI scans. Comparable quantification of radiomic features in all directions was obtained by resampling MR images to isotropic voxels of 1.0 mm using B-spline interpolation. Finally, MR intensity values were discretized into a fixed bin width of five intensity values to allow quantification of texture. All image pre-processing steps were performed using the open-source package PyRadiomics[17].

Radiomic features

Radiomic features were extracted using PyRadiomics[17] for each separate delineation strategy. Features were divided into the categories shape, intensity and texture. These features were extracted from the original image, the image with a wavelet filter and the image with a Laplacian of Gaussian (LoG) filter. A wavelet filter was used to examine different spatial frequencies of the image in 8 decompositions, where a LoG filter determines different texture coarseness (4 levels, sigma of 0.5, 1.0, 1.5 and 2.0 mm). A total of 1184 radiomic features were extracted for each delineation.

Stable features were assessed by intraclass correlation coefficient (ICC) and Mann-Whitney *U* test for each separate delineation strategy used for model construction. First, radiomic features were considered to be stable when ICC between the radiomic features extracted from the experienced radiologist and the appropriate tumor delineation (*Non-experienced*, *GTV*, *Largest Diameter*, *Spherical ø4mm* and *Spherical øBest-Fit*) was higher than 0.75. For the *Experienced* model, ICC was calculated between features extracted from the *Experienced* reader and *Non-experienced* reader. ICC calculated stable features were assessed by Mann-Whitney *U* test to exclude differences of magnetic field strength. Features without significant differences (p-value ≥ 0.05) were considered stable. Finally, collinearity between the remaining stable features was assessed by Pearson correlation (>0.9), removing the features with the largest collinearity. The stable features for each separate delineation strategy were used as input for the prediction model.

Construction of the radiomics models

Features were standardized per delineation strategy, using zero mean and unit variance, to obtain scalar homogeneity in each approach. Then, recursive feature elimination[18] was used to select a feature

subset by iteratively removing the feature with the weakest importance score. The remaining feature subset was used for analysis by the logistic regression classifier to predict HPV tumor status and subsequent model testing.

For the prediction model, the cohort was divided into a training (60%, n = 91) and test (40%, n = 62) subset, stratified by magnetic field strength and HPV status of the tumor. Hyperparameters for classification were optimized using 1000 iterations of Bayesian hyperparameter optimization on the training subset. During this step, fourfold cross-validation was applied to calculate the minimal loss function. Then, the optimal hyperparameters were applied on the unseen test set to evaluate prediction performance. A detailed description of the workflow can be found in our previous publication[5]. The radiomic pipeline is summarized in Fig. 2.

The impact of tumor delineation variability on the prediction performance of HPV was investigated using three methods:

Method 1: Separate model construction and testing for each delineation strategy

Prediction models were built (trained and validated) and tested on each tumor delineation separately (*Experienced, Non-experienced, GTV, Largest diameter, Spherical ⌀4mm and Spherical ⌀BestFit)*, resulting in six separate models. To prevent artificial inflation of model performance, all models were forced to select the same number of features as selected in the experienced model.

Method 2: Testing the experienced model using the alternative delineations

Performance of the prediction model that was trained and validated using *Experienced* delineations was tested on the test subset using each of the six tumor delineation strategies.

Method 3: Testing the experienced model without shape and size features using the alternative delineations

As spherical or 2D delineations do not reliably represent shape and size features, the *Experienced* model was trained and validated without shape and size features and tested using the six alternative delineations.

Statistical analysis

An independent *t*-test was applied to calculate differences in age for both HPV status groups. Fisher exact test was applied to the other clinical variables. A p-value below 0.05 was considered as significant. Spatial agreement between the six delineation strategies was calculated by using the Dice Similarity Coefficient (DSC)[19] and Hausdorff Distance (HD)[20].

Performance of the prediction models was evaluated by area under the curve (AUC), sensitivity and specificity. Median values, with its 95% confidence interval (95% CI) were calculated using 500 iterations of bootstrap (with replacement) using the test set.

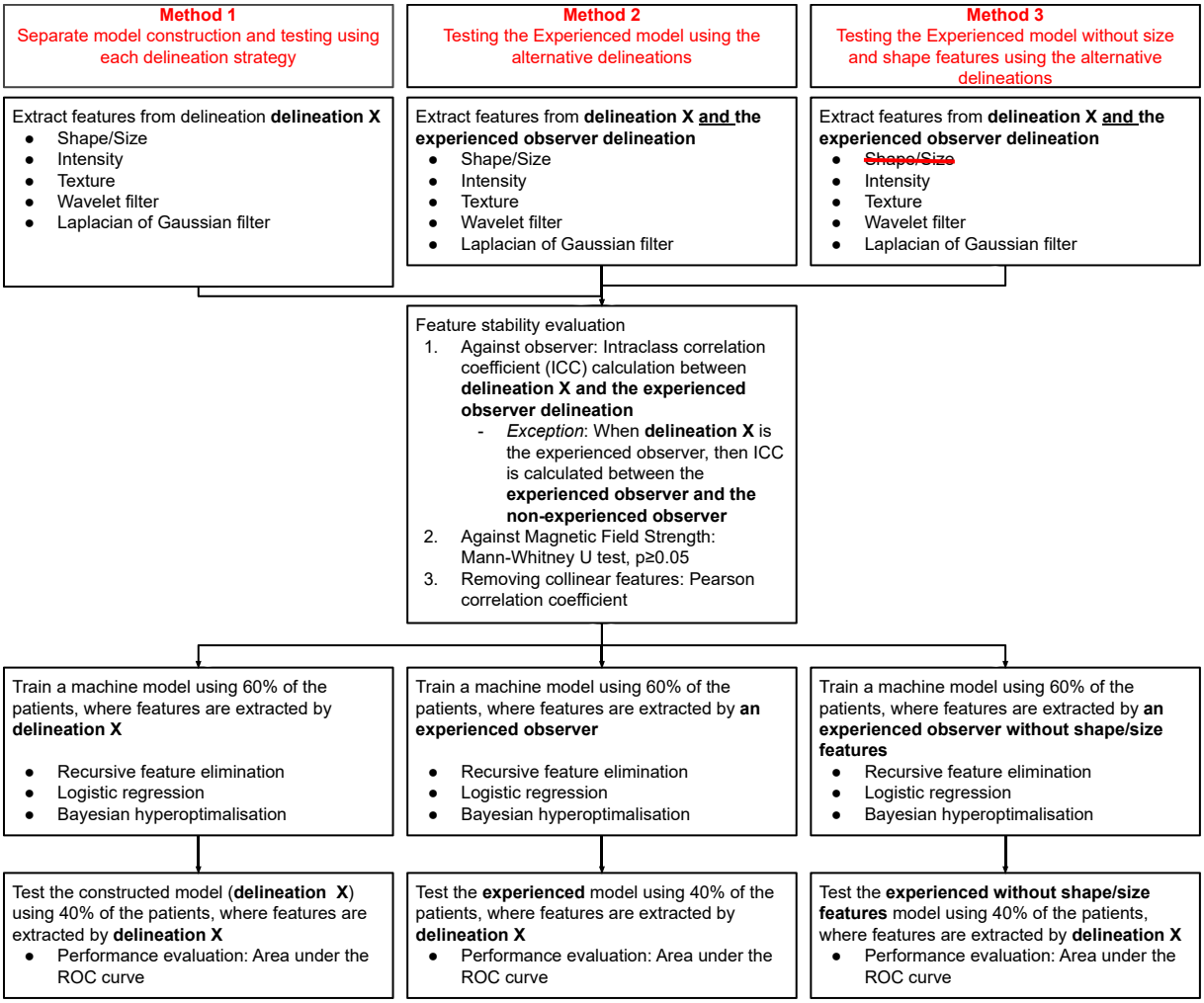

Fig. 2. A flowchart describing the radiomic workflow of the three methods. Delineation X can be one of the six delineation strategies, including the experienced observer, non-experienced observer, gross tumor volume (GTV), largest diameter on the single slice, a sphere with a diameter of 4 mm or a sphere with a diameter best fitted in the tumor volume.

Results

Patient demographics

Patient demographics are summarized in Table 1. The patients show an equal distribution for HPV tumor classification (n = 77 HPV negative tumors, n = 76 HPV positive tumors). Younger (p = 0.007), non-smoking patients (p < 0.001) with a high *T*-classification (p=<0.0001) or tumor not located in the soft palate (p = 0.017) were more likely to have HPV positive tumors. Other cancer subsites and gender were not significantly different between HPV negative and positive tumors. *N*-classification was slightly higher in HPV positive compared to HPV negative tumors with near significance (p = 0.051).

Time recordings

The *Non-experienced* observer delineated a tumor with a median of 34 min [range: 25–65], and was checked and corrected in a median of 9 min [range: 6–14] by the *Experienced* observer. The time required to place a ROI with a diameter of 4 mm or user-determined diameter was 1.5 and 3 min, respectively. *Largest Diameter* delineations were automatically extracted, and therefore, obtained within seconds. Time recordings of *GTV* delineations were not available, since those were previously delineated for radiotherapy purposes.

Tumor delineation agreement

Agreement between tumor volumes was calculated with DSC and HD, see Table A.1 and Table A.2. The *Experienced* and *Non-experienced* observer show reasonable similarity with a mean DSC of 0.84 and mean HD of 18.7 mm. *GTV* tumor delineation shows a lower similarity with *Experienced* observer (DSC: 0.43, HD: 183.3 mm).

Logistic regression prediction model

Prediction performances of all models for the three methods are summarized in Table 2, ROC curves are visualized in Fig. 3.

Method 1: Separate model construction and testing for each delineation strategy

0.3 to 6.5% of the total features were defined as stable (see Table 3), resulting in 77, 10, 20, 4 and 13 radiomic features as input for the *Experienced/Non-experienced*, *GTV*, *Largest Diameter*, *Spherical ⌀4mm* and *Spherical ⌀BestFit* model, respectively.

The model built and tested based on *Largest Diameter* delineation shows higher performance, higher specificity and similar sensitivity (AUC/Sens/Spec: 0.84/0.75/0.84) compared to the standard *Experienced* model (AUC/Sens/Spec: 0.76/0.76/0.71). Prediction performance of the *Spherical ⌀4mm* delineations model was comparable to standard *Experienced* delineation model with slightly lower sensitivity and similar specificity (AUC/Sens/Spec: 0.77/0.59/0.71). Performance of models based on *Non-experienced* (AUC/Sens/Spec: 0.68/0.69/0.55), *GTV* (AUC/Sens/Spec: 0.71/0.69/0.58) and *Spherical ⌀BestFit* (AUC/Sens/Spec: 0.64/0.59/0.62) delineations were considerably lower than the standard *Experienced* model.

Table A.3 summarizes the selected features for each model. The models based on *Experienced* and *Largest Diameter* delineation include shape/size features (sphericity and maximum 2D diameter respectively), as well as textural features. Models based on the other delineations included only textural features.

Method 2: Testing the experienced model using alternative delineations

The standard *Experienced* model shows the highest performance when tested on expert radiologist tumor delineations (AUC/Sens/Spec: 0.76/0.76/0.71). Overall performance and specificity were considerably lower when the *Experienced* model was tested using the *Non-experienced* (AUC/Sens/Spec: 0.63/0.76/0.50) delineations. Test performance approached randomness when tested with the remaining delineations. Sensitivity and specificity for testing with the 2D or spherical delineations were 0 and 1 or vice versa.

Method 3: Testing the experienced model without shape and size features using the alternative delineations

Of the extracted 1184 radiomic features, 14 features belong to the shape and size group. Those 14 features were excluded when shape and size features were omitted. Of the remaining 1170 features, 71 (6.1%) features were considered as stable (see Table 3).

Performance of the *Experienced* model without shape and size features was comparable to the standard Experienced model *with* shape and size features (AUC/Sens/Spec: 0.76/0.75/0.65 vs 0.76/0.76/0.71). This performance is similar to the *Largest Diameter* model (AUC/Sens/Spec: 0.76/0.69/0.69). Prediction performances increased when the *Experienced model without shape and size features* was tested using *Non-experienced* delineations (AUC/Sens/Spec: 0.82/0.76/0.80). Performance of this model using G*TV*, *Spherical ⌀BestFit* or *Spherical ⌀4mm* delineations was considerably lower, as summarized in Table 2.

Table 1
Patient characteristics for the total cohort and subgroups stratified by HPV status. Summaries are given as number of patients and % of the total group between parentheses. Median and interquartile range (IQR) are used to summarize continuous variables. [a]Independent *t*-test, [b]Fisher's exact test and [c]Chi-square test. Values were statistic significant (marked with an asterisk) if p-value was below 0.05 (p < 0.007 after Bonferroni correction).

	Total cohort	HPV negative	HPV positive	P-value
Patients, *n*	153	77	76	–
Age, *median y* (IQR)	61 (56–66)	63 [57–67]	59 [55–65]	0.007[a]*
Sex, *n male* (%)	96 (63)	54 (70)	42 (55)	0.067[b]
Smoking, *n* (%)	114 (75)	72 (94)	42 (55)	<0.001[b]*
T-stage, *n* (%)				<0.001[b]*
T1 + T2	78 (51)	25 (32)	53 (70)	–
T3 + T4	75 (49)	52 (68)	23 (30)	–
N-stage (N > 0), *n* (%)	127 (83)	59 (77)	68 (89)	0.051[b]
Subsite of cancer				0.406[c]
Tonsillar tissue	88 (58)	42 (55)	46 (60)	0.514[b]
Soft palate	13 (8)	11 (14)	2 (3)	0.017[b]*
Base of tongue	48 (31)	20 (26)	28 (37)	0.166[b]
Posterior wall	4 (3)	4 (5)	0 (0)	0.120[b]

*Note: HPV indicates Human Papillomavirus.

Discussion

This study shows that less labor-intensive, easily applicable, delineations might substitute labor-intensive experienced delineations in the application of radiomics models to predict HPV status. Moreover, some of these alternative delineation strategies seem to increase model performance compared to standard expert delineations.

In contrast to our expectations, all delineations (except *Spherical ⌀BestFit)* show good prediction performance, regardless of delineation precision. This suggest that each separate delineation capture information with regard to tumor biology in a different matter.

The model based and tested on largest tumor diameter delineations appeared to outperform the standard experienced delineation based model. This may be explained by the effect of interpolation on the radiomic features. Interpolation is recommended as necessary preprocessing step to correct for pixel size and slice thickness variance for 3D volumes. This interpolation to isotropic voxels induces smoothing effects that might remove relevant feature information from 3D delineations that will be present in (unsmoothed) 2D tumor delineations [21]. Additional experiments (see appendix B) supports this hypothesis, as performance of a model based on 3D tumor volumes delineated by an experienced observer (AUC: 0.74) increases when interpolation was omitted (AUC: 0.81).

Table 2
Performances (expressed in AUC, sensitivity and specificity) of the models of all three methods in predicting human papillomavirus (HPV) status of the tumor. Confidence intervals were calculated from 500 times bootstrapping. Stable features were calculated between features extracted from the experienced delineation ([a]) and the listed delineations ([b]).

Method		1			2			3		
Model construction specifics	Stable feature selection based on	Experienced and listed delineations [a,b]			Experienced and Non-experienced delineations			Experienced and Non-experienced delineations		
	Features removed	None			None			Shape and size features		
	Model construction based on	Listed delineations*			Experienced			Experienced		
Model testing results	**Delineation**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**
	Experienced[a]	0.76 [0.76–0.77]	0.76 [0.75–0.77]	0.71 [0.70–0.72]	0.76 [0.76–0.77]	0.76 [0.75–0.77]	0.71 [0.70–0.72]	0.76 [0.76–0.77]	0.75 [0.74–0.76]	0.65 [0.64–0.66]
	Non-experienced[b]	0.68 [0.68–0.69]	0.69 [0.68–0.70]	0.55 [0.54–0.56]	0.63 [0.63–0.64]	0.76 [0.76–0.77]	0.50 [0.49–0.51]	0.82 [0.81–0.82]	0.76 [0.75–0.77]	0.80 [0.79–0.81]
	GTV[b]	0.71 [0.70–0.72]	0.69 [0.68–0.70]	0.58 [0.56–0.59]	0.53 [0.52–0.54]	0.33 [0.32–0.34]	0.61 [0.60–0.62]	0.70 [0.69–0.70]	0.75 [0.74–0.76]	0.65 [0.64–0.66]
	Largest Diameter[b]	0.84 [0.83–0.85]	0.75 [0.74–0.76]	0.84 [0.83–0.85]	0.53 [0.52–0.54]	0	1	0.76 [0.75–0.76]	0.69 [0.68–0.70]	0.69 [0.68–0.70]
	Spherical ⌀4mm[b]	0.77 [0.76–0.77]	0.59 [0.58–0.60]	0.71 [0.70–0.72]	0.39 [0.38–0.40]	1	0	0.58 [0.57–0.58]	0.56 [0.55–0.57]	0.50 [0.49–0.51]
	Spherical ⌀BestFit[b]	0.64 [0.64–0.65]	0.59 [0.58–0.60]	0.62 [0.60–0.63]	0.52 [0.51–0.53]	1	0	0.67 [0.66–0.68]	0.59 [0.58–0.60]	0.68 [0.67–0.69]

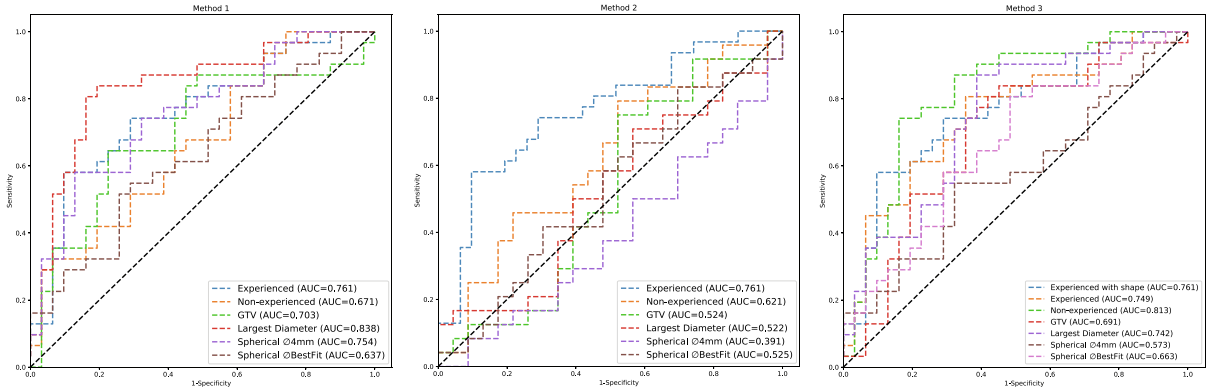

Fig. 3. Receiver operating characteristic (ROC) curves of the three methods. Performances of the test set for each individual delineation are assessed by the area under the curve (AUC).

Table 3
The number of features after each stability check for each observer versus the experienced delineation model. The number of stable features are given, with the percentage of the total number of features between parentheses.

Experienced vs observer delineation\Stability check	Non-experienced[a] (%)	GTV (%)	Largest Diameter (%)	Spherical ⌀4mm (%)	Spherical ⌀BestFit (%)	Experienced without shape and size features (%)
None	1184 (100)	1184 (100)	1184 (100)	1184 (100)	1184 (100)	1170 (100)
for delineation (ICC > 0.75)	926 (78.2)	241 (20.4)	483 (40.8)	90 (7.6)	310 (26.2)	913 (78.0)
for magnetic field strength (mwu ≥ 0.05)	240 (20.3)	34 (2.9)	68 (5.7)	11 (0.9)	64 (5.4)	231 (19.7)
for collinear features (Pearson > 0.9)	77 (6.5)	10 (0.8)	20 (1.7)	4 (0.3)	13 (1.1)	71 (6.1)

Note: [a]Stable features for the *Experienced* model *with* shape and size features were also selected by this comparison.
ICC represents Interclass correlation Coefficient; Mwu, Mann-Whitney U Test.

Poor model performance was observed when the standard experienced model was applied to the test subset using the alternative delineations. This poor test performance might be explained by the reduced ability of the "faster" delineations to adequately quantify the sphericity feature (see appendix Table A.3) that is part of the experienced model. This does not rule out that applying the experienced model using alternative delineations may be useful in other predictive models that only rely on textural features.

Removal of shape and size features (method 3) did not change the performance when the model was constructed and tested using the expert radiologist delineations. As expected, prediction performances were considerably better when this experienced model (constructed *without* shape and size features) was tested with the alternative delineations compared to the standard experienced model (constructed

with shape and size features (method 2)). Taken together, this implies that the loss of shape and size features might be adequately compensated with textural features without losing predictive properties.

To make radiomics clinically applicable, substitution of the labor-intensive time-consuming delineations is desirable. This study shows that easy delineation strategies needed a shorter time to perform the delineation (*Non-experienced delineation* vs *Spherical* delineation: 34 min vs 3 min). While no direct comparison can be made for the 2D delineation, it can be safely assumed that delineating only a single slice requires less time compared to the full 3D tumor delineation. Taking prediction performance and ease of implementation into account, the largest diameter seems to be the most preferable alternative delineation strategy.

Evidently, the findings of this study are only applicable to models predicting HPV in OPSCC. Other delineation strategies may be more applicable for radiomics models trained to predict other outcome variables or applied to other tumor types. Besides tumor delineation and the studied outcome parameter, each step of the radiomic pipeline shows large variations, limiting reproducible and repeatable results[7–9,22]. Preselected choices in image acquisition, tumor delineation, feature selection and/or machine learning model construction parameters directly affect the radiomic pipeline and therefore the set of predictive features. Though all these variances, direct and reliable comparison between studies is limited.

A good example of this are the contrary results between findings of this study and Lang *et al.*[23] regarding the superiority of 2D delineations over 3D tumor volumes in the prediction of HPV status. Significant differences within the methodology (e.g. MR images vs CT images, machine learning model vs deep learning model, feeding one vs multiple 2D slices in the model) impede critical evaluation.

As our study aimed to find suitable delineation alternatives to full tumor delineations by an experienced observer, observer variability of model performance was not assessed. Observer variability of delineations should be addressed in future studies, or studies aiming to adopt this alternative delineation approach. It is obvious that observer variability is less of an issue in the proposed faster delineations compared to full tumor volume delineations as tumor margins are not delineated. Another important limitation of this study is the bias introduced by interdependency of delineations. The single slice delineations are calculated from the expert 3D delineations, which may inflate the performance of single slice delineations compared to the 3D delineations. Furthermore, the results presented for the single slice delineations do not represent the real-world scenario of an observer manually selecting and delineating the largest tumor diameter from the image. Additionally, expert and non-expert delineations are not totally independent, as the expert delineations are basically the corrected non-expert delineations. Future research should take these limitations into account by evaluating independently acquired manual delineations.

Besides the easy implementation of radiomics in the clinical workflow, the alternative delineations would also benefit standardization of radiomics analysis. Reliable automatic segmentation of tumors would be the best solution to time and labor-intensive delineations while eliminating interobserver bias[10,11]. Multiple studies investigated the potential of deep learning in auto segmentation in head and neck cancer patients, where substantial overlap (DSC $>$ 0.74) between the manual and automatic delineations was shown[24,25]. Other studies proposed multi-task deep learning to combine automatic segmentations with models predictive of treatment outcome[26] or HPV status[23]. However, to our knowledge, no reliable automatic tools for the delineations of complex oropharyngeal tumors based on MR images are available at this point in time, and therefore automatic delineations are not included in this study.

As mentioned earlier, various factors can influence robustness and stability of individual features and should be used to select the most suitable feature for every radiomics model. Feature stability across delineations was used as a selection criterion in this study, where features were defined as stable when agreement between the experienced radiologist and the appropriate delineation was high. By selecting features with only high agreement, features prognostic for HPV status might be eliminated since they were different across full tumor and single slice delineation. Additionally, feature robustness can be influenced by the MRI scanner used and circumstances under which the MRI scan was performed[22]. Evidently, this could not be addressed in this single center study, and should be addressed in future projects.

Recently, advances have been made to increase performance of radiomics models by improving image quality using AI techniques. For instance, Chen *et al.* have improved the predictive performance of a radiomics model by denoising CT images using Generative Adversarial Networks[27]. These techniques could also be employed to improve the quality of MRI images and/or the similarity of MRI image acquired from different scanners. By improving predictive performance of radiomics models, these technique might also increase performance of the alternative delineation strategies proposed in this study.

Conclusions

In conclusion, this study shows that alternative delineations with low labor/time consumption can substitute labor and time intensive full tumor delineations in the application of a model that predicts HPV status in OPSCC. These faster delineations may improve adoption of radiomics in the clinical setting. Evidently, the findings in this paper are only relevant to the radiomics model predicting HPV status used in this paper, future research should evaluate whether these alternative delineations are valid in other radiomics models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2022.07.004.

References

[1] Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. J Oral Pathol Med 2020;49:977–85. https://doi.org/10.1111/jop.13089.
[2] Yuan Y, Ren J, Shi Y, Tao X. MRI-based radiomic signature as predictive marker for patients with head and neck squamous cell carcinoma. Eur J Radiol 2019;117:193–8. https://doi.org/10.1016/j.ejrad.2019.06.019.
[3] Freuhwald-Pallamar J, Hesselink JR, Mafee MR, Holzer-Freuhwald L, Czerny C, Mayerhoefer ME. Texture-Based Analysis of 100 MR Examinations of Head and Neck Tumors – Is It Possible to Discriminate Between Benign and Malignant Masses in a Multicenter Trial? *NMR Biomed.* 2013:26:195-202. https://doi.org/ 10.1055/s-0041-106066.
[4] Zheng Y-M, Xu W-J, Hao D-P, Liu X-J, Gao C-P, Tang G-Z, et al. A CT-based radiomics nomogram for differentiation of lympho-associated benign and malignant lesions of the parotid gland. Eur Radiol 2021;31(5):2886–95.
[5] Bos P, Brekel MWM, Gouw ZAR, Al-Mamgani A, Waktola S, Aerts HJWL, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. Head Neck 2021;43(2):485–95.
[6] Romeo V, Cuocolo R, Ricciardi C, Ugga L, Cocozza S, Verde F, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. Anticancer Res 2020;40(1):271–80.
[7] Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology 2018;288(2):407–15.
[8] Pfaehler E, Zhovannik I, Wei L, Boellaard R, Dekker A, Monshouwer R, et al. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. Phys Imaging Radiat Oncol 2021;20:69–75.
[9] Liu R, Elhalawani H, Radwan Mohamed AS, Elgohari B, Court L, Zhu H, et al. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. Clin Transl Radiat Oncol 2020;21:11–8.

[10] Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: From machine learning to deep learning. Physica Med 2021;83:9–24.
[11] Balagurunathan Y, Mitchell R, El Naga I. Requirements and reliability of AI in the medical context. Physica Med 2021;83:72–8. https://doi.org/10.1016/j.ejmp.2021.02.024.
[12] Harari PM, Song S, Tomé WA. Emphasizing Conformal Avoidance Versus Target Definition for IMRT Planning in Head-and-Neck Cancer. Int J Radiat Oncol Biol Phys 2010;77:950–8. https://doi.org/10.1016/j.ijrobp.2009.09.062.
[13] Zhang X, Zhong L, Zhang B, Zhang Lu, Du H, Lu L, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: Evaluation with two disease groups. Cancer Imaging 2019;19(1). https://doi.org/10.1186/s40644-019-0276-7.
[14] Sepehri S, Tankyevych O, Iantsen A, Visvikis D, Hatt M, Cheze Le Rest C. Accurate tumor delineation vs rough volume of interest analysis for 18F-FDG PET/CT Radiomics-based prognostic modeling in Non-Small Cell Lung Cancer. *Front Oncol.* 2021:11:726865. Doi: 10.3389/fonc.2021.726865.
[15] Henneman R, Van Monsjou HS, Verhagen CVM, van Velthuysen MF, ter Haar NT, Osse EM, et al. Incidence changes of human papillomavirus in oropharyngeal squamous cell carcinoma and effects on survival in the Netherlands Cancer Institute, 1980–2009. Anticancer Res 2015;35:4015–22.
[16] Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: a toolbox for intensity based medical image registration. IEEE Trans Med Imaging 2010;29:196–205. https://doi.org/10.1109/TMI.2009.2035616.
[17] van Griethuysen JJM, Fedorov A, Parmar CPG, Hosny A, Aucoin N, Narayan V, et al., Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017:77:e104–e108. Doi: 10.1158/0008-5472.CAN-17-0339.
[18] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using Support vector machines. Machine Learning 2002;46:389–422. https://doi.org/10.1023/A:1012487302797.
[19] Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.
[20] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern analysis and machine intelligence 1993;15:850–63. https://doi.org/10.1109/CVPR.1992.223209.
[21] Park S-H, Lim H, Bae BK, Hahm MH, Chong GO, Jeong SY, et al. Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer. Cancer Imaging 2021;21(1). https://doi.org/10.1186/s40644-021-00388-5.
[22] Sun M, Baiyasi A, Liu X, Shi X, Li Xu, Zhu J, et al. Robustness and reproducibility of radiomics in T2 weighted images from magnetic resonance image guided linear accelerator in a phantom study. Physica Med 2022;96:130–9.
[23] Lang DM, Peeken JC, Combs SE, Wilkens JJ, Bartzsch S. Deep learning based HPV status prediction for oropharyngeal cancer patients. Cancers 2021;13(4):786. https://doi.org/10.3390/cancers13040786.
[24] Fontaine P, Andrearczyk V, Oreiller V, Castelli J, Jreige M, Prior JO, et al. Fully automatic head and neck cancer prognosis prediction in PET/CT. *International workshop on multimodal learning for clinical decision Support.* Springer 2021;59–68. https://doi.org/10.1007/978-3-030-89847-2_1.
[25] V. Andrearczyk V. Oreiller M. Jreige M. Vallieres J. Castelli H. Elhalawani et al. Overview of the HECKTOR challenge at MICCAI 2020: Automatic head and neck tumor segmentation in PET/CT 2020 Springer Cham.
[26] Andrearczyk V, Fontaine P, Oreiller V, Castelli J, Jreige M, Prior JO, et al. Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer. In: Predictive Intelligence In Medicine. Cham: Springer; 2021. https://doi.org/10.1007/978-3-030-87602-9_14.
[27] Chen J, Bermejo I, Dekker A, Wee L. Generative models improve radiomics performance in different tasks and different datasets: An experimental study. Physica Med 2022;98:11–7. https://doi.org/10.1016/j.ejmp.2022.04.008.