HHS Public Access

Author manuscript

Phys Med Biol. Author manuscript; available in PMC 2018 July 07.

Published in final edited form as:

Phys Med Biol. 2017 July 07; 62(13): 5327–5343. doi:10.1088/1361-6560/aa73cc.

Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal cancer based on FDG-PET image characteristics

Michael R. Folkert[1], **Jeremy Setton**[1], **Aditya P. Apte**[2], **Milan Grkovski**[2], **Robert J. Young**[3], **Heiko Schöder**[3], **Wade L. Thorstad**[4], **Nancy Y. Lee**[1], **Joseph O. Deasy**[2], and **Jung Hun Oh**[2]

[1]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[2]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[3]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[4]Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO 63110, USA

Abstract

In this study, we investigate the use of imaging feature-based outcomes research ("radiomics") combined with machine learning techniques to develop robust predictive models for the risk of all-cause mortality (ACM), local failure (LF), and distant metastasis (DM) following definitive chemoradiation therapy (CRT). One hundred seventy four patients with stage III-IV oropharyngeal cancer (OC) treated at our institution with CRT with retrievable pre- and post-treatment 18F-fluorodeoxyglucose positron emission tomography (FDG-PET) scans were identified. From pre-treatment PET scans, 24 representative imaging features of FDG-avid disease regions were extracted. Using machine learning-based feature selection methods, multiparameter logistic regression models were built incorporating clinical factors and imaging features. All model building methods were tested by cross validation to avoid overfitting, and final outcome models were validated on an independent dataset from a collaborating institution. Multiparameter models were statistically significant on 5-fold cross validation with the area under the receiver operating characteristic curve (AUC)=0.65 (p=0.004), 0.73 (p=0.026), and 0.66 (p=0.015) for ACM, LF, and DM, respectively. The model for LF retained significance on the independent validation cohort with AUC=0.68 (p=0.029) whereas the models for ACM and DM did not reach statistical significance, but resulted in comparable predictive power to the 5-fold cross validation with AUC=0.60 (p=0.092) and 0.65 (p=0.062), respectively. In the largest study of its kind to date, predictive features including increasing metabolic tumor volume, increasing image heterogeneity, and increasing tumor surface irregularity significantly correlated to mortality, LF, and DM on 5-fold cross validation in a relatively uniform single-institution cohort. The LF model also retained significance in an independent population.

Corresponding author: Jung Hun Oh, PhD, Department of Medical Physics, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, Tel: (646)-888-8017, Fax: (212)-717-3010, ohj@mskcc.org.

Keywords

radiomics; FDG-PET; oropharyngeal cancer

Introduction

For patients with head-and-neck squamous cell cancer (HNSCC), contrast-enhanced volumetric imaging with computed tomography (CT) and/or magnetic resonance imaging (MRI) is the standard of care for clinical workup (Pfister *et al* 2000), and the use of [18F] fluoro-2-deoxy-D-glucose (FDG) based positron emission tomography (PET) imaging has become increasingly integrated into initial workup, treatment planning, and monitoring (Kubicek *et al* 2010, Heron *et al* 2008, MacManus *et al* 2009, Thomas *et al* 2014, Nestle *et al* 2009, Schöder *et al* 2009, Ong *et al* 2008).

Multiple studies based on FDG-PET imaging have reported correlations of simple standardized uptake value (SUV) measures and volume-based measurements, such as metabolic tumor volume (MTV) and total lesion glycolysis (TLG), with oncologic outcomes in HNSCC (Allal *et al* 2002, Kubicek *et al* 2010, Higgins *et al* 2012, Lim *et al* 2012, Romesser *et al* 2012, Tang *et al* 2012, Abd El-Hafez *et al* 2013, Kikuchi *et al* 2014, Romesser *et al* 2014). To a limited degree, more complex analytics that explore the association of imaging features with oncologic outcomes have also been investigated, as these may be less dependent on the dose and timing of radiotracer administration (El Naqa *et al* 2009, Kwon *et al* 2014, Apostolova *et al* 2014). Although it appears that these individual imaging metrics are predictive to some degree for patient outcomes, they may not have sufficient predictive power to be clinically useful. The development of multiparametric models has been proposed as a method to improve prediction of clinical outcomes and select patients who could benefit from dose reduction or intensification strategies (El Naqa *et al* 2009).

In this study, we investigate the use of imaging feature-based outcomes research (“radiomics”) combined with machine learning techniques to develop predictive models for the risk of mortality, local failure (LF), and distant metastasis (DM) in a subset of HNSCC patients with stage III-IV oropharyngeal cancer (OC) following definitive chemoradiation therapy. By incorporating multiple imaging metrics, these multiparametric models are tested to determine whether they may have greater predictive power than any of their individual components.

Materials and methods

Patients

From 12/2002 to 3/2009, all stage III-IV OC patients treated at our institution with definitive concurrent chemoradiation therapy with retrievable pre- and post-treatment FDG-PET/CT scans were identified. To ensure relative uniformity in technique and quality of scans for subsequent analysis, any patient with an FDG-PET/CT scan performed outside of our institution was excluded unless a repeat FDG-PET/CT scan was obtained at our institution

prior to initiation of treatment. Patients with metastatic disease at presentation, noted on staging or treatment planning imaging, were also excluded as were those patients who were managed surgically. This resulted in 174 assessable patients. Our Institutional Review Board approved this retrospective study and all patients provided informed consent.

Image acquisition

Pre-treatment FDG-PET imaging was performed as part of the standard treatment planning process for definitive concurrent chemoradiation therapy. The mean uptake time was 72±17 minutes. Procedures for FDG-PET/CT imaging at our institution have been previously described (Lim *et al* 2012, Romesser *et al* 2014). In brief, patients are instructed to fast for a minimum of 6 hours, with water intake permitted and encouraged. Prior to administration of 18F-FDG (dose range, 12–15 mCi), a blood glucose level of <200 mg/dL is confirmed. 18F-FDG is injected intravenously, followed by an uptake period during which patients drink diluted oral contrast. Low-dose CT (120–140 kV, 80 mA) and PET scans are then obtained for the torso (3 min/bed position, thoracic inlet to upper thigh) with the arms up, followed by dedicated images of the head and neck (5 min/bed position) with the arms down. Intravenous contrast is also administered for radiotherapy planning scans. PET images were acquired on either a hybrid PET/CT GE or Siemens system, normalized and corrected for scatter, randoms, attenuation, decay and dead time, and reconstructed using the ordered-subsets expectation maximization (OSEM) algorithm (2 iterations and 8 subsets for Siemens scanner, and 2 iterations for GE scanners with 28, 21, and 20 subsets for Discovery LS, ST, and STE, respectively).

Image analysis

Pre-treatment attenuation-corrected FDG-PET scan images were converted to the Computational Environment for Radiotherapy Research (CERR) format. CERR is an open source radiotherapy research toolkit designed to facilitate developing and sharing research results for radiotherapy planning; it is MATLAB-based software and provides a common file type for the creation of multi-institutional treatment plan databases for various types of research studies, including dose-volume outcomes analyses and radiomics studies (Deasy *et al* 2003). Bounding boxes were generated for the primary lesion by creating cuboidal structures that encompassed the individual FDG-avid elements. For the purposes of this study, SUV was defined as the decay-corrected measurement of activity per unit volume of tissue (MBq/ml) adjusted by the total administered activity (MBq) and divided by the patient's weight (kg) measured on the date of the scan. A threshold of 42% of the maximum SUV value (SUV_{max}) was then applied to define a region of interest (ROI) for further analysis, and the volume of ROI was defined as MTV (Erdi *et al* 1997).

A total of 24 representative features of the FDG-avid regions, defined as ROIs, were then extracted for each image. These features include common statistical features such as the SUV_{max} and quantizations of the intensity-volume histogram distribution of SUV values over the defined ROI volume. Additionally, more complex shape and textural features that take morphological features and second-order gray-level co-occurrence matrix (GLCM)-based features of the analyzed ROI into account were included (Lian *et al* 2016, Huang *et al*

2016, Sutton *et al* 2016, Leijenaar *et al* 2013). These are further described in the following sections.

Statistical features

First-order statistical features were derived from the distribution of voxel values over the analyzed ROI intensity-volume histogram. They include kurtosis, skewness, slope, and the minimum, maximum, median, and average value of the SUV. Kurtosis is a measure of the flatness or "peakedness" of the intensity-volume histogram, whereas skewness is a measure of the asymmetry of the intensity-volume histogram. Slope is the change in volume over the change in the SUV threshold used to generate the ROI volume.

Shape features

Shape features are related to the morphology of the ROI itself and in this study include eccentricity, solidity, extent, and Euler number. Eccentricity is a measure of "non-circularity" defined as the ratio of the minor axis to the major axis of the best fitted ellipsoid to the analyzed ROI, with an eccentricity of 0 or 1 corresponding to a linear or a perfectly round ROI, respectively. Solidity is derived by calculating the proportion of pixels of the ROI to the largest possible convex hull polygon structure of the ROI; the convex hull is the best-fitting polygon that encloses all the pixels of the ROI. An ROI of the same shape and volume as the convex hull would have a solidity of 1 whereas an irregular ROI would have solidity < 1. Extent is similar to solidity except that a rectangular prism or cuboid is used, instead of a convex hull, to measure the proportion. The Euler number is an integral value that indicates the number of connected objects in the ROI minus the number of holes.

Texture features

The GLCM was constructed by summing up the co-occurrence frequencies for each matrix of 13 directions across a 3D image with 16 gray-scale levels. Texture features quantify the voxel-to-voxel interaction within the ROI, and include homogeneity, entropy, contrast, and coherence (also known as energy in Haralick texture features) (Haralick *et al* 1973, Tesar *et al* 2008). These features are independent of the ROI position, orientation, and size and reflect the distribution of tumor metabolic uptake while minimizing the potential contribution of variations in administered dose and time to FDG-PET image acquisition (El Naqa *et al* 2009). Homogeneity is a measurement of the similarity in intensity of each voxel and its neighboring voxels whereas contrast is a measurement of the variability for the difference in intensity between each voxel and its neighbors. Entropy is a measurement of the randomness of the intensity level over the voxels in the ROI, and coherence measures the uniformity of the intensity level in the ROI. As these texture features are dependent on the intensity relationships between neighboring voxels, noise or artifacts caused in the image acquisition process may greatly impact on the values of imaging features. To remove or minimize noise or artifacts, a smoothing method was applied: for a voxel (v) in the ROI, a sphere with a user-defined radius (from 0.5 cm to 1.5 cm with a step of 0.5 cm) with the voxel (v) being centered was created. An intensity value averaged from all voxels in the sphere was placed in the voxel (v). This procedure was performed for all voxels in the ROI. Image features were extracted from the smoothed image. The post-smoothing was applied after reconstruction.

Statistical analysis

All outcomes were measured from the start of radiation therapy to the time of event. All-cause mortality (ACM) time was defined as the time to death from any cause. LF time was defined as the time to any LF in the high-dose region of radiation therapy treatment. DM time was defined as the time to first clinical or pathological evidence of distant disease recurrence. Patients were censored at the date of last follow-up if death did not occur.

Univariate and multivariate survival analyses were performed using Cox proportional-hazards regression for ACM whereas competing-risks analysis, based on Fine and Gray's proportional sub-hazards model, was used for LF and DM (Fine and Gray 1999). After univariate analysis on clinical variables, multivariate analysis was performed with variables with p-values < 0.05. For each endpoint, Kaplan-Meier analysis with log-rank test was performed to investigate the difference of survival rates between a lower risk group and a higher risk group (Kaplan and Meier 1958, Wilson 1927). Statistical analysis was performed using Stata/MP version 12 (Stata Corporation, College Station, TX).

Model development

Multiparameter logistic regression models were built incorporating clinical factors and imaging features. For the best-fitting model selection, leave-one-out cross validation (LOOCV) with forward feature selection was performed. A model that occurs with the most frequency during the LOOCV process was chosen as a final model for each endpoint. Models were characterized by the area under the receiver operating characteristic (ROC) curve (AUC) and p-values were computed using the Spearman's correlation coefficient (Borkowf 2000, Hanley and McNeil 1982).

Cross validation

For an unbiased model estimate, a 5-fold cross validation method was iterated 30 times (Myles *et al* 1997). At each iteration, univariate logistic regression analysis was performed, and features with p-values < 0.05 were used in multiparameter logistic regression as shown in the above section, resulting in an AUC computed between the predicted and original outcomes. After the whole process, AUC values were averaged. In addition, ROC curve analysis was performed to determine an optimal cutoff value using Youden's index based on which sensitivity and specificity were computed.

Validation on an independent dataset

Additionally, final outcome models (as shown in the above section) were tested on an independent cohort, consisting of 65 patients with stage III-IV OC treated with IMRT-based chemoradiation therapy at another institution (Washington University School of Medicine, St. Louis, MO, USA) from 7/2003 to 11/2009. PET images were acquired on a Siemens Biograph Duo scanner or a Siemens Biograph 40 scanner. Patients were instructed to fast for a minimum of 4 hours. Prior to administration of 18F-FDG (dose range, 10–15 mCi), a blood glucose level of <200 mg/dL was confirmed. A spiral CT scan was obtained at approximately 60 min postinjection and noncontrast CT images were obtained for attenuation correction and fusion with PET images. After that, emission images were obtained. For segmentation, a threshold of 42% of the maximum SUV value was used, and

the same set of image features used in this study was extracted. For more information about this dataset, see Garsa *et al* (2013).

Results

Primary cohort patients

We analyzed 174 HNSCC patients with stage III-IV oropharyngeal cancer following definitive chemoradiation therapy who were treated at our institution. Among them, 48 (27.6%) patients died, and 12 (6.9%) and 33 (19.0%) patients had LF and DM, respectively. Characteristics of the primary study cohort are provided in Table 1. Median follow-up time was 55 months (range: 6–112 months). The majority (87.4%) of the patients were male, and most (69%) were current or former smokers (20.7% and 48.3%, respectively). The average age was 57 years (range: 27–84 years). The OC subsite was the tonsils in 47.1% of patients, base of the tongue in 48.9% of patients, and the soft palate or posterior pharyngeal wall in 4% of patients. The overall stage was IV in 78.7% of patients and III in 21.3% of patients; 39.1% of patients had T3 or T4 disease, and the majority had N2 disease (68.4%); only a small portion of patients in our cohort presented with N3 disease (4.6%). The median value for MTV was 11.2 cc (range: 2.2–60.9 cc).

All (100%) patients were treated with definitive concurrent chemoradiation therapy. The median dose to the tumor was 70 Gy (range: 66–70 Gy), and the median dose to the lower neck was 50.4 Gy (range: 50–70 Gy). Concurrent chemotherapy consisted of cisplatin alone in 56.1% of patients, cetuximab alone in 10.4% of patients, carboplatin and 5-fluorouracil in 12.1% of patients, cisplatin and bevacizumab in 13.3% of patients, and other multidrug regimens in the remaining 8.1% of patients.

For pre-treatment scans, the median uptake time was 67 minutes. To investigate whether the variability in uptake time has impact on the results, patients were split into two groups with a cutoff of 67 in uptake time and four texture features were compared using Wilcoxon rank-sum test. No significant differences were found between the two groups with coherence (p=0.396), contrast (p=0.880), entropy (p=0.445), and homogeneity (p=0.855).

The four most common sizes of voxel were 3.91×3.91×4.25 mm (n=56; 32.2%), 4.69×4.69×3.27 mm (n=46; 26.4%), 5.33×5.33×4.00 mm (n=28; 16.1%), and 5.15×5.15×2.40 mm (n=20; 11.5%). Using a Kruskal-Wallis test, the comparison of texture features between patients with the four different voxel sizes resulted in non-significance for coherence (p=0.278), contrast (p=0.707), entropy (p=0.547), and homogeneity (p=0.778).

We performed a Kruskal-Wallis test to investigate whether there are significant differences in texture features between different PET/CT scanners. No significant differences were found with coherence (p=0.271), contrast (p=0.783), entropy (p=0.479), and homogeneity (p=0.855). These results imply that the impact of variability in uptake time, voxel size, and scanner on texture features is not significant in this cohort.

Predictive factors and models

For each endpoint, the predictive power of individual clinical factors (T, N, overall stage, smoking status, location, Karnofsky performance status [KPS], age, sex, and biologically equivalent dose with alpha/beta=10 [BED_{10}]) and extracted imaging features were assessed using logistic regression. When a smoothing sphere with a radius of 0.5 cm was used, better power was achieved than that without the smoothing method. For instance, contrast and homogeneity features showed statistical significance in LF with AUC=0.69 (p=0.035) and AUC=0.70 (p=0.026), respectively, using the smoothing method, whereas there was no significant texture feature without the smoothing method. Thus, for all tests, we used imaging features extracted after smoothing the ROI with a sphere with a radius of 0.5 cm. Table 2 shows single factor AUC and p-value results that had statistical significance (p<0.05) for at least one endpoint, resulting from univariate logistic regression analysis. For comparison, SUV_{max} and SUV_{mean} were also displayed, which did not show statistical significance for all three endpoints. KPS, T stage, extent, skewness, MTV, and TLG had significant correlations with all three endpoints. The best predictors were T stage (AUC=0.67, p<0.001) and skewness (AUC=0.67, p<0.001) for ACM whereas MTV was significantly associated with both LF and DM with AUC=0.81 (p=0.001) and 0.66 (p=0.004), respectively. BED_{10} did not reach statistical significance for all three endpoints with AUC=0.49 (p=0.799), 0.51 (p=0.731), and 0.51 (p=0.490) for ACM, LF, and DM, respectively.

For the multiparameter logistic regression models, ACM was correlated to kurtosis and MTV; LF was correlated to homogeneity and MTV; DM was correlated to solidity, kurtosis, and MTV. The models are given in Table 3. The 5-fold cross validation used for an unbiased model estimate resulted in AUC=0.65 (standard deviation [SD]=0.02; p=0.004), 0.73 (SD=0.04; p=0.026), and 0.66 (SD=0.04; p=0.015) for ACM, LF, and DM, respectively, showing statistical significance for all three endpoints.

Validation on an independent cohort

Characteristics of the independent validation cohort are provided in Table 4 (Garsa *et al* 2013). All patients in the independent cohort had stage III-IV OC and were treated with definitive concurrent chemotherapy. Those patients who were also surgically managed were excluded, resulting in 65 evaluable patients. Among them, 31 (47.7%) patients died, and 10 (15.4%) and 11 (16.9%) had LF and DM, respectively. Similar to the primary study cohort, the majority were male (78.5%), and the average age was 58 years (range: 38–78 years). Smoking status was unknown in 50.8% of patients, but among those whose information was available, 81.3% were identified as smokers. The majority of patients were stage IV (86.2%); many had advanced primary tumors, with 69.2% presenting with T3 (20.0%) or T4 (49.2%) disease. Most (63.1%) patients presented with N2 disease although there was a larger proportion of patients in the independent cohort with N3 disease (7.7%) than the primary study cohort (4.6%). The median value for MTV was 18.7 cc (range: 3.5–64.7 cc). The models for ACM, LF, and DM were tested on this independent cohort. As shown in Table 3, significant predictive power was retained in LF with AUC=0.68 (p=0.029) whereas the models for ACM and DM showed borderline significance with AUC=0.60 (p=0.092) and 0.65 (p=0.062), respectively.

Survival analysis

For ACM, Cox proportional-hazards regression was performed whereas for LF and DM, Fine and Gray's proportional sub-hazards models were performed with each single clinical variable, and, using variables with $p<0.05$, multivariate models were tested. Smoking history (69% were long-time smokers) was not statistically significant. KPS was most commonly associated with outcomes ($p<0.05$ for ACM, LF, and DM) as shown in Table 5. Other clinical factors significantly associated with outcomes included T stage with DM and ACM ($p=0.001$ and 0.010, respectively) and N stage with DM ($p=0.001$).

It should be noted that MTV was chosen in all three models as shown in Table 3. To investigate the significance of MTV further, patients were split into two groups by median MTV for each endpoint, and a Kaplan-Meier analysis with log-rank test was performed. Statistically significant differences were found for all three endpoints with $p=0.034$, 0.025, and 0.026 for ACM, LF, and DM, respectively. Additionally, patients' MTV values were sorted in ascending order and grouped into three groups of equal size; those patients in the middle group were removed. In the comparison between one-third of patients with smaller MTV and one-third of patients with larger MTV, statistically significant differences were found for all three endpoints with $p=0.014$, 0.006, and 0.037 for ACM, LF, and DM, respectively (Figure 1A). Similarly, for each endpoint, predicted outcomes obtained using the predictive models shown in Table 3 were sorted, and one-third of patients in the middle were removed. In the comparison between the riskiest one-third of patients and the least risky one-third of patients, statistically significant differences were found for all three endpoints with $p<0.001$, 0.006, and <0.001 for ACM, LF, and DM, respectively (Figure 1B).

Discussion

In this study, we tested the association of FDG-PET imaging intensity, shape, and textural features with oncologic outcomes, both as individual factors and as models based on multiple factors. The strongest observed predictive power was obtained through the use of multiparametric models. Note that MTV was chosen in all three models as shown in Table 3. Models with MTV alone achieved reasonable performance with AUC=0.62 ($p=0.016$), 0.81 ($p=0.001$), and 0.62 ($p=0.028$) for ACM, LF, and DM, respectively (see Table 2). On 5-fold cross validation, multiparameter models were statistically significant with AUC=0.65 ($p=0.004$), 0.73 ($p=0.026$), and 0.66 ($p=0.015$) for ACM, LF, and DM, respectively. This is also observed in Figure 1, showing more separation in Kaplan-Meier curves between a lower risk group and a higher risk group when final models in Table 3 were used as compared with MTV alone.

The utility of FDG-PET imaging for modeling in HNSCC has been a subject of thorough investigation. Wong *et al* (2002) and Allal *et al* (2002, 2004) previously reported on the association of high SUV with poor outcomes in patients with head and neck cancer. Further studies have demonstrated that factors that incorporate volume and metabolic information of the tumor, such as gross tumor volume (GTV) and MTV, correlate better with ultimate outcomes (Lim *et al* 2012, Romesser *et al* 2014, Romesser *et al* 2012). This observation has been validated in independent datasets (Tang *et al* 2012). TLG, a function of the MTV and the mean SUV of the defined MTV, has also been identified as an independent predictive

factor for disease-free and disease-specific survival (Abd El-Hafez *et al* 2013, Dibble *et al* 2012, Larson *et al* 1999, Lim *et al* 2012). These volume-based factors have been shown to be predictive and independent from p16 and p53 status (Kikuchi *et al* 2014). Textural features have been used to delineate tumor volumes by assisting with discrimination between normal and malignant tissue (Yu *et al* 2009a, Yu *et al* 2009b), and a recent study by Kwon *et al* (2014) identified an FDG-PET "heterogeneity factor" based on the change in tumor volume over the change in threshold (mathematically similar to the "slope" function described in our study) that significantly correlated with overall survival in patients with oral cavity cancer.

Multiparametric analysis of medical images has been explored as a modeling tool, incorporating multiple image features within an imaging modality or over multiple imaging modalities, in multiple clinical indications (Apostolova *et al* 2014, Cheng *et al* 2013, El Naqa *et al* 2009, Garzon *et al* 2011, Pinker *et al* 2014). In terms of FDG-PET for HNSCC, multiparametric analysis using textural features has been investigated to a limited degree; Cheng *et al* (2013) studied pre-treatment textural features in addition to TLG in a cohort of 70 patients with stage III-IV OC, and developed a risk stratification model incorporating TLG and uniformity (coherence). They found that the model incorporating both TLG and uniformity had a stronger association with oncologic outcomes (progression-free, disease-free, and overall survival) than the individual factors. Apostolova *et al* (2014) reported on the association of a shape-related factor "asphericity" with progression-free and overall survival and observed that this spatial irregularity in uptake in the primary tumor correlated with outcomes, especially when combined with MTV.

In this study, a cuboidal ROI was defined by a single physician to enclose the FDG-avid region with a generous margin, and then an FDG-PET SUV threshold of 42% was applied based on the work of Erdi *et al* (1997) to generate an ROI for further analysis. However, use of the optimal threshold is still controversial (Burger *et al* 2016). Several studies have suggested different threshold values including 40% (Miller and Grigsby 2002, El Naqa *et al* 2009), 42% (Burger *et al* 2013, Wu *et al* 2014), and 50% (Cheebsumon *et al* 2012, Frings *et al* 2014). According to the European Association of Nuclear Medicine (EANM) guidelines (Boellaard *et al* 2015), 41% and 50% threshold values are recommended. A drawback of thresholding is the uncertainty regarding the optimal threshold value (Jeraj *et al* 2015). It seems that the optimal threshold depends on several factors including tumor size, tumor site, PET image size, reconstruction and acquisition parameters, patient biology, etc. Another disadvantage of the thresholding method is the tendency to overestimate the lesion when tumor is small (Foster *et al* 2014). In contrast, manual segmentation also has its drawback regarding segmentation time, labor, and operator variation. The change of feature values for different threshold values was investigated. The comparison of FDG-PET SUV thresholds of 42% and 40% resulted in Spearman's correlation coefficients > 0.91 ($p < 0.001$) for all texture features and MTV, suggesting that the differences in texture features and MTV between 40% and 42% thresholds are minimal.

Hatt *et al* (2015) found that the correlation between MTV and two texture features including entropy and dissimilarity tends to decrease with increasing MTV, and therefore tumor volume and texture features can provide complementary information for large tumors (>10 cc). In this study, we used four texture features. In comparison of these texture features

between lesions with MTV > 10 cc and MTV ≤ 10 cc, we found that absolute Spearman's correlation in entropy and coherence decreased from 0.57 and 0.44 to 0.18 and 0.12, respectively, which is in accordance with Hatt *et al*'s finding. In contrast, absolute Spearman's correlation in contrast and homogeneity increased from 0.10 and 0.18 to 0.36 and 0.36, respectively. However, the correlation of the two features with MTV was unable to be compared with Hatt *et al*'s work, since they did not investigate the correlation between MTV and the two features.

A question raised by the findings of this study is the mechanism by which the imaging features are associated with clinical tumor behavior, and how they may relate to tumor heterogeneity. Henriksson *et al* (2007) previously showed that heterogeneous uptake of FDG within HNSCC can discriminate between patches of active cells and areas with greater amounts of necrosis and stromal tissue in tumor xenografts. HNSCC is known to contain heterogeneous populations of cancer cells, some of which may demonstrate stem cell-like properties (Prince *et al* 2007), and may be associated with other forms of malignant behavior, including metastatic potential and chemotherapy and radiation insensitivity (Leith and Michelson 1990, Zhang *et al* 2013). Intratumoral genetic heterogeneity, defined by one investigator as the mutant-allele tumor heterogeneity (MATH) value above the median (32 units in the cited study), has also been associated with poor survival outcome in patients with HNSCC (Mroz *et al* 2013). It is not yet known whether the imaging features investigated in this study have a biological and/or histopathological correlate to provide a mechanistic explanation for their predictive power.

Strengths of our study include: the relatively large cohort size, with all patients centrally pathologically confirmed and staged, and treated with consistent and reviewed radiation treatment plans; the use of an objective thresholding method to define the analyzed volume reduced potential bias from individually contoured ROIs; all imaging data was centrally processed and reviewed; and most importantly, the predictive models were demonstrated to be robust by validation on an independent dataset at a separate institution, demonstrating that the LF model is transportable to another institution despite differences in patient cohorts. However, the current study has several limitations that must be taken into consideration. In addition to the inherent biases present in any retrospective analysis, further selection bias may have been introduced at several steps. For example, while limiting the study patient population to those who had undergone FDG-PET/CT based pre-treatment imaging at our institution improved our technical access and quality of imaging data for analysis, it prevented the use of an unselected consecutive cohort. FDG-PET images were acquired on two separate PET/CT platforms, which may introduce technical issues in image processing and harmonization. The patient cohort was treated before standardized human papillomavirus (HPV) testing was performed, and very limited data pertaining to the patients' HPV status (p16 or HPV DNA) was available. While prior studies have suggested that complex image features may be predictive regardless of HPV status (Kikuchi *et al* 2014), this is a potential confounder that must be addressed in future studies.

Conclusions

In the largest study of its kind to date, predictive models constructed using FDG-PET intensity, shape, and textural features significantly correlated to mortality, LF, and DM in a relatively uniform single-institution cohort as well as on cross validation. Additionally, the LF model retained significance in an independent population whereas the models for ACM and DM did not reach statistical significance, but resulted in reasonable predictive performance. Such models could assist in patient selection for dose reduction following identification of low-risk patients for LF, or treatment intensification with additional adjuvant chemotherapy following identification of high-risk patients for distant failure. However, understanding of the biological basis of image features, and validation with other datasets will be necessary before these models can be clinically implemented.

Acknowledgments

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748 and a research grant from Varian Medical Systems.

References

Abd El-Hafez YG, Moustafa HM, Khalil HF, Liao CT, Yen TC. Total lesion glycolysis: a possible new prognostic parameter in oral cavity squamous cell carcinoma. Oral Oncol. 2013; 49:261–8. [PubMed: 23036774]

Allal AS, Dulguerov P, Allaoua M, Haenggeli CA, El-Ghazi el A, Lehmann W, Slosman DO. Standardized uptake value of 2-[(18)F] fluoro-2-deoxy-D-glucose in predicting outcome in head and neck carcinomas treated by radiotherapy with or without chemotherapy. J. Clin. Oncol. 2002; 20:1398–404. [PubMed: 11870185]

Allal AS, Slosman DO, Kebdani T, Allaoua M, Lehmann W, Dulguerov P. Prediction of outcome in head-and-neck cancer patients using the standardized uptake value of 2-[18F]fluoro-2-deoxy-D-glucose. Int. J. Radiat. Oncol. Biol. Phys. 2004; 59:1295–300. [PubMed: 15275712]

Apostolova I, Steffen IG, Wedel F, Lougovski A, Marnitz S, Derlin T, Amthauer H, Buchert R, Hofheinz F, Brenner W. Asphericity of pretherapeutic tumour FDG uptake provides independent prognostic value in head-and-neck cancer. Eur. Radiol. 2014; 24:2077–87. [PubMed: 24965509]

Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, Verzijlbergen FJ, Barrington SF, Pike LC, Weber WA, Stroobants S, Delbeke D, Donohoe KJ, Holbrook S, Graham MM, Testanera G, Hoekstra OS, Zijlstra J, Visser E, Hoekstra CJ, Pruim J, Willemsen A, Arends B, Kotzerke J, Bockisch A, Beyer T, Chiti A, Krause BJ. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur. J. Nucl. Med. Mol. Imaging. 2015; 42:328–54. [PubMed: 25452219]

Borkowf CB. A new nonparametric method for variance estimation and confidence interval construction for Spearman's rank correlation. Comput. Statist. Data Anal. 2000; 34:219–41.

Burger IA, Casanova R, Steiger S, Husmann L, Stolzmann P, Huellner MW, Curioni A, Hillinger S, Schmidtlein CR, Soltermann A. 18F-FDG PET/CT of non-small cell lung carcinoma under neoadjuvant chemotherapy: background-based adaptive-volume metrics outperform TLG and MTV in predicting histopathologic response. J. Nucl. Med. 2016; 57:849–54. [PubMed: 26823566]

Burger IA, Vargas HA, Donati OF, Andikyan V, Sala E, Gonen M, Goldman DA, Chi DS, Schöder H, Hricak H. The value of 18F-FDG PET/CT in recurrent gynecologic malignancies prior to pelvic exenteration. Gynecol. Oncol. 2013; 129:586–92. [PubMed: 23369941]

Cheebsumon P, Boellaard R, de Ruysscher D, van Elmpt W, van Baardwijk A, Yaqub M, Hoekstra OS, Comans EF, Lammertsma AA, van Velden FH. Assessment of tumour size in PET/CT lung cancer studies: PET- and CT-based methods compared to pathology. EJNMMI Res. 2012; 2:56. [PubMed: 23034289]

Cheng NM, Fang YH, Chang JT, Huang CG, Tsan DL, Ng SH, Wang HM, Lin CY, Liao CT, Yen TC. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. J. Nucl. Med. 2013; 54:1703–9. [PubMed: 24042030]

Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med. Phys. 2003; 30:979–85. [PubMed: 12773007]

Dibble EH, Alvarez AC, Truong MT, Mercier G, Cook EF, Subramaniam RM. 18F-FDG metabolic tumor volume and total glycolytic activity of oral cavity and oropharyngeal squamous cell cancer: adding value to clinical staging. J. Nucl. Med. 2012; 53:709–15. [PubMed: 22492732]

El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, Chaudhari S, Yang D, Schmitt M, Laforest R, Thorstad W, Deasy JO. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit. 2009; 42:1162–71. [PubMed: 20161266]

Erdi YE, Mawlawi O, Larson SM, Imbriaco M, Yeung H, Finn R, Humm JL. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer. 1997; 80:2505–9. [PubMed: 9406703]

Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J. Am. Stat. Assoc. 1999; 94:496–509.

Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. Comput. Biol. Med. 2014; 50:76–96. [PubMed: 24845019]

Frings V, van Velden FH, Velasquez LM, Hayes W, van de Ven PM, Hoekstra OS, Boellaard R. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. Radiology. 2014; 273:539–48. [PubMed: 24865311]

Garsa AA, Chang AJ, Dewees T, Spencer CR, Adkins DR, Dehdashti F, Gay HA, Thorstad WL. Prognostic value of 18F-FDG PET metabolic parameters in oropharyngeal squamous cell carcinoma. J. Radiat. Oncol. 2013; 2:27–34. [PubMed: 24563726]

Garzon B, Emblem KE, Mouridsen K, Nedregaard B, Due-Tonnessen P, Nome T, Hald JK, Bjornerud A, Haberg AK, Kvinnsland Y. Multiparametric analysis of magnetic resonance images for glioma grading and patient survival time prediction. Acta Radiol. 2011; 52:1052–60. [PubMed: 21969702]

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]

Haralick RM, Shanmuga K, Dinstein I. Textural features for image classification. IEEE Trans. Syst. Man. Cybern. 1973; Smc3:610–21.

Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, Hindie E, Martineau A, Pradier O, Hustinx R, Perdrisot R, Guillevin R, El Naqa I, Visvikis D. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J. Nucl. Med. 2015; 56:38–44. [PubMed: 25500829]

Henriksson E, Kjellen E, Wahlberg P, Ohlsson T, Wennerberg J, Brun E. 2-Deoxy-2-[18F] fluoro-D-glucose uptake and correlation to intratumoral heterogeneity. Anticancer Res. 2007; 27:2155–9. [PubMed: 17695498]

Heron DE, Andrade RS, Beriwal S, Smith RP. PET-CT in radiation oncology: the impact on diagnosis, treatment planning, and assessment of treatment response. Am. J. Clin. Oncol. 2008; 31:352–62. [PubMed: 18845994]

Higgins KA, Hoang JK, Roach MC, Chino J, Yoo DS, Turkington TG, Brizel DM. Analysis of pretreatment FDG-PET SUV parameters in head-and-neck cancer: tumor SUVmean has superior prognostic value. Int. J. Radiat. Oncol. Biol. Phys. 2012; 82:548–53. [PubMed: 21277108]

Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, Liang C, Tian J, Liang C. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer. Radiology. 2016; 281:947–57. [PubMed: 27347764]

Jeraj R, Bradshaw T, Simoncic U. Molecular imaging to plan radiotherapy and evaluate its efficacy. J. Nucl. Med. 2015; 56:1752–65. [PubMed: 26383148]

Kaplan EL, Meier P. Nonparametric-estimation from incomplete observations. J. Am. Stat. Assoc. 1958; 53:457–81.

Kikuchi M, Koyasu S, Shinohara S, Usami Y, Imai Y, Hino M, Itoh K, Tona R, Kanazawa Y, Kishimoto I, Harada H, Naito Y. Prognostic value of pretreatment F-fluorodeoxyglucose positron emission tomography/CT volume-based parameters in patients with oropharyngeal squamous cell carcinoma with known p16 and p53 status. Head Neck. 2014; 37:1524–31. [PubMed: 24890445]

Kubicek GJ, Champ C, Fogh S, Wang F, Reddy E, Intenzo C, Dusing RW, Machtay M. FDG-PET staging and importance of lymph node SUV in head and neck cancer. Head Neck Oncol. 2010; 2:19. [PubMed: 20637102]

Kwon SH, Yoon JK, An YS, Shin YS, Kim CH, Lee DH, Jo KS, Lee SJ. Prognostic significance of the intratumoral heterogeneity of F-FDG uptake in oral cavity cancer. J. Surg. Oncol. 2014; 110:702–6. [PubMed: 24975131]

Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, Casilla C, Fazzari M, Srivastava N, Yeung HW, Humm JL, Guillem J, Downey R, Karpeh M, Cohen AE, Ginsberg R. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. Clin. Positron imaging. 1999; 2:159–71. [PubMed: 14516540]

Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, Hoekstra CJ, Boellaard R, Dekker AL, Gillies RJ, Aerts HJ, Lambin P. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013; 52:1391–7. [PubMed: 24047337]

Leith JT, Michelson S. Tumor radiocurability: relationship to intrinsic tumor heterogeneity and to the tumor bed effect. Invasion Metastasis. 1990; 10:329–51. [PubMed: 2265986]

Lian C, Ruan S, Denœux T, Jardin F, Vera P. Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction. Med. Image Anal. 2016; 32:257–68. [PubMed: 27236221]

Lim R, Eaton A, Lee NY, Setton J, Ohri N, Rao S, Wong R, Fury M, Schoder H. 18F-FDG PET/CT metabolic tumor volume and total lesion glycolysis predict outcome in oropharyngeal squamous cell carcinoma. J. Nucl. Med. 2012; 53:1506–13. [PubMed: 22895812]

MacManus M, Nestle U, Rosenzweig KE, Carrio I, Messa C, Belohlavek O, Danna M, Inoue T, Deniaud-Alexandre E, Schipani S, Watanabe N, Dondi M, Jeremic B. Use of PET and PET/CT for radiation therapy planning: IAEA expert report 2006–2007. Radiother. Oncol. 2009; 91:85–94. [PubMed: 19100641]

Miller TR, Grigsby PW. Measurement of tumor volume by PET to evaluate prognosis in patients with advanced cervical cancer treated by radiation therapy. Int. J. Radiat. Oncol. Biol. Phys. 2002; 53:353–9. [PubMed: 12023139]

Mroz EA, Tward AD, Pickering CR, Myers JN, Ferris RL, Rocco JW. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. Cancer. 2013; 119:3034–42. [PubMed: 23696076]

Myles AJ, Murray AF, Wallace AR, Barnard J, Smith G. Estimating MLP generalisation ability without a test set using fast, approximate leave-one-out cross-validation. Neural. Comput. Appl. 1997; 5:134–51.

Nestle U, Weber W, Hentschel M, Grosu AL. Biological imaging in radiation therapy: role of positron emission tomography. Phys. Med. Biol. 2009; 54:R1–25. [PubMed: 19060363]

Ong SC, Schöder H, Lee NY, Patel SG, Carlson D, Fury M, Pfister DG, Shah JP, Larson SM, Kraus DH. Clinical utility of 18F-FDG PET/CT in assessing the neck after concurrent chemoradiotherapy for Locoregional advanced head and neck cancer. J. Nucl. Med. 2008; 49:532–40. [PubMed: 18344440]

Pfister DG, Ang K, Brockstein B, Colevas AD, Ellenhorn J, Goepfert H, Hicks WL Jr, Hong WK, Kies MS, Lydiatt W, McCaffrey T, Mittal BB, Ridge JA, Schuller DE, Shah JP, Spencer S, Trotti A 3rd, Urba S, Weymuller EA Jr, Wheeler RH 3rd, Wolf GT, and National Comprehensive Cancer N. NCCN practice guidelines for head and neck cancers. Oncology. 2000; 14:163–94. [PubMed: 11195409]

Pinker K, Bogner W, Baltzer P, Karanikas G, Magometschnigg H, Brader P, Gruber S, Bickel H, Dubsky P, Bago-Horvath Z, Bartsch R, Weber M, Trattnig S, Helbich TH. Improved differentiation

of benign and malignant breast tumors with multiparametric 18fluorodeoxyglucose positron emission tomography magnetic resonance imaging: a feasibility study. Clin. Cancer Res. 2014; 20:3540–9. [PubMed: 24963052]

Prince ME, Sivanandan R, Kaczorowski A, Wolf GT, Kaplan MJ, Dalerba P, Weissman IL, Clarke MF, Ailles LE. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. Proc. Natl. Acad. Sci. U S A. 2007; 104:973–8. [PubMed: 17210912]

Romesser PB, Lim R, Spratt DE, Setton J, Riaz N, Lok B, Rao S, Sherman EJ, Schoder H, Lee NY. The relative prognostic utility of standardized uptake value, gross tumor volume, and metabolic tumor volume in oropharyngeal cancer patients treated with platinum based concurrent chemoradiation with a pre-treatment [F] fluorodeoxyglucose positron emission tomography scan. Oral Oncol. 2014; 50:802–8. [PubMed: 25043882]

Romesser PB, Qureshi MM, Shah BA, Chatburn LT, Jalisi S, Devaiah AK, Subramaniam RM, Truong MT. Superior prognostic utility of gross and metabolic tumor volume compared to standardized uptake value using PET/CT in head and neck squamous cell carcinoma patients treated with intensity-modulated radiotherapy. Ann. Nucl. Med. 2012; 26:527–34. [PubMed: 22610386]

Schöder H, Fury M, Lee N, Kraus D. PET monitoring of therapy response in head and neck squamous cell carcinoma. J. Nucl. Med. 2009; 50:74S–88S. [PubMed: 19380408]

Sutton EJ, Dashevsky BZ, Oh JH, Veeraraghavan H, Apte AP, Thakur SB, Morris EA, Deasy JO. Breast cancer molecular subtype classifier that incorporates MRI features. J. Magn. Reson. Imaging. 2016; 44:122–9. [PubMed: 26756416]

Tang C, Murphy JD, Khong B, La TH, Kong C, Fischbein NJ, Colevas AD, Iagaru AH, Graves EE, Loo BW Jr, Le QT. Validation that metabolic tumor volume predicts outcome in head-and-neck cancer. Int. J. Radiat. Oncol. Biol. Phys. 2012; 83:1514–20. [PubMed: 22270174]

Tesar L, Shimizu A, Smutek D, Kobatake H, Nawano S. Medical image analysis of 3D CT images based on extension of Haralick texture features. Comput. Med. Imaging Graph. 2008; 32:513–20. [PubMed: 18614335]

Thomas CM, Pike LC, Hartill CE, Baker S, Woods E, Convery DJ, Greener AG. Specific recommendations for accurate and direct use of PET-CT in PET guided radiotherapy for head and neck sites. Med. Phys. 2014; 41:041710. [PubMed: 24694130]

Wilson EB. Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. 1927; 22:209–12.

Wong RJ, Lin DT, Schöder H, Patel SG, Gonen M, Wolden S, Pfister DG, Shah JP, Larson SM, Kraus DH. Diagnostic and prognostic value of [(18)F]fluorodeoxyglucose positron emission tomography for recurrent head and neck squamous cell carcinoma. J. Clin. Oncol. 2002; 20:4199–208. [PubMed: 12377963]

Wu X, Pertovaara H, Korkola P, Dastidar P, Järvenpää R, Eskola H, Kellokumpu-Lehtinen PL. Correlations between functional imaging markers derived from PET/CT and diffusion-weighted MRI in diffuse large B-cell lymphoma and follicular lymphoma. PLoS One. 2014; 9:e84999. [PubMed: 24454777]

Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. IEEE Trans. Med. Imaging. 2009a; 28:374–83. [PubMed: 19244009]

Yu H, Caldwell C, Mah K, Poon I, Balogh J, MacKenzie R, Khaouam N, Tirona R. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. Int. J. Radiat. Oncol. Biol. Phys. 2009b; 75:618–25. [PubMed: 19683403]

Zhang XC, Xu C, Mitchell RM, Zhang B, Zhao D, Li Y, Huang X, Fan W, Wang H, Lerma LA, Upton MP, Hay A, Mendez E, Zhao LP. Tumor evolution and intratumor heterogeneity of an oropharyngeal squamous cell carcinoma revealed by whole-genome sequencing. Neoplasia. 2013; 15:1371–8. [PubMed: 24403859]

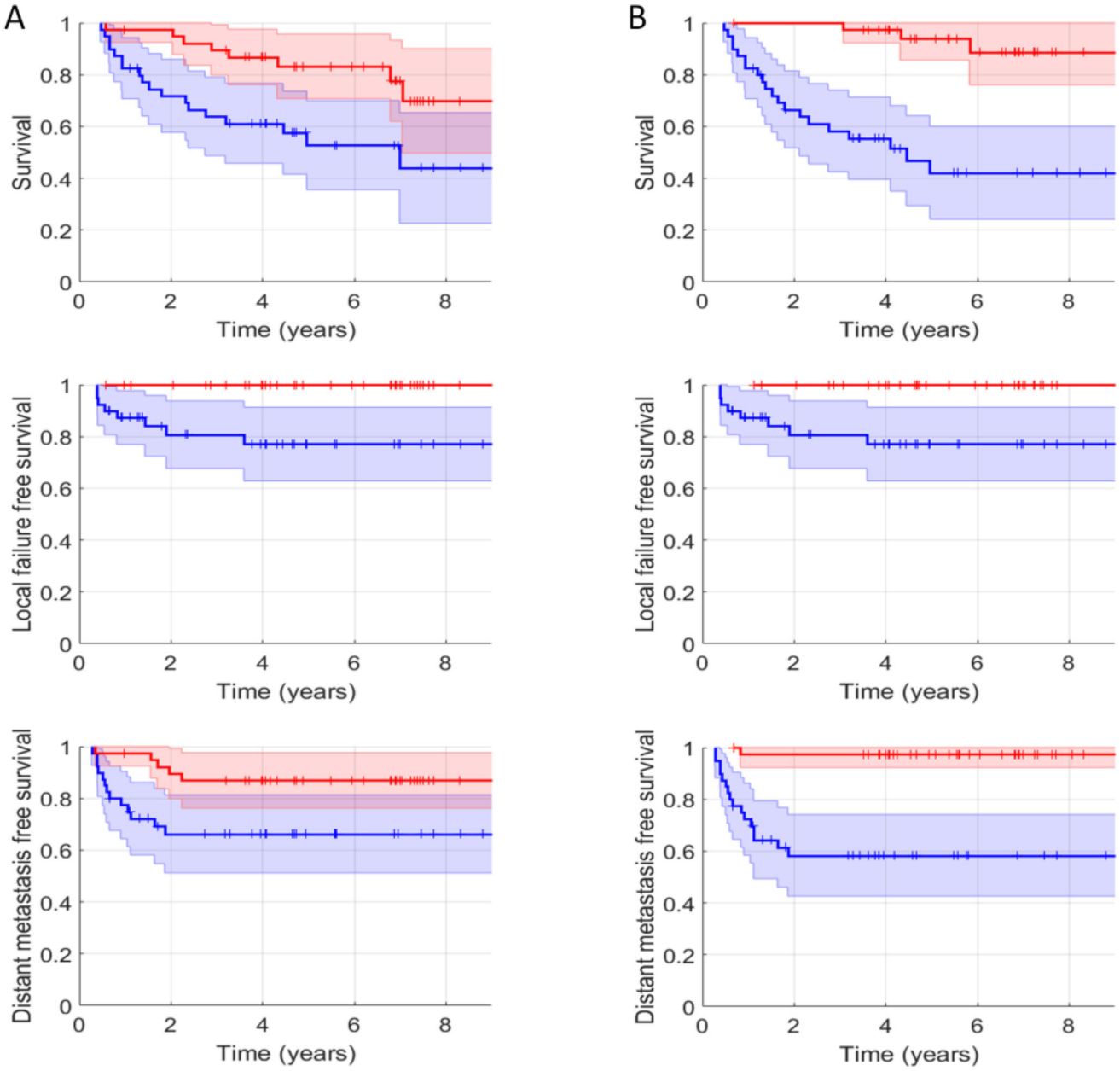

Figure 1.

Kaplan-Meier analysis with log-rank test for all-cause mortality, local failure, and distant metastasis in the primary study cohort. (A) Patients' MTV values and (B) predicted outcomes were sorted in ascending order and grouped into three groups with equal size; those patients in the middle group were removed. (A) One-third of patients with smaller MTV and one-third of patients with larger MTV and (B) the riskiest one-third of patients and the least risky one-third of patients were compared. Shaded areas indicate 95% confidence interval.

Table 1

Patient characteristics of the primary study cohort.

Metrics		N (%)
Total number of patients		174
Median follow-up (range)		55 months (6–112)
Mean age (range in years)		57 years (27–84)
Gender		
	Male	152 (87.4%)
	Female	22 (12.6%)
Median Karnofsky performance status Smoking status		90
	Current	36 (20.7%)
	Former	84 (48.3%)
	Never	54 (31%)
Site		
	Tonsil	82 (47.1%)
	Base of tongue	85 (48.9%)
	Other (soft palate and posterior pharyngeal wall)	7 (4%)
Histology		
	Squamous cell carcinoma	174 (100%)
T stage		
	1	34 (19.5%)
	2	72 (41.4%)
	3	36 (20.7%)
	4	32 (18.4%)
N stage		
	0	7 (4%)
	1	40 (23%)
	2	119 (68.4%)
	3	8 (4.6%)
Overall stage		
	III	37 (21.3%)
	IV	137 (78.7%)
Treatment status		
	Definitive	174 (100%)
Median dose to primary tumor		70 Gy
	(range in Gy)	(67.8–70)
Median dose to lower neck		50.4 Gy
	(range in Gy)	(50–70)
Chemotherapy type		
	Cisplatin	97 (55.8%)
	Cetuximab	18 (10.3%)
	Carboplatin + 5-FU	21 (12.1%)

Metrics		N (%)
	Cisplatin + Bevacizumab	23 (13.2%)
	Other	15 (8.6%)
PET/CT system (Pre-treatment)		
	GE Discovery LS	60 (34.5%)
	GE Discovery ST	46 (26.5%)
	GE Discovery STE	18 (10.3%)
	Siemens Biograph	50 (28.7%)

Table 2

Univariate logistic regression analysis results with AUC values and *p*-values resulting from Spearman's correlation coefficient test for metrics with statistical significance for at least one endpoint. Note that SUV_{max} and SUV_{mean} were added for comparison despite being not significant. The metrics in bold indicate those with statistical significance for all the three endpoints.

Metrics	ACM		LF		DM	
	AUC	*p*-value	AUC	*p*-value	AUC	*p*-value
Age	0.61	0.030	0.53	0.759	0.59	0.103
KPS	0.66	< 0.001	0.79	< 0.001	0.63	0.007
Stage	0.57	0.038	0.53	0.641	0.58	0.052
T stage	0.67	< 0.001	0.72	0.011	0.65	0.005
N stage	0.58	0.047	0.58	0.269	0.60	0.024
Contrast	0.55	0.306	0.69	0.035	0.55	0.374
Homogeneity	0.52	0.675	0.70	0.026	0.53	0.591
Solidity	0.59	0.060	0.63	0.155	0.62	0.039
Extent	0.62	0.016	0.70	0.025	0.64	0.011
Skewness	0.67	< 0.001	0.72	0.013	0.66	0.005
Kurtosis	0.65	0.002	0.67	0.054	0.65	0.008
MTV	0.62	0.016	0.81	0.001	0.62	0.028
TLG	0.65	0.002	0.78	0.001	0.66	0.004
SUV Max	0.58	0.094	0.57	0.415	0.58	0.164
SUV Mean	0.59	0.070	0.59	0.303	0.58	0.146

KPS=Karnofsky performance status; MTV=metabolic tumor volume; TLG=total lesion glycolysis; SUV=standardized uptake value; ACM=all-cause mortality; LF=local failure; DM=distant metastasis; AUC=area under the receiver operating characteristic curve.

Table 3

Validation of multiparameter logistic regression models based on “probability of endpoint=1/(1+exp(-Y))”. In the 5-fold cross validation column, the performance shows average AUC and p-values, repeating 5-fold cross validation 30 times and the Independent validation column shows the performance when the predictive models were applied to an independent cohort. The sensitivity and specificity were computed using an ROC analysis. The parenthesis indicates standard deviation.

Endpoint	Predictive model	Performance metric	5-fold cross validation	Independent validation
ACM	Y=0.0481×MTV-0.735×Kurtosis+0.25	AUC	0.65(0.02) (p=0.004)	0.60 (p=0.092)
		SEN	0.67(0.15)	0.58
		SPE	0.60(0.16)	0.62
LF	Y=0.0977×MTV-6.19×Homogeneity-2.31	AUC	0.73(0.04) (p=0.026)	0.68 (p=0.029)
		SEN	0.65(0.06)	0.67
		SPE	0.85(0.06)	0.70
DM	Y=0.0244×MTV-5.57×Solidity-1.16×Kurtosis+6.32	AUC	0.66(0.04) (p=0.015)	0.65 (p=0.062)
		SEN	0.62(0.16)	0.64
		SPE	0.66(0.18)	0.80

ACM=all-cause mortality; LF=local failure; DM=distant metastasis; MTV=metabolic tumor volume; AUC=area under the receiver operating characteristic (ROC) curve; SEN=sensitivity; SPE=specificity.

Table 4

Patient characteristics of the independent validation cohort.

Metrics		N (%)
Total number of patients		65
Median follow-up (range)		28 months (2–83)
Mean age (range in years)		58 years (38–78)
Gender		
	Male	51 (78.5%)
	Female	14 (21.5%)
Smoking status		
	Yes	26 (40.0%)
	No	6 (9.2%)
	Unknown	33 (50.8%)
T stage		
	1	2 (3.1%)
	2	16 (24.6%)
	3	13 (20.0%)
	4	32 (49.2%)
	Unknown	2 (3.1%)
N stage		
	0	8 (12.3%)
	1	11 (16.9%)
	2	41 (63.1%)
	3	5 (7.7%)
Overall stage		
	III	9 (13.8%)
	IV	56 (86.2%)

Table 5

Multivariate Cox proportional-hazards regression for ACM and multivariate Fine and Gray's proportional sub-hazards models for LF and DM.

	Metrics	*p*-value	Hazard ratio	95% CI
ACM	Age	0.774	1.00	0.97–1.02
	KPS	< 0.001	0.90	0.85–0.95
	T stage	0.010	1.49	1.10–2.01
	Stage	0.066	2.46	0.94–6.40
	Metrics	*p*-value	Sub-hazard ratio	95% CI
LF	KPS	< 0.001	0.83	0.77–0.89
	T stage	0.223	1.55	0.77–3.13
	Metrics	*p*-value	Sub-hazard ratio	95% CI
DM	KPS	0.022	0.94	0.89–0.99
	T stage	0.001	1.73	1.25–2.40
	N stage	0.001	2.62	1.52–4.50

ACM=all-cause mortality; LF=local failure; DM=distant metastasis; KPS=Karnofsky performance status.