Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Radiomics Model to Predict Early Progression of Nonmetastatic Nasopharyngeal Carcinoma after Intensity Modulation Radiation Therapy: A Multicenter Study

Richard Du, MSc • Victor H. Lee, MD, FRCR • Hui Yuan, MBBS, PhD • Ka-On Lam, MBBS, FRCR • Herbert H. Pang, PhD • Yu Chen, MD • Edmund Y. Lam, PhD • Pek-Lan Khong, MD, FRCR • Anne W. Lee, MD, FRCR • Dora L. Kwong, MD, FRCR • Varut Vardhanabhuti, MBBS, FRCR, PhD

From the Departments of Diagnostic Radiology (R.D., H.Y., P.L.K., V.V.) and Clinical Oncology (V.H.L., K.O.L., A.W.L., D.L.K.) and the School of Public Health (H.H.P.), Li Ka Shing Faculty of Medicine, The University of Hong Kong, Room 406, Block K, Queen Mary Hospital, Pok Fu Lam Road, Hong Kong SAR; Department of Radiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China (Y.C.); and Department of Electrical and Electronic Engineering, Faculty of Engineering, The University of Hong Kong, Hong Kong SAR (E.Y.L.). Received November 20, 2018; revision requested January 14, 2019; revision received April 4; accepted May 7. **Address correspondence to** V.V. (e-mail: *varv@hku.hk*).

Supported by the Research Grants Council, University Grants Committee (AoE M-06/08), University Research Committee, University of Hong Kong (201611159216), and the Lee Shau Kee Foundation.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2019; 1(4):e180075 • https://doi.org/10.1148/ryai.2019180075 • Content codes: IN HN

Purpose: To examine the prognostic value of a machine learning model trained with pretreatment MRI radiomic features in the assessment of patients with nonmetastatic nasopharyngeal carcinoma (NPC) who are at risk for 3-year disease progression after intensity-modulated radiation therapy and to explain the radiomics features in the model.

Materials and Methods: A total of 277 patients with nonmetastatic NPC admitted between March 2008 and December 2014 at two imaging centers were retrospectively reviewed. Patients were allocated to a discovery or validation cohort based on where they underwent MRI (discovery cohort, $n = 217$; validation cohort, $n = 60$). A total of 525 radiomics features extracted from contrast material–enhanced T1- or T2-weighted MRI studies and five clinical features were subjected to radiomic machine learning modeling to predict 3-year disease progression. Feature selection was performed by analyzing robustness to resampling, reproducibility between observers, and redundancy. Features for the final model were selected with Kaplan-Meier analysis and the log-rank test. A support vector machine was used as the classifier for the model. To interpret the pattern learned from the model, Shapley additive explanations (SHAP) was applied.

Results: The final model yielded an area under the receiver operating characteristic curve of 0.80 in both the discovery (95% bootstrap confidence interval: 0.80, 0.81) and independent validation (95% bootstrap confidence interval: 0.73, 0.89) cohorts. Analysis with SHAP revealed that tumor shape sphericity, first-order mean absolute deviation, T stage, and overall stage were important factors in 3-year disease progression.

Conclusion: These results add to the growing evidence of the role of radiomics in the assessment of NPC. By using explanatory techniques, such as SHAP, the complex interaction of features learned by the model may be understood.

Supplemental material is available for this article.

Nasopharyngeal carcinoma (NPC) is endemic in southeast Asia (1). Despite good overall survival after treatment, approximately one-third of patients still experience relapse (2,3). Current treatment stratification is primarily based on TNM staging of the disease. With the advent of intensity-modulated radiation therapy (IMRT), studies have shown that this technique can yield excellent local-regional control (>90%) and can enable better sparing of adjacent organs from unnecessary radiation (4–8). This led to pretreatment T stage becoming less predictive of local control and survival after IMRT (9). Alternative pretreatment strategies for prognosis of NPC after treatment are needed to provide better risk stratification for NPC.

Commonly used imaging modalities in the staging of NPC are MRI, CT, and PET/CT. Many studies found an association between clinical outcomes and quantitative measurement based on MRI and PET/CT findings with parameters such as primary tumor volume, nodal volume, and standard uptake values (9–13). To enable a more quantitative approach in prediction, several groups have studied the role of quantitative analysis of medical images, which led to the development of radiomics. Radiomics is the process of extracting large amounts of image-based features, known as radiomic features, from routine diagnostic scans. The underlying hypothesis is that radiomic features that quantify tumor shape, image intensity, and texture may reflect the characteristics of disease that are important in clinical decision making (14–16). Recent studies have shown that pretreatment radiomic features are prognostic of survival in patients with NPC (17–19). Despite evidence

Abbreviations

AUC = area under the receiver operating characteristic curve, GLCM = gray level co-occurrence matrix, GLRLM = gray level run length matrix, ICC = intraclass correlation coefficient, IMRT = intensity-modulated radiation therapy, NPC = nasopharyngeal carcinoma, PCC = Pearson correlation coefficient, PFS = progression-free survival, SHAP = Shapley additive explanations

Summary

A machine-learning radiomic model based on pretreatment MRI findings has potential in the identification of patients with nonmetastatic nasopharyngeal carcinoma who are at risk for early disease progression after primary treatment.

Key Points

- We trained and developed a machine-learning model based on pretreatment MRI radiomic features to predict 3-year disease progression in patients with nonmetastatic nasopharyngeal carcinoma after primary treatment.
- The radiomic model achieved an area under the receiver operating characteristic curve of 0.80 in discriminating patients with disease progression within 3 years in both the discovery cohort and the independent validation cohort.
- By using the explanatory machine learning framework Shapley additive explanations, we identified tumor shape sphericity and first-order mean absolute deviation as important factors in driving the risk of 3-year disease progression in patients.

of the benefit of radiomics in the assessment of NPC, the clinical utility of radiomic-based prediction models remains unclear. This is mainly due to the lack of validation of this model outside the discovery cohort and the reliance on complex machine learning to yield accurate predictions. The lack of understanding of predictions made by the models led to skepticism about its clinical application. In recent years, much effort has been made in explanatory machine learning to improve the interpretability of traditional complex black box machine learning models (20,21). In a recent study, Lundberg et al (22) explained predictions made by a clinical machine learning model by using the Shapley additive explanations (SHAP) framework, potentially increasing understanding and usability of their model.

The purpose of this study was to investigate the prognostic value of radiomics in the assessment of patients with nonmetastatic NPC. We hypothesized that radiomic features from pretreatment MRI were associated with the survival of NPC and, when modeled with machine learning and SHAP, could yield accurate and explainable prediction of disease progression in patients with NPC.

Materials and Methods

This study is a multicenter retrospective study of patients with newly diagnosed NPC who were admitted between March 1, 2008, and December 31, 2014, at the University of Hong Kong (H1) and Queen Mary Hospital (H2). Institutional ethics review board approval was obtained for this study, and informed consent was waived owing to the retrospective nature of the study. Data from 122 patients in this study were previously reported (23). This prior study examined the prognostic importance of MRI-based morphologic parameters and PET/CT-based parameters of the primary tumor and lymph nodes, whereas in the current study we analyzed MRI radiomics of primary tumors. The results may not be directly comparable, as there were more patients in this cohort, with longer median follow-up and updated survival data.

Patient Cohort

A total of 277 patients were eligible for and were included in this study based on the following criteria: *(a)* They had histologically confirmed NPC. *(b)* They had no evidence of distant metastases at diagnosis. *(c)* They underwent pretreatment contrast material–enhanced T1- and T2-weighted MRI. *(d)* They underwent a standard treatment regimen that consisted of IMRT and concurrent or adjuvant chemotherapy with or without induction based on the TNM classification (7th edition American Joint Committee on Cancer/Union for International Cancer Control). *(e)* They had at least 3 years of follow-up data for survival analysis. A total of 217 patients were examined at H1, and 60 patients were examined at H2. Patients from H1 served as the discovery cohort, and patients from H2 served as the validation cohort. Patient characteristics for both cohorts are summarized in Table 1. Significant differences between the cohorts were found for patient age, overall stage, and N stage. No significant differences were found for the other parameters.

All patients underwent a standard treatment that has been described in a previous study (24). In general, stage I disease was treated with IMRT alone, while stage II disease was treated with IMRT with or without concurrent chemotherapy. Stage III or IV disease was treated with concurrent chemoradiotherapy with adjuvant chemotherapy. The clinical outcome of this study was progression-free survival (PFS), which was defined as the time (in months) from the first day of treatment to the date of disease progression (local-regional recurrences or distant metastases), death, or last follow-up.

For imaging, all patients underwent diagnostic head and neck contrast-enhanced T1- and T2-weighted MRI with a 3-T imager. Detailed description of acquisition and imaging parameters used in both centers is given in Table E1 (supplement).

To ensure there is no bias in the model toward the validation cohort, all feature selection and modeling were performed in the discovery cohort only. The validation cohort was used for validation of the final selected models only.

Feature Extraction

The study workflow is summarized in Figure 1. The volume of interest of this study was the primary NPC tumor. Tumor segmentation of the primary tumor was performed and reviewed by two board-certified radiologists in consensus (H.Y., V.V.; 5 and 11 years of experience, respectively). When retropharyngeal lymph node was inseparable from the primary tumor, it was included in the region of interest because a clear distinction between the two structures is difficult, as has been previously documented (25,26). Segmentation was conducted separately on the contrast-enhanced T1- and T2-weighted MR images.

To ensure spatial consistency in texture analysis across the images, all images were resampled spatially into $1 \times 1 \times 4$ mm

resolution. Radiomic features based on the study by Aerts et al (27) were selected and evaluated. For each MRI sequence, four subbands of Coiflet wavelet transforms were performed, yielding a total of 10 images per patient (four subbands and original image per MRI sequence). For each image, 11 first-order intensity features and 41 texture features (gray level co-occurrence matrix [GLCM], gray level run length matrix [GLRLM], and neighborhood gray level difference matrix) were extracted. Also, five shape features were extracted from the contrast-enhanced T1-weighted sequence, leading to a total of 525 features per patient. A detailed description of feature extraction and a list of features extracted are provided in Table E2 (supplement).

Table 1: Summary of Patient Characteristics in Both Cohorts

Characteristic	Discovery Cohort (n = 217)	Validation Cohort (n = 60)	P Value
Sex	…	…	.579*
Male	155 (71.4)	40 (66.7)	…
Female	62 (28.6)	20 (33.3)	…
Age (y)†	…	…	…
Male	51 (43–60)	53 (41–62)	.248‡
Female	49 (40–68)	60 (53–74)	.003‡
Both	50 (42–59)	55 (47–65)	.007‡
Overall stage	…	…	<.001*
I	18 (8.3)	0 (0)	…
II	48 (22.1)	5 (8.3)	…
III	86 (36.6)	42 (70.0)	…
IV	65 (30.0)	13 (21.7)	…
T stage	…	…	.14*
T1	61 (28.1)	10 (16.7)	…
T2	42 (19.4)	9 (15.0)	…
T3	80 (36.9)	31 (51.7)	…
T4	34 (15.7)	10 (16.7)	…
N stage	…	…	<.001*
N0	43 (19.8)	8 (13.3)	…
N1	86 (39.6)	10 (16.7)	…
N2	54 (24.9)	39 (65.0)	…
N3	34 (15.7)	3 (5.0)	…
3-year PFS	…	…	.329*
Yes	187 (86.2)	48 (80.0)	…
No	30 (13.8)	12 (20.0)	…
WHO histologic type	…	…	.797*
I	2 (0.9)	1 (1.7)	…
II	5 (2.3)	2 (3.3)	…
III	210 (96.8)	57 (95)	…

Note.—Unless otherwise indicated, data are numbers of patients, and data in parentheses are percentages. PFS = progression-free survival, WHO = World Health Organization.
* P value was determined with the Pearson contingency χ^2 test. $P > .05$ suggests no significant difference between the proportion of subjects in the two cohorts.
† Data in parentheses are the interquartile range.
‡ P value was determined with the independent samples t test. $P < .05$ suggests a significant difference between the age in the two cohorts.

Feature Selection

The feature selection process involved three steps. First, robustness was tested based on different image resampling. Second, interobserver variability was assessed by segmentation between readers. Third, redundant features were eliminated by conducting hierarchal clustering analysis. For resampling, Pearson linear correlation coefficient (PCC) was calculated between features extracted with image resampling and features extracted without image resampling. Features that were highly correlated after resampling suggested that texture information measured from the feature remained similar. The features with PCC less than 0.9 were excluded from analysis. For interobserver variability of the features, a subset of 30 randomly selected patients was independently delineated by three different board-certified radiologists (H.Y., Y.C., and V.V.; 5, 8, and 11 years of experience, respectively). Two-way random effects one-rater intraclass correlation coefficient (ICC) for absolute agreement was calculated between features extracted from each segmented volume of interest. ICC measures the degree of agreement between measurements made by two or more readers. According to the guidelines, an ICC above 0.75 was indicative of good agreement and was selected as the cutoff for subsequent analysis (28).

The remaining robust features were subjected to hierarchical cluster analysis to identify similar and redundant feature groups. Both contrast-enhanced T1- and T2-weighted features were considered for hierarchical clustering. All radiomic features were standardized prior to clustering analysis. Once the cluster groups were identified, univariate Kaplan-Meier analysis was performed for each feature in each group. The median was used as the cutoff for survival groups. The feature that was the most significant based on the log-rank test was selected as the representative feature of the group (lowest P value). In addition to the radiomic features, five clinical features, including patient age, sex, T stage, N stage, and overall stage, were also analyzed. In consideration of sample size of staging classification subgroups, the best groupings of staging classification based on the log-rank test were used. Radiomic and clinical features that were associated with PFS were selected for the final classification modeling.

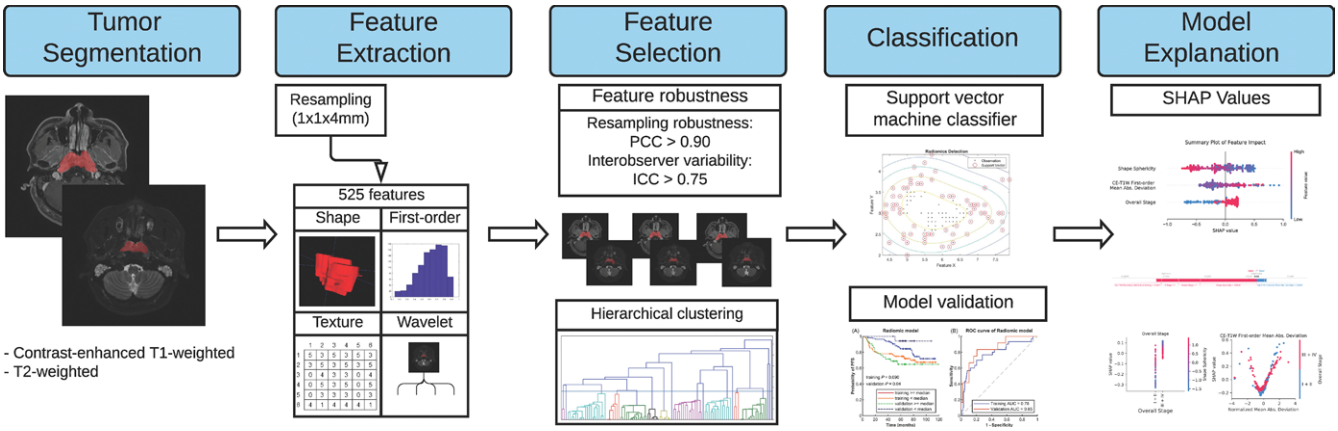


Figure 1: Workflow of the radiomic analysis in this study. Abs. = absolute, AUC = area under the ROC curve, CE-T1W = contrast-enhanced T1-weighted, GLCM = gray level co-occurrence matrix, ICC = intraclass correlation coefficient for absolute agreement, PCC = Pearson correlation coefficient, PFS = progression-free survival, ROC = receiver operating characteristic curve, SHAP = Shapley additive explanations, wavelet-LL = low-low band wavelet transforms.

Classification and Evaluation

Selected features were subjected to classification modeling to predict early 3-year disease progression. A support vector machine with a Gaussian kernel was selected as the classifier for the model. The performance of the model was evaluated internally with the discovery cohort and externally with the validation cohort. The area under the receiver operating characteristic curve (AUC) was used as the main performance measure of the model. Sensitivity, specificity, and positive and negative predictive values in predicting 3-year disease progression were also calculated. For the AUC, 95% bias-corrected and accelerated bootstrap confidence intervals were calculated for 1000 bootstrap samples. The Brier score was used to measure model calibration by assessing the mean square difference between the predicted probability and the actual outcome. The lower the Brier score, the more accurately the predictions are calibrated (29).

To enable understanding and explanation of the underlying decision rule learned by the machine learning support vector machine radiomic model, SHAP values were calculated for each prediction. The SHAP method is derived from game theory and measures how much each feature of a model contributes to the increase or decrease of the probability of a single output (ie, the risk of early disease progression in this case) (21). This is achieved by perturbating the features and modeling them with the prediction of the trained model with a linear modeling method, such as logistic regression, hence allowing a simple linear interpretation of the impact of each feature. A feature's impact and the interaction between features were investigated.

Software and Tools

Manual segmentation was performed by a radiologist using the open-source ITK-SNAP 3.6 software *(http://www.itksnap.org)* (30). All image processing and radiomic feature extraction were conducted in Python using the open-source pyradiomics package (31). For statistical analysis, hierarchal clustering, and machine learning analysis, the open-source Python packages SciPy, scikit-learn, and SHAP were used (21,32,33).

Results

Feature Selection

A total of 530 features were evaluated in this study (525 radiomic features, five clinical features). Of the 525 radiomic features, only 114 were found to be robust against image domain resolution resampling (PCC $>$ 0.9). A further six features were excluded because they were not reproducible against interobserver variability (ICC $<$ 0.75). PCC and ICC values of each feature are given in Appendix E1 (supplement). The remaining 108 robust radiomic features were subjected to hierarchical cluster analysis to identify similar and redundant feature groups. The cophenetic correlation coefficient of the cluster was found to be 0.844. The cophenetic correlation coefficient measures the goodness of fit of the cluster, and a value greater than 0.75 is considered good. A dendrogram of the cluster is shown in Figure 2. A total of 17 feature cluster groups were identified based on the dendrogram.

Kaplan-Meier analysis was performed for each feature in each group. The feature that was most significant based on the *P* value of the log-rank test was selected as being representative of the group (Fig 2). For radiomic features, shape sphericity (P = .032), contrast-enhanced T1-weighted first-order mean absolute deviation (P = .036), contrast-enhanced T1-weighted low-low band wavelet transforms GLRLM gray level nonuniformity normalized (P = .019), and contrast-enhanced T1-weighted low-low band wavelet transforms GLCM sum entropy (P = .051) were found to be significant or nearly significant in separating survival. For staging classification, the best groupings were as follows: T1 + T2 and T3 + T4 (P = .056), N0 + N1 + N2 and N3 (P = .088), and I + II and III + IV (P = .038). Overall stage and T stage were selected for the subsequent model, as it was found to be significant and nearly significant. ICC, survival groups, and Kaplan-Meier analysis of the selected features are shown in Table 2 and Figure 3. No other representative radiomic or clinical features were found to be significantly associated with PFS (P $>$ .15). A list of log-rank tests of all features can be found in Appendix E2 (supplement).

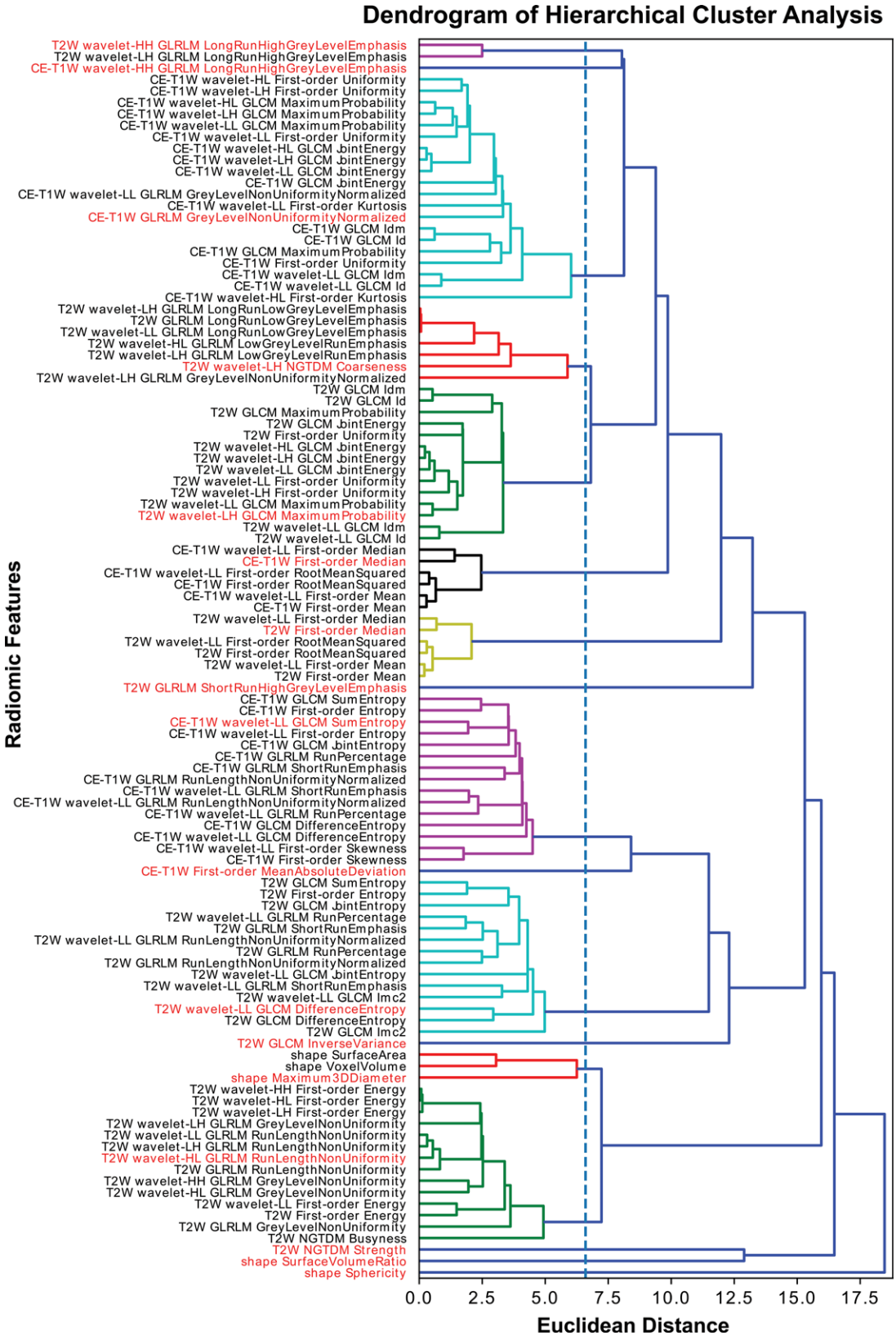


Figure 2: Dendrogram of hierarchical cluster of radiomic features. The dotted line indicates the threshold used for separation of the cluster groups. Representative features for each cluster group are highlighted in red. CE = contrast-enhanced, GLCM = gray level co-occurrence matrix, GLRLM = gray level run length matrix, T1W = T1-weighted, T2W = T2-weighted, wavelet-HH = high-high band wavelet transforms, wavelet-HL = high-low band wavelet transforms, wavelet-LH = low-high band wavelet transforms, wavelet-LL = low-low band wavelet transforms.

Classification Model and Performance

Three different Gaussian kernel support vector machine models were trained with the six features mentioned in Table 3 to predict 3-year disease progression. Performance of the models is summarized in Table 3. The best performance was found in the model trained with both radiomic and clinical features, which achieved an AUC of 0.80 in both the discovery cohort (95% bootstrap confidence interval: 0.80, 0.81) and the validation cohort (95% bootstrap confidence interval: 0.73, 0.89). The receiver operating characteristic curve plot and the Kaplan-Meier plot of the model are shown in Figure 4. By using median output probability as the threshold for separating the survival group, the model was found to be significantly associated with PFS in the discovery cohort ($P <$ .001) and nearly significantly associated with PFS in the validation cohort (P = .057). The performance of the radiomic-only model (AUC = 0.71 [95% bootstrap confidence interval: 0.71, 0.72] and 0.76 [95% bootstrap confidence interval: 0.58, 0.92] in the discovery and validation cohorts, respectively) was found to be significantly higher than the performance of the clinical-only model (AUC = 0.57 [95% bootstrap confidence interval: 0.55, 0.57] and 0.55 [95% bootstrap confidence interval: 0.53, 0.55] in the discovery and validation cohorts, respectively). For model calibration, the Brier score was 0.101 and 0.150 in the discovery and validation cohorts, respectively.

Table 2: Kaplan-Meier Analysis and Interobserver Variability of the Selected Radiomic and Clinical Features for Machine Learning Modeling

Feature Type and Risk Group	No. of Patients	Threshold	*P* Value*	ICC
Clinical features				
T stage				
Low risk	103	T1 + T2	.057	NA
High risk	114	T3 + T4	…	…
Overall stage				
Low risk	66	I + II	.038	NA
High risk	151	III + IV	…	…
Radiomic features				
Shape sphericity				
Low risk	109	≥0.641	.032	0.86
High risk	108	<0.641	…	…
CE T1W first-order mean absolute deviation				
Low risk	108	<49.4	.033	0.91
High risk	109	≥49.4	…	…
CE T1W wavelet LL GLCM sum entropy				
Low risk	108	<6.98	.051	0.93
High risk	109	≥6.98	…	…
CE T1W wavelet LL GLRLM GLNUN				
Low risk	109	≥0.016	.019	0.99
High risk	108	<0.016	…	…

Note.—Risk groups were determined visually with the Kaplan-Meier plot. CE T1W = contrast-enhanced T1-weighted, GLCM = gray level co-occurrence matrix, GLNUN = gray level nonuniformity normalized, GLRLM = gray level run length matrix, ICC = intraclass correlation coefficient, wavelet LL = low-low band wavelet transforms.
* *P* values were determined with the log-rank test.

Model Explanation with SHAP

The SHAP values of each feature for each prediction were calculated. For each prediction, a positive SHAP value indicates an increase in the risk of early disease progression and vice versa. SHAP values of each prediction are summarized in Figure 5. As seen in the plot, shape sphericity was found to be the most important risk factor, with a decrease in sphericity corresponding to an increase in risk. The contrast-enhanced T1-weighted first-order mean absolute deviation was also an important factor in the prediction of the radiomic model. T stage and overall stage were found to be the clearest indicators of risk, with advanced stage III + IV and T3 + T4 tumor clearly indicative of higher risk of disease progression. Figure 6 shows the dependence plots of SHAP values with features. A clear interaction can be observed between sphericity and overall stage. Figure 6a shows that the range of SHAP values at high sphericity in stage I + II tumors was lower than that in stage III + IV tumors, meaning high sphericity in stage I + II tumors does not impact prediction as much as in stage III + IV tumors. This is also represented in Figure 6c, where for stage I + II tumors, a low sphericity score corresponds to decreased risk, whereas a low sphericity score for stage III + IV tumors resulted in higher risk, as demonstrated by the reverse in trends in Figure 6c. A similar but less profound trend was observed for contrast-enhanced T1-weighted first-order mean absolute deviation, with this feature having a similar impact, independent of overall staging (Fig 6d). Despite being significantly associated with PFS in the Kaplan-Meier plot, the model did not find contrast-enhanced T1-weighted low-low band wavelet transforms GLCM sum entropy and gray level nonuniformity normalized to be important for prediction of 3-year disease progression.

Discussion

In this study, we evaluated the prognostic value of radiomic features extracted from pretreatment MRI examinations in the assessment of patients at risk for early disease progres-

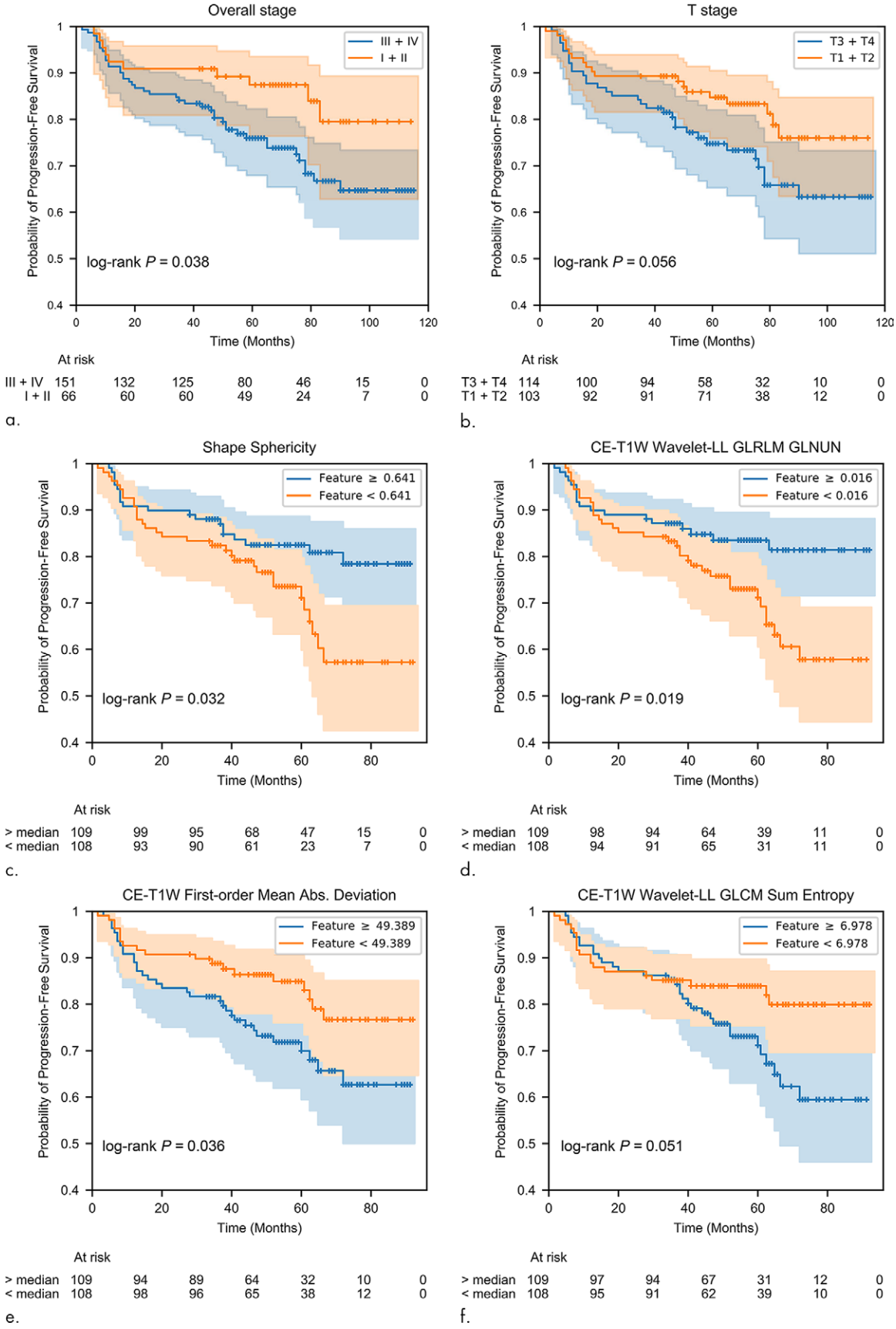


Figure 3: Kaplan-Meier plots of features that are associated with progression-free survival. Log-rank *P* value and risk table are given in each plot. **(a)** Overall stage. **(b)** T stage. **(c)** Shape sphericity. **(d)** Contrast-enhanced T1-weighted low-low band wavelet transforms gray level run length matrix gray level nonuniformity normalized. **(e)** Contrast-enhanced T1-weighted first-order mean absolute deviation. **(f)** Contrast-enhanced T1-weighted low-low band wavelet transforms gray level co-occurrence matrix sum entropy.

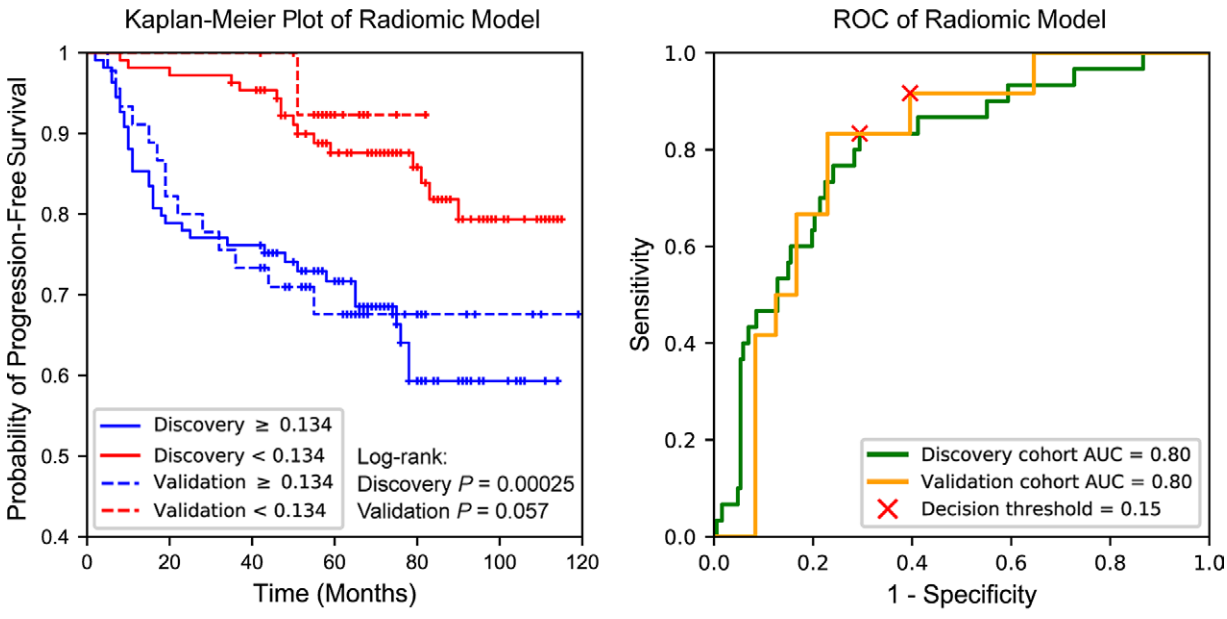
sion. We developed and validated a machine learning model based on a combination of clinical and radiomic features that can predict 3-year disease progression in patients with a diagnosis of NPC after primary treatment. The results showed that the model could discriminate patients who had 3-year disease progression equally well in both the discovery cohort and the independent validation cohort. Interestingly, despite a nearly significant association found in the validation cohort, the model was able to separate PFS under Kaplan-Meier analysis. The significantly higher proportion of advanced-stage tumor in the validation cohort may have resulted in the difference in the log-rank test results in the validation cohort. However, the differences did not affect the ability of the model to discriminate 3-year disease progression, as demonstrated by the high separation at around 36 months in the Kaplan-Meier plot. Despite reasonable discriminability in the model, the positive predictive values were found to be low in both cohorts. This was due to low prevalence of disease (ie, low incidence of disease recurrence) in the population, especially after IMRT (34). The high negative predictive value in the results indicated the model might be more useful in excluding disease recurrence; however, further studies are needed to confirm these assertions.

For feature selection, we decided to only select features that were associated with PFS under the log-rank test. Features that were associated with survival would make

Table 3: Performance of Radiomic and Clinical Models in the Prediction of 3-year Disease Progression

Model	AUC	Sensitivity	Specificity	PPV	NPV
Radiomic and clinical					
Discovery cohort	0.80	0.83	0.71	0.31	0.96
Validation cohort	0.80	0.92	0.52	0.32	0.96
Radiomic only*					
Discovery cohort	0.71	0.67	0.49	0.32	0.94
Validation cohort	0.76	0.92	0.77	0.29	0.95
Clinical only†					
Discovery cohort	0.57	0.53	0.57	0.17	0.89
Validation cohort	0.55	0.75	0.49	0.22	0.84

Note.—AUC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value.
* Features used for the radiomic-only model are shape sphericity, contrast-enhanced T1-weighted first-order mean absolute deviation, contrast-enhanced T1-weighted low-low band wavelet transforms gray level co-occurrence matrix sum entropy, and contrast-enhanced T1-weighted low-low band wavelet transforms gray level run length matrix gray level nonuniformity normalized.
† Features used for the clinical-only model are T stage and overall stage.



a. b.

Figure 4: Performance of the radiomic model. **(a)** Kaplan-Meier plot of the radiomic model. Survival group for both the discovery cohort and the validation cohort is given. The median output probability of the discovery cohort was used as the threshold for separation of groups for both discovery and validation cohorts. **(b)** Receiver operating characteristic (ROC) curve of discovery and validation cohort. Optimal decision threshold was selected by the maximum Youden index of the ROC of the discovery cohort. AUC = area under the ROC curve.

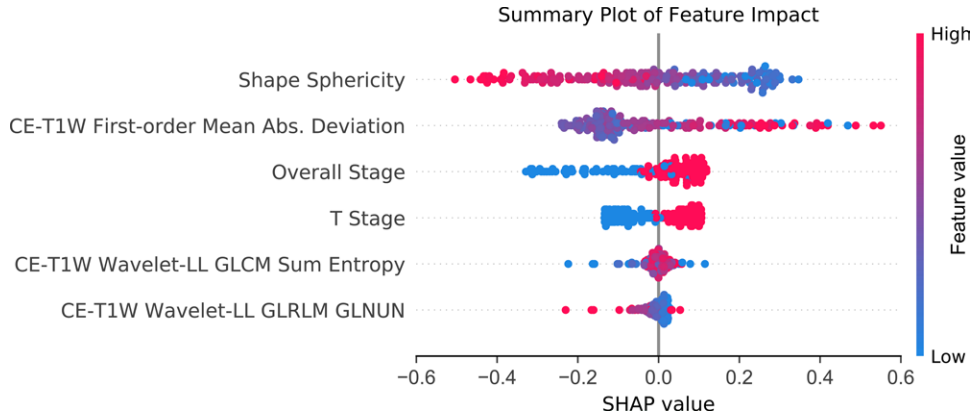


Figure 5: Summary plot of feature impact on the decision of the radiomic model and interaction between features in the model. A positive Shapley additive explanations (SHAP) value indicates an increase in risk and vice versa. For overall stage and T stage, the high value corresponds to stage III + IV and T3 + T4 groups. Each point corresponds to a prediction in a patient. CE-T1W = contrast-enhanced T1-weighted, GLCM = gray level co-occurrence matrix, GLNUN = gray level nonuniformity normalized, GLRLM = gray level run length matrix, wavelet-LL = low-low band wavelet transforms.

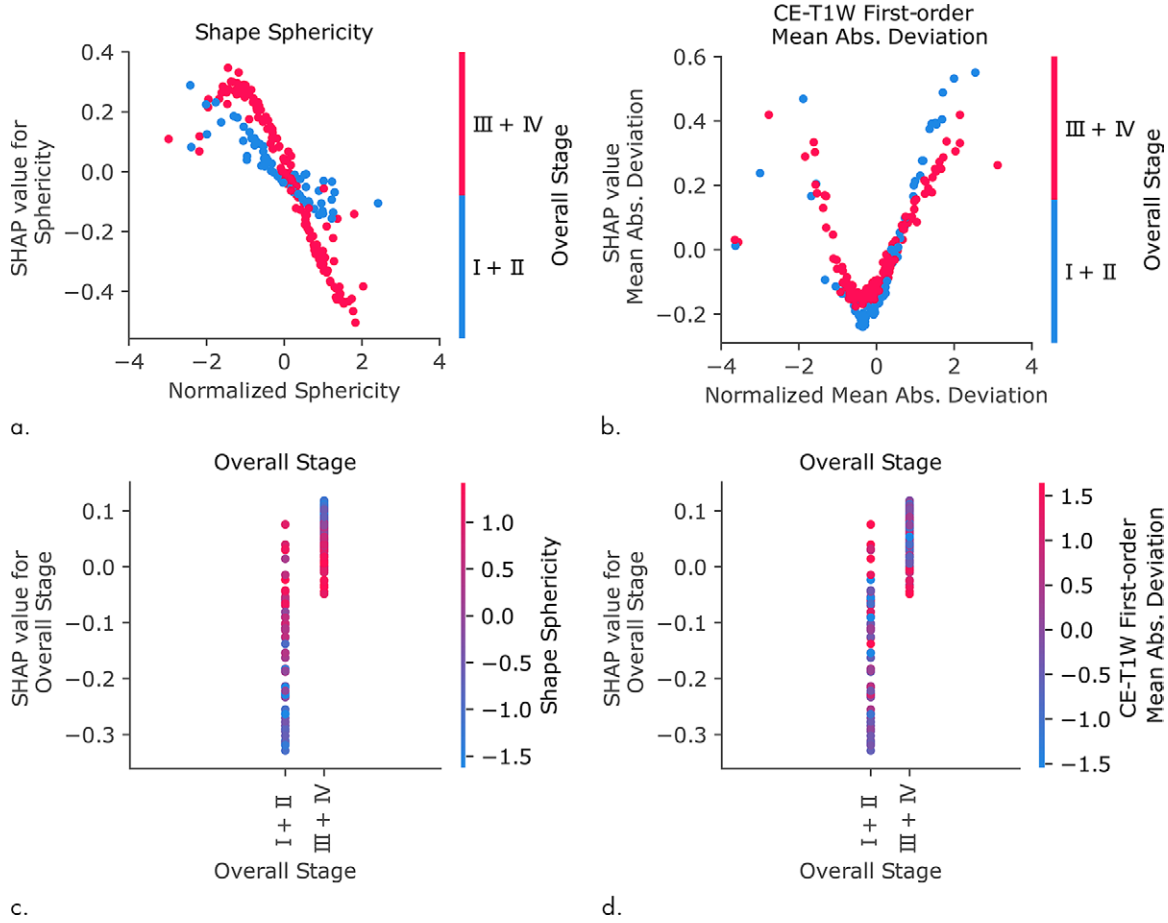


Figure 6: Shapley additive explanations (SHAP) dependence plot of the features in the model. Each point on the plot corresponds to a prediction in a patient. **(a)** Dependence plot of shape sphericity shows how SHAP values vary with varying sphericity and the interaction with overall staging. **(b)** Dependence plot shows how SHAP values vary with contrast-enhanced T1-weighted first-order mean absolute deviation and the interaction with overall staging. **(c)** Dependence plot shows how SHAP values vary with overall staging and the interaction with shape sphericity. **(d)** Dependence plot shows how SHAP values vary with overall staging and the interaction with contrast-enhanced T1-weighted first-order mean absolute deviation.

a better and more interpretable model. One major concern with a radiomic-based prediction model is the interpretability of models. Radiomic models that rely on complex machine learning algorithms often have low clinical utility, as clinicians are unable to understand or explain certain predictions made by the model. To allow for interpretability, clinical models are often restricted to simple linear modeling methods, such as logistic regression, which often leads to lower accuracy. To address this issue, we applied the recently proposed SHAP values to interpret the model. The advantage of SHAP is that it can uncover patterns learned by complex prediction models without being restricted to simple modeling methods. The analysis identified shape sphericity as the most important factor in predicting 3-year disease progression. Sphericity is a measure of tumor compactness, with high sphericity indicating a compact spherical tumor. A previous study on head and neck cancer also yielded similar findings, in that tumor compactness measured with CT was prognostic of survival (27). While there are no clear pathologic findings that relate to the compactness of a tumor, a possible explanation could be that complex nonspherical geometry could correspond to a more invasive and infiltrative tumor. Contrast-enhanced T1-weighted first-order mean absolute deviation was also an important risk factor. Mean absolute deviation is a description of the distribution of intensities. A high mean absolute deviation means there is high contrast between high and low intensity in a tumor, which could reflect contrasting mixture of viable and necrotic tissues or microscopic-level tumoral heterogeneity. T stage and overall stage were also found to be clear indicators of risk in the model, and these findings were consistent with conventional knowledge. One important note is that complexity of the tumor geometry measured with sphericity could be confounded with tumor staging. However, an interesting observation in the model was that despite the fact that a decrease in sphericity generally increases the risk of early progression, low sphericity in early stage I + II tumors was indicative of lower risk when compared with low sphericity in early stage III + IV tumors. Early stage tumors tend to conform to the shape of the nasopharynx, which is irregular in nature and might explain the low sphericity value (further illustrated in Fig E1 [supplement]). In one previous study, a radiomic model based on a combination contrast-enhanced T1-weighted texture and TNM staging achieved an AUC equivalent to 0.78 in discriminating 3-year PFS on a holdout test set, which was consistent with our results (18). However, the radiomic features selected in the model were not selected in models developed in two previous MRI NPC radiomic studies (18,19). A major difference between the present study and the previous studies was that we analyzed robustness against interobserver variability and redundancy. The features selected in the two studies were removed due to robustness or were redundant against the selected features in the model.

It is important to note that patterns explained by SHAP values do not directly explain the underlying sample characteristics but instead explain the pattern learned by the machine learning model. Outliers that deviate from the general trend of the data could not be explained. The results of this study demonstrate that explanatory machine learning techniques such as SHAP may offer value in understanding the prediction made by radiomic or clinical models that rely on machine learning.

There were several limitations to this study. First, we were unable to investigate the association between Epstein-Barr virus and PFS. Because an elevated level of Epstein-Barr virus DNA at diagnosis is indicative of risk in patients with NPC, it could be a potentially useful predictor of early disease progression (35). Owing to the retrospective nature of our study, this was not performed for every patient. Second, we were unable to investigate test-retest or time-dependent variability of radiomic features. Given one of the important predictors in the model was measured on contrast-enhanced T1-weighted MR images, images acquired with different postcontrast times could potentially affect the value and robustness of the features. Finally, because of the retrospective nature of the cohorts, the resolution across examinations was heterogeneous. We decided to resample all scans to the same resolution to provide consistency in texture analysis. Reducing image resolution in texture analysis means that coarser texture was measured. In consideration of the multicenter application of the model, we selected only those features that were highly consistent before and after resampling. This meant that only features that were highly linearly correlated when they were finely and coarsely measured were selected, which limited the number of features in this study.

In conclusion, the results of this study add to the growing evidence of the use of radiomics in tumor diagnosis and risk assessment. By using SHAP values, we were able to uncover an interaction pattern learned by the radiomic model between tumor shape and staging, hence improving explainability and potentially aiding the clinical utility of the model. Further studies on test-retest and time-dependent variability and validation in a larger prospective cohort are needed to determine the true value of radiomics in the assessment of NPC.

Author contributions: Guarantors of integrity of entire study, R.D., V.H.L., V.V.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, R.D., V.H.L., H.Y., V.V.; clinical studies, R.D., V.H.L., H.Y., K.O.L., Y.C., P.L.K., A.W.L., D.L.K.; statistical analysis, R.D., V.H.L., K.O.L., H.H.P., A.W.L., V.V.; and manuscript editing, R.D., V.H.L., K.O.L., H.H.P., Y.C., E.Y.L., P.L.K., A.W.L., V.V.

Disclosures of Conflicts of Interest: **R.D.** disclosed no relevant relationships. **V.H.L.** disclosed no relevant relationships. **H.Y.** disclosed no relevant relationships. **K.O.L.** disclosed no relevant relationships. **H.H.P.** disclosed no relevant relationships. **Y.C.** disclosed no relevant relationships. **E.Y.L.** disclosed no relevant relationships. **P.L.K.** disclosed no relevant relationships. **A.W.L.** disclosed no relevant relationships. **D.L.K.** disclosed no relevant relationships. **V.V.** disclosed no relevant relationships.

References

1. Tang LL, Chen WQ, Xue WQ, et al. Global trends in incidence and mortality of nasopharyngeal carcinoma. Cancer Lett 2016;374(1):22–30.
2. Chua DTT, Wei WI, Wong MP, Sham JST, Nicholls J, Au GKH. Phase II study of gefitinib for the treatment of recurrent and metastatic nasopharyngeal carcinoma. Head Neck 2008;30(7):863–867.
3. Hsieh JCH, Hsu CL, Ng SH, et al. Gemcitabine plus cisplatin for patients with recurrent or metastatic nasopharyngeal carcinoma in Taiwan: a multicenter prospective Phase II trial. Jpn J Clin Oncol 2015;45(9):819–827.

4. Lee AWM, Ma BBY, Ng WT, Chan ATC. Management of nasopharyngeal carcinoma: current practice and fduture perspective. J Clin Oncol 2015;33(29):3356–3364.
5. Peng G, Wang T, Yang KY, et al. A prospective, randomized study comparing outcomes and toxicities of intensity-modulated radiotherapy vs. conventional two-dimensional radiotherapy for the treatment of nasopharyngeal carcinoma. Radiother Oncol 2012;104(3):286–293.
6. Lai SZ, Li WF, Chen L, et al. How does intensity-modulated radiotherapy versus conventional two-dimensional radiotherapy influence the treatment results in nasopharyngeal carcinoma patients? Int J Radiat Oncol Biol Phys 2011;80(3):661–668.
7. Zhang MX, Li J, Shen GP, et al. Intensity-modulated radiotherapy prolongs the survival of patients with nasopharyngeal carcinoma compared with conventional two-dimensional radiotherapy: a 10-year experience with a large cohort and long follow-up. Eur J Cancer 2015;51(17):2587–2595.
8. Chua MLK, Wee JTS, Hui EP, Chan ATC. Nasopharyngeal carcinoma. Lancet 2016;387(10022):1012–1024.
9. Liu LT, Chen QY, Tang LQ, et al. Advanced-stage nasopharyngeal carcinoma: restaging system after neoadjuvant chemotherapy on the basis of MR imaging determines survival. Radiology 2017;282(1):171–181.
10. Lin J, Xie G, Liao G, et al. Prognostic value of 18F-FDG-PET/CT in patients with nasopharyngeal carcinoma: a systematic review and meta-analysis. Oncotarget 2017;8(20):33884–33896.
11. Hsieh TC, Hsieh CY, Yang TY, et al. [18F]-Fluorodeoxyglucose positron emission tomography standardized uptake value as a predictor of adjuvant chemotherapy benefits in patients with nasopharyngeal carcinoma. Oncologist 2015;20(5):539–545.
12. Aktan M, Kanyilmaz G, Yavuz BB, Koc M, Eryilmaz MA, Adli M. Prognostic value of pre-treatment ^{18}F-FDG PET uptake for nasopharyngeal carcinoma. Radiol Med 2017 Nov 25 [Epub ahead of print].
13. Liu J, Mao Y, Li Z, et al. Use of texture analysis based on contrast-enhanced MRI to predict treatment response to chemoradiotherapy in nasopharyngeal carcinoma. J Magn Reson Imaging 2016;44(2):445–455.
14. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–446.
15. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563–577.
16. Neri E, Del Re M, Paiar F, et al. Radiomics and liquid biopsy in oncology: the holons of systems medicine. Insights Imaging 2018;9(6):915–924.
17. Zhang B, Ouyang F, Gu D, et al. Advanced nasopharyngeal carcinoma: pre-treatment prediction of progression based on multi-parametric MRI radiomics. Oncotarget 2017;8(42):72457–72465.
18. Zhang B, Tian J, Dong D, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res 2017;23(15):4259–4269.
19. Ouyang FS, Guo BL, Zhang B, et al. Exploration and validation of radiomics signature as an independent prognostic biomarker in stage III-IVb nasopharyngeal carcinoma. Oncotarget 2017;8(43):74869–74879.
20. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? explaining the predictions of any classifier. ArXiv 1602.04938 [preprint]. https://arxiv.org/abs/1602.04938. Posted February 16, 2016. Revised August 9, 2016. Accessed November 14, 2018.
21. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, et al, eds. Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, 2017; 4765–4774.
22. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2(10):749–760.
23. Yuan H, Ai QY, Kwong DL, et al. Cervical nodal volume for prognostication and risk stratification of patients with nasopharyngeal carcinoma, and implications on the TNM-staging system. Sci Rep 2017;7(1):10387.
24. Lee VHF, Kwong DLW, Leung TW, et al. Prognostication of serial post-intensity-modulated radiation therapy undetectable plasma EBV DNA for nasopharyngeal carcinoma. Oncotarget 2017;8(3):5292–5308.
25. Chua DT, Sham JS, Kwong DL, et al. Volumetric analysis of tumor extent in nasopharyngeal carcinoma and correlation with treatment outcome. Int J Radiat Oncol Biol Phys 1997;39(3):711–719.
26. Shen C, Lu JJ, Gu Y, Zhu G, Hu C, He S. Prognostic impact of primary tumor volume in patients with nasopharyngeal carcinoma treated by definitive radiation therapy. Laryngoscope 2008;118(7):1206–1210.
27. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1):4006.
28. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–163 [Published correction appears in J Chiropr Med 2017;16(4):346.].
29. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.
30. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116–1128.
31. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–e107.
32. Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/. Published 2001. Accessed April 3, 2019.
33. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2012;12:2825–2830.
34. Au KH, Ngan RKC, Ng AWY, et al. Treatment outcomes of nasopharyngeal carcinoma in modern era after intensity modulated radiotherapy (IMRT) in Hong Kong: a report of 3328 patients (HKNPCSG 1301 study). Oral Oncol 2018;77:16–21.
35. Li WF, Zhang Y, Huang XB, et al. Prognostic value of plasma Epstein-Barr virus DNA level during posttreatment follow-up in the patients with nasopharyngeal carcinoma having undergone intensity-modulated radiotherapy. Chin J Cancer 2017;36(1):87.