scientific reports

Check for updates

OPEN

Interpretable survival network for progression risk analysis of multimodality imaging biomarkers in poor-prognosis head and neck cancers

Lise Wei✉, Madhava P. Aryal, Choonik Lee, Jennifer L. Shah, Michelle L. Mierzwa & Yue Cao

This study explores the predictive utility of multi-time point, multi-modality quantitative imaging biomarkers (QIBs) and clinical factors in patients with poor-prognosis head and neck cancers (HNCs) using interpretable machine learning. We examined 93 patients with p16 + oropharyngeal squamous cell carcinoma or locally advanced p16- HNCs enrolled in a phase II adaptive radiation dose escalation trial. FDG-PET and multiparametric MRI scans were conducted before radiation therapy and at the 10th fraction (2 weeks). A survival network analyzed MRI and PET-derived biomarkers such as gross tumor volume (GTV), blood volume (BV), and metabolic tumor volume (MTV_{50}), along with clinical factors to predict local (LF) and distant failures (DF). Feature attributions and interactions were assessed using Expected Gradients (EG) and Expected Hessian (EH). Through rigorous cross-validation, the model for predicting LF, incorporating biomarkers like p16 status and radiation boost, achieved a c-index of 0.758. Similarly, the DF prediction model showed a c-index of 0.695. The analysis of feature attributions and interactions enhanced understanding of important features and complex factor interplays, potentially guiding more personalized and intensified treatment approaches for HNC patients.

Keywords HNCs, Functional imaging biomarker, Interpretable machine learning

Head and neck squamous cell carcinomas (HNSCCs) is the sixth most common cancer worldwide, with a 30% increase expected by 2030[1], and primarily linked to usage of tobacco and alcohol, and human papillomavirus (HPV) infection[2]. The patients with HPV-associated oropharyngeal squamous cell carcinoma (OPSCC) typically exhibit a more favorable prognosis than those with non-HPV one[2]. Furthermore, the patients with early stage HPV-associated OPSCC have an excellent prognosis with "standard chemoradiation therapy (CRT)"[3]. However, up to 20% of the patients are present with advanced stage III of HPV-associated OPSCC and have high recurrence and metastasis rates. The locoregional recurrence rate is as high as 50% during the first 2 years after diagnosis in the patients with advanced HNSCCs including HPV-associated stage III OPSCC and non-HPV related HNSCCs[4]. The response-based adaptive intensification of CRT could improve locoregional control in the patients with local advanced poor prognosis HNSCCs.

Multi-modality imaging serves as a crucial tool for response assessment and for prediction of patient outcomes. Several studies of MRI and PET were carried out to explore predictive imaging biomarkers for locoregional and distant progression of HNSCC[5–19]. However, most of these studies are limited by their reliance on a single imaging modality and at a single time point, which could miss complementary information from different modalities, which could be particularly important in locally advanced diseases, as well as temporal dynamic responses of the cancer. In addition, conventional statistical methods used in most of the studies could miss interactive information from different predictors. Furthermore, characteristics and behaviors of imaging biomarkers could be substantially different between high and low risk patients for locoregional and distant progression, which are not differentiated in most of the early studies. All these together could weaken the predictive power of the model. Recently, there has been considerable growth in the use of machine learning (ML) techniques for analyzing imaging and/or clinical data in HNSCC[20–24]. A cox model is commonly used for multivariate time-to-event analysis, which is a semiparametric model that can calculate partial likelihood without

Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA. ✉email: liswei@med.umich.edu

knowing the baseline hazard function. However, it assumes that the log-risk of failure is a linear combination of predictive covariates, which may not be able to reveal complicated underlying relationships of the covariates. Despite the substantial progress, how to integrate multi-modality images and clinical data and how to interpret complicated machine learning findings are largely unaddressed.

Our study seeks to bridge existing gaps in the field by constructing time-to-event prediction models using multi-time points (pre-CRT and at 2-week during RT), multi-modality (PET and multiparametric MR) imaging biomarkers and clinical data (e.g., p16 status) for LF and DF progression analysis in the patients with locally advanced poor prognosis HNSCCs. Taking a step away from conventional statistical methods, we utilize sophisticated deep learning-based survival models to investigate complicated relationships between outcomes and biomarkers. In contrast to most deep learning-based studies, we not only focus on data modeling but also decipher the model based upon feature attribution, interaction and relation with outcomes.

Methods

Patients and treatment

Ninety-three patients [median age of 63; females 10, p16+ (64%)] with advanced HNSCC were enrolled in a randomized phase II functional imaging-based adaptive RT trial[4]. The trial was approved by the Institutional Review Board of the University of Michigan and written informed consent was obtained from all enrolled patients. All procedures were performed in accordance with the relevant guidelines and regulations. Fifty-nine patients had p16 + T4/N3 squamous cell carcinoma of oropharynx and 34 had locally advanced p16– HNSCC. All patients are M0 stage. There are 79 T4N0-3, 8 T3N2b-3, 3 T3N0-2a, 2 T0-2N3, and 1 T2N1. P16 status was evaluated by immunohistochemistry. The clinical trial was reported elsewhere[4]. The patients who had 1 cc or greater volume of a union of the persisting low blood volume (BV) (defined below) and the persisting low apparent diffusion coefficient (ADC) (defined below) in the gross tumor volume (GTV) pre-RT to after 10 fractions of 2 Gy were randomized to a standard arm of RT (70 Gy in 35 fractions) or an experimental arm, and otherwise to an observation arm (70 Gy in 35 fractions). In the experimental arm, the union received 2.5 Gy per fraction for the last 15 of 35 fractions. The patients had concurrent weekly cisplatin 40 mg/m2, or carboplatin (AUC = 2) for cisplatin ineligibility. After completion of CRT, patients were followed up every 2–3 months per standard care for oncologic outcomes as well as toxicity. The local failure free survival (LF) was defined from the start of RT to the date of local progression. LF times were censored at the earlier death or last follow-up. Distant failure free survival (DF) was defined as the time interval from the start of RT to the date of DF and censored at the earlier death or last follow-up the same as LF.

The median value of primary GTVs is 49.6 cc (range 4.2–595.2 cc, SD 68.7 cc). Twenty-one patients (23%) (8 p16+) had LF, 9 patients had regional progression and 26 patients (28%) (11 p16+) had DF. All cases with LF were confirmed pathologically, and distant metastases were diagnosed pathologically or by overt radiographic presentation. For the patients who did not have any progression, median follow-up was 33 months (range 6–83 months). Basic patient characteristics are given in supplementary Table 1.

Image acquisition, registration and quantitative analysis

The patients had FDG-PET/CT and functional MRI scans 2–4 weeks prior to RT and at fraction 10 (20 Gy) (2wk). All MRI scans were performed on a 3T scanner (Skyra, Siemens Healthineers) using individual patient RT immobilization devices to acquire anatomic, diffusion weighted (DW), and dynamic contrast enhanced (DCE) T1-weighted image series. DW images were acquired with spatial resolution of $1.2 \times 1.2 \times 4.8$ mm and b-values of 50 and 800 s/mm^2 by either a 2D spin-echo single shot echo-planar pulse sequence or a readout segmentation of long variable echo-trains (RESOLVE) pulse sequence that reduced geometric distortion[25]. The DCE image volumes were acquired using a 3D gradient echo pulse sequence in a sagittal orientation with voxel size $1.5 \times 1.5 \times 2.5$ mm during an injection of one standard dose of Gd-DTPA. Post-Gd T1-weighted images were acquired in the axial plane with spatial resolution of $0.875 \times 0.875 \times 3.3$ mm by a 2D fast spin echo sequence with fat saturation.

BV maps were quantified from DCE-MRI using the modified Tofts model implemented in an in-house software (imFIAT), which was validated using a digital reference object[26]. ADC maps were calculated from DW images with b-values of 50 and 800 to mitigate the perfusion effect. The BV and ADC maps pre-RT were reformatted onto post-Gd T1-weighted images pre-RT using coordinates in DICOM headers. The FDG-PET/CT pre-RT and 2wk and MR images at 2wk were co-registered to pre-RT post-Gd T1-weighted images using rigid body transformation and mutual information. Target displacement errors, including image mis-registration and geometric distortion in ADC maps, between image series were assessed and reported previously[27]. Reproducibility of BV maps was 16%, which was reported previously[28].

Image metrics of each tumor (primary or nodal) pre-RT and at 2wk that were considered in this analysis included (1) mean ADC and mean BV in each GTV, (2) low BV subvolume of (BV < 7.64 ml/100 g, TV_{LBV}) and low ADC subvolume (ADC < 1.2 um^2/ms, TV_{LADC}) of each GTV defined previously in[29], (3) metabolic tumor volume of 50% of maximum of standardized uptake value (SUV) of FDG (MTV_{50}) and mean SUV in MTV_{50} and GTV. In addition, two peak values of a bi-distribution of ADC values in a primary GTV (mu1 and mu2) were included to account for the ADC heterogeneous distribution in the large primary tumor volume[30]. For DF analysis, the image metrics were summed or averaged over all primary and nodal tumor volumes for volume-related or intensity-related metrics, respectively, except mu1 and mu2.

Modeling for local and distant progression prediction

To achieve a better prediction of tumor progression, inter- and intra- variable interactions should be included, which cannot be addressed by univariate analysis. Katzman et al. proposed neural network based nonlinear methods for estimating the hazard function by the weights of the network[31]. The neural network can model

nonlinear multivariate relationships even though still using the loss function in the Cox model, and could be superior to the linear combination of the covariates in the Cox model. We used a survival neural network (Survnet) for multivariate analysis. In the model development, we included all image metrics and clinical factors of p16 status and boost initially. Then, we analyzed attribution scores of each input variable and excluded the variable that had little contributions to the model (described below). The input to the network is patient data and the output is the hazard function and optimized by the average negative log partial likelihood similar as the original Cox model. The input features were normalized to be in the range of 0–1. The survival network output was then used to calculate Cox partial likelihood. AdamW was used as the optimizer with learning rate of 0.0005 and weight decay of 0.01. Dropout of 0.1 was used to avoid overfitting. Both models used 10 times 5-fold cross-validation to avoid overfitting. A P-value < 0.05 was considered significant. Kaplan-Meier analysis was performed as well. We used repeated cross-validation to evaluate the performance of the model (10 times 5-fold). We used lasso Cox model as the benchmark model for the LF/DF prediction, in which lasso can shrink coefficients of non-important features[32], and the linear relationship of the features and endpoints were assumed. We applied the same repeated cross validation for the benchmark model as well.

Model interpretation analysis and ablation study

To understand the underlying mechanism of the model, we investigated which input variables contributed most to the model prediction as well as interactions between variables. To do so, we ranked attribution/interaction scores of each variable by using the expected integrated gradient/Hessian (EG/EH)[33,34]. For a model represented by a function $f(x) : \mathcal{R}^d \mapsto \mathcal{R}^1$, where d is the dimension of the feature vector, the integrated gradient (IG) attribution for feature i is:

$$\varphi_i = \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial f\left(x' + \alpha\ (x - x')\right)}{\partial x_i} d\alpha \tag{1}$$

where x is the sample to be explained and x′ is a baseline value. A positive sign of φ_i denotes feature i as a positive contributing variable to f, e.g., a high-risk factor to progression in our case, and the magnitude of φ_i indicates the degree of contribution of feature i to f. There is some randomness in the selection of the baseline. Erion et al.[17] proposed an extension of IG called Expected Gradients (EG), which samples many baseline inputs from the training set and calculate the average IG[35].

Since φ_i itself is also a differentiable function, Expected Gradients can be applied to φ_i to explain the degree to which feature j impacted the importance of feature i[34]:

$$\Gamma_{i,j}(x) = \varphi_j(\varphi_i(x)) = \left(x_i - x_i'\right)\left(x_j - x_j'\right) \times \int_{\beta=0}^{1}\int_{\alpha=0}^{1} \alpha\,\beta\, \frac{\partial^2 f\left(x' + \alpha\,\beta\ (x - x')\right)}{\partial x_i \partial x_j} d\alpha\, d\beta\,, \tag{2}$$

where $i \neq j$. The magnitude of $\Gamma_{i,j}$ indicates the degree of the interaction between features i and j while a positive sign of $\Gamma_{i,j}$ denotes a positive direction of interaction of the two features.

To build accurate, stable and generalizable model, irrelevant, low-contribution, or correlative biomarker metrics need to be identified and excluded from the model. To do so, an ablation study was performed by ranking variable attribution scores by Eq. (3) first. Then, the model was recalculated while removing the variables with low attribution scores one by one. A final model that had a few variables with top ranked attributions was achieved with the best c-index.

Results

Survnet and feature attribution, interaction

LF prediction

Survnet achieved c-index of 0.697 (CI 0.653-0.741) when using all the input image metrics at two time points (pre-RT and at 2wk) and 2 clinical variables of p16 status and boost status. The feature contributions to prediction of LF were ranked by their attribution scores, see Fig. 1. The top six variables were p16 status, boost, MTV50 at 2wk, mean BV pre-RT, mean ADC pre-RT, and GTV at 2wk. P16+ tumor and radiation boost had low risk for LF. Large values of MTV50 at 2wk, mean ADC pre-RT, and GTV at 2wk were high risk for LF. A high value of mean BV pre-RT was low risk for LF.

Considering possibly overfitting of the model with too many correlative, irrelevant or noisy input variables, we eliminated the variables with low attribution scores one-by one and reconstructed new models. The LF prediction model reached an optimal performance with c-index of 0.77 when only six variables (p16 status, boost, MTV_{50} at 2wk, mean BV pre-RT, mean ADC pre-RT, and GTV 2wk) with high attribution scores were included Supplementary Table 2. Including any more variables caused performance decline of the model for LF prediction.

The interactions of the six variables in the optimal LF prediction model are illustrated in Fig. 2 (left). The p16 status interacted with boost, MTV_{50} at 2wk, mean BV pre-RT, GTV at 2wk and mean ADC pre-RT with the degree of interaction in a descending order, which are represented as normalized interaction scores and listed in Supplementary Table 3. Note that the interaction between p16 status and boost was the strongest one, which was 3.5 folds and 33.3 folds stronger than ones between p16 status and MTV_{50} at 2wk and between p16 status and mean ADC pre-RT, respectively. The p16 status interacted with the first four variables in a positive direction, and with the fifth one in a negative direction. The positive interaction between p16 status and boost indicates that boost has a positive effect (low LF) on the p16+primary tumors but a less apparent effect on the p16- ones, see Fig. 3b, given the overall negative attribution for p16 status as shown in Fig. 3a. To visualize the attribution

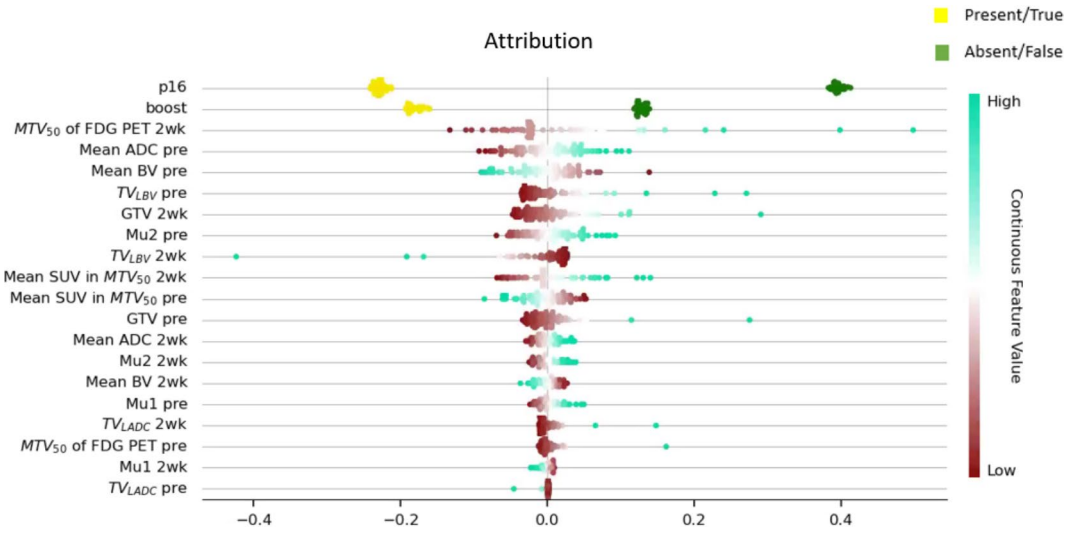

Fig. 1. Feature attributions for prediction of local progression in the Survnet model. The feature attributions are ranked from high to low listed along the left vertical axis. The horizontal axis represents the feature attribution score. The colorbar denotes the magnitude of a feature value as cyan is for a high value and brown is for a low value for continuous features and present/true is gold and absent/false is green for binary features. A feature has cyan/gold color on the negative side of the attribution score indicates low risk for progression and otherwise high risk for progression.

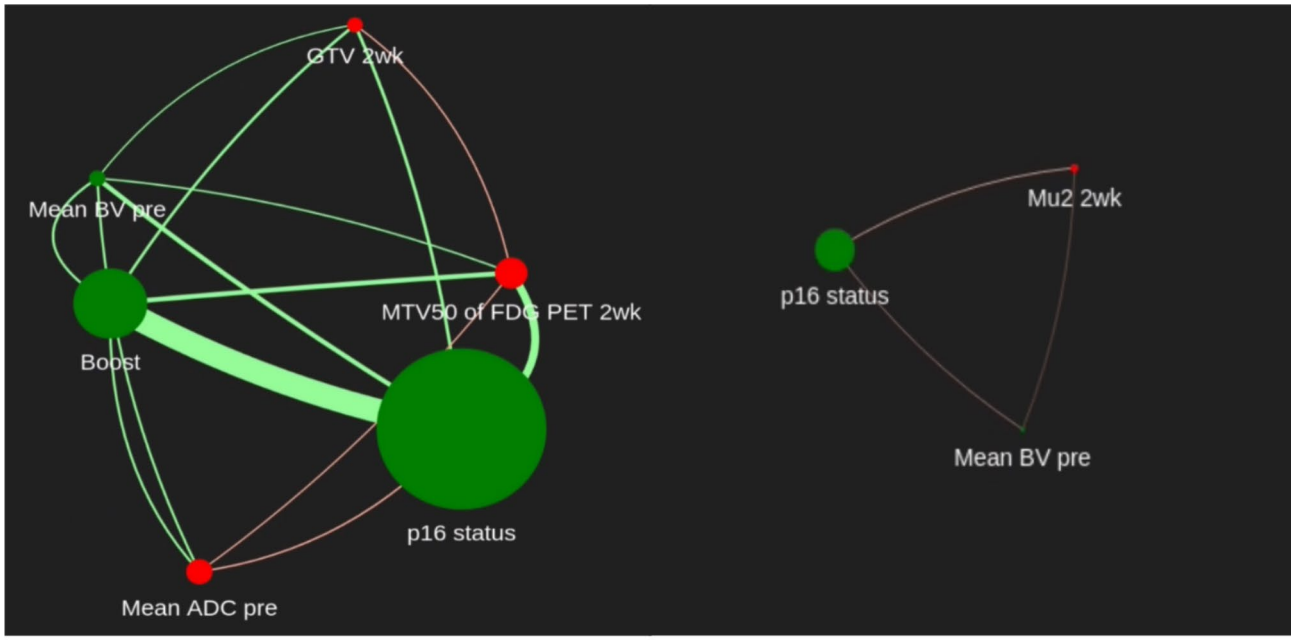

Fig. 2. Interactions between different features for LF (left)/DF (right). The size of the nodes represents the importance of the feature (red means the larger the feature is, the higher risk of LF). The thickness of the edge represents the interaction strength between features. The thicker the edge is, the stronger interaction between the two features. Red edge means that the increase of one feature value will make the other feature less important for the outcome. Green edge is the opposite.

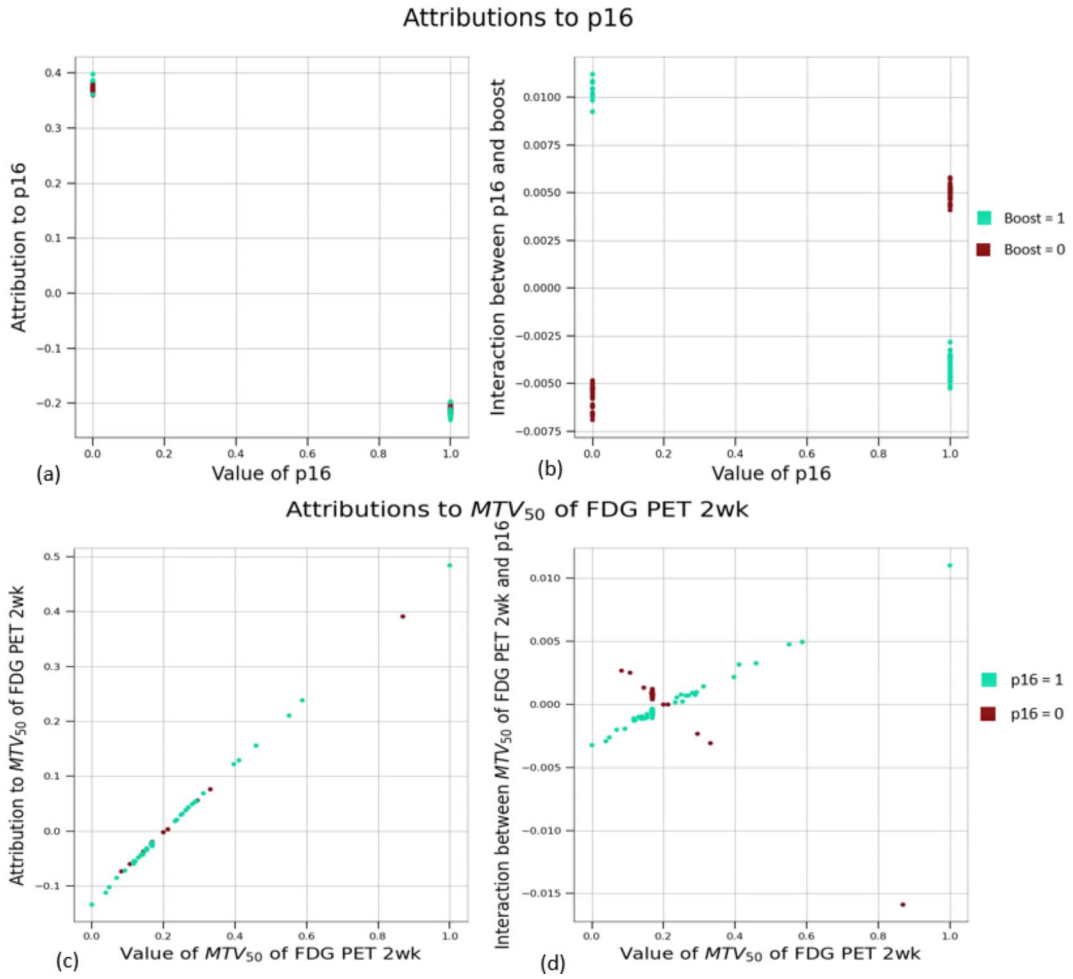

Fig. 3. (**a**): Expected gradient (attribution score) vs. value of p16 status. (**b**): Expected Hessian (feature interaction) between p16 status and Boost vs. value of p16 status. (**c**): Expected gradient (attribution score) vs. value of MTV_{50} at 2wk. (**d**): Expected Hessian (interaction score) between MTV_{50} at 2wk and p16 status vs. value of MTV_{50} at 2wk.

and interaction values better, we plotted histograms of them in supplementary Fig. 1. The positive interactions between p16 status and the first three image metrics indicate that large values of MTV_{50} at 2wk, mean BV pre-RT and GTV at 2wk are associated with high LF in p16 + primary tumors but not in p16- ones. Figure 3c shows that large values of MTV_{50} at 2wk in p16 + primary tumors are associated with higher LF risk, and the positive interaction between MTV_{50} at 2wk and p16 is shown in see Fig. 3d. The negative interaction between p16 status and ADC pre-RT indicates that higher mean ADC pre-RT in p16- primary tumors are associated with high LF but not in p16 + ones. Also, radiation boost had an observable but minor interaction with MTV_{50} at 2wk, which is likely related to the imbalance between boost and non-boost cohorts of the p16 + oropharynx patients.

DF prediction

Survnet achieved c-index 0.680 (CI 0.645–0.715) for DF prediction when using all the variables, for which the attribution scores of all variables are illustrated in Fig. 4. An ablation study showed that the model using three variables with top ranked attribution scores (p16 status, mean BV pre-RT, and mu2 at 2wk) achieved c-index of 0.695 (CI 0.659–0.731), similar to one using all the variables (Supplementary table 2). The p16+ tumors with a high value of mean BV pre-RT and a low value of mean ADC had a low risk for DF. The interactions of p16 status with mean BV pre-RT and mu2 at 2wk in the 3-variable DF prediction model are illustrated in Fig. 2 (right). Their normalized interaction scores to one between p16 status and boost for LF are small and given Supplementary table 3. The p16 status interacted with mean BV pre-RT and mu2 at 2wk in a negative direction, indicating that large values of mean BV pre-RT and mu2 at 2wk in the p16- patients are associated with high risk for DF.

Comparison with p16 alone, adding TN stage, and benchmark Lasso Cox models and final model performance

Performances of the optimal Survnet models were compared with the p16 alone and final lasso Cox models and are summarized in Table 1. C-indices of the Survnet models, 0.77 (CI 0.73–0.79) for LF and 0.69 (CI 0.64–0.72)

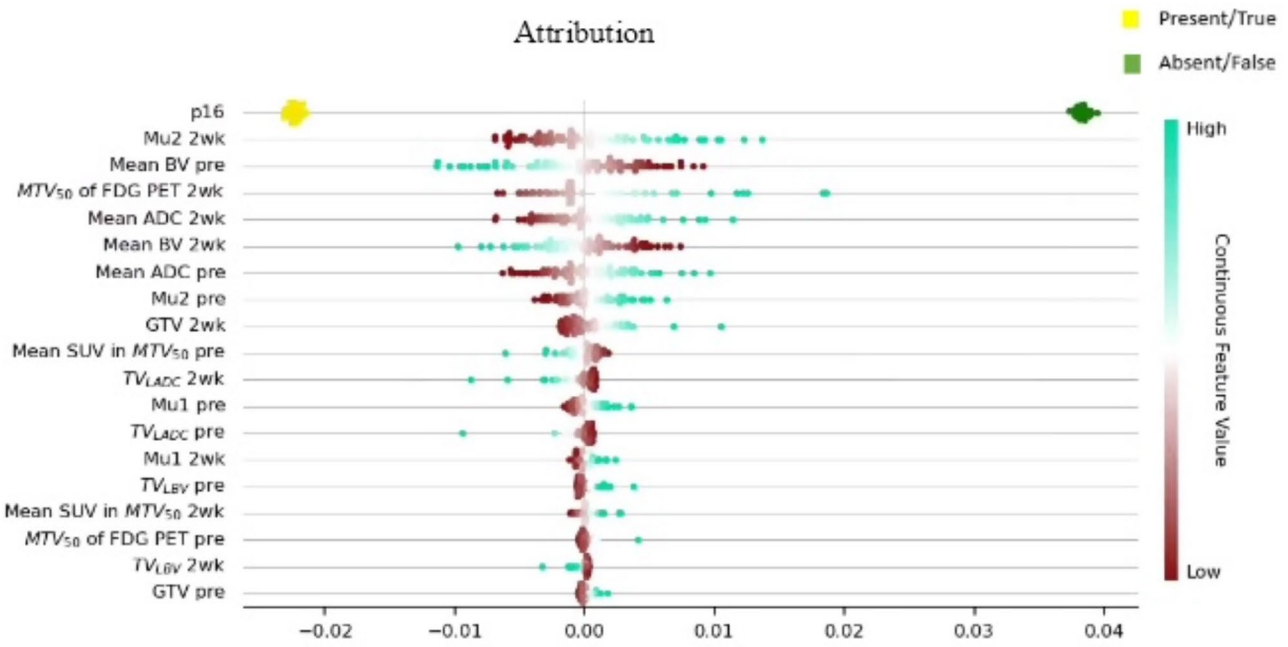

Fig. 4. Feature attributions for prediction of distant progression in the Survnet model. The feature attributions are ranked from high to low listed along the left vertical axis. The horizontal axis represents the feature attribution score. The colorbar denotes the magnitude of a feature value as cyan is for a high value and brown is for a low value for continuous features and present/true is gold and absent/false is green for binary features. A feature has cyan/gold color on the negative side of the attribution score indicates low risk for progression and otherwise high risk for progression.

c-index	LF	DF
P16 status	0.69 (CI 0.65–0.73)	0.66 (CI 0.64–0.69)
P16 + TN stage	0.69 (CI 0.65–0.73)	0.64 (CI 0.61–0.68)
Lasso Cox	0.71 (CI 0.67–0.74)	0.63 (CI 0.60–0.66)
Survnet	0.77 (CI 0.73–0.79)	0.69 (CI 0.64–0.72)
t test p value for Survnet vs. p16 alone	0.003	0.16
t test p value for Survnet vs. Cox	0.03	0.04

Table 1. Final Lasso Cox and Survnet prediction results.

for DF, are significantly better than those of the lasso Cox models [0.71 (CI 0.67–0.74) for LF and 0.63 (CI 0.60–0.66) for DF]. Although Survnet LF model is significantly better than that of p16 alone model [0.69 (CI 0.65–0.73)] with p value of 0.003, Survnet DF model is not significantly better than p16 alone [0.66 (CI 0.64–0.69)]. Adding clinical TN stage information didn't improve the performance of the p16 alone model, and thus they were not further analyzed.

Finally, the patients were stratified to high and low risk groups of LF and DF using the median values of outputs obtained from the optimal Survnet LF and DF prediction models, respectively. Kaplan-Meyer curves showed significant differences between the high and low risk subgroups with $p = 0.00031$ and $p = 0.014$ for LF and DF, respectively, see Fig. 5. There are 13 p16 + patients in high-risk group and 46 p16 + in low-risk group for LF. There are 16 p16 + patients in high-risk group and 43 p16 + in low-risk group for DF.

The statistical power of the final LF/DF models for risk stratification was assessed through a power analysis using the package powerSurvEpi in R. Based on a sample size of 93, a significance level of 0.05, the study achieved a power of 0.88 for LF and 0.64 for DF. The power level of 0.88 indicates that the LF prediction is sufficiently powered to stratify high and low risk patients for LF. However, the power of 0.64 showed that DF model is

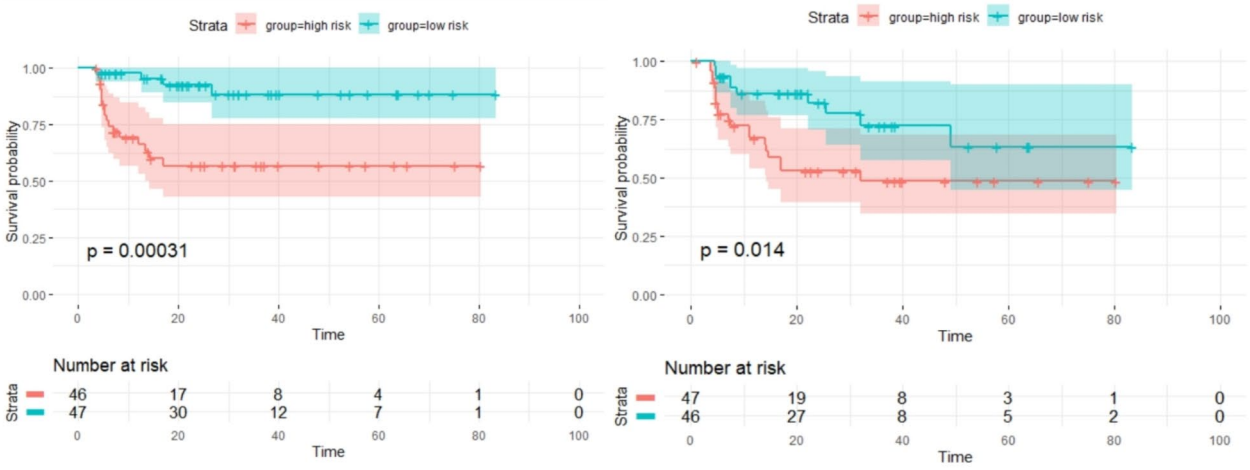

Fig. 5. Kaplan-Meyer plot for LF (left) and DF (right) stratifying high/low risk groups.

underpowered for stratifying patients for DF risk. Future studies with a larger sample size are warranted to increase the power and further validate these findings, especially for DF prediction.

Discussion

In this study, we investigated multi-time point, multi-modality imaging for prediction of LF and DF in poor-prognosis HNCs using a survival neural network that allows nonlinear interactions of imaging metrics and clinical factors. Applying the Expected Gradient and Expected Hessian methods to investigate feature attribution and interaction, we identified top features attributed to the predictive models and revealed the interactions of p16 status with boost and the imaging metrics. By eliminating non-attributive features, we obtained the optimal prediction models with improved accuracy compared with ones using all features. Also, we demonstrated that the machine learning based model can be more transparent and interpretable regarding attribution features and feature interactions. This kind of progression risk modeling could assist in patient stratification in individualized adaptive local and systematic treatments in patients with poor prognosis HNCs.

The identified important features attributed to the Survnet-based prediction models by the EG method are consistent with clinical and imaging studies[27,29,36–38]. As expected from the clinical trial report, p16 status and radiation boost of functional MRI defined high risk subvolumes of the tumor were the two top variables for local progression risk[4]. Interestingly, the top image metrics selected in the optimal LF prediction model included one metric for each modality/contrast as MTV_{50} at 2wk from FDG PET, mean BV pre-RT from DCE MRI, and mean ADC pre-RT for DW-MRI. Note that the metrics corresponding to the selected three imaging metrics but measured at different time points were ranked low (see Fig. 1), suggesting that the Survnet could differentiate high correlations of a metric at two different time points. This feature selection is driven by feature attributions in the model, which can be used to eliminate irrelevant and noisy features to reduce overfitting risks of the model. The performance in prediction of the DF model falls short compared to the LF model. Although Survnet DF model shows significant improvement over Lasso Cox model, it is just slightly better than p16 alone model. This disparity may arise from the presence of more confounding factors in the DF model relative to the LF one. In the case of LF, local features such as GTV and MTV_{50} play a significant role in its predictive accuracy as we found, which are not that predictive in the DF. This also means currently we still haven't found a good prediction model or biomarker for DF other than p16 status. Further research is warranted in the future to improve DF prediction.

The effect of p16 status on imaging metrics is highly expected due to different biology and morphology of p16 + and p16- HNCs[30,39], but is well under-investigated. With this insight, p16 status and imaging metrics could interactively affect the prediction of tumor progression risks. The Survnet al.lows a non-linear interaction of input variables, which could improve the prediction power of the metrics compared with no variable interactions. Also, it is hard to consider the non-linear interaction in standard statistical models, e.g., Cox proportional-hazards model. Although the interaction of p16 status with MTV_{50} of FDG or mean BV in primary HNCs has not been reported before, the interaction between p16 status and ADC metrics has been explored and showed a trend of high mean ADC values in p16- tumors compared to p16 + tumors[29,40,41], which is consistent with the non-linear interaction between these two features found in the Survnet model. Also, the interaction between p16 status and RT boost found in the Survnet model is consistent with the clinical trial report and suggests that the boosting strategy in the trial is more effective for poor prognosis p16 + tumors than p16- ones. We used boost as a predictor since it is randomized in this cohort of patients. However, once boost becomes a clinical practice, it will not be used as a predictor.

The interaction between p16 status and mean ADC pre-RT in LF and the interactions between p16 status and mean BV pre-RT/mu2 at 2wk in DF are less pronounced compared to the interaction between p16 and boost in LF (supplementary Table 3). While these interactions hint at the potential correlation of these features, their lower interaction scores suggest caution in interpreting these results too strongly. Further validation with larger cohorts is necessary, particularly for interactions with notably lower scores.

Our results have the potential to be used to guide the treatment strategies for poor prognosis HNC. For p16 + patients, though a favorable prognosis for LF and DF is found, we can further identify those p16 + patients who have higher risk of LF accurately using our model. Patients can be either monitored closely or treated adaptively with escalated doses. For p16- subgroup, since our model can predict the DF with a c-index of ~ 0.7, identifying high risk patient and use systemic treatment is possible to improve the prognosis for these patients. With more data accumulated in the future, there is still a considerable room for improving the performance of outcome prediction. Finally, our current findings warrant with further validation in a larger, independent data set.

Conclusion

Multi-modality, multi-time points QIBs were used in this study to investigate CRT outcomes (LF/DF) in poor prognosis HNC patients. The Survnet model can predict LF/DF with good accuracy. Important QIBs for LF/DF were recognized and the feature interactions were analyzed to reveal different response patterns for different subgroups of patients. These results could be used to stratify patients for more personalized and optimal treatment.

Data availability

The datasets generated during and/or analyzed during the current study and the corresponding scripts will be made public through GitHub soon after published. Before releasing, they can be available from the corresponding author on request.

Received: 30 June 2024; Accepted: 21 November 2024
Published online: 03 December 2024

References

1. Ferlay, J. et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* **144**, 1941–1953 (2019).
2. Johnson, D. E. et al. Head and neck squamous cell carcinoma. *Nat. Rev. Dis. Primers* **6**, 1–22 (2020).
3. Cooper, J. S. et al. Postoperative concurrent radiotherapy and chemotherapy for high-risk squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **350**, 1937–1944 (2004).
4. Mierzwa, M. L. et al. Randomized phase II study of physiologic MRI-directed adaptive radiation boost in Poor prognosis head and neck cancer. *Clin. Cancer Res.* **28**, 5049–5057 (2022).
5. Xie, P. et al. 18F-FDG PET or PET-CT to evaluate prognosis for head and neck cancer: a meta-analysis. *J. Cancer Res. Clin. Oncol.* **137**, 1085–1093 (2011).
6. Machtay, M. et al. Pretreatment FDG-PET standardized uptake value as a prognostic factor for outcome in head and neck cancer. *Head Neck J. Sci. Specialties Head Neck*. **31**, 195–201 (2009).
7. Zschaeck, S. et al. Prognostic value of baseline [18F]-fluorodeoxyglucose positron emission tomography parameters MTV, TLG and asphericity in an international multicenter cohort of nasopharyngeal carcinoma patients. *PloS ONE* **15**, e0236841 (2020).
8. Kim, S. et al. Diffusion-weighted magnetic resonance imaging for predicting and detecting early response to chemoradiation therapy of squamous cell carcinomas of the head and neck. *Clin. Cancer Res.* **15**, 986–994 (2009).
9. Lambrecht, M. et al. Integrating pretreatment diffusion weighted MRI into a multivariable prognostic model for head and neck squamous cell carcinoma. *Radiother. Oncol.* **110**, 429–434 (2014).
10. Hatakenaka, M. et al. Pretreatment apparent diffusion coefficient of the primary lesion correlates with local failure in head-and-neck cancer treated with chemoradiotherapy or radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **81**, 339–345 (2011).
11. Khattab, H. M., Montasser, M. M., Eid, M., Kandil, A. & Desouky, S. E. D. Diffusion-weighted magnetic resonance imaging (DWMRI) of head and neck squamous cell carcinoma: could it be an imaging biomarker for prediction of response to chemoradiation therapy. *Egypt. J. Radiol. Nuclear Med.* **51**, 1–14 (2020).
12. Bernstein, J. M. et al. Tumor plasma flow determined by dynamic contrast-enhanced MRI predicts response to induction chemotherapy in head and neck cancer. *Oral Oncol.* **51**, 508–513 (2015).
13. Cao, Y. et al. Early prediction of outcome in advanced head-and-neck cancer based on tumor blood volume alterations during therapy: a prospective study. *Int. J. Radiat. Oncol. Biol. Phys.* **72**, 1287–1290 (2008).
14. Wang, P., Popovtzer, A., Eisbruch, A. & Cao, Y. An approach to identify, from DCE MRI, significant subvolumes of tumors related to outcomes in advanced head-and-neck cancer a. *Med. Phys.* **39**, 5277–5285 (2012).
15. Awan, M. J. et al. Post-treatment PET/CT and p16 status for predicting treatment outcomes in locally advanced head and neck cancer after definitive radiation. *Eur. J. Nucl. Med. Mol. Imaging* **44**, 988–997 (2017).
16. Spector, M. E. et al. Matted nodes as a predictor of distant metastasis in advanced-stage III/IV oropharyngeal squamous cell carcinoma. *Head Neck* **38**, 184–190 (2016).
17. Mowery, Y. M. et al. Early 18F-FDG-PET response during radiation therapy for HPV-related oropharyngeal cancer may predict disease recurrence. *Int. J. Radiat. Oncol. Biol. Phys.* **108**, 969–976 (2020).
18. Ng, S. P. et al. Changes in apparent diffusion coefficient (ADC) in serial weekly MRI during radiotherapy in patients with head and neck cancer: results from the PREDICT-HN study. *Curr. Oncol.* **29**, 6303–6313 (2022).
19. Riaz, N. et al. Precision radiotherapy: reduction in radiation for oropharyngeal cancer in the 30 ROC trial. *JNCI J. Natl. Cancer Inst.* **113**, 742–751 (2021).
20. Wang, X. & Li, B. Deep learning in head and neck tumor multiomics diagnosis and analysis: review of the literature. *Front. Genet.* **12**, 42 (2021).
21. Suh, C. H. et al. Oropharyngeal squamous cell carcinoma: radiomic machine-learning classifiers from multiparametric MR images for determination of HPV infection status. *Sci. Rep.* **10**, 1–10 (2020).
22. Martens, R. M. et al. Early response prediction of multiparametric functional MRI and 18F-FDG-PET in patients with head and neck squamous cell carcinoma treated with (chemo) radiation. *Cancers* **14**, 216 (2022).

23. Naser, M. A. et al. Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET-CT imaging data. *medRxiv* (2021).
24. Kann, B. H. et al. Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. *J. Clin. Oncol.* **38**, 1304–1311 (2020).
25. Zhao, M. et al. Readout-segmented echo-planar imaging in the evaluation of sinonasal lesions: a comprehensive comparison of image quality in single-shot echo-planar imaging. *Magn. Reson. Imaging* **34**, 166–172 (2016).
26. Cao, Y., Li, D., Shen, Z. & Normolle, D. Sensitivity of quantitative metrics derived from DCE MRI and a pharmacokinetic model to image quality and acquisition parameters. *Acad. Radiol.* **17**, 468–478 (2010).
27. Teng, F. et al. Adaptive boost target definition in high-risk head and neck cancer based on multi-imaging risk biomarkers. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 969–977 (2018).
28. Aryal, M. P. et al. Real-time quantitative assessment of accuracy and precision of blood volume derived from DCE-MRI in individual patients during a clinical trial. *Tomography* **5**, 61–67 (2019).
29. Cao, Y. et al. Predictive values of MRI and PET derived quantitative parameters for patterns of failure in both p16 + and p16–High risk head and neck cancer. *Front. Oncol.* **9**, 1118 (2019).
30. Cao, Y. et al. Diffusion MRI correlation with p16 status and prediction for tumor progression in locally advanced head and neck cancer. *Front. Oncol.* **13**:998186 .
31. Katzman, J. L. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018).
32. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).
33. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. International conference on machine learning: PMLR; 3319-28. (2017).
34. Janizek, J. D., Sturmfels, P. & Lee, S-I. Explaining explanations: axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.* **22**, 1–54 (2021).
35. Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M. & Lee, S-I. Learning explainable models using attribution priors. (2019).
36. Li, H. et al. Association of human papillomavirus status at head and neck carcinoma subsites with overall survival. *JAMA Otolaryngol. Head Neck Surg.* **144**, 519–525 (2018).
37. Paidpally, V. et al. FDG-PET/CT imaging biomarkers in head and neck squamous cell carcinoma. *Imaging Med.* **4**, 633 (2012).
38. Knegjens, J. L. et al. Tumor volume as prognostic factor in chemoradiation for advanced head and neck cancer. *Head Neck* **33**, 375–382 (2011).
39. Touska, P. & Connor, S. Imaging of human papilloma virus associated oropharyngeal squamous cell carcinoma and its impact on diagnosis, prognostication, and response assessment. *Br. J. Radiol.* **95**, 20220149 (2022).
40. Piludu, F. et al. Multiparametric MRI evaluation of oropharyngeal squamous cell carcinoma. A mono-institutional study. *J. Clin. Med.* **10**, 3865 (2021).
41. Ravanelli, M. et al. Correlation between human papillomavirus status and quantitative MR imaging parameters including diffusion-weighted imaging and texture features in oropharyngeal carcinoma. *Am. J. Neuroradiol.* **39**, 1878–1883 (2018).

Acknowledgements

NIH/NCI grant U01CA183848 and NIH/NCI grant R01CA184153.

Author contributions

L.W. analyzed and wrote the manuscript. M.A. contributed to the data collection and preprocessing. C.L. contributed to the data collection and organization. J.S. and M.M. contributed to the patient enrollment. M.M. and Y.C. designed the clinical trial. Y.C. provided overall study guidance. All authors have involved in manuscript modification, reading and approving the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80815-2.

Correspondence and requests for materials should be addressed to L.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.