scientific reports

Check for updates

OPEN

Identification of CT-based non-invasive radiomic biomarkers for overall survival prediction in oral cavity squamous cell carcinoma

Xiao Ling[1], Gregory S. Alexander[2], Jason Molitoris[1], Jinhyuk Choi[3], Lisa Schumaker[4], Ranee Mehra[4]✉, Daria A. Gaykalova[5,6,7]✉ & Lei Ren[1]✉

This study addresses the limited non-invasive tools for Oral Cavity Squamous Cell Carcinoma (OSCC) survival prediction by identifying Computed Tomography (CT)-based biomarkers to improve prognosis prediction. A retrospective analysis was conducted on data from 149 OSCC patients, including CT radiomics and clinical information. An ensemble approach involving correlation analysis, score screening, and the Sparse-L1 algorithm was used to select functional features, which were then used to build Cox Proportional Hazards models (CPH). Our CPH achieved a 0.70 concordance index in testing. The model identified two CT-based radiomics features, Gradient-Neighboring-Gray-Tone-Difference-Matrix-Strength (GNS) and normalized-Wavelet-LLL-Gray-Level-Dependence-Matrix-Large-Dependence-High-Gray-Level-Emphasis (HLE), as well as stage and alcohol usage, as survival biomarkers. The GNS group with values above 14 showed a hazard ratio of 0.12 and a 3-year survival rate of about 90%. Conversely, the GNS group with values less than or equal to 14 had a 49% survival rate. For normalized HLE, the high-end group (HLE > − 0.415) had a hazard ratio of 2.41, resulting in a 3-year survival rate of 70%, while the low-end group (HLE ≤ − 0.415) had a 36% survival rate. These findings contribute to our knowledge of how radiomics can be used to predict the outcome so that treatment plans can be tailored for patients people with OSCC to improve their survival.

Oral cavity Squamous Cell Carcinoma (OSCC) is an aggressive site among malignancies of the Head and Neck Squamous Cell Carcinoma (HNSCC) with a poor prognosis. Despite improvements in surgical techniques and adjuvant therapies, therapies the 5-year overall survival rate hovers between 30 and 50%, depending on the stage and recurrence status of the disease[1]. It also has a substantial impact on public health worldwide[2–4], with millions of new cases reported annually. For patients with resectable disease, surgery is the standard of care with adjuvant treatments recommended depending on pathologic features, and individualized risk of recurrence. For patients with more advanced stages or adverse pathologic features radiotherapy with or without the addition of cisplatin-based chemotherapies are often recommended. Adjuvant radiotherapy comes with significant detriments to quality of life, which are exacerbated by the use of concurrent chemotherapy. It is, therefore, crucial to identify patients who are at a higher risk of poor survival to increase the therapeutic window with the use of prognostic and predictive biomarkers. Additionally, it may help to identify patients who are at higher risk of distant metastatic spread who may benefit from novel systemic agents that may reduce the risk of spread. By utilizing these biomarkers, patients' cancer trajectory can be better estimated, enabling medical professionals

[1]Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore, MD, USA. [2]Department of Radiation Oncology, Thomas Jefferson University, Philadelphia, PA, USA. [3]Department of Breast Surgery, Kosin University Gospel Hospital, Busan, Republic of Korea. [4]Marlene and Stewart Greenebaum Comprehensive Cancer Center, University of Maryland School of Medicine, Baltimore, MD, USA. [5]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [6]Department of Otorhinolaryngology-Head and Neck Surgery, Marlene & Stewart Greenebaum Comprehensive Cancer Center, University of Maryland Medical Center, Baltimore, MD, USA. [7]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA. ✉email: rmehra@som.umaryland.edu; dgaykalova@som.umaryland.edu; lren@som.umaryland.edu

to tailor treatment plans accordingly. Early access to personalized therapies for high-risk patients can positively impact their prognosis and improve their overall outcomes.

Histological biopsies are commonly used for OSCC diagnosis and prognosis evaluation[5]. Plus, biological fluids-based measurements collected from saliva, blood, serum, and plasma have also been explored as diagnostic and prognostic biomarkers for OSCC. Numerous studies[6–8] have validated the dysregulation of specific miRNAs in OSCC, including miRNA-184[6], miR-31[9], and miR-27b[10], which are associated with malignant transformation and disease progression. Additionally, elevated levels of biomarkers such as CA125[11,12], tissue polypeptide antigen, and Cyfra 21-1 have been observed in OSCC patients compared to control groups. Molecular biomarkers including p16[13], EGFR[14], TP53[15], and Cyclin D1[16] have also demonstrated their ability to distinguish between patients and control groups and show disparities in overall survival based on cutoff values.

While these biomarkers evaluated in pre-clinical settings hold promise in enhancing disease detection, prognosis, and personalized treatment, several challenges persist. The technical complexity involved in extracting and analyzing biomarkers, such as H&E staining, the tissue microarray, and sequencing, often results in high costs and the need for specialized resources. As a result, their applicability could be limited in certain contexts. Issues related to reproducibility and standardization across different laboratories, as well as risks of false positives and negatives, further complicate their utility. Additionally, the validation process of a biomolecule-based assay, from discovery to clinical application, is typically laborious and time-consuming, with many potential markers failing to demonstrate efficacy in diverse populations[17]. Even more, minimizing the total package time are essential to prevent adverse effects on patients in need of adjuvant therapy, allowing for timely initiation of appropriate treatments. Furthermore, some biomarkers, despite indicating the presence of a disease, may not provide actionable insights for treatment strategies, thus limiting their clinical utility[18,19]. Finally, the use of genetic and other biomolecule biomarkers raises ethical, legal, and social considerations[20]. That explains why those biomarkers did not reach the clinical setting yet and were not FDA-approved.

Over recent decades, Imaging biomarkers (IBs) have demonstrated their capacity to furnish accurate prognostic information for treatment outcomes across a variety of diseases trajectories[21–24], including cancer[25]. The pervasive use of CT, PET, and MRI biomarkers in cancer research can be attributed to their non-invasive nature and their ability to portray the multifaceted nature of heterogeneous tumors[26]. Moreover, imaging provides unique 3D information about the tumor. Studies have shown the clinical utility of CT and PET in predicting treatment outcomes for rectal and lung cancer patients[27,28]. The rationale behind this approach is that these images capture crucial information about the tumor phenotype and microenvironment[29]. Recent investigations also hint at the potential role of PET imaging in identifying cancer-associated genetic mutations[30] and the prospective use of radiomics-based imaging biomarkers for outcome predictions in head and neck cancer[31]. To the best of our knowledge, only a few studies[32–35] have identified MRI radiomics features that could be exploited as prognostic tools for overall survival in OSCC. Study[36] found a significant association between Rad-score (linear combination of three PET-based radiomics features) and overall survival. Study[37] suggests that 18F-FDG metabolic tumor volume (MTV) and total glycolytic activity (TGA) could potentially serve as PET markers for overall survival in OSCC. However, the identification of reliable CT-based imaging markers for overall survival (OS) in OSCC patients, who receive various treatment modalities, remains elusive. This study aims to identify CT imaging markers for OS prediction in OSCC patients by employing a novel machine-learning (ML) framework.

A single institution academic health network, serving a diverse population, possesses a diverse cohort of oral cavity squamous cell carcinoma cases. This study endeavors to identify imaging biomarkers associated with overall survival in oral cavity squamous cell carcinoma patients. We employed a retrospective study design using high-dimensional radiomics data and clinical follow-up information. The primary endpoint of the study was overall survival. We used the Cox Proportional Hazard model (CPH) to achieve the goal. Given the high dimensionality of the imaging data, the feature selection methods, in conjunction with Best Subset Selection (BSS) strategy, were deployed to avoid overfitting and to select a parsimonious set of candidate factors. In addition, we also investigated the variations in overall survival outcomes by stratifying individuals based on different levels of two identified radiomics biomarkers, respectively. Lastly, the final CPH model was summarized in a nomogram to facilitate the treatment decision.

Results

Feature selection

In this study, we employed three independent feature selection methods to create a refined input dataset for Cox proportional hazard modeling, addressing various issues posed by high-dimensional data. By pruning features based on the Pearson correlation coefficient and Cramer's V score, we reduced the initial 1092 radiomics features to 79 and the 7 categorical features to 6. Subsequently, a score screening procedure further narrowed down the radiomics features to 17. Finally, the SparseL1 algorithm[38] was fine-tuned to select seven features, resulting in an active input dataset size of 13. The feature selection process effectively mitigated multicollinearity, reduandancy, and computational complexity. The final tenfold concordance index demonstrated the efficacy of this approach.

Best subset selected CPH model

In the tenfold cross-validation of the final Cox model, the results in Table 1 consistently demonstrated a strong model fit for overall survival and absolute survival, as evidenced by the log-likelihood ratio and score test.

This CI, an indicator of predictive accuracy, demonstrates the effectiveness of this prognostic model. The p-values for the continuous variables Gradient-NGTDM-Strength (GNS), Wavelet-LLL-GLDM-LargeDependenceHighGrayLevelEmphasis (HLE), and Stage demonstrated varying levels of significance across the cross-validation folds. GNS (mean p-value = 0.015, SD = 0.013), HLE (mean p-value = 0.023, SD = 0.012), and Stage (mean p-value = 0.019, SD = 0.016) all showed significance in the model to varying degrees. The covariate ETOH

	1	2	3	4	5	6	7	8	9	10	Mean	SD
A. Overall survival												
GNS	0.010	0.010	0.009	0.012	0.009	0.011	0.012	0.010	0.050	0.017	0.015	0.013
HLE_s	0.013	0.032	0.007	0.023	0.029	0.032	0.011	0.045	0.024	0.011	0.023	0.012
Stage	0.007	0.004	0.011	0.028	0.008	0.047	0.009	0.034	0.007	0.039	0.019	0.016
$ETOH_{no}$	0.065	0.023	0.117	0.094	0.146	0.200	0.036	0.046	0.081	0.031	0.084	0.057
LRT	2e−06	9e−07	2e−06	3e−05	6e−06	1e−04	1e−06	5e−06	2e−05	1e−05	3e−06	6e−06
Score	2e−06	3e−06	6e−06	4e−05	2e−05	3e−04	2e−06	6e−06	3e−05	2e−05	6e−06	1e−05
CI_{train}	0.729	0.747	0.726	0.717	0.732	0.705	0.739	0.731	0.726	0.722	0.727	0.011
CI_{test}	0.652	0.537	0.741	0.800	0.658	0.867	0.569	0.650	0.764	0.750	0.699	0.103
B. Absolute survival												
GNS	0.038	0.074	0.067	0.017	0.048	0.027	0.057	0.043	0.036	0.037	0.044	0.018
HLE_s	0.008	0.019	0.033	0.079	0.044	0.012	0.021	0.032	0.014	0.041	0.030	0.021
Stage	0.034	0.027	0.029	0.056	0.084	0.086	0.051	0.127	0.054	0.065	0.061	0.031
$ETOH_{no}$	0.548	0.136	0.294	0.074	0.117	0.110	0.402	0.197	0.267	0.417	0.256	0.158
LRT	4e−06	7e−06	3e−06	1e−06	3e−05	2e−05	9e−05	4e−05	1e−05	8e−05	3e−05	3e−05
Score	1e−05	1e−05	1e−05	8e−06	7e−05	4e−05	2e−04	7e−05	4e−05	2e−04	6e−05	6e−05
CI_{train}	0.763	0.762	0.752	0.769	0.749	0.747	0.742	0.743	0.745	0.734	0.751	0.011
CI_{test}	0.673	0.636	0.478	0.467	0.740	0.829	0.850	0.767	0.737	0.882	0.706	0.144

Table 1. Tenfold Cross-validation *p*-value mean and standard deviation of CPH models and variable.

represents alcohol usage status, with three categories: "Alcohol user" (category 1, reference level), "Alcohol non-user" (category 2), and "Unknown" (category 3). The average *p*-value for the tenfold cross-validation of the non-user categroy is 0.084, with a standard deviation of 0.057 This suggests that alcohol usage may have a significant influence on the model, although the effect might not be as robust as the continuous variables, given there are 18 missing values in ETOH. Table 2 shows the final CPH model parameter estimates and goodness of fit statistics. For instance, the hazard ratios in column *H.ratio* provide insight into the risk effect of each factor. For the clinical ETOH effects, a hazard ratio of 0.54 indicates that non-users have a hazard of 0.54 times that of users. In terms of odds, the probability of death occurring (P) can be calculated by P = H.ratio / (1 + H.ratio). For alcohol users, there is a 65% chance of death, while for non-users, there is a 35% chance of death. The 95% confidence interval (CI) for the effect of non-users lies between 0.29 and 1.01, indicating an acceptable variability in the hazard ratio of 0.54. We also investigated the association between stage, treated as a ordinal variable, and survival time. The hazard ratio for stages is 1.37, indicating a 37% increase in risk when moving from one stage to the next.

Texture analysis

By standardizing HLE to HLE_s (zero mean with a standard deviation of 1), the Hazard ratio of 1.29 for HLE_s indicates that the chance of death increases by 1.29 times for patients with one standard deviation higher HLE_s compared to the previous HLE_s. In other words, for each unit increase in HLE_s, there is a 14% increase in the chance of death compared to the previous one. The lower bound of 95% confidence interval [1.05, 1.58] is also greater than 1, further suggesting that HLE_s is a significant risk factor for overall survival in OSCC. On the contrary, GNS has both a hazard ratio of less than 1 and a 95% confidence interval. The hazard ratio of 0.94 suggests a 52% chance of death for a patient with one unit increase in GNS, compared to a 48% chance of death for a patient with no increase in GNS. Interestingly, even when we include the stage as a covariate in the multivariate analysis, we still identify these two significant radiomics features. This suggests that there are distinct survival differences associated with these features, as demonstrated by the Kaplan–Meier estimator.

The NGTDM contains information about the average grey level of a voxel within a pixel neighborhood, and it also stores information regarding spatial changes in intensity. Strength is one of the five key perceptual attributes of texture, which indicates the ability of elements to be easily distinguished. GNS serves as an indicator of the distinctiveness of the tumor textures. A high GNS value corresponds to a strong texture. In study[39], a significant correlation was found between GNS and coarseness. GLDM quantifies the amount of local variation present in an image[40]. HLE measures the distribution of large dependence with higher gray level intensities, which has been associated with the heterogeneity of tumors[41,42].

Stratification analysis

The Hazard Ratio in Fig. 1 illustrates the impact of two radiomics features HLE and GNS, on each stage, stratified by ETOH and cancer stage while keeping other covariates fixed at their sample means. The KM curves are stratified based on drinking status (drinker and non-drinker) across different cancer stages. Each curve represents the estimated hazard ratio for individuals in a specific group (e.g., drinkers with stage 1 OSCC, and non-drinkers with stage 2 OSCC). The vertical axis represents the hazard ratio, while the horizontal axes represent the radiomics features values. The left plot demonstrates that a higher Hazard Ratio is linked to advanced stage status, with a positive correlation between HLE and Hazard Ratio when other factors are held constant. Alcohol users (ETOH = 1) exhibit a consistently higher hazard ratio across all stages than non-users (ETOH = 2). In the

Factor	Coef.	H.ratio	Se	95% CI	*p* value
A. Overall survival					
Radiomics					
HLE_s	0.259	1.29	0.103	[1.05, 1.58]	0.014
GNS	−0.062	0.94	0.024	[0.90, 0.98]	0.009
Clinical					
Stage	0.313	1.37	0.121	[1.08, 1.73]	0.009
ETOH:2	−0.611	0.54	0.319	[0.29, 1.01]	0.055
ETOH:3	−0.280	0.76	0.443	[0.32, 1.80]	0.527
Goodness of fit					
LRT test	38.58				3e−07
Wald test	33.80				4e−04
Score test	36.94				6e−07
B. Absolute survival					
Radiomics					
HLE_s	0.274	1.32	0.116	[1.05, 1.65]	0.019
GNS	−0.066	0.94	0.031	[0.88, 0.99]	0.032
Stage	0.343	1.41	0.169	[1.01, 1.96]	0.043
Clinical					
ETOH:2	−0.456	0.63	0.357	[0.31, 1.28]	0.202
ETOH:3	−1.356	0.26	1.029	[0.03, 1.94]	0.188
Goodness of fit					
LRT test	32.82				4e−06
Wald test	26.35				8e−05
Score test	30.92				1e−05

Table 2. Final Cox model fitting on the sample data. 95% CI is calculated by exp(coef ± 1.96 × se). Both tests yielded significant *p*-values across all folds, with average *p*-values of 3e−06 and 6e−06 for the log-likelihood ratio and the score test, respectively, suggesting that our model is highly significant and provides a good fit to the data. The training concordance index (CI) remained stable and high across all iterations, with an average value of 0.727 (SD = 0.011). The testing CI has a mean value of 0.699 and a standard deviation of 0.103. In a simlar vien, the average training CI for absolute survival is 0.751 and testing CI is 0.706.

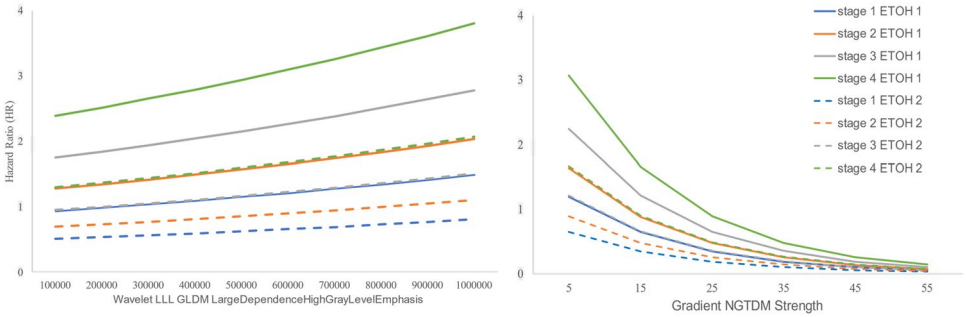

Figure 1. Hazard Ratio functions of radiomics features at diagnosis on each state transition, stratified by ETOH (solid versus dashed lines) and cancer stage (various colors).

right plot, we observe a negative relationship between GNS and Hazard Ratio. The hazard ratio shows a steadier change for non-users at each.

Several trends are notable in this comprehensive dataset reflecting the clinical characteristics of two stratified cohorts, respectively, designated as HLE_l vs HLE_h and GNS_l vs GNS_h. The low-end group (GNS_l) consists of patients with a GNS value less than or equal to 14, while the high-end group (GNS_h) comprises patients with GNS values greater than 14. Similarly, for the normalized HLE (HLE_s) feature, we formed two groups: the low-end group (HLE_l) with an HLE_s value less than or equal to −0.415 and the high-end group (HLE_h) with HLE_s values greater than −0.415. This stratification allows us to identify differences in outcomes and treatment response between the groups, as well as explore the potential predictive value of these radiomics features. Predominantly male participation is observed in both cohorts, with the low-end cohort significantly outweighing the high-end cohort. The mean age at diagnosis is almost identical in both cohorts. As for lifestyle habits, low-end cohorts have

a slightly higher percentage of smokers. Regarding alcohol consumption, a larger 73% of GNS_h participants were alcohol users compared to 45% in GNS_l. The alcohol consumption rate was distributed evenly in HLE groups.

Figure 2 displays the Kaplan–Meier curves for each feature. Stratification by GNS reveals a significant difference both in overall survival and absolute survival between the two groups. The group with GNS greater than 14 shows a flat overall survival curve with only two events out of 24 risks and absolue survival curve with only 1 event out of 23, suggesting that this group generally has a good prognosis with a high overall survival probability over the follow-up period. In contrast, the median overall survival for the group with GNS less than 14 is approximately 37 months, with 64 events out of 125. The 3-year overall and absolute survival rate for the group with GNS greater than 14 are 90% and 96%, compared to 49% and 57% in the group with GNS less than 14. The median overall survival for the HLE_l group is 154 months, compared to 10 months for the HLE_h group. The 3-year overall and absoulte survival rate are 60% and 75% for the former group and 23% and 49% for the latter.

Radiomics-based nomogram

Based on our model, we developed a nomogram in Fig. 3 that visually represents the CPH model presented in Table 2. This nomogram allows the estimation of overall survival for OSCC patients after treatment. To use the nomogram, one simply needs to input the values of four variables (HLE, GNS, stage, and ETOH) and mark them on their respective axes. Connecting these marked values with vertical lines to the top scale (points scale)

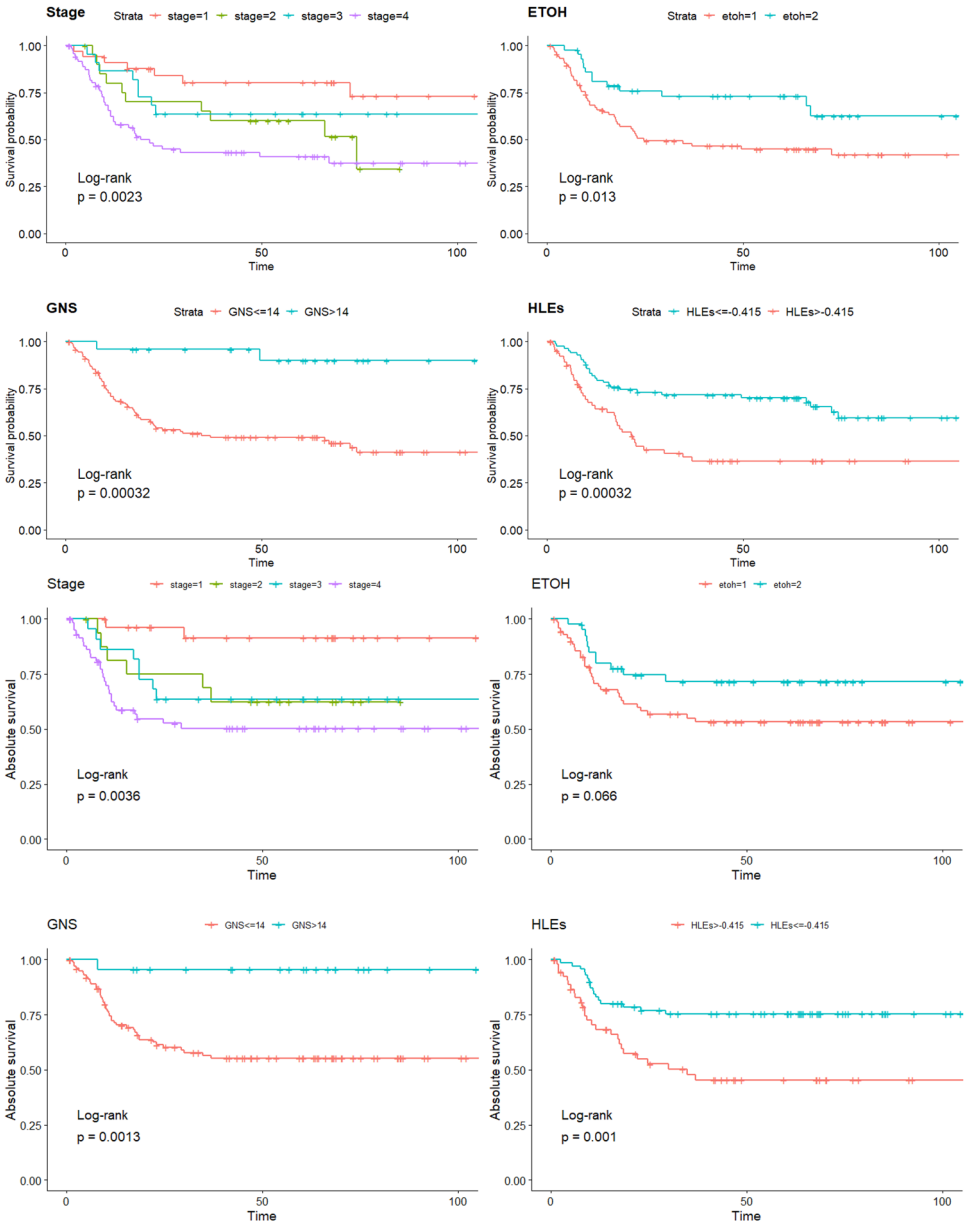

Figure 2. The top four KM curves represent overall survival rates stratified by each factor, while the bottom four KM curves depict the absolute survival stratified by each factor.

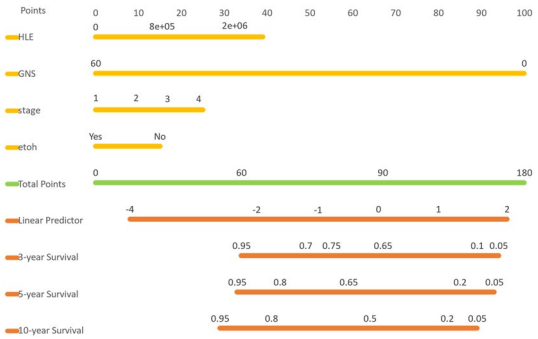

Figure 3. Nomogram from the fitted Cox model.

determines the points for each variable. Adding these points together and marking them on the total points axis provides the total points. By connecting the position of total points with the corresponding survival probability, one can estimate the overall outcomes based on the Linear Predictor.

The purpose of the nomogram in thIs study is to provide a practical and user-friendly tool for estimating overall survival in OSCC patients after treatment. By integrating multiple prognostic factors into a graphical representation, the nomogram allows healthcare professionals to easily assess individual patient outcomes and make informed decisions regarding treatment strategies. The benefit of using a nomogram lies in its ability to incorporate complex statistical models into a visually intuitive format, enabling personalized risk prediction. It offers improved prognostic accuracy, individualized treatment planning, and enhanced communication between healthcare providers and patients. The nomogram serves as a valuable addition to clinical practice by facilitating shared decision-making and promoting precision medicine approaches in the management of OSCC.

Discussion

The present study aimed to evaluate the prognostic value of radiomics features in oral cavity squamous cell carcinoma (OSCC) patients. Our findings demonstrate that radiomics analysis of pre-treatment CT scans can provide valuable insights into the factors influencing survival and serve as prognostic biomarkers in this patient population.

Key findings of our study include two significant hazard ratios, 1.29 and 0.94, between two radiomics features and overall survival. The concordance-index (CI) showed a stable and high average value of 0.7, indicating good predictive accuracy of the prognostic model. These results highlight the potential of medical imaging, particularly radiomics, as a non-invasive and quantitative method for treatment prognosis. Methodological considerations were also addressed in our study. We standardized the voxel spacing across patients by resampling CT images to ensure accurate feature calculation. Additionally, gray-level normalization was applied to enhance the comparability of features and improve their robustness against variations in different settings. These steps are crucial for reliable and reproducible radiomics analysis. The feature selection process in our study involved sequential approaches consisting of Correlation Analysis, Score screening, and the SparseL1 algorithm. This process effectively reduced the dimensionality of the feature space while retaining a significant amount of the prognostic information present in the original data. This approach helps to mitigate issues such as bias, overfitting, and multicollinearity that can arise in high-dimensional data analysis. Best subset selection modeling techniques were utilized to identify optimal Cox proportional hazards models, leading to the identification of biomarkers associated with survival in OSCC patients.

The Cox proportional hazards modeling revealed several significant radiomics features associated with survival in OSCC patients. The continuous variables, Gradient-NGTDM-Strength (GNS) and Wavelet-LLL-GLDM-LargeDependenceHighGrayLevelEmphasis (HLE), showed varying levels of significance in the model, indicating their potential as prognostic biomarkers.

The categorical variable, alcohol consumption (ETOH = 2), also demonstrated some degree of influence on the model. The performance of our Cox model was assessed using log-likelihood ratio and score tests, which consistently yielded small *p*-values across all folds in the tenfold cross-validation. The concordance index (CI), a measure of predictive accuracy, remained stable and high, indicating the effectiveness of our model. These results suggest that our Cox model provides robust predictive performance for survival in OSCC patients.

While our study highlights the potential of radiomics in OSCC prognostication, it is important to acknowledge its limitations. The relatively small sample size and the nature of the survival study may impact the stability of the validated radiomics features in our model. To support a low-biased and variance survival model with four effects, it is recommended to have at least 40 events in each training set, requiring a sample size containing 67 events if 60% is allocated for the training set. This criterion restricts the degrees of freedom our model can reach, potentially affecting the prognostic ability of the underlying radiomics. Additionally, the current analysis focused on extracting radiomics features from a single imaging modality, i.e. CT. Future studies are warranted to investigate radiomics features from multiple modalities, such as CT and MRI, which opens up the potential to improve the prediction accuracy further. Last, the significant association between certain radiomics features and overall survival suggests that imaging features may reflect some of the underlying molecular characteristics of the tumors. Future investigations are warranted to integrate genetic TP53 mutations[15] and P16 overexpression[43]

and radiomics data to characterize squamous cell carcinoma of the head and neck and provide an alternative non-invasive, multi-modal approach to OSCC outcome predictions.

In conclusion, our study demonstrated the potential of radiomics as an effective tool to predict treatment response in OSCC patients. Incorporating radiomics analysis into clinical practice could improve decision support and enhance patient stratification, reducing both over-treatment and under-treatment to improve outcomes. The findings from the study pave the way for future investigations through a larger clinical trial to further evaluate the clinical efficacy of radiomics biomarkers for overall survival prediction for OSCC patients.

Methods

Endpoints of interest and study cohorts

This retrospective cohort study examines a group of oral cavity squamous cell carcinoma (OSCC) patients who underwent contrast-enhanced CT scans at the institution between 2006 and 2017. The sample size consisted of 149 patients. We collected six clinical attributes, including age at diagnosis, gender, tobacco use, alcohol consumption, stage, and race, summarized in Table 3. Table 1 presents a comprehensive summary of the clinical factors observed in this cohort. The mean age at the diagnosis was 62, ranging from 29 to 98 years' old. Patients were categorized into four stages (I, II, III, and IV) based on the pathological assays of tumor specimens. Smoking and alcohol status were self-reported and coded as 1 for yes and 2 for no. The missing values for smoking and alcohol status were hard-coded as 3 due to their substantial representation within the dataset. Six treatment modalities, including chemoradiotherapy (CRT), chemotherapy (CT), surgery (Sx), radiotherapy plus surgery (RT + Sx), CRT + Sx, and CT + Sx, were administered to patients as their initial treatment. The endpoint in this study was overall survival (OS), defined as the time from the date of diagnosis (determined by the diagnostic scan or biopsy) to the date of death or last follow-up day. As of November 04, 2019, a total of 66 patients died after treatment. The average survival time among all 149 patients was 40 months (ranging from 1 to 154 months). Among the patients who died, the average survival time was 19 months (ranging from 2 to 154 months), whereas among patients alive at the last follow-up the average survival time was 58 months (ranging from 1 to 137 months). We also compared overall survival to absolute survival defined as the proportion of patients who were alive after primary surgery.

Gender	
Male	79
Female	70
Race	
EA	133
AA	16
Smoking	
Yes	104
No	34
Unknown	11
Alcohol	
Yes	88
No	43
Unknown	18
AJCC stage	
I	33
II	21
III	22
IV	73
Treatment	
CRT	6
CT	2
Sx	87
Sx + RT	27
Sx + CRT	26
Sx + CT	1
Status	
Alive	83
Dead	66

Table 3. Clinical characteristics summary.

Data preparation and overall workflow

This study aims to enhance the prognostic and predictive value of radiomics for OSCC patients by extracting1092 radiomics features from pre-treatment CT scans. The primary objective is to identify prognostic and predictive biomarkers within these CT scans that can facilitate the assessment of treatment effectiveness at the individual patient level. This will enable the selection of tailored treatment strategies and ultimately lead to improved patient outcomes. The workflow outlining our approach is illustrated in Fig. 4**.** In this workflow, the tumor volume serves as the region of interest (ROI) from which all radiomics features are computed (as shown in Fig. 5). The contouring of the ROI was performed manually by experienced Radiation Oncologists using the Varian Medical System Eclipse software environment. These features underwent a selection process to minimize redundancy and were combined with clinical data. A CPH model, optimized via the best subset approach and tenfold cross-validation, was then applied. The model's predictive performance was evaluated using the concordance index. All statistical analyses were performed using R programming language, with a significance level (alpha) set at 0.05 for all tests. All procedures included in this application have been approved previously by the institutional review boards (IRB) of the University of Maryland School of Medicine Institutional Review Board.

Pre-processing

It is worth noting that a previous study[44] highlights that radiomics features are sensitive to voxel size. Therefore, maintaining consistent voxel sizes across patients is crucial to obtain accurate and reliable radiomics feature calculations. In this study, CT resolution sizes varied from $0.3 \times 0.3 \times 0.5$ to $1.3 \times 1.3 \times 5$. We resampled all CT images to the resolution of $1 \times 1 \times 1$ mm^3 using the basis spline algorithm (Bspline) to interpolate the HU values in the resampled voxels. Correspondingly, we used the nearest neighbor algorithm to resample the tumor contours to the same resolution. The subsequent procedure, gray level normalization, is critical in improving the comparability and robustness of radiomics features across different settings and variations among patients. As demonstrated in a previous study[44], gray level normalization reduces the variance and enhances the robustness of radiomics features, particularly regarding varying discretization levels. Therefore, normalization was applied to scale the Hounsfield Unit (HU) values to a uniform range across patients. This was achieved by subtracting the mean and dividing the voxel values by the standard deviation. Conventionally, this process yields an ROI with intensities approximately in the range [−3, 3] after removing outliers outside three standard deviations. These resulting values are then further scaled using a normalized scale parameter, such as a value of 100, resulting in an approximate range of [−300, 300]. Finally, the intensities within the ROI were discretized using a unified bin-width of 5, starting from the minimum normalized HU value of 0. In this study, we selected a bin width of 5 to ensure an adequate number of bins (between 1 and 400) for capturing more granular textural information[45]. This

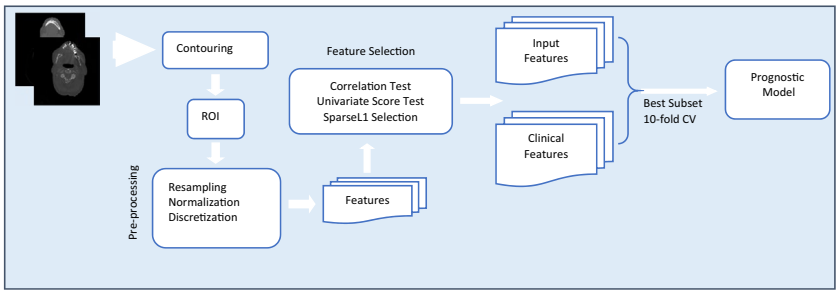

Figure 4. Image feature extraction and outcome prediction workflow.

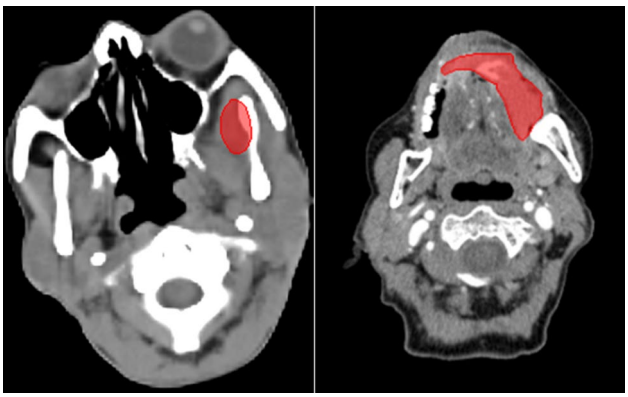

Figure 5. ROI (red) in the left oral cavity.

discretization step assigns a new value to each voxel using the formula floor $(\frac{originintesity}{5}) + 1$. This discretization approach offers the advantage of noise suppression and improved robustness of radiomics features.

Feature extraction

Features extracted from medical images carry the phenotypic characteristics of a tumor, as shown in Fig. 6. A typical medical image features data set involves measurements of tens of thousands of voxel intensities for a single tumor sample; usually, the number of features (p) is $\gg$ sample size (n). In this study, feature extraction was performed in each set of images using the Python library PyRadiomics[46]. Imaging Biomarker Standardization Initiative (IBSI)[47] described features were extracted in six families, including shape-based, first-order statistics, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM)[48], grey level size zone matrix (GLSZM)[49,50], gray level dependence matrix (GLDM)[51] and neighborhood grey tone difference matrix (NGTDM)[39]. Additionally, features are calculated from wavelet, Laplacian of Gaussian (Log), square, square root, logarithm, exponential, and gradient-filtered images, making the total number of features 1092.

Feature selection

Motivation and overview

In the field of time-to-event data analysis, the Lasso-Cox model[52] has gained popularity for its efficiency and simplicity in modeling high-dimensional data[53]. This model allows for simultaneous regression and variable selection, making it a preferred choice in handling high-dimensional data. However, this algorithm often falls short of providing accurate predictive results. The regularization parameter in the Lasso-Cox model emphasizes penalization on large coefficients, leading to potential bias and an overly simplistic model that may underfit the data. To mitigate this issue and improve prediction accuracy, an alternative model called the Elastic-Net was introduced[54]. The Elastic-Net model incorporates both ℓ_1 and ℓ_2 penalties in ordinary least squares estimation and has been extended to handle time-to-event data[55]. However, the increased complexity of the Elastic-Net model poses challenges in searching for optimal hyperparameters, requiring additional validation or resampling techniques. This increases the computational complexity and potentially results in solutions trapped in local optima. Another challenge in radiomics data is high multicollinearity, which occurs when there is a high correlation between two or more measurements in the data. For example, sphericity, minor axis length, max axis length and elongation are variables that exhibit strong multicollinearity in the sense that elongation is merely the inverse of spherical disproportion, and elongation is the ratio of the minor axis length to the max axis length. Multicollinearity can lead to the phenomenon where a variable is not deemed significant when correlated features are also present in the model. While regularization modeling techniques can partially mitigate multicollinearity, they may struggle when dealing with highly correlated variables.

Regression modeling often suffers from bias, overfitting, and numerically unstable estimation in cases where the number of predictors (p) is much larger than the number of samples (n). Therefore, feature selection plays a crucial role in this study. From a practical standpoint, it is desirable to build parsimonious prognostic models that are both effective and easy to use for healthcare professionals. This challenge also extends to Cox regression for time-to-event data. Studies[56–58] have emphasized the importance of developing parsimonious Cox models. We propose a feature selection approach to optimize the Cox Proportional Hazard model, consisting of several steps outlined in Table 4. Firstly, we pruned the highly correlated features.

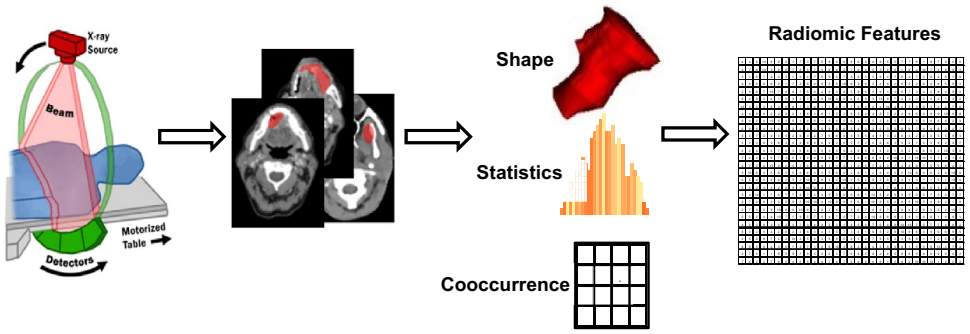

Figure 6. The typical workflow of radiomics feature extraction.

1	Perform Pearson correlation analysis
2	Perform univariate score tests
3	Apply SparseL1 to constrain the degree of freedom

Table 4. Feature selection procedure.

Next, we selected a subset of features based on their relationship with overall survival. Finally, SparseL1[38] recommended a final active subset, effectively eliminating redundant features to ensure computational feasibility for the Best Subset Selection strategy.

Pearson correlation analysis

Given that radiomics features exhibit strong multicollinearity, relying solely on regularization modeling can introduce bias. There is now a substantial body of research on mitigating multicollinearity, such as Principal Component Analysis (PCA), Sparse PCA[59], and Kernel PCA (KPCA)[60,61]. To mitigate multicollinearity, we first employed Pearson's correlation coefficient to detect linear dependencies among radiomics features. The Pearson correlation ranges from − 1 to 1, with a value of 0 indicating no linear correlation. In the medical field, a Pearson's score of 0.7 suggests a moderate agreement between two features based on previous studies[62,63]. Features with a Pearson's score exceeding 0.7 were pruned, resulting in 79 features for subsequent analysis. The result of pruning is illustrated in Fig. 7. Figure 7 uses a color scheme where white represents no correlation, blue represents a perfect negative correlation, and red represents a perfect positive correlation. The left diagonal map illustrates the correlation coefficients prior to the feature selection process, revealing the initial relationships between features. The right diagonal map presents the correlation coefficients after highly correlated features have been removed, demonstrating the outcome of the feature selection process. The left diagonal map initially revealed numerous red and blue shades, indicating strong positive and negative correlations, respectively, among the data. By comparison, the right diagonal map is lighter, indicating these highly correlated features were subsequently removed from the data. To detect correlation among categorical features, we utilized Cramer's V value. Figure 8 is a graphical representation of the Cramer's V between categorical variables in the data, where white (Cramer's V = 0) represents no correlation and red (Cramer's V = 1) represents a perfect correlation. Feature T and Stage were found to be correlated. We dropped the T variable and retained smoke and ETOH, considering that numerous studies have identified smoking and drinking as risk factors. These correlation measures provide insights into the relationships among radiomics features and aid in addressing multicollinearity to ensure more robust and accurate modeling in our study. After pruning, 1013 radiomics features and 1 categorical feature were effectively eliminated, resulting in a total of 85 features for subsequent analysis.

Univariate score test

The univariate Cox score is the most straightforward method for identifying features associated with variability in survival time in time-to-event data analysis[57]. Our focus is on reducing the number of radiomics features. The screening procedure consists of two steps: fitting 79 univariate Cox proportional hazards models for all radiomics features and using the score test statistic to assess the strength of association between each feature and the outcome. We prioritize *p*-values over setting a threshold for the statistic value. Features with a score test *p*-value less than or equal to 0.05 were considered significantly associated with the outcome and retained for subsequent analysis, resulting in 17 features for the next step.

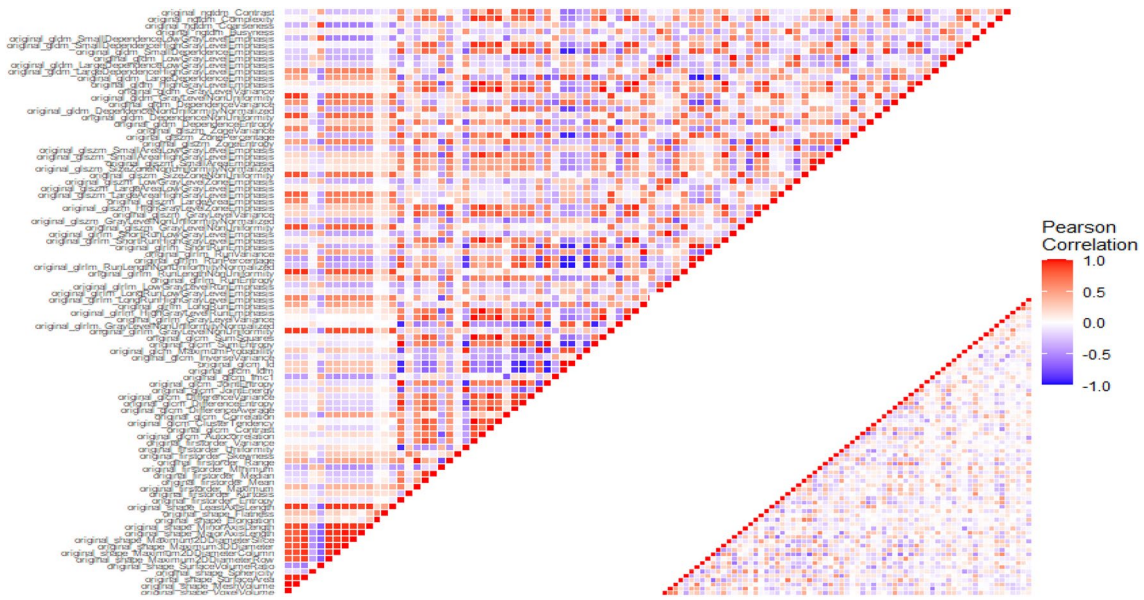

Figure 7. Comparison of Correlation Coefficient Heatmaps: On the left, the diagonal heatmap illustrates the pairwise correlations among features before pruning. On the right, the diagonal heatmap demonstrates the correlations after pruning. The color scale represents the strength of the correlation, with blue indicating negative correlation, red indicating positive correlation, and white representing no correlation. We created the correlation heatmap using the 'ggplot2' package in R version 4.3.1 (Wickham, 2016).

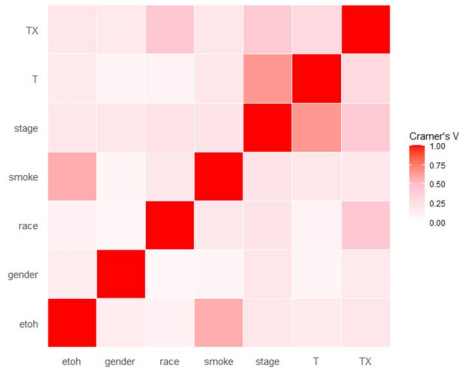

Figure 8. Cramér's V for clinical features.

SparseL1 selection

To fit a CPH model, it is critical to choose an appropriate degree of freedom that balances the complexity and accuracy of the model. According to[64,65], a useful heuristic is to limit the number of predictors used in the fitting should be at most 15% of the events in the training sample. This criterion is corroborated by the simulations study in[66] that the prediction error is lower in CPH models that satisfy this condition. In our study, we observed 66 out of 149 events. Therefore, our target degree of freedom in the final model should be at most 4. Subsequently, we employed the Best Subset Selection (BSS) Modeling strategy in order to identify the best Cox Proportional Hazard (CPH) models based on their Concordance Index. BSS is widely recognized as a highly effective strategy for identifying the best parsimonious model, surpassing other strategies such as stepwise selection, forward selection, backward elimination, and Lasso. However, the computational cost associated with BSS limits its practical usage compared to other techniques. Considering fitting 2, 3, 4 degree-of-freedom CPH with an input data of 17 radiomics features plus 6 clinical features, BSS needs to estimate 41,262 coefficients at least. Exhaustively evaluating all possible subsets is computationally infeasible. Therefore, we need to reduce computational efforts by limiting the number of input variables before BSS. In this study, we employed a variation of Principal Component Analysis (PCA) known as the SparseL1 algorithm[38] to constrain the input data. The SparseL1 algorithm approximates the solution vector v by solving the NP-hard problem:

$$\min_{v,\alpha} |X - \alpha v| + \lambda |v|$$

where $\alpha^{n\times 1}$ represents the representation vector of n observations and $v^{1\times p}$ is a sparse vector in which each coordinate corresponds to a specific radiomics feature. By enforcing regularization, SparseL1 encourages many coordinates to be zero, effectively eliminating redundant features. SparseL1 is less sensitive to outliers compared to PCA, ensuring robust and consistent solutions. Additionally, the sparsity of the subspace can be adjusted using a single parameter λ, which serves as a controller for the number of inputs. Using this algorithm, we were able to reduce the previous 17 radiomics features to 7 features.

Cox modelling and best subset selection

After feature selection, a multivariate Cox proportional hazards model was utilized to model the prognosis for individual patients. The Cox proportional hazards model is a commonly used approach for analyzing time-to-event data and assessing the effects of predictors on survival time. The Cox proportional hazards modeling is concerned with estimating the coefficients in the linear model:

$$\ln \frac{h(t|x_i)}{h_0(t)} = \beta x_i$$

where $h(t|x_i)$ is the cumulative hazard function for subject i with m variables, under the assumption the hazard ratio comparing any two observations remains constant over time. The coefficients β were estimated by solving maximizing the partial likelihood function:

$$\mathcal{L}(\beta) = \prod_{r\in E} \frac{\exp(\beta x)}{\sum_{i\in \overline{E}_r} \exp(\beta x^i)}$$

where E is the set of indices of dead patients and $\overline{E}_r$ is the set of the indices of alive patients at the time t_r. We employ the Concordance index as the primary criterion for Best Subset Selection:

$$\max_{\beta} \sum_{i=1}^{n} P\left(\beta x_i > \beta x_j | T_i < T_j\right) s.t. |\beta|_0 \le k$$

where $|\beta|_0$ is the $\updownarrow_0$-norm of β. Finally, a tenfold cross-validation procedure was used to assess how the selected model will generalize to a new data set. The data was split into 10 folds, ensuring an even distribution of status

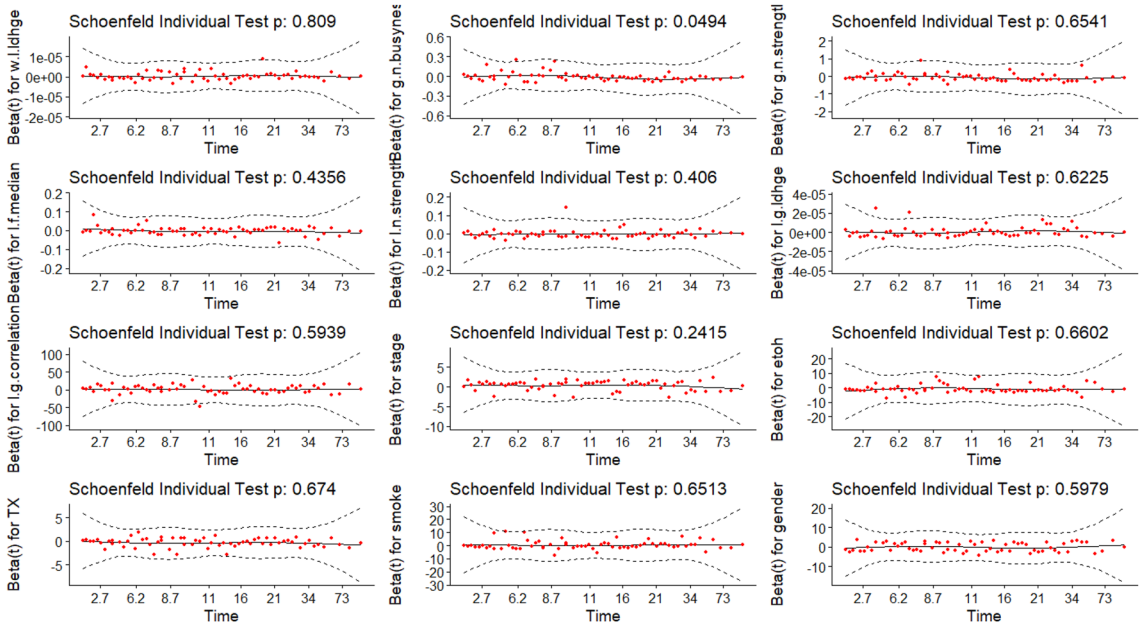

Figure 9. Hazard Proportional Assumption Test. Global Schoenfeld Test p-value = 0.594.

and race within each fold. The training set model, represented by $\widehat{\beta}_{train}$ was used to predict the risk factors of the testing set, and the concordance index was calculated and averaged across all folds.

To validate the assumption, Schoenfeld tests were conducted, and a graphical examination was performed by observing changes in the effect over time. In Fig. 9, the solid line represents a smoothing spline fit to the plot, with the dashed lines depicting the 95% confidence band around the fit. Notably, all splines remain within this band without any discernible pattern over time, indicating no changes in the effects. Additionally, Schoenfeld tests at both individual and global levels do not provide sufficient evidence to reject the proportional hazards assumption. Thus, strong statistical support is obtained for the assumption of hazard proportionality across all effects. CPH models are fitted by considering all combinations of 13 features. The Likelihood Ratio Test, the Score Test, and the number of significant covariates of $2^{13} = 8192$ CPH models are compared with each other.

Ethical approval

The study protocol complied with the Declaration of Helsinki and was approved by the institutional review board of the University of Maryland School of Medicine (approval no. 00007145). All patients provided written informed consent prior to enrollment.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to institutional policy but are available from the corresponding author on reasonable request.

Received: 14 August 2023; Accepted: 21 November 2023
Published online: 08 December 2023

References

1. Licitra, L., Locati, L. & Bossi, P. Head and neck cancer. *Ann. Oncol.* **15**, iv267–iv273 (2004).
2. Hunter, K. D., Parkinson, E. K. & Harrison, P. R. Profiling early head and neck cancer. *Nat. Rev. Cancer* **5**, 127–135 (2005).
3. Bettendorf, O., Piffko, J. & Bankfalvi, A. Prognostic and predictive factors in oral squamous cell cancer: Important tools for planning individual therapy?. *Oral Oncol.* **40**, 110–119 (2004).
4. Leemans, C. R., Braakhuis, B. J. & Brakenhoff, R. H. The molecular biology of head and neck cancer. *Nat. Rev. Cancer* **11**, 9–22 (2011).
5. Fuller, C. *et al.* Adjunctive diagnostic techniques for oral lesions of unknown malignant potential: Systematic review with meta-analysis. *Head Neck* **37**, 755–762 (2015).
6. Menini, M. *et al.* Salivary micro-RNA and oral squamous cell carcinoma: A systematic review. *J. Personal. Med.* **11**, 101 (2021).
7. Al Rawi, N. *et al.* The role of differentially expressed salivary microRNA in oral squamous cell carcinoma. A systematic review. *Arch. Oral Biol.* **125**, 105108 (2021).
8. D'Souza, W. & Kumar, A. microRNAs in oral cancer: Moving from bench to bed as next generation medicine. *Oral Oncol.* **111**, 104916 (2020).
9. Liu, C.-J., Lin, S.-C., Yang, C.-C., Cheng, H.-W. & Chang, K.-W. Exploiting salivary miR-31 as a clinical biomarker of oral squamous cell carcinoma. *Head Neck* **34**, 219–224 (2012).
10. Momen-Heravi, F., Trachtenberg, A., Kuo, W. & Cheng, Y. Genomewide study of salivary microRNAs for detection of oral cancer. *J. Dent. Res.* **93**, 86S–93S (2014).
11. Nagler, R., Bahar, G., Shpitzer, T. & Feinmesser, R. Concomitant analysis of salivary tumor markers—a new diagnostic tool for oral cancer. *Clin. Cancer Res.* **12**, 3979–3984 (2006).

12. Balan, J. J. *et al.* Analysis of tumor marker CA 125 in saliva of normal and oral squamous cell carcinoma patients: A comparative study. *J. Contemp. Dent. Pract.* **13**, 671–5 (2012).
13. Gillison, M. *et al.* Analysis of the effect of p16 and tobacco pack-years (py) on overall (OS) and progression-free survival (PFS) for patients with oropharynx cancer (OPC) in radiation therapy oncology group (RTOG) protocol 9003. *J. Clin. Oncol.* **28**, 5510–5510 (2010).
14. Grandis, J. R. & Tweardy, D. J. TGF-α and EGFR in head and neck cancer. *J. Cell. Biochem.* **53**, 188–191 (1993).
15. Poeta, M. L. *et al.* TP53 mutations and survival in squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **357**, 2552–2561 (2007).
16. Michalides, R. *et al.* Overexpression of cyclin D1 correlates with recurrence in a group of forty-seven operable squamous cell carcinomas of the head and neck. *Cancer research* **55**, 975–978 (1995).
17. Redston, M. *et al.* Abnormal TP53 predicts risk of progression in patients with Barrett's esophagus regardless of a diagnosis of dysplasia. *Gastroenterology* **162**, 468–481 (2022).
18. Flaherty, K. T. *et al.* The molecular analysis for therapy choice (NCI-MATCH) trial: Lessons for genomic trial design. *JNCI J. Natl. Cancer Inst.* **112**, 1021–1029 (2020).
19. Blucher, A. S., Mills, G. B. & Tsang, Y. H. Precision oncology for breast cancer through clinical trials. *Clin. Exp. Metastasis* **39**, 71–78 (2022).
20. Prudente, S., Dallapiccola, B., Pellegrini, F., Doria, A. & Trischitta, V. Genetic prediction of common diseases. Still no help for the clinical diabetologist!. *Nutr. Metab. Cardiovasc. Dis.* **22**, 929–936 (2012).
21. Galbán, C. J. *et al.* Computed tomography–based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. *Nat. Med.* **18**, 1711–1715 (2012).
22. Jiang, T., Kambadakone, A., Kulkarni, N. M., Zhu, A. X. & Sahani, D. V. Monitoring response to antiangiogenic treatment and predicting outcomes in advanced hepatocellular carcinoma using image biomarkers, CT perfusion, tumor density, and tumor size (RECIST). *Invest. Radiol.* **47**, 11–17 (2012).
23. Zhang, Z. *et al.* Patient-specific deep learning model to enhance 4D-CBCT image for radiomics analysis. *Phys. Med. Biol.* **67**(8), 085003 (2022).
24. Zhang, Z. *et al.* 4D radiomics: Impact of 4D-CBCT image quality on radiomic analysis. *Phys. Med. Biol.* **66**(4), 045023. https://doi.org/10.1088/1361-6560/abd668 (2021).
25. O'connor, J. P. *et al.* Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol.* **14**, 169–186 (2017).
26. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **5**, 1–9 (2018).
27. Nie, K. *et al.* Incremental value of radiomics in 5-year overall survival prediction for stage II–III rectal cancer. *Front. Oncol.* **12**, 779030 (2022).
28. Le, V.-H., Kha, Q.-H., Hung, T. N. K. & Le, N. Q. K. Risk score generated from CT-based radiomics signatures for overall survival prediction in non-small cell lung cancer. *Cancers* **13**, 3616 (2021).
29. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
30. Watson, K. S. *et al.* Lung cancer screening and epigenetics in African Americans: The role of the socioecological framework. *Front. Oncol.* **9**, 87 (2019).
31. Andrearczyk, V., Oreiller, V., Hatt, M. & Depeursinge, A. *Head and Neck Tumor Segmentation and Outcome Prediction* (Springer, UK, 2022).
32. Corti, A. *et al.* MRI-based radiomic prognostic signature for locally advanced oral cavity squamous cell carcinoma: Development, testing and comparison with genomic prognostic signatures. *Biomarker Res.* **11**, 69 (2023).
33. Mossinelli, C. *et al.* The role of radiomics in tongue cancer: A new tool for prognosis prediction. *Head Neck* **45**, 849–861 (2023).
34. Mes, S. W. *et al.* Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures. *Eur. Radiol.* **30**, 6311–6321 (2020).
35. Wang, F. *et al.* Magnetic resonance imaging-based radiomics features associated with depth of invasion predicted lymph node metastasis and prognosis in tongue cancer. *J. Magn. Reson. Imaging* **56**, 196–209 (2022).
36. Liu, Z. *et al.* Radiomics-based prediction of survival in patients with head and neck squamous cell carcinoma based on pre-and post-treatment 18F-PET/CT. *Aging (Albany NY)* **12**, 14593 (2020).
37. Dibble, E. H. *et al.* 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.* **53**, 709–715 (2012).
38. Ling, X. & Brooks, J. P. L1-norm regularized L1-norm best-fit line problem. *arXiv preprint* arXiv:2010.04684 *(2020).*
39. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Trans. Syst. Man Cybern.* **19**, 1264–1274 (1989).
40. Haralick, R. M., Shanmugam, K. & Dinstein, I. H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 610–621 (1973).
41. Guo, R. *et al.* Magnetic susceptibility and R2*-based texture analysis for evaluating liver fibrosis in chronic liver disease. *Eur. J. Radiol.* 111155 (2023).
42. Granata, V. *et al.* Preliminary report on computed tomography radiomics features as biomarkers to immunotherapy selection in lung adenocarcinoma patients. *Cancers* **13**, 3992 (2021).
43. Ang, K. K. *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *N. Engl. J. Med.* **363**, 24–35 (2010).
44. Shafiq-ul-Hassan, M. *et al.* Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys.* **44**, 1050–1062 (2017).
45. Larue, R. T. *et al.* Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: A comprehensive phantom study. *Acta Oncol.* **56**, 1544–1553 (2017).
46. van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, E104–E107 (2017).
47. Zwanenburg, A. *et al.* The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
48. Galloway, M. M. Texture analysis using gray level run lengths. *Comput. Graphics Image Process.* **4**, 172–179 (1975).
49. Thibault, G., Fertil, B., Navarro, C. *et al.* Texture indexes and gray level size zone matrix: Application to cell nuclei classification in proceedings of the pattern recognition and information processing 2009. in *International Conference on Pattern Recognition and Information Processing (PRIP'09)* 140–145.
50. Thibault, G., Angulo, J. & Meyer, F. Advanced statistical matrices for texture characterization: Application to cell classification. *IEEE Trans. Biomed. Eng.* **61**, 630–637 (2013).
51. Sun, C. & Wee, W. G. Neighboring gray level dependence matrix for texture classification. *Comput. Vis. Graphics Image Process.* **23**, 341–352 (1983).
52. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996).
53. Bertsimas, D., King, A. & Mazumder, R. Best subset selection via a modern optimization lens. (2016).
54. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
55. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1 (2011).

56. Salerno, S. & Li, Y. High-dimensional survival analysis: Methods and applications. *Ann. Rev. Stat. Its Appl.* **10**, 25–49 (2023).
57. Witten, D. M. & Tibshirani, R. Survival analysis with high-dimensional covariates. *Stat. Methods Med. Res.* **19**, 29–51 (2010).
58. Lang, M. *et al.* Automatic model selection for high-dimensional survival analysis. *J. Stat. Comput. Simul.* **85**, 62–76 (2015).
59. Brooks, J. P., Dulá, J. H. & Boone, E. L. A pure L1-norm principal component analysis. *Comput. Stat. Data Anal.* **61**, 83–98 (2013).
60. Bui, A. T., Im, J.-K., Apley, D. W. & Runger, G. C. Projection-free kernel principal component analysis for denoising. *Neurocomputing* **357**, 163–176 (2019).
61. Ling, X., Bui, A. & Brooks, P. Kernel l1-norm principal component analysis for denoising. *Optim. Lett.* https://doi.org/10.1007/s11590-023-02051-3 (2023).
62. Akoglu, H. User's guide to correlation coefficients. *Turk. J. Emerg. Med.* **18**, 91–93 (2018).
63. Chan, Y. Biostatistics 104: Correlational analysis. *Singap. Med. J.* **44**, 614–619 (2003).
64. Harrell, F. E. Jr., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
65. Harrell, F. E. Regression modeling strategies. *Bios* **330**, 14 (2017).
66. Smith, L. R., Harrell, F. & Muhlbaier, L. H. Problems and potentials in modeling survival. *Medical Effectiveness Research Data Methods (Summary Report), AHCPR Pub* 151–159 (1992).

Acknowledgements

Xiao Ling, and Lei Ren was supported by the National Institutes of health (NIH) grants R01EB032680, R01CA279013, R01EB028324, and U54CA273956. Daria A. Gaykalova was supported by a Research Scholarship Grant, RSG-21-020-01-MPC from the American Cancer Society, and R01DE027809 from the National Institute of Health.

Author contributions

X.L. was responsible for the Methodology, Software, Formal analysis, and the Writing of the Original draft. G.S.A., J.M., and J.C. contributed to Resources, Writing, Reviewing, and Editing. L.S. assisted with Resources, Data curation, Writing, Reviewing and Editing. R.M. and D.A.G. conducted the Investigation, with D.A.G. also handling Validation, Conceptualization, and Writing, Reviewing and Editing. L.R. provided Resources, Conceptualization, Supervision, Project administration, Writing, Reviewing and Editing, and Funding acquisition. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.M., D.A.G. or L.R.

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.