Radiomics-based model for prediction of TGF-β1 expression in head and neck squamous cell carcinoma

Kai Qin  
Departments of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

Yi Cheng  
Departments of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

Li Li  
Departments of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030

Chengxia Liu  
Departments of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030

Feng Yang  
Departments of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

Jie Rao  
Departments of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030

Qianxia Li (liqianx110@163.com)  
Departments of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

Research Article

Keywords: TCGA, TCIA, head and neck squamous cell carcinoma, Radiomics, TGF-β1 expression, prediction model

Posted Date: May 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2920092/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Purpose: TGF-β1 is an immunosuppressive gene that regulates a variety of activities relating to immune responses. However, the association between TGF-β1 expression and the survival rate of HNSCC patients remains unclear. This study is to explore that whether there is a connection between TGF-β1 expression and patients’ survival in HNSCC, and whether the TGF-β1 expression in HNSCC patients can be non-invasively predicted by CT-Based Radiomics.

Materials and Methods: Transcriptional profiling data and clinical information were obtained from TCGA database, and then grouped basing on Cutoff value of TGF-β1 expression. 139 HNSCC patients (112 for training and 27 for validation) were selected basing on the completeness of enhanced arterial phase CT images. 3D Slicer software is used for image segmentation, and PyRadiomics package for extraction of radiomic features. The optimal features for establishing the corresponding gradient enhancement prediction models were obtained using mRMR_RFE algorithm and Repeat_LASSO algorithm. Conclusively, comprehensive performances of two models, including diagnostic efficacy, calibration and clinical practicability, were compared.

Results: 483 patients were classified into two groups (high expression (n=333) and low expression (n=150)) basing on the cut-off of TGF-β1 expression (5.208), and then used for survival analysis. Kaplan-Meier curve showed that TGF-β1, as an independent risk factor, significantly decreased patients’ survival (p=0.001). For construction of grid enhancement prediction models, we respectively obtained two features-glrlm and ngtdm-and three radiation features-glrlm, first order_10percentile and gldm-using mRMR_RFE algorithm and Repeat_LASSO algorithm. The two established models showed strong predictive potentials in both training cohort and validation cohort. In training set, ROC curve shows that AUC of mRMR_RFE_GBM model is 0.911 and Repeat_LASSO_GBM model is 0.733. And it is statistically significant that AUC of mRMR_RFE_GBM model (0.911) is higher than Repeat_LASSO_GBM model (0.733). Likewise, in validation set, AUC of mRMR_RFE_GBM model is 0.849 and Repeat_LASSO_GBM model is 0.72. And the difference between two models in AUC value is not statistically significant (p=0.212). In addition, calibration curve shows high consistency between the predictive result and real value, and DCA diagram shows its good clinical practicability. Moreover, whether in training set or in validation set, there is no statistical difference in AUC values between mRMR_RFE_GBM model and LASSO_GBM model (p=0.443, p=0.912), indicating that the two models both fit well.

Conclusion: TGF-β1 is an independent risk factor and significantly associated with poor prognosis. mRMR_RFE_GBM model and Repeat_LASSO_GBM model based on CT-Based Radiomics features can effectively and non-invasively predict TGF-β1 expression in HNSCC. Considering the efficacy of prediction, mRMR_RFE_GBM model is better for clinical application.

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC), which initiates from mucous epithelium of mouth, pharynx and larynx, is a malignant tumor with high incidence [1–2]. Currently, standard treatment for HNSCC is combination of surgery and chemoradiotherapy in clinics, but the survival rate of patients within 5 years is unsatisfied yet and only up to 34% [1, 3]. Classical prognostic indicators of HNSCC such as human papillomavirus (HPV) status, p16, PD-L1 expressions and clinical stages [4] can no longer meet the clinical needs of precision medicine. Thus, it is necessary to further explore new prognostic indicators for individualized stratified treatment of patients.

In fact, in addition to traditional treatments, immunotherapy has been widely used for HNSCC. However, patients really benefited from this treatment are still low, only 15–20% [5]. Therefore, it is essential to explore the mechanisms of immune resistance in immune microenvironment and provide some evidence for HNSCC treatment. TGF-β family is an important immunosuppressive gene in HNSCC and associated with poor prognosis [6]. The dysfunction of TGF-β signal promotes tumor progression and metastasis via regulating proliferation of epithelial cells, inhibiting cell apoptosis and inducing genomic instability of tumor cells [7]. In HNSCC patients, PD-L1 and TGF-β are usually highly expressed. PD-L1 can inhibit activation of lymphocytes and induce their apoptosis by binding with PD-1 receptor on the surface of lymphocytes, thus leading to immune escape of tumor cells. TGF-β not only drives tumorigenesis but also mediates the primary drug resistance against PD-L1 monoclonal antibody. Therefore, blocking dual signals of PD-L1 and TGF-β simultaneously can produce a synergistic anti-tumor effect and improve the response rate of PD-1 inhibitors [8]. Although as a member of TGF-β family, TGF-β1 is highly expressed in the majority of HNSCCs [9], whether it is an independent factor that could predict the survival of HNSCC patients remains unclear.

Radiomics is a high-throughput ‘image sequencing’ that can acquire a large number of image parameters by non-invasive, dynamic and quantitative detection on tumor features [10]. Radiomics have been considered as an effective technology in HNSCC to instruct early diagnosis and classification, evaluate tumor heterogeneity, and identify cell constituents in tumor microenvironment [11]. In addition, radiomics has great potential in overcoming limitations of traditional tumor markers [12] because it provides complete three-dimensional information about tumors and non-invasive repetitive analysis can be carried out based on follow-up images. Hence, in this study, we determined the prognostic value of TGF-β1 in HNSCC by CT-Based Radiomics and explored its potential molecular mechanism and the relationship with constituents of immune cells via integrated bioinformatics analysis.

MATERIALS & METHODS

Sources of data & images

Transcriptome profiling data and clinical information of 528 HNSCC patients were collected from TCGA database (https://portal.gdc.cancer.gov/ ). 211 arterial enhanced CT images are obtained from TCIA-HNSC (https://www.cancerimagingarchive.net/ ). All data and images are anonymous and public, so they are exempted from ethics and informed consent after being approved by the unit ethics committee. TCGA data is applied for prognosis analysis, and TCIA images are applied for identification of radiomic features and model establishment. The criteria for samples selection: preoperative samples, complete clinical data
Analysis of TGF-β1 expression and patients’ survival

R package ‘survminer’ is used for acquiring the cut-off of TGF-β1 expression. RNA seq data in FRKM format from UCSC XENA (https://xenabrowser.net/datapages/) was processed by Toil [13]. TGF-β1 expression in HNSC tissue and normal tissue was compared by R package ‘ggplot2’. Patients’ survival between two groups was estimated by Kaplan-Meier analysis. Prognostic value of variables, including sex, age, HPV status, nerve invasion, grade, TNM stage, chemoradiotherapy, and primary tumor site and gene expression was evaluated by univariate and multivariate Cox regression analysis. The link of TGF-β1 expression and the prognosis of patients in different subgroups of covariates was analyzed using univariate cox regression method. The interaction between TGF-β1 and other covariates was analyzed by likelihood ratio test. In addition, clinical features, infiltrated immune cells and enriched genes or pathways in different TGF-β1 expression groups were further analyzed.

Cohort design, Tumor segmentation & radiomic feature determination

The scheme is shown in Fig. 1B. 139 images obtained form TCIA database were randomly grouped (training or validation) according to the ratio of 8:2. Images segmentation was performed by 3D Slicer software (version 4.10.2; https://www.slicer.org/). An imaging physician (10 years’ working experience in radiology department) can independently delineate the tumor focus layer by layer manually even is blind to the patient’s clinical data and diagnosis results. Pyradiomics (https://pyradiomics.readthedocs.io/en/latest/index.html) which contains 107 features is used for radiomics analysis, and a total of four features including volume and shape features, texture features and wavelet transform features were obtained. The eigenvalues in training set were standardized by z-score using the preProcess function in R package ‘caret’, then obtained mean and standard deviation were used to standardize the characteristic values in validation set.

Evaluation of consistency

The consistency of image omics features was evaluated by intraclass correlation efficiency (ICC) based on two doctors’ description VOI [14, 15]. The former described the cases and 20 samples were randomly selected by “random number table”. Another (8 years working experience in radiology department) described the cases again and determined radiomics features for consistency evaluation. Generally, ICC ≥ 0.8 is considered as good consistency, 0.51 ~ 0.79 is medium, and less than 0.50 is poor.

Determination of radiomic features

Over-tted model with redundant features will affect the result for predicting TGF-β1 expression. Thus, in order to eliminate redundant features or select optimal features, maximum correlation-redundancy (mRMR) algorithm and recursive feature elimination (RFE) algorithm are used [16]. Firstly, features are ranked basing on their importance by using maximum correlation minimum redundancy algorithm (“mRMR” package in R language), that maximizing mutual information (MI) with classification labels and minimizing MI with other features. Then, the features that contribute least to the model are eliminated repeatedly using recursive feature elimination (RFE) algorithm until remaining features reaches required number. Finally, 20 features obtained by mRMR algorithm and 20 features selected by RFE algorithm are used for identification of paired and correlated features.

Establishment of mRMR_RFE_GBM model

Gradient Boosting Machine (GBM) is to rank the predicted probability scores using the “gbm” method. GBM algorithm is to adjust the super-parameters for model establishment using a set of weak classifiers (usually decision trees) and cross-validation method. mRMR_RFE_GBM model was established using GBM algorithm and basing on selected radiomic features [17].

Establishment of Repeat_LASSO_GBM model

Repeat lasso (Least Absolute Shrinkage and Selection Operator) is to perform 1000 times of Lasso regression on the radiomic features using the “glmnet” package, and then select top N features with the highest frequency as the final subset [18]. In this study, repeat_LASSO_GBM model was established using GBM algorithm and basing on the radiomic features with top two frequencies.

Evaluation of mRMR_RFE_GBM model and Repeat_LASSO_GBM model

The predictive efficacy of mRMR_RFE_GBM model and Repeat_LASSO_GBM model is respectively evaluated by ROC curves. The calibration degree of the image omics prediction model was evaluated by calibration curve using Hosmer-Lemeshow goodness-of-fit test. The comprehensive performance of image ensemble prediction model is determined by Brier score (measures package). Furthermore, TGF-β1 expression respectively predicted by mRMR_RFE_GBM model and Repeat_LASSO_GBM model were compared using Wilcoxon test. The AUC value in training and validation set were compared using Delong test.

Statistical analysis
R packages and SPSS software are used for statistical analysis. All data were presented as means ± standard deviations or as medians and interquartile ranges. The differences between training and validation set in gender, age and other baseline characteristics are evaluated using independent sample t test, Wilcoxon test and χ² test. Univariate and multivariate Cox regression analysis are utilized for risk factor evaluation. R’ survival’ is used for Kaplan-Meier survival analysis. All hypothesis tests were bilateral, and p < 0.05 is significantly different.

RESULTS

Clinical features

483 HNSCC patients were divided into two groups based on cut-off of TGF-β1 expression (5.208). The clinical information of patients is shown in Table 1. There were no significant differences in age, sex, histological grade, TNM staging and chemoradiotherapy therapy among two groups (P = 0.847). There were significant differences in HPV status, nerve invasion and primary tumor location between two groups.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Total (n = 483)</th>
<th>Low (n = 150)</th>
<th>High (n = 333)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.343</td>
</tr>
<tr>
<td>Female</td>
<td>128 (27)</td>
<td>35 (23)</td>
<td>93 (28)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>355 (73)</td>
<td>115 (77)</td>
<td>240 (72)</td>
<td></td>
</tr>
<tr>
<td>Age, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.847</td>
</tr>
<tr>
<td>~59</td>
<td>211 (44)</td>
<td>67 (45)</td>
<td>144 (43)</td>
<td></td>
</tr>
<tr>
<td>60~</td>
<td>272 (56)</td>
<td>83 (55)</td>
<td>189 (57)</td>
<td></td>
</tr>
<tr>
<td>HPV_status, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Negative</td>
<td>68 (14)</td>
<td>17 (11)</td>
<td>51 (15)</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>30 (6)</td>
<td>19 (13)</td>
<td>11 (3)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>385 (80)</td>
<td>114 (76)</td>
<td>271 (81)</td>
<td></td>
</tr>
<tr>
<td>Perineural_invasion, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>NO</td>
<td>181 (37)</td>
<td>64 (43)</td>
<td>117 (35)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>141 (29)</td>
<td>56 (37)</td>
<td>85 (26)</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>161 (33)</td>
<td>30 (20)</td>
<td>131 (39)</td>
<td></td>
</tr>
<tr>
<td>Grade, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.011</td>
</tr>
<tr>
<td>G1/G2</td>
<td>348 (72)</td>
<td>96 (64)</td>
<td>252 (76)</td>
<td></td>
</tr>
<tr>
<td>G3/G4/GX</td>
<td>135 (28)</td>
<td>54 (36)</td>
<td>81 (24)</td>
<td></td>
</tr>
<tr>
<td>T-stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.328</td>
</tr>
<tr>
<td>T1/T2</td>
<td>173 (36)</td>
<td>59 (39)</td>
<td>114 (34)</td>
<td></td>
</tr>
<tr>
<td>T3/T4/TX/Unknown</td>
<td>310 (64)</td>
<td>91 (61)</td>
<td>219 (66)</td>
<td></td>
</tr>
<tr>
<td>N-stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.476</td>
</tr>
<tr>
<td>NO</td>
<td>164 (34)</td>
<td>47 (31)</td>
<td>117 (35)</td>
<td></td>
</tr>
<tr>
<td>N1/N2/N3/NX/Unknown</td>
<td>319 (66)</td>
<td>103 (69)</td>
<td>216 (65)</td>
<td></td>
</tr>
<tr>
<td>M-stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.257</td>
</tr>
<tr>
<td>M0</td>
<td>174 (36)</td>
<td>48 (32)</td>
<td>126 (38)</td>
<td></td>
</tr>
<tr>
<td>M1/MX/Unknown</td>
<td>309 (64)</td>
<td>102 (68)</td>
<td>207 (62)</td>
<td></td>
</tr>
<tr>
<td>Chemotherapy, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.251</td>
</tr>
<tr>
<td>NO</td>
<td>322 (67)</td>
<td>94 (63)</td>
<td>228 (68)</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>161 (33)</td>
<td>56 (37)</td>
<td>105 (32)</td>
<td></td>
</tr>
<tr>
<td>Radiotherapy, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.309</td>
</tr>
<tr>
<td>NO</td>
<td>234 (48)</td>
<td>67 (45)</td>
<td>167 (50)</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>249 (52)</td>
<td>83 (55)</td>
<td>166 (50)</td>
<td></td>
</tr>
<tr>
<td>Primary_tumor_site, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Larynx</td>
<td>109 (23)</td>
<td>40 (27)</td>
<td>69 (21)</td>
<td></td>
</tr>
<tr>
<td>Oral Cavity</td>
<td>297 (61)</td>
<td>69 (46)</td>
<td>228 (68)</td>
<td></td>
</tr>
<tr>
<td>Oropharynx/Hypopharynx</td>
<td>77 (16)</td>
<td>41 (27)</td>
<td>36 (11)</td>
<td></td>
</tr>
</tbody>
</table>

HNSCC: Head and Neck squamous cell carcinoma, TGF-β1: The Transforming growth factor-β1, HPV: human papilloma virus.

The correlation analysis of high TGF-β1 expression and poor prognosis

The Toil process transformation analysis showed that patients have higher TGF-β1 expression than normal tissue, and the median expression difference was 1.784 (P < 0.001). The correlation of TGF-β1 expression and survival of patients was analyzed using Log-Rank test. Kaplan-Meier survival curve showed that the median survival time of patients in low expression group was 69.43 months, and patients in high expression group is 46.46 months. These data demonstrated that high TGF-β1 expression is significantly correlated with poor prognosis (p < 0.01).
The potential prognostic factors were identified using the methods of univariate and multivariate Cox regression. The results of univariate analysis showed that TGF-β1 is a risk factor for overall survival (HR = 1.876, 95%CI = 1.335–2.635, P < 0.001), other hazardous variables include nerve invasion (HR = 2.207, 95%CI 1.552–3.139, P < 0.001), T stage (HR = 1.842, 95%CI 1.348–2.516, P < 0.001), sex (HR = 0.001) 95%CI is 1.362–2.623, P < 0.001) and radiotherapy (HR = 0.477, 95%CI is 0.36–0.633, P < 0.001) (Table 2). The results of multivariate analysis showed that TGF-β1 (HR = 1.773, 95%CI = 1.231–2.555, P = 0.002), nerve invasion (HR = 1.676, 95%CI = 1.155–2.433, P = 0.007), t N stage (HR = 1.943, 95%CI 1.37–2.755, P < 0.001) and radiotherapy (HR = 0.364, 95%CI 0.265–0.501, P < 0.001) all are independent risk factors for overall survival (Table 2).

### Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Univariate analysis</th>
<th>Multivariate analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR (95% CI)</td>
<td>P value</td>
</tr>
<tr>
<td>TGFβ1: High vs. Low</td>
<td>1.876(1.335–2.635)</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>Gender: Male vs. Female</td>
<td>0.737(0.549–0.989)</td>
<td>0.042*</td>
</tr>
<tr>
<td>Age: 60 ~ vs. ~59</td>
<td>1.262(0.952–1.674)</td>
<td>0.106</td>
</tr>
<tr>
<td>HPV_status: Positive vs. Negative</td>
<td>0.349(0.105–1.161)</td>
<td>0.086</td>
</tr>
<tr>
<td>HPV_status: Unknown vs. Negative</td>
<td>1.117(0.726–1.717)</td>
<td>0.615</td>
</tr>
<tr>
<td>Perineural_invasion: Unknown vs. NO</td>
<td>1.836(1.263–2.67)</td>
<td>0.001*</td>
</tr>
<tr>
<td>Perineural_invasion: YES vs. NO</td>
<td>2.207(1.552–3.139)</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>Grade: G3/G4/GX vs. G1/G2</td>
<td>0.913(0.673–1.24)</td>
<td>0.561</td>
</tr>
<tr>
<td>T_stage: T3/T4/TX/Unknown vs. T1/T2</td>
<td>1.842(1.348–2.516)</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>N_stage: N1/N2/N3/NX/Unknown vs. NO</td>
<td>1.897(1.372–2.623)</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>M_satge: M1/MX/Unknown vs. M0</td>
<td>1.345(0.989–1.828)</td>
<td>0.059</td>
</tr>
<tr>
<td>Chemotherapy: YES vs. NO</td>
<td>0.989(0.735–1.33)</td>
<td>0.939</td>
</tr>
<tr>
<td>Radiotherapy: YES vs. NO</td>
<td>0.477(0.36–0.633)</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>Primary_tumor_site: Oral Cavity vs. Larynx</td>
<td>1.146(0.822–1.598)</td>
<td>0.422</td>
</tr>
<tr>
<td>Primary_tumor_site: Oropharynx/Hypopharynx vs. Larynx</td>
<td>0.832(0.504–1.373)</td>
<td>0.471</td>
</tr>
</tbody>
</table>

*reflected the significant difference with the P value < 0.05.

The comparison analysis in high and low TGF-β1 expression groups

In subgroups’ analysis, TGF-β1 is a risk factor in cohort less than 60 years old (HR = 1.891, 95%CI = 1.111–3.218, p = 0.019) as well as over 60 years old (HR = 1.878, 95%CI = 1.235–2.865, p = 0.003). And no interaction was observed between TGF-β1 expression and age (p = 0.97), as well as TGF-β1, different HPV status, nerve invasion and primary tumor site subgroup. In addition, the correlation between TGF-β1 expression and clinical features was analyzed by Spearman grade correlation coefficient. The heatmap showed that TGF-β1 was significantly correlated with tumor grade and nerve invasion (p < 0.01). The violin chart showed that the number of infiltrated CD8 T cells, naive B cells and M0 macrophages were significantly decreased in high TGF-β1 expression group (p < 0.001). GO enrichment analysis suggested that genes which differs in high/low TGF-β1 expression groups were significantly enriched in pathways relating to DNA-binding transcription factor binding, GTP enzyme binding and transcription auxiliary regulator activity. Likewise, KEGG enrichment analysis indicated that these genes were significantly enriched in tumor necrosis factor signaling pathway and others related to cell cycle or apoptosis.

The comparison of clinical characteristics and consistency in HNSCC cohort

There is no difference in clinical characteristics between training and validation set (P > 0.05) (Table 3). The ICC data showed that there are 96 radiomics features beyond 0.8, which approximately accounts for 89.7% of all features, and the median ICC value is 0.926. The ICC value of radiomics features which identified by two methods are all above 0.8 (Table S1), which shows good consistency in HNSCC cohort.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Total (n = 139)</th>
<th>Train (n = 112)</th>
<th>Validation (n = 27)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGF-β1, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.983</td>
</tr>
<tr>
<td>Low</td>
<td>44 (32)</td>
<td>36 (32)</td>
<td>8 (30)</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>95 (68)</td>
<td>76 (68)</td>
<td>19 (70)</td>
<td></td>
</tr>
<tr>
<td>Age, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.207</td>
</tr>
<tr>
<td>~59</td>
<td>64 (46)</td>
<td>55 (49)</td>
<td>9 (33)</td>
<td></td>
</tr>
<tr>
<td>60~</td>
<td>75 (54)</td>
<td>57 (51)</td>
<td>18 (67)</td>
<td></td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.582</td>
</tr>
<tr>
<td>Female</td>
<td>34 (24)</td>
<td>29 (26)</td>
<td>5 (19)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>105 (76)</td>
<td>83 (74)</td>
<td>22 (81)</td>
<td></td>
</tr>
<tr>
<td>HPV_status, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.735</td>
</tr>
<tr>
<td>Negative</td>
<td>15 (11)</td>
<td>13 (12)</td>
<td>2 (7)</td>
<td></td>
</tr>
<tr>
<td>Positive/Unknown</td>
<td>124 (89)</td>
<td>99 (88)</td>
<td>25 (93)</td>
<td></td>
</tr>
<tr>
<td>Grade, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.439</td>
</tr>
<tr>
<td>G1/G2</td>
<td>97 (70)</td>
<td>76 (68)</td>
<td>21 (78)</td>
<td></td>
</tr>
<tr>
<td>G3/G4/GX</td>
<td>42 (30)</td>
<td>36 (32)</td>
<td>6 (22)</td>
<td></td>
</tr>
<tr>
<td>T_stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.759</td>
</tr>
<tr>
<td>T1/T2</td>
<td>42 (30)</td>
<td>35 (31)</td>
<td>7 (26)</td>
<td></td>
</tr>
<tr>
<td>T3/T4/TX/Unknown</td>
<td>97 (70)</td>
<td>77 (69)</td>
<td>20 (74)</td>
<td></td>
</tr>
<tr>
<td>N_stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.996</td>
</tr>
<tr>
<td>N0</td>
<td>54 (39)</td>
<td>43 (38)</td>
<td>11 (41)</td>
<td></td>
</tr>
<tr>
<td>N1/N2/N3/NX/Unknown</td>
<td>85 (61)</td>
<td>69 (62)</td>
<td>16 (59)</td>
<td></td>
</tr>
<tr>
<td>M_stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.77</td>
</tr>
<tr>
<td>M0</td>
<td>66 (47)</td>
<td>52 (46)</td>
<td>14 (45)</td>
<td></td>
</tr>
<tr>
<td>M1/MX/Unknown</td>
<td>73 (53)</td>
<td>60 (54)</td>
<td>13 (48)</td>
<td></td>
</tr>
<tr>
<td>Radiotherapy, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.464</td>
</tr>
<tr>
<td>NO</td>
<td>68 (49)</td>
<td>57 (51)</td>
<td>11 (41)</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>71 (51)</td>
<td>55 (49)</td>
<td>16 (59)</td>
<td></td>
</tr>
<tr>
<td>Primary_tumor_site, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.061</td>
</tr>
<tr>
<td>Larynx</td>
<td>34 (24)</td>
<td>24 (21)</td>
<td>10 (37)</td>
<td></td>
</tr>
<tr>
<td>Oral Cavity</td>
<td>84 (60)</td>
<td>73 (65)</td>
<td>11 (41)</td>
<td></td>
</tr>
<tr>
<td>Oropharynx/Hypopharynx</td>
<td>21 (15)</td>
<td>15 (13)</td>
<td>6 (22)</td>
<td></td>
</tr>
<tr>
<td>Chemotherapy, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.945</td>
</tr>
<tr>
<td>NO</td>
<td>96 (69)</td>
<td>78 (70)</td>
<td>18 (67)</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>43 (31)</td>
<td>34 (30)</td>
<td>9 (33)</td>
<td></td>
</tr>
<tr>
<td>Perineural_invasion, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.286</td>
</tr>
<tr>
<td>NO</td>
<td>48 (35)</td>
<td>36 (32)</td>
<td>12 (44)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>49 (35)</td>
<td>39 (35)</td>
<td>10 (37)</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>42 (30)</td>
<td>37 (33)</td>
<td>5 (19)</td>
<td></td>
</tr>
<tr>
<td>OS, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>0.249</td>
</tr>
<tr>
<td>Alive</td>
<td>88 (63)</td>
<td>74 (66)</td>
<td>14 (52)</td>
<td></td>
</tr>
<tr>
<td>Dead</td>
<td>51 (37)</td>
<td>38 (34)</td>
<td>13 (48)</td>
<td></td>
</tr>
</tbody>
</table>
Identification of radiomics features for constructing models

We intersectedaly analyzed the top 20 features obtained by mRMR algorithm and another 20 features obtained by RFE algorithm to obtain three mutual features including original_glrlm_RunVariance, original_first_order_10percentile and original_gldm_dependence_nonuniformity_normalized. The three features for constructing mRMR_RFE_GBM model were shown in Fig. 3. The two optimal features with top frequencies including original_glrlm_RunVariance and original_ngtdm_Complexity were obtained using the repeat LASSO algorithm. And the two selected features for constructing Repeat_LASSO_GBM model were shown in Fig. 4.

Evaluation of mRMR_RFE_GBM model and Repeat_LASSO_GBM model

The predictive efficacy of mRMR_RFE_GBM model (Fig. 6) and Repeat_LASSO_GBM model (FIGURE S1) were evaluated by ROC curve. In training set, ROC curve shows that AUC value of mRMR_RFE_GBM model is 0.911 and Repeat_LASSO_GBM model is 0.733. In validation set, AUC value of mRMR_RFE_GBM model is 0.849 and Repeat_LASSO_GBM model is 0.72. In comparison of two cohorts (training set and validation set), we found that AUC value of mRMR_RFE_GBM model in training set (p = 0.443, p = 0.912) are not statistically different with that in validation set. Likewise, AUC value of Repeat_LASSO_GBM model in training set (p = 0.443, p = 0.912) are also not statistically different with that in validation set. These data suggested that both two models fit well. In comparison of two models (mRMR_RFE_GBM model and Repeat_LASSO_GBM model), we observed that AUC value of mRMR_RFE_GBM model is significantly higher than Repeat_LASSO_GBM model in training cohort (P < 0.001), while no difference was observed in validation cohort (p = 0.212). In addition, the calibration curve shows that the predictive result of TGF-β1 expression using radiomics-based model is consistent with real value. DCA diagram shows the models were in good clinical practicability. Finally, in the mRMR_RFE_GBM model, we not only observed that there is a significant difference in distribution of rad score between training set and validation set, but also observed that higher rad score is presented in high TGF-β1 expression group (FIGURE S2-D). However, in the Repeat_LASSO_GBM model, although no difference in distribution of rad score between high and low TGF-β1 expression groups (FIGURE S2A) was observed in validation set (P > 0.05), it was significantly different in training set (P < 0.001) (FIGURE S2B). In summary, mRMR_RFE_GBM model and Repeat_LASSO_GBM model both have great potentials in predicting TGF-β1 expression, and mRMR_RFE_GBM model is relatively better in clinical practicality basing on their predictive performance.

DISCUSSION

High recurrence and metastasis are two major causes that [1] reduced the overall median survival time (OS) in advanced HNSCC patients [19]. Hence, it is significant to improve the precancerous diagnosis and the judgment of prognosis in HNSCC. Previous studies have suggested that the immunohistochemical p16 status of OPSCC and the expression level of CPS are the diagnostic biomarkers in HNSCC, and TMB and TILs are the prognostic biomarkers for immunotherapy [4]. However, it still has limitations in detection of these reported biomarkers and is necessary to identify novel biomarkers or technologies. TGF-β is highly expressed in many types of cancers and involves in a number of important biological events including extracellular matrix synthesis, cell growth, and cell differentiation [20], thus leading to tumor progression, invasion and metastasis [31]. Furthermore, TGF-β has been regarded as the marker that associated with prognosis of HNSCC patients [26–28], and which has been confirmed by KIM et al. [29]. As the most abundant type in TGF-β family [21], TGF-β1 has been reported that it is related to the poor prognosis of many malignancies [21–25], including HNSCC. For example, the TGF-β1 expression in interstitial tissue of HNSCC is significantly increased [9], and the elevated TGF-β1 level in plasma is negatively correlated with prognosis of HNSCC patients with cetuximab treatment [30]. In this study, we verified that TGF-β1 is an independent prognostic marker in HNSCC and its expression is negatively correlated with prognosis of patients (p < 0.01). Therefore, non-invasive prediction of TGF-β1 expression in HNSCC will be helpful for clinical application.

In radiomics studies, the major imaging mode of HNSCC is CT imaging [32], followed by MRI [32–34], and the rare is PET-CT or PET-MR [35, 36]. The most commonly feature is gray level co-occurrence matrix (GLCM), followed by gray level run-length matrix (GLRLM), gray level size-area matrix (GLZM) and gray level distance-area matrix (GLDM) [32]. For example, Francesco Mungai et al. reported that GLRLM can be applied to construct a model for predicting human papillomavirus in oropharyngeal squamous cell carcinoma [37]. Wenwu et al. found that GLCM can distinguish the differentiation degree of HNSCC based on CT radiation characteristics [38]. In our study, 20 radiomics features were obtained using the mRMR and RFE algorithms. The optimal feature subset obtained by mRMR_RFE algorithm includes one first-order feature and two second-order features. The optimal feature subset obtained by Repeat_LASSO algorithm all are second-order features. In all these features, the gray run-length matrix (GLRLM) is the best one for both mRMR_RFE_GBM model and Repeat_LASSO_GBM model.

The radiomics studies of HNSCC involves in many aspects including prediction of the side effects in anti-tumor therapy, discrimination of pathological features, stages and risk stratification, classification of molecular subtypes and identification of prognostic markers [32]. For example, the random forest classifier model based on CT radiomics features can significantly distinguish the differentiation degree of HNSCC[38]. Tanzhu et al. predicted the mutation status of HPV and TP53 in HNSCC patients by using the random forest classifier model constructed by CT radiomics characteristics [39]. Dang M et al. used the subset size forward selection algorithm to establish a radiomics model, and the accuracy of this model in predicting p53 status was 81.3% [40]. In this study, we constructed two GBM models-mRMR_RFE_GBM model and Repeat_LASSO_GBM model-through utilizing these features respectively, including GLRLM, NGTDM and GLDM. Moreover, we found that gradient-enhanced CT imaging features can effectively predict TGF-β1 expression in HNSCC, which is significantly associated with prognosis of patients.

Although the non-invasive prediction model based on enhanced CT-based radiomics works well, there still have some limitations. Firstly, data obtained from public datasets may be unstable in image quality, thus affecting the result of prognosis analysis. Secondly, TGF-β1 was the only marker used for prognosis.
In this study, other radiomics biomarkers such as FAT1, KMT2D, NSD1 and CD8 + T cells deserve further study. Thirdly, only 128 enhanced CT images were included in this study, whereas the predictive potential of MRI sequences and PET-CT images remains unclear. Therefore, larger sample size and more prognostic biomarkers are necessary to the stability of prognostic model, which provides a direction for radiomics studies of HNSCC in the future.

In summary, TGF-β1 expression is negatively related to prognosis of HNSCC patients. CT radiomics basing on gradient enhancement algorithm provide an image substitute that can effectively and non-invasively predict TGF-β1 expression in HNSCC. Furthermore, mRMR_RFE_GBM model and the Repeat_LASSO_GBM model both have great potential in predicting TGF-β1 expression, while mRMR_RFE_GBM model is better than latter when considering their predictive performances.

Declarations

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

KQ and QX L contributed to conception and design of the study. LL, CX L collected and analyzed the required CT data. YC performed the statistical analysis. FY and J R collected and assembled the total data. KQ and QX L wrote and reviewed the manuscript. All authors contributed to manuscript revision and approved the submitted version.

Funding:

This work was supported by the National Natural Science Foundation of China (82173311).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. We declare that the datasets supporting the findings of this study are available in TCGA database (https://portal.gdc.cancer.gov/), and TCIA (https://www.cancerimagingarchive.net/). The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Acknowledgments

The authors thank 51runse (www.51runse.cn) for the English language editing during the preparation of this manuscript.

ETHICS STATEMENT

All data and images from TCGA and TCIA are anonymous and public, so they are exempted from ethics and informed consent after being recognized by the unit ethics committee in accordance with the national legislation and the institutional requirements.

References


Figures
Figure 1

Figure 2

The comparison analysis of high vs low TGF-β1 expression groups. (A) The expression of TGF-β1 in tumor and normal tissues. (B) Kaplan–Meier survival analysis. (C) Correlation analysis of TGF-β1 and clinical covariates. (D) Enrichment analysis of top ten KEGG pathways. (E) Enrichment analysis of top thirty GO biological processes. (F) Analysis of infiltrated immune cells.

Figure 3

Screening of radiomics features by repeat mRMR_RFE_GBM algorithm. A&B: Bar chart of training cohort (B) and validation cohort (A). C: Ranking imaging features based on their predictive potential.
Figure 4
Screening of radiomics features by LASSO-GBM algorithm. A&B: Bar chart of training cohort (B) and validation cohort (A). C: Ranking imaging features based on their predictive potential.

Figure 5
A: Plot of ten-fold cross-validation for determining the optimal lambda (tuning parameter). B: Plot of non-zero coefficients or image features against the L1 norm penalty. C: Four top features based on their frequencies were identified using the method of 1000 lasso regression.
Figure 6

Comprehensive analysis of mRMR_RFE_GBM model for predicting TGF-β1 expression

(A&B) ROC curve of mRMR_RFE_GBM model in training set (A) and validation set (B). (C&D) Recall curve of mRMR_RFE_GBM model in training set (C) and validation set (D). (E&F) Calibration curve of mRMR_RFE_GBM model in (E) training set and (F) validation set. (G& H) Decision curve of the model in (G) training set and (H) validation set. Gray heavy lines: ideal performance; Dotted lines: real performance; Solid lines: corrected performance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TCGATCIANHSCC.xlsx
- SupplementaryFIGURE1ReceiveroperatingcharacteristiccurvesrecallcurvecalibrationcurvesanddecisioncurveanalysisoftheRepeatLASSOGBMmodelforpred
- SupplementaryFIGURE2DifferencebetweengroupsofpredictedvaluesofRepeatLASSOGBMandmRMRRFEGBMmodelsinhighbroandlowTGF1expressiongroups..
- SupplementaryTABLE1CCbetweentrainingandvalidationgroupsexlxs