CT-based deep multi-label learning prediction model for outcome in patients with oropharyngeal squamous cell carcinoma

Ma, Baoqiang; Guo, Jiapan; Zhai, Tian Tian; van der Schaaf, Arjen; Steenbakkers, Roel J.H.M.; van Dijk, Lisanne V.; Both, Stefan; Langendijk, Johannes A.; Zhang, Weichuan; Qiu, Bingjiang

Published in:
Medical Physics

DOI:
10.1002/mp.16465

Publication date:
2023

Received: 24 December 2022 | Revised: 23 April 2023 | Accepted: 1 May 2023

DOI: 10.1002/mp.16465

RESEARCH ARTICLE

MEDICAL PHYSICS

CT-based deep multi-label learning prediction model for outcome in patients with oropharyngeal squamous cell carcinoma

Baoqiang Ma[1] | **Jiapan Guo**[1,2,3] | **Tian-Tian Zhai**[4] | **Arjen van der Schaaf**[1] | **Roel J. H. M. Steenbakkers**[1] | **Lisanne V. van Dijk**[1,5] | **Stefan Both**[1] | **Johannes A. Langendijk**[1] | **Weichuan Zhang**[6] | **Bingjiang Qiu**[1,2] | **Peter M. A. van Ooijen**[1,2] | **Nanna M. Sijtsema**[1]

[1]Department of Radiation Oncology, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands

[2]Machine Learning Lab, Data Science Centre in Health (DASH), University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands

[3]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, the Netherlands

[4]Department of Radiation Oncology, Cancer Hospital of Shantou University Medical College, Shantou, China

[5]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Centre, Houston, Texas, USA

[6]Institute for Integrated and Intelligent Systems, Griffith University, Queensland, Australia

Correspondence
Baoqiang Ma, Department of Radiation Oncology, University of Groningen, UMCG (Fonteinstraat 18), Hanzeplein 1, 9713 GZ, Groningen, the Netherlands
Email: b.ma@umcg.nl

Funding information
China Scholarship Council, Grant/Award Number: 202006020054; University of Groningen

Abstract

Background: Personalized treatment is increasingly required for oropharyngeal squamous cell carcinoma (OPSCC) patients due to emerging new cancer subtypes and treatment options. Outcome prediction model can help identify low or high-risk patients who may be suitable to receive de-escalation or intensified treatment approaches.

Purpose: To develop a deep learning (DL)-based model for predicting multiple and associated efficacy endpoints in OPSCC patients based on computed tomography (CT).

Methods: Two patient cohorts were used in this study: a development cohort consisting of 524 OPSCC patients (70% for training and 30% for independent testing) and an external test cohort of 396 patients. Pre-treatment CT-scans with the gross primary tumor volume contours (GTVt) and clinical parameters were available to predict endpoints, including 2-year local control (LC), regional control (RC), locoregional control (LRC), distant metastasis-free survival (DMFS), disease-specific survival (DSS), overall survival (OS), and disease-free survival (DFS). We proposed DL outcome prediction models with the multi-label learning (MLL) strategy that integrates the associations of different endpoints based on clinical factors and CT-scans.

Results: The multi-label learning models outperformed the models that were developed based on a single endpoint for all endpoints especially with high AUCs $\geq$ 0.80 for 2-year RC, DMFS, DSS, OS, and DFS in the internal independent test set and for all endpoints except 2-year LRC in the external test set. Furthermore, with the models developed, patients could be stratified into high and low-risk groups that were significantly different for all endpoints in the internal test set and for all endpoints except DMFS in the external test set.

Conclusion: MLL models demonstrated better discriminative ability for all 2-year efficacy endpoints than single outcome models in the internal test and for all endpoints except LRC in the external set.

Baoqiang Ma and Jiapan Guo contributed equally.

KEYWORDS
computed tomography, deep learning, head and neck cancer, multi-label learning, oropharyngeal squamous cell carcinoma, outcome prediction

1 | INTRODUCTION

Oropharyngeal squamous cell carcinoma (OPSCC) is a sub-type of head and neck cancers (HNCs). Many patients with OPSCC are treated primarily with (chemo) radiotherapy. The 5-year overall survival (OS) rate of human papillomavirus (HPV) related (HPV-positive) OPSCC patients is significantly higher than that of the HPV-negative cases (75–80% vs. 45–50%).[1] The OPSCC subtypes and treatment options have increased thus that there is more need for personalized treatment. Outcome prediction models may stratify patients with favorable and unfavorable outcome, which may help selecting patients to investigate innovative de-escalation or intensified treatment approaches.[2–4]

The tumor-node-metastasis (TNM) staging system, translated in overall tumor staging (including HPV-status in the AJCC^{8th} edition[5]) is widely used for prognostic evaluation and risk stratification of OPSCC.[6,7] However, this staging is not specific enough for a more personalized response evaluation. Ang et al. classified OPSCC patients as a low, intermediate, or high risk of death using HPV status, smoking combined with tumor stage, and nodal stage.[8] Prediction models based on clinical parameters and radiomics features have shown good performance in the prediction efficacy endpoints, such as overall survival (OS) and local control (LC), or complications like xerostomia in HNCs[9–12] and OPSCC patients.[13,14] However, these descriptive image features are restricted as they translate image characteristic in quantitative hand-crafted radiomics values. In contrast, convolutional neural networks (CNNs) can use information from the entire image, not compressing it to single values, and may therefore be more capable of obtaining the comprehensive and predictive information from the clinical images, and thus potentially improving OPSCC outcome predictions.[15]

Recently, CNNs have been successfully applied in image classification tasks[16,17] and showed the potential of predicting complications and prognostic outcomes in HNC.[18–20] For OPSCC specifically, Cheng et al.[21] and Fujima et al.[22] proposed fully automated CNNs-based models using (positron emission tomography) PET images for the prediction of OS and local tumor control, respectively. Moreover, the winner in HECKTOR 2021 challenge[23] used FDG-PET/CT images, GTVt contours, and clinical parameters together to build a DenseNet[24] for (progression-free survival) PFS prediction.[25]

Multi-label learning is a technique that predicts all the relevant labels for a given example by exploiting the label correlations and the input feature information.[26] Its advantage over single-label learning is that it can capture the correlations between different labels and exploit them for better prediction performance.[27] To our best knowledge, no previous studies have applied multi-label learning in the outcome prediction of OPSCC. Inspired by strong associations between different outcome endpoints of the same patient,[28] we propose a CNN-based multi-label learning (MLL) approach, which integrates these associations for further improving the prediction ability in OPSCC outcome prediction. The proposed approach first utilized the 3D Resnet[16] to extract image features from the gross primary tumor volume (GTVt) contoured on the contrast-enhanced planning CT. Second, all clinical parameters were encoded and then concatenated with the extracted image features. Finally, a multilayer perceptron (MLP) was applied to further extract combined features to enable multi outcome endpoints prediction at 2 years after treatment. The performance of the MLL model was compared with that of a single-label model.

2 | MATERIALS AND METHODS

2.1 | Data and preprocessing

The development cohort included a total of 524 OPSCC patients from the publicly available TCIA (The Cancer Imaging Archive) OPC-Radiomics set collected from Princess Margaret Cancer Centre, Toronto, Canada.[29] The OPC-radiomic set was randomly divided into a training set ($n = 367$) and an independent internal (hold out) test set ($n = 157$). Pre-treatment CT imaging with a slice thickness of 2.5 or 2.0 mm and GTVt contours used for treatment planning were available for each patient. Detailed descriptions of the OPC-Radiomics dataset can be found in [45].[30] Furthermore, a total of 396 OPSCC patients collected from the University Medical Centre Groningen, the Netherlands (UMCG) were used as the external test set. A detailed description of the UMCG OPC dataset (the external test set) can be found in Supplementary Part 1. GTVt masks were generated with pixel values 1 inside 0 outside the GTVt contour. All CT images and the corresponding GTVt masks were resampled to a resolution of $1\times1\times2$ mm^3. Subsequently, the CT images and GTVt masks were cropped to a size of $180 \times 180 \times 92$ voxels ($180 \times 180 \times 184$ mm^3) with respect to the center of mass of the GTVt contour to include the complete tumour volume for all patients. Finally, the CT intensity values were truncated to a range between -1024 and 1800,[31] and then normalized to [0, 1].

The candidate clinical predictors include categorical variables: gender (female vs. male), T-stage (T4 vs. T1-3), N-stage (N3 vs. N2 vs. N0-N1), WHO PS (1-3 vs. 0), smoking status (non-smoker vs. ex-smoker vs. current), HPV status (unknown vs. positive vs. negative), and one continuous variable: age. Outcome endpoints consist of 2-year local control (LC), regional control (RC), locoregional control (LRC), distant metastasis-free survival (DMFS), disease-specific survival (DSS), overall survival (OS), and disease-free survival (DFS).

The events of 2-year LC or RC were defined as recurrent or residual tumor within or around the primary site or the regional nodes within 2 years after radiotherapy. The events for 2-year LRC include the events of LC and RC. Distant metastasis was the event in the 2-year DMFS endpoint. Events for 2-year DSS and 2-year OS were defined as the death related to the tumor and death by any cause, respectively. Events for 2-year DFS were all events mentioned above. Positive (label: 1) cases for the 2-year outcome endpoints represent the patients having endpoint events within 2 years (their follow up can be any time) and negative (label: 0) cases had no events for at least 2 years (their follow up $\geq$ 2 years). Patients with follow up $<$ 2 years and no events were censored. Follow-up was executed according to the National Comprehensive Cancer Network guidelines[32] in OPC-Radiomics set. For patients in UMCG OPC follow-up data were acquired every 3 months in the first year after treatment and then every 6 months.

2.2 | Model architecture

We constructed a deep MLL model for multi-prognostic outcome prediction in OPSCC patients, as illustrated in Figure 1a. First, image features from the CT cube including tumor volume as well as the GTVt mask were extracted by the 3D ResNet. The extracted 32 image features were then concatenated with the 32 features encoded from seven clinical parameters using a dense layer. Finally, a multi-layer perceptron (MLP) was used to predict 2-year outcome for multiple endpoints with a focus on one single endpoint.

In detail, the 3D ResNet consists of one convolutional layer, one maxpooling layer with stride 2, eight residual blocks, and one average pooling layer, which converts 512 feature maps to 512 image features. Then, a dense layer converted 512 features to 32 features. Each residual block includes two consecutive convolution layers and one residual connection summing up the input and output of the residual block. Solid curves mean direct summing up while dash curves mean downsampling the input by a factor of two before summing it up with the output. Every convolutional layer in the 3D ResNet was followed by one batch normalization layer. Relu activation functions were added after either the batch normalization layer or the residual connection operation as shown in Figure 1b. MLP was composed of one input layer and three dense connected layers which have 64, 32, 16, and 7 nodes corresponding to 7 outcome endpoints, respectively. The first two dense layers were followed by Relu functions, while the last layer was followed by a sigmoid function.

2.3 | Label encoding and loss definition

A multi-tasking strategy using associations among endpoints was deployed, in which the feature extraction was optimized for all endpoints with a focus on one single endpoint. We denote by N the number of patients in the training set and E the number of different outcome endpoints. The training set can be represented by $T = \{(\mathbf{V}_1, \mathbf{y}_1, \mathbf{c}_1), (\mathbf{V}_2, \mathbf{y}_2, \mathbf{c}_2), \ldots, (\mathbf{V_N}, \mathbf{y_N}, \mathbf{c_N})\}$, where $\mathbf{V_i}$ is the 3D volumetric data and clinical data of the i-th patient, $\mathbf{y_i} \equiv [y_i^1, y_i^2 \ldots; y_i^E]$ is the corresponding 2-year outcome endpoints label vector with the label of the j-th endpoint $y_i^j \in \{0, 1\}$, and $\mathbf{c_i} = [c_i^1, c_i^2 \ldots; c_i^E]$ $(c_i^j \in \{0, 1\})$ is the censoring labels. The censoring label $c_i^j = 1$ if follow-up $<$ 2 years and $y_i^j = 0$, otherwise $c_i^j = 0$. When $c_i^j = 1$, the j-th endpoint of the i-th patient will not be included in the loss function in the training process. For the training criterion, we used a combined loss function which was denoted as $\mathcal{L}$.

$$\mathcal{L} = a \sum_{j=1}^{E} \mathcal{L}_j + \beta \mathcal{L}_e$$

where $\mathcal{L}_j$ is the loss between the prediction and the label of each endpoint in E endpoints while $\mathcal{L}_e$ is the loss of one single enhanced endpoint. This enhanced endpoint, whose predictive performance gets a higher weight in the loss function than other endpoints in the MLL model training process, can be selected as any one of LC, RC, LRC, DMFS, DSS, OS and DFS. $\mathcal{L}_e$ enables the models to focus more on the prediction of the enhanced endpoint than other endpoints, and $\mathcal{L}_j$ is applied to utilize the associations among different endpoints for improving prediction. Both two loss functions are composed of a binary cross entropy loss and a Dice loss and they were weighted with parameters α and β, respectively. The Dice loss was used to solve the problem of the class imbalance.

2.4 | Implementation details

For training the models, we augmented the data with random rotations (-45 to 45 degree) and flipping randomly along the vertical and horizontal directions. The models were trained with a batch size of 14 and an Adam optimizer with an initial learning rate of 0.001. The learning rate will decrease by 10 times if the training loss does not decrease in 10 consecutive training epochs. The α in the loss function was set to 1. We trained two

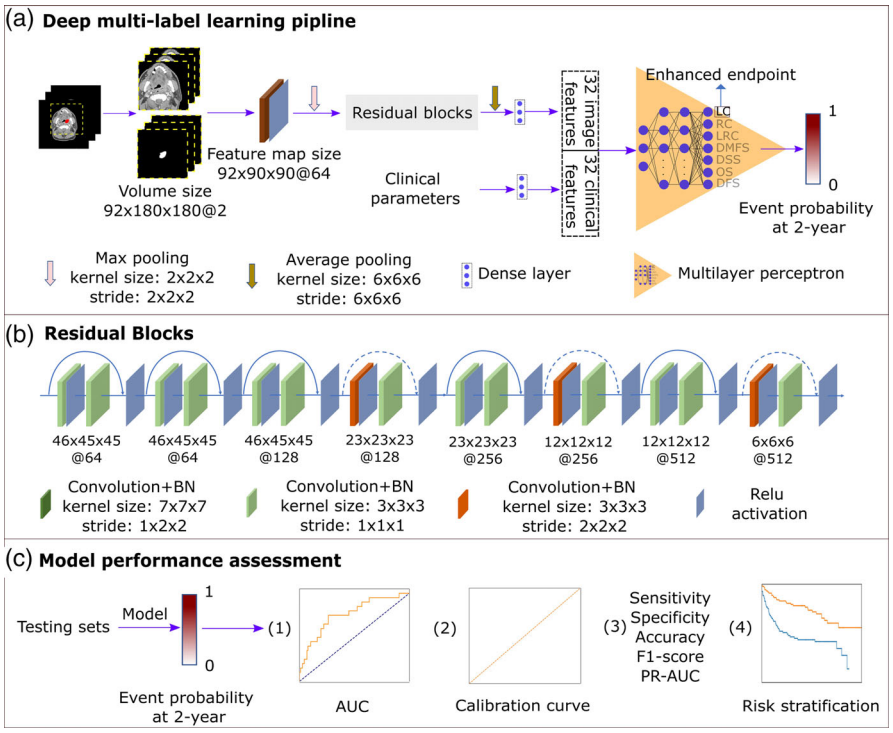


FIGURE 1 The schematic overview of the deep multi-label learning pipeline for multiple 2-year prognostic outcome endpoints prediction in OPSCC patients. (a) Deep multi-label learning pipeline. (b) The detailed structure of residual blocks. (c) The assessment methods of models' performance. BN, batch normalization layer. H×W×D@N stands for the height, width, depth, and number of the feature maps obtained after each residual block.

MLL models, MLL1 and MLL2, by empirically setting β to 5 and 17, respectively, corresponding to a weight of the enhanced endpoint of 50% and 75% in the total loss function, to illustrate the influence of the weight of the enhanced endpoint on the model performance. To avoid over-fitting, the early stopping was set to 20 epochs, meaning that the training process stops if the validation set loss does not decrease in 20 consecutive training epochs. Finally, we obtained the model weights with the lowest validation loss. Additionally in each outcome prediction task, we trained the model for five times and ensembled five models' outputs on the test set by their average. In each of the five training cycles, for the original training set, a new random selection of 30% of the patients for the validation set was performed and the other 70% of the patients were used for training. During this process, the rate of the positive/negative/censored samples was kept equal to the original training set.

2.5 | Assessment of model performance

The evaluation of model performance is illustrated in Figure 1c. The receiver operating characteristic (ROC)-area under the curve (AUC) was employed to evaluate the discriminative power of MLL models. The 95% confidence interval (CI) of AUCs was calculated according to 1000 bootstrapping samples in test sets. The goodness-of-fit of the MLL models was assessed with calibration curves which show the observed actual 2-year event rates versus the MLL predicted 2-year event rates. Additionally, the discriminative ability of the MLL models was evaluated by sensitivity, specificity, and balanced accuracy, with the final predictions determined by the cut-off which is the probability value obtaining the highest S-score in the training set (see Supplementary Part 2). F1-score and PR-AUC (Precision-Recall Area Under the Curve) were also evaluated for MLL models, and F1-score was calculated using the cut-off of the probability obtaining the highest F1-score in the training set. Finally, these cut-off values determined by the S-score were used to stratify patients into a high-risk group (predicted probability by MLL $>$ cut-off) and a low-risk group (predicted probability by MLL $\leq$ cut-off) for each endpoint. Kaplan-Meier curves[33] were generated to investigate LC, RC, LRC, DMFS, DSS, OS, and DFS rates for the high and low-risk groups.

To demonstrate the benefit of MLL, we compared its performance with single-label learning-based models, clinical and image models. SLL models have the same inputs as MLL models but can only predict one outcome endpoint by training with the loss of only this endpoint. The clinical and image models were obtained by training SLL models with clinical data and image data only as input, respectively.

2.6 | Statistical analysis

Statistical analyses were performed using scikit-learn v0.24.2 package in Python v3.7.4. The AUCs of clinical, image, SLL, and MLL models in each outcome endpoint were compared using the independent-sample *t*-test. Hosmer-Lemeshow (HL) tests [48] were performed to evaluate the goodness-of-fit of the points in calibration curves with the ideal calibration line. The log-rank tests[34] were applied to compare the differences between the high and low-risk patient groups in the Kaplan-Meier cures. Categorical and normally distributed clinical variables between different cohorts were compared using chi-square tests and independent-sample *t*-test, respectively. Two-tailed *p*-values < 0.05 were considered statistically significant.

2.7 | Other strategies of improving the MLL model

Because the dataset we used is quite imbalanced for some endpoints such as LC and RC as shown in Table S2, in additional to the Dice loss, oversampling the training set may help further solve the imbalanced classes problem. Thus, oversampling was used in the training of MLL models to investigate whether this strategy (MLL + oversampling) can obtain higher predictive performance for outcome endpoints than the MLL models.

Additionally, some studies demonstrated that combing radiomics features and deep learning achieved better outcome prediction performance in HNC cancer[18] and OPSCC cancer.[35] Thus, we built radiomics models for each endpoint as described in Supplementary Part 3. Then, we incorporated radiomics features from the radiomic model into the MLL models by concatenating them with the seven clinical features of the MLL model so that the MLL model will extract 32 clinical and radiomics features using a fully connected layer (Figure 1). This strategy (MLL + radiomics) was compared with the MLL model for each endpoint.

3 | RESULTS

The patient characteristics of the data sets are listed in Table S1. No significant differences were observed in the characteristics between the patients in training and internal independent test set. However, gender, smoking status, HPV status, and T-stage were significantly different between the training and external test sets. Higher proportions of female (32.1% vs. 18.0%), current smoker (49.2% vs. 32.4%), HPV negative (49.2% vs. 24.3%), and T4 stage patients (45.2% vs. 21.2%) were present in the external test set than the training set.

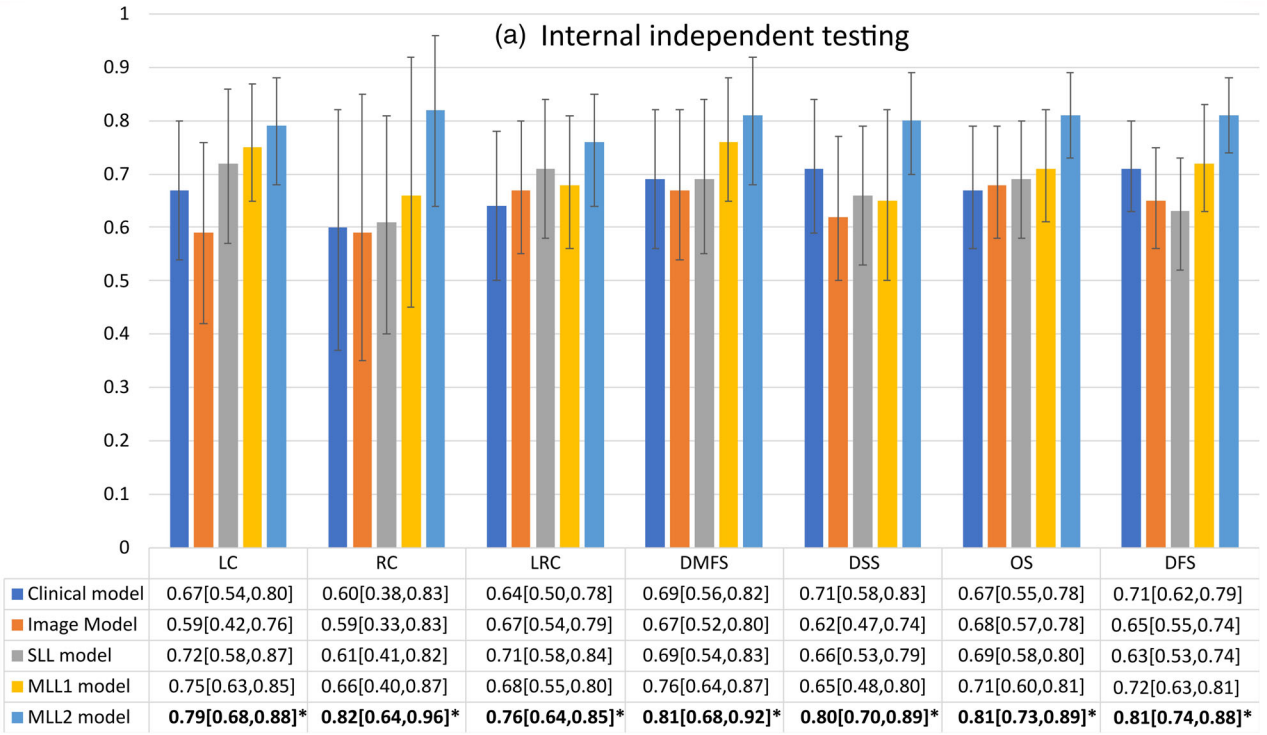
The number of patients with and without events as well as the censored patients and the outcome rate at 2-year follow-up for each endpoint of HPV negative and positive patients were shown in Table S2 for the three datasets. For HPV-negative and HPV-positive patients, respectively, the 2-year rate of each outcome endpoint among three sets are comparable.

Figure 2 shows the AUC [95% CI] values of different models in the test sets. In the internal test set, the SLL models for LC, RC, LRC, and OS, which utilize both clinical and image data, showed higher AUC than the clinical or image only models (Figure 2a). The AUCs for five endpoints (LC: 0.75 RC: 0.66, DMFS: 0.76, OS: 0.71, and DFS: 0.72) were higher in the MLL1 models than in the other non-MLL models. The MLL2 models obtained significantly highest AUCs in all endpoints especially AUCs > 0.80 in RC, DMFS, DSS, OS and DFS. In the external test set, the MLL1 model had significantly higher AUCs of 0.57 and 0.66 for RC and DMFS, respectively, than other non-MLL models (Figure 2B). The MLL2 model achieved significantly highest AUCs in six of the seven endpoints (LC: 0.73, RC: 0.65, DMFS: 0.69, DSS: 0.78, OS: 0.72, and DFS: 0.65) in the external test set. Figure 3 displays receiver operator characteristic (ROC) curves of the different models for 2-year DMFS predictions, in which the MLL2 model obtained highest AUC values in both test sets.

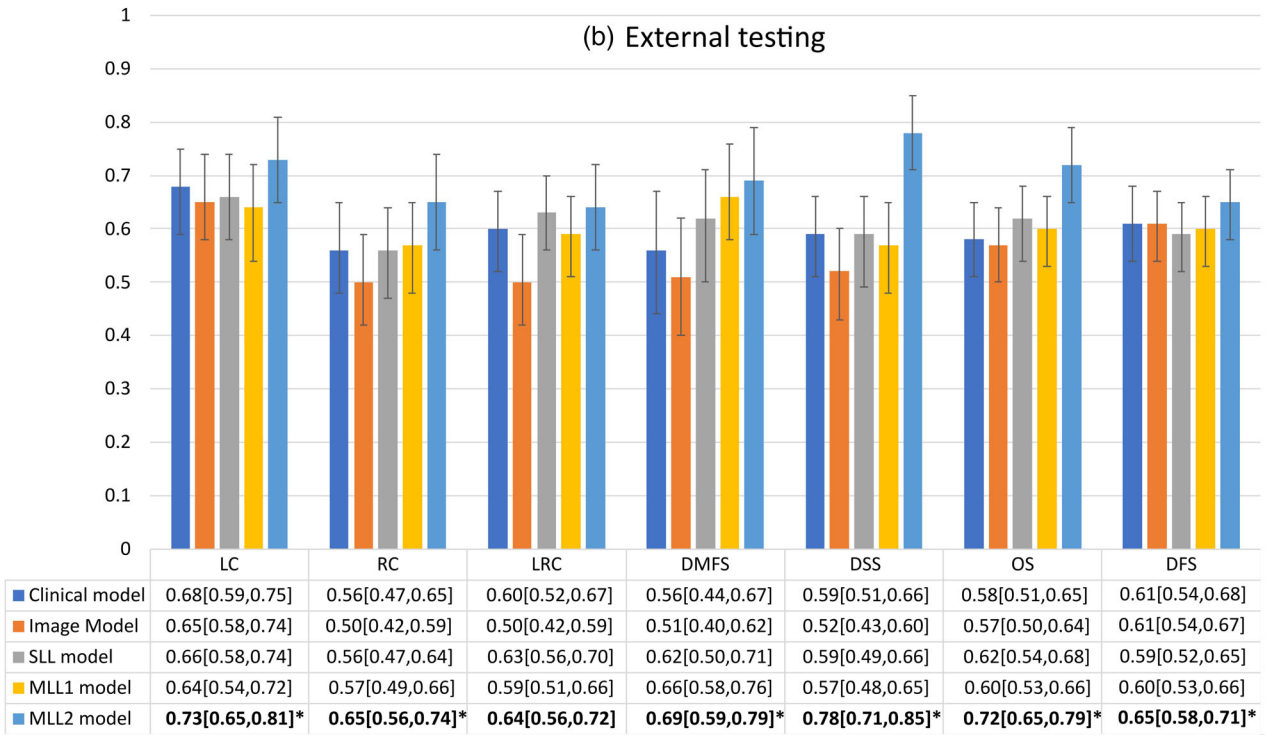
Because MLL2 models achieved higher AUCs than MLL1 models, all other performance analysis of MLL models were based on MLL2 models instead of MLL1 models. The AUCs results of radiomics models and MLL2 models trained by only imaging data were shown in Figure S4 in which MLL2 models (image only) generally showed better AUCs than radiomics models in most endpoints in both internal and external test sets. Additionally, the AUCs comparison of MLL2, MLL2 + oversampling and MLL2 + radiomics models is also displayed in Figure S4. Generally, MLL2 + oversampling and MLL2 + radiomics models cannot improve the AUCs of most endpoints compared to MLL2 models in both internal and external test sets.

Additionally, we investigated sensitivity, specificity, balanced accuracy, F1-score, and PR-AUC values of MLL2 models in the internal and external test sets (Figure 4 and Figure S2), respectively. Balanced sensitivity and specificity values and balanced accuracies ≥ 0.70 were obtained for all endpoints except DSS in the internal test set and for DSS in the external test set and either high sensitivity or specificity values were obtained for the other endpoints. High F1-score of 0.62 and PR-AUC of 0.65 was observed in DFS only and not in the other endpoints in the internal test set.

Figure 5 and Figure S3 display the calibration curves of MLL2 models for all endpoint events in the internal independent test set and external test set, respectively. In Figure 5, most circles were located near the ideal calibration lines (slope: 1 and intercept: 0). A good calibration ($p > 0.05$ with HL tests) was obtained for all endpoints in the internal independent test set. In



	LC	RC	LRC	DMFS	DSS	OS	DFS
Clinical model	0.67[0.54,0.80]	0.60[0.38,0.83]	0.64[0.50,0.78]	0.69[0.56,0.82]	0.71[0.58,0.83]	0.67[0.55,0.78]	0.71[0.62,0.79]
Image Model	0.59[0.42,0.76]	0.59[0.33,0.83]	0.67[0.54,0.79]	0.67[0.52,0.80]	0.62[0.47,0.74]	0.68[0.57,0.78]	0.65[0.55,0.74]
SLL model	0.72[0.58,0.87]	0.61[0.41,0.82]	0.71[0.58,0.84]	0.69[0.54,0.83]	0.66[0.53,0.79]	0.69[0.58,0.80]	0.63[0.53,0.74]
MLL1 model	0.75[0.63,0.85]	0.66[0.40,0.87]	0.68[0.55,0.80]	0.76[0.64,0.87]	0.65[0.48,0.80]	0.71[0.60,0.81]	0.72[0.63,0.81]
MLL2 model	**0.79[0.68,0.88]***	**0.82[0.64,0.96]***	**0.76[0.64,0.85]***	**0.81[0.68,0.92]***	**0.80[0.70,0.89]***	**0.81[0.73,0.89]***	**0.81[0.74,0.88]***



	LC	RC	LRC	DMFS	DSS	OS	DFS
Clinical model	0.68[0.59,0.75]	0.56[0.47,0.65]	0.60[0.52,0.67]	0.56[0.44,0.67]	0.59[0.51,0.66]	0.58[0.51,0.65]	0.61[0.54,0.68]
Image Model	0.65[0.58,0.74]	0.50[0.42,0.59]	0.50[0.42,0.59]	0.51[0.40,0.62]	0.52[0.43,0.60]	0.57[0.50,0.64]	0.61[0.54,0.67]
SLL model	0.66[0.58,0.74]	0.56[0.47,0.64]	0.63[0.56,0.70]	0.62[0.50,0.71]	0.59[0.49,0.66]	0.62[0.54,0.68]	0.59[0.52,0.65]
MLL1 model	0.64[0.54,0.72]	0.57[0.49,0.66]	0.59[0.51,0.66]	0.66[0.58,0.76]	0.57[0.48,0.65]	0.60[0.53,0.66]	0.60[0.53,0.66]
MLL2 model	**0.73[0.65,0.81]***	**0.65[0.56,0.74]***	**0.64[0.56,0.72]**	**0.69[0.59,0.79]***	**0.78[0.71,0.85]***	**0.72[0.65,0.79]***	**0.65[0.58,0.71]***

FIGURE 2 Comparison of AUC values [95% CI] of different models on the internal independent test set (a) and the external test set (b). * means the AUCs of this model are significantly higher ($p < 0.05$ by independent-sample *t*-test) than that of all other models.

Figure S3, the models developed for distant metastasis (DM) ($p = 0.51$), death ($p = 0.14$), and any recurrence and death (ARD) ($p = 0.59$) calibrated well on the external test set.

Kaplan-Meier curves of LC, DMFS, and OS and all endpoints in the two test sets stratified by MLL2 models are shown in Figure 6 and Figure S4, respectively. For the internal independent test set, differences between high

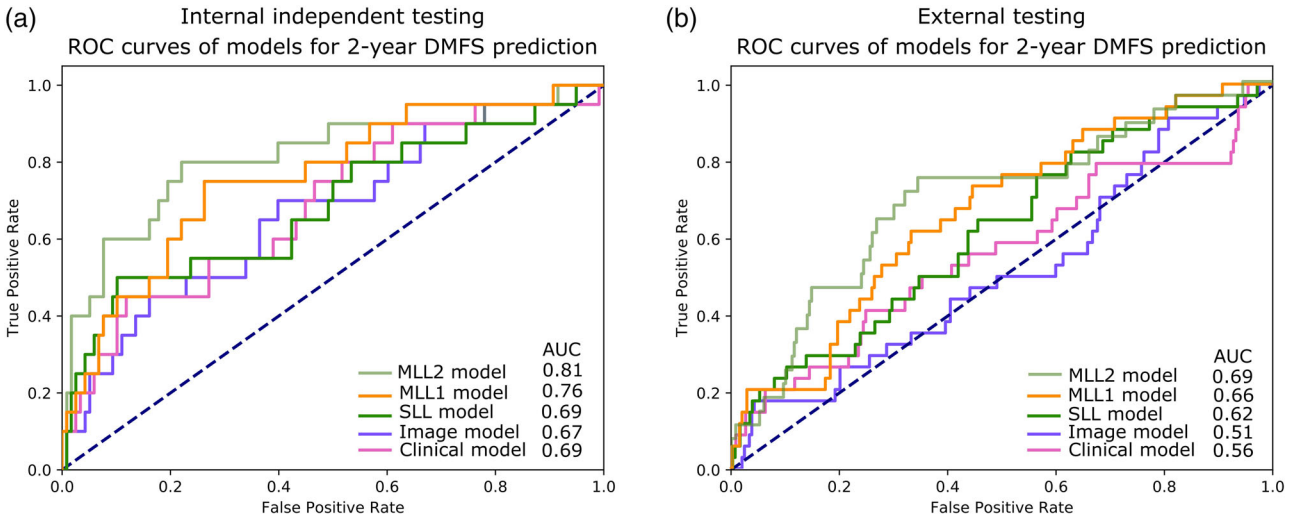
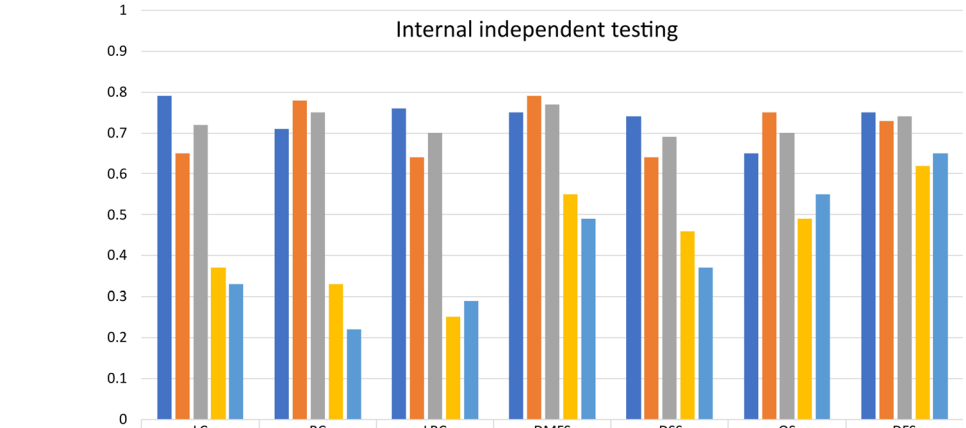


FIGURE 3 Comparison of ROC curves of different models for 2-year DMFS prediction on the internal independent test set (a) and the external test set (b).



	LC	RC	LRC	DMFS	DSS	OS	DFS
Sensitivity	0.79	0.71	0.76	0.75	0.74	0.65	0.75
Specificity	0.65	0.78	0.64	0.79	0.64	0.75	0.73
Balanced accuracy	0.72	0.75	0.70	0.77	0.69	0.70	0.74
F1-score	0.37	0.33	0.25	0.55	0.46	0.49	0.62
PR-AUC	0.33	0.22	0.29	0.49	0.37	0.55	0.65

FIGURE 4 Sensitivity, specificity and balanced accuracy, F1-score, and PR-AUC values of MLL2 models in the internal independent test set.

and low-risk groups are significant for all endpoints. For the external test set, significant differences were found for all endpoints except DMFS.

Overall, these results indicate that our proposed MLL model is effective in discriminating and stratifying patients into different risk groups.

4 | DISCUSSION

In this study, we proposed a deep learning model based on a multi-label learning strategy for 2-year outcome prediction in OPSCC patients. In the internal independent test set, MLL2 models obtained significantly better discriminative performance (AUC values) in the prediction of all endpoints than the clinical, image, and SLL models and showed good calibration for all endpoints. Furthermore, the models had the capability to effectively stratify patients into low and high-risk groups for all endpoints. In the external test set, MLL models still obtained significantly highest AUCs in all endpoints except LRC, good calibration in DM, Death and ARD, and good risk stratification for most endpoints.

Although MLL1 models achieved higher AUCs in five of seven endpoints than clinical, image, and SLL models in the internal test set (Figure 2), they obtained higher AUCs in only RC and DMFS in the external test set. One possible reason is that the MLL model

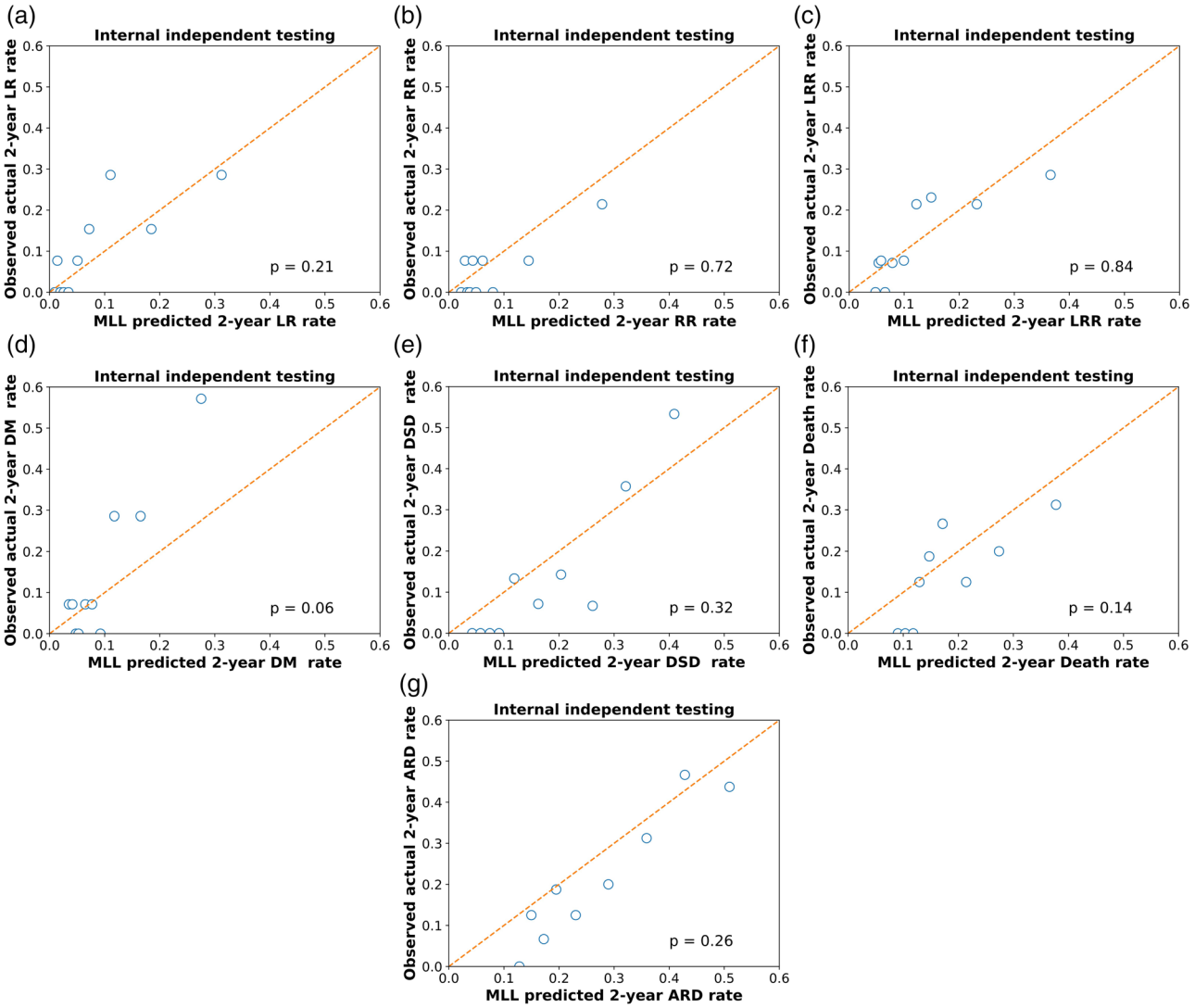


FIGURE 5 Calibration curves of MLL2 models on the internal independent test set. *p*-value was from the HL test. ARD, any recurrence and death; DM, distant metastasis; DSD, disease-specific death; LR, local recurrence; LRR, loco-regional recurrence; RR, regional recurrence.

depends on associations between different outcomes. These associations could be different between institutions due to different treatment, follow-up procedures, and HPV-positive patients' percentage, which decreased the performance of MLL models in external testing. Another possible reason is that our MLL1 model did not give enough attention to the enhanced endpoint while focusing too much on optimizing the correlated endpoints. The MLL2 models that used a larger weight for the enhanced endpoint in the loss function performed significantly better than the other models in all endpoints except LRC in the external test set (Figure 2). This shows that a suitable weight of the enhanced endpoint is essential for training MLL models. Additionally, MLL model validation in the internal test set showed better performance (higher AUCs) than in the external test set. This is most probably related to the differences in clinical characteristics (gender, smoking status, HPV status, and T-stage shown in Table S1), CT quality, GTVt delineation, and the treatments between the internal and external patient cohorts.

Many papers have demonstrated the prognostic values of clinical features for OPSCC such as HPV status, age, gender, T-stage, N-stage, and smoking status.[8,14,36–44] They were also used as candidate predictors in our work. As shown in Figure 2A, high AUC values > 0.60 were obtained by clinical models for almost all outcome endpoints in the internal test set. The AUCs of 0.64 and 0.67 for LRC and OS are comparable to the reported AUCs of 0.61 and 0.71 in,[14] which built and tested clinical models for LRC and OS prediction using 177 OPSCC patients in one center.

Compared with radiomics features, CNNs may extract more comprehensive and representative high-level features which may improve the outcome prediction of OPSCC. From Figure S1, we could observe that our

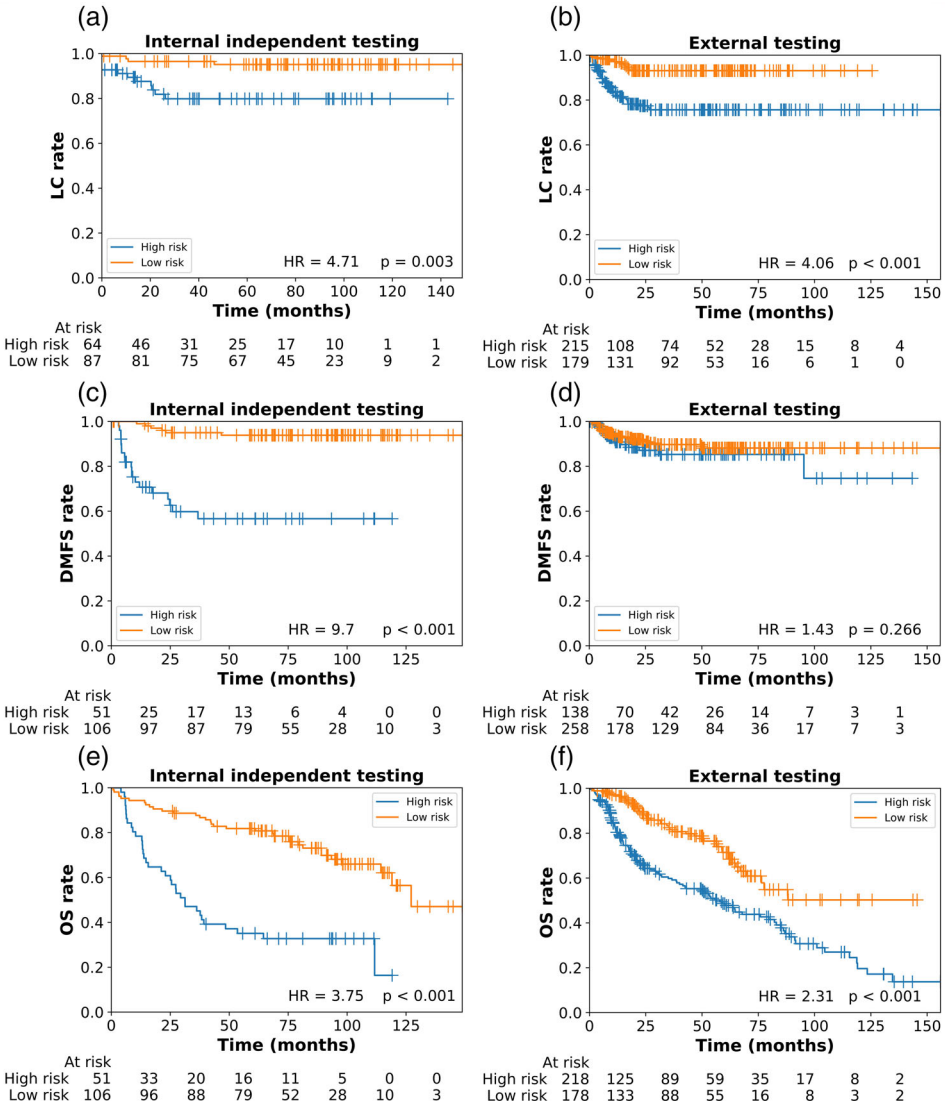


FIGURE 6 Kaplan-Meier curves of high and low risk groups of LC, DMFS, and OS determined by MLL2 models. HR, the hazard ratio between high and low-risk groups. *p* values are from the log-rank test.

MLL2 models trained by using image data only showed better prediction performance for most endpoints than traditional radiomics models in both internal and external test sets. In other relevant studies, CNNs were already used to acquire highly representative image features from PET- with/without CT-images, which showed good prediction ability for OS,[21] local failure,[22] and PFS[25,45] of OPSCC. Pang et al. proposed an advanced combination of training loss with oversampling to train a 3D ResNet18 based on pre-treatment CT and GTV, which achieved the state-of-art AUCs of 0.91, 0.78, and 0.70 for DMFS, LRC, and OS prediction in HNC patients, respectively.[46] Our obtained highest AUCs of 0.76 and 0.82 for LRC and OS, respectively, in the internal testing set as shown in Figure 2A are comparable to this study while our AUC of 0.81 for DMFS prediction is worse. In Pang's study, heat maps were obtained by Grad-CAM[47] to visualize the contribution of each part of image to the outcome prediction. Similarly, the Grad-CAM results of MLL2 models for LC and RC prediction are shown in Figure S5, which shows that image features extracted from primary tumor and peri-tumor regions contributed mostly to the outcome prediction, while other regions such as lymph nodes did not contribute much to the outcome prediction.

This study has some limitations. First, the training set was imbalanced. From Table S2, control and survival rates of all endpoints are larger than 50% in the training set, which probably led to a better prediction for patients without events than for the patients with events. As a result, the positive samples may not have been identified well, which can be reflected by the low F1-scores and PR-AUCs, which focus on the positive samples (Figure 4). The low positive sample rate in the internal

and external test sets for each endpoint (Table S2) also made the range of 95% CI of AUCs large as shown in Figure 2, especially in the internal test set due to its lower positive sample rates than the external test set. Second, patients in the training set came from only one center, which may restrict the MLL in the learning of a more general association between outcome endpoints. Third, the β values in the loss function were set empirically instead of selecting an optimal value. Finally, GTV contour of lymph nodes, which may provide additional predictive information for outcomes were not used.

In the future, MLL models could be used as pre-trained models and finetuned on the external test sets for the prognostic outcome prediction. Additionally, advanced deep learning methods such as attention mechanisms and larger, balanced and multicenter training datasets could be used to improve MLL models. Third, the GTV contour of lymph nodes could help MLL models extract more image features from lymph node for outcome prediction. Finally, learnable weights of all endpoints in the loss function can be used. The learned weights may reveal the contributions of each endpoint to the training of the prediction model for the enhanced endpoint.

In conclusion, we designed a new multilabel learning (MLL) model predicting multiple endpoints simultaneously based on planning CT and clinical data to improve the prediction of OPSCC outcomes compared to single label CNNs. The MLL models showed better discriminative performance for all 2-year outcome predictions than the other models in the internal test set. Furthermore, the models showed good calibration. In the external test set, the model performance was lower due to differences between the patient cohorts, but the MLL models still outperformed the other models for most endpoints.

ACKNOWLEDGMENTS

Baoqiang gratefully acknowledges the financial support for his PhD study provided by the China Scholarship Council (CSC) (202006020054) and University of Groningen.

CONFLICTS OF INTEREST STATEMENT

The authors have no conflict to disclose.

REFERENCES

1. O'Sullivan B, Huang SH, Su J, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol*. 2016;17(4):440-451. doi:10.1016/S1470-2045(15)00560-4
2. Bahig H, Yuan Y, Mohamed ASR, et al. Magnetic resonance-based response assessment and dose adaptation in human papilloma virus positive tumors of the oropharynx treated with radiotherapy (MR-ADAPTOR): an R-IDEAL stage 2a-2b/Bayesian phase II trial. *Clin Transl Radiat Oncol*. 2018;13:19-23. doi:10.1016/j.ctro.2018.08.003
3. Heukelom J, Hamming O, Bartelink H, et al. Adaptive and innovative Radiation Treatment FOR improving Cancer treatment outcomE (ARTFORCE);a randomized controlled phase II trial for individualized treatment of head and neck cancer. *BMC Cancer*. 2013;13(1). doi:10.1186/1471-2407-13-84
4. van Dijk LV, Frank SJ, Yuan Y, et al. Proton image-guided radiation assignment for therapeutic escalation via selection of locally advanced head and neck cancer patients [PIRATES]: a Phase I safety and feasibility trial of MRI-guided adaptive particle radiotherapy. *Clin Transl Radiat Oncol*. 2022;32:35-40. doi:10.1016/j.ctro.2021.11.003
5. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93-99. doi:10.3322/caac.21388
6. Würdemann N, Wagner S, Sharma SJ, et al. Prognostic impact of AJCC/UICC 8th edition new staging rules in oropharyngeal squamous cell carcinoma. *Front Oncol*. 2017;7(JUN). doi:10.3389/fonc.2017.00129
7. Mizumachi T, Homma A, Sakashita T, Kano S, Hatakeyama H, Fukuda S. Confirmation of the eighth edition of the AJCC/UICC TNM staging system for HPV-mediated oropharyngeal cancer in Japan. *Int J Clin Oncol*. 2017;22(4):682-689. doi:10.1007/s10147-017-1107-0
8. Ang KK, Harris J, Wheeler R, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 2010;363(1):24-35. doi:10.1056/NEJMoa0912217
9. Deist TM, Dankers FJWM, Valdes G, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. *Med Phys*. 2018;45(7):3449-3459. doi:10.1002/mp.12967
10. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep*. 2017;7(1). doi:10.1038/s41598-017-10371-5
11. van Dijk LV, Thor M, Steenbakkers RJHM, et al. Parotid gland fat related magnetic resonance image biomarkers improve prediction of late radiation-induced xerostomia. *Radiother Oncol*. 2018;128(3):459-466. doi:10.1016/j.radonc.2018.06.012
12. Zhai TT, Langendijk JA, van Dijk LV, et al. The prognostic value of CT-based image-biomarkers for head and neck cancer patients treated with definitive (chemo-)radiation. *Oral Oncol*. 2019;95:178-186. doi:10.1016/j.oraloncology.2019.06.020
13. Haider SP, Zeevi T, Baumeister P, et al. Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma. *Cancers (Basel)*. 2020;12(7):1778. doi:10.3390/cancers12071778
14. Bos P, van den Brekel MWM, Gouw ZAR, et al. Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models. *Eur J Radiol*. 2021;139:109701. doi:10.1016/j.ejrad.2021.109701
15. Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H. From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. *IEEE Signal Process Mag*. 2019;36(4):132-160. doi:10.1109/MSP.2019.2900993
16. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 2016. December 2016. doi:10.1109/CVPR.2016.90
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2015.
18. Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. *Sci Rep*. 2019;9(1). doi:10.1038/s41598-019-39206-1
19. Men K, Geng H, Zhong H, Fan Y, Lin A, Xiao Y. A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the RTOG 0522 Clinical Trial. *Int J Radiat Oncol Biol Phys*. 2019;105(2):440-447. doi:10.1016/j.ijrobp.2019.06.009

20. Ma B, Guo J, Biase ADe, et al. Self-supervised multi-modality image feature extraction for the progression free survival prediction in head and neck cancer. *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer; 2021:308-317.
21. Cheng NM, Yao J, Cai J, et al. Deep learning for fully automated prediction of overall survival in patients with oropharyngeal cancer using FDG-PET imaging. *Clin Cancer Res*. 2021;27(14). doi:10.1158/1078-0432.CCR-20-4935
22. Fujima N, Andreu-Arasa VC, Meibom SK, et al. Prediction of the local treatment outcome in patients with oropharyngeal squamous cell carcinoma using deep learning analysis of pretreatment FDG-PET images. *BMC Cancer*. 2021;21(1). doi:10.1186/s12885-021-08599-6
23. Andrearczyk V, Oreiller V, Boughdad S, et al. Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images. *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer; 2021:1-37.
24. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ, Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017:4700-4708.
25. Wahid KA, He R, Dede C, et al. Combining tumor segmentation masks with PET/CT images and clinical data in a deep learning framework for improved prognostic prediction in head and neck squamous cell carcinoma. Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Cham: Springer International Publishing, 2022:300-307.
26. Shi C, Kong X, Yu PS, Wang B, Multi-label ensemble learning *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 6913 LNAI. 2011. doi:10.1007/978-3-642-23808-6_15
27. Zhou Z-H, Zhang M-L. Multi-label Learning. 2017:875-881
28. Machtay M, Paulus R, Moughan J, et al. Defining local-regional control and its importance in locally advanced non-small cell lung carcinoma. *J Thorac Oncol*. 2012;7(4):716-722.
29. Kwan JYY, Su J, Huang S, et al. Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma. *Cancer Imaging Arch*. 2019.
30. Kwan J, Yee Y, et al. Radiomic biomarkers to refine risk models for distant metastasis in HPV-related oropharyngeal carcinoma. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1107-1116.
31. Patrick S, Praveen Birur N, Gurushanth K, Shubhasini Raghavan A, Gurudath S. Comparison of gray values of cone-beam computed tomography with hounsfield units of multislice computed tomography: an in vitro study. *Indian J Dent Res*. 2017;28(1):66. doi:10.4103/ijdr.IJDR_415_16
32. Network NCC. NCCN guidelines. *Color Cancer, Version I*. 2010;2010.
33. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481. doi:10.1080/01621459.1958.10501452
34. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*. 1966;50(3).
35. Meng M, Bi L, Feng D, Kim J. Radiomics-enhanced deep multitask learning for outcome prediction in head and neck cancer. *arXiv Prepr arXiv221105409*. 2022.
36. Kühn JP, Schmid W, Körner S, et al. Hpv status as prognostic biomarker in head and neck cancer—which method fits the best for outcome prediction? *Cancers (Basel)*. 2021;13(18):4730. doi:10.3390/cancers13184730
37. Langius JAE, Bakker S, Rietveld DHF, et al. Critical weight loss is a major prognostic indicator for disease-specific survival in patients with head and neck cancer receiving radiotherapy. *Br J Cancer*. 2013;109(5):1093-1099. doi:10.1038/bjc.2013.458
38. Xiao R, Pham Y, Ward MC, et al. Impact of active smoking on outcomes in HPV+ oropharyngeal cancer. *Head Neck*. 2020;42(2):269-280. doi:10.1002/hed.26001
39. Ward MJ, Thirdborough SM, Mellows T, et al. Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer. *Br J Cancer*. 2014;110(2):489-500. doi:10.1038/bjc.2013.639
40. Chen SY, Last A, Ettyreddy A, et al. 20 pack-year smoking history as strongest smoking metric predictive of HPV-positive oropharyngeal cancer outcomes. *Am J Otolaryngol - Head Neck Med Surg*. 2021;42(3):102915. doi:10.1016/j.amjoto.2021.102915
41. Rios Velazquez E, Hoebers F, Aerts HJWL, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. *Radiother Oncol*. 2014;113(3):324-330. doi:10.1016/j.radonc.2014.09.005
42. Fakhry C, Zhang Q, Nguyen-Tân PF, et al. Development and validation of nomograms predictive of overall and progression-free survival in patients with oropharyngeal cancer. *J Clin Oncol*. 2017;35(36):4057-4065. doi:10.1200/JCO.2016.72.0748
43. de França GM, da Silva WR, Medeiros CKS, Júnior JF, de Moura Santos E, Galvão HC. Five-year survival and prognostic factors for oropharyngeal squamous cell carcinoma: retrospective cohort of a cancer center. *Oral Maxillofac Surg*. 2021;26(2):261-269. doi:10.1007/s10006-021-00986-4
44. Yin LX, D'Souza G, Westra WH, et al. Prognostic factors for human papillomavirus–positive and negative oropharyngeal carcinomas. *Laryngoscope*. 2018;128(8):E288-E296. doi:10.1002/lary.27130
45. Bourigault E, McGowan DR, Mehranian A, Papież BW. Multimodal PET/CT tumour segmentation and prediction of progression-free survival using a full-scale UNet with attention. *arXiv Prepr arXiv211103848*. 2021.
46. Pang S, Field M, Dowling J, Vinod S, Holloway L, Sowmya A. Training radiomics-based CNNs for clinical outcome prediction: challenges, strategies and findings. *Artif Intell Med*. 2022;123:102230. doi:10.1016/j.artmed.2021.102230
47. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336-359. doi:10.1007/s11263-019-01228-7

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ma B, Guo J, Zhai T-T, et al. CT-based deep multi-label learning prediction model for outcome in patients with oropharyngeal squamous cell carcinoma. *Med Phys*. 2023;50:6190-6200. https://doi.org/10.1002/mp.16465