European Radiology
https://doi.org/10.1007/s00330-023-09884-7

European Radiology
ESR EUROPEAN SOCIETY OF RADIOLOGY

CHEST

Using clinical and radiomic feature–based machine learning models to predict pathological complete response in patients with esophageal squamous cell carcinoma receiving neoadjuvant chemoradiation

Jin Wang[1] · Xiang Zhu[1] · Jian Zeng[2] · Cheng Liu[3] · Wei Shen[3] · Xiaojiang Sun[1] · Qingren Lin[1] · Jun Fang[1] · Qixun Chen[2] · Yongling Ji[1]

Received: 25 January 2023 / Revised: 25 March 2023 / Accepted: 22 April 2023

Abstract

Objective This study aimed to build radiomic feature-based machine learning models to predict pathological clinical response (pCR) of neoadjuvant chemoradiation therapy (nCRT) for esophageal squamous cell carcinoma (ESCC) patients.

Methods A total of 112 ESCC patients who underwent nCRT followed by surgical treatment from January 2008 to December 2018 were recruited. According to pCR status (no visible cancer cells in primary cancer lesion), patients were categorized into primary cancer lesion pCR (ppCR) group ($N=65$) and non-ppCR group ($N=47$). Patients were also categorized into total pCR (tpCR) group ($N=48$) and non-tpCR group ($N=64$) according to tpCR status (no visible cancer cells in primary cancer lesion or lymph nodes). Radiomic features of pretreatment CT images were extracted, feature selection was performed, machine learning models were trained to predict ppCR and tpCR, respectively.

Results A total of 620 radiomic features were extracted. For ppCR prediction models, radiomic model had an area under the curve (AUC) of 0.817 (95% CI: 0.732–0.896) in the testing set; and the combination model that included rad-score and clinical features had a great predicting performance, with an AUC of 0.891 (95% CI: 0.823–0.950) in the testing set. For tpCR prediction models, radiomic model had an AUC of 0.713 (95% CI: 0.613–0.808) in the testing set; and the combination model also had a great predicting performance, with an AUC of 0.814 (95% CI: 0.728–0.881) in the testing set.

Conclusion This study built machine learning models for predicting ppCR and tpCR of ESCC patients with favorable predicting performance respectively, which aided treatment plan optimization.

Clinical relevance statement This study significantly improved the predictive value of machine learning models based on radiomic features to accurately predict response to therapy of esophageal squamous cell carcinoma patients after neoadjuvant chemoradiation therapy, providing guidance for further treatment.

Key Points

- *Combination model that included rad-score and clinical features had a great predicting performance.*
- *Primary tumor pCR predicting models exhibit better predicting performance compared to corresponding total pCR predicting models.*

Keywords Humans · Esophageal squamous cell carcinoma · Neoadjuvant therapy · Area under the curve · Lymph nodes

Abbreviations

AUC	Area under the curve
BMI	Body mass index
CT	Computed tomography
CTV	Clinical target volume
EC	Esophageal cancer

J Wang, X Zhu, and J Zeng contributed equally and should be considered co-first authors

✉ Jin Wang
wangjin@zjcc.org.cn

✉ Qixun Chen
chenqixun64@163.com

✉ Yongling Ji
jiyl@zjcc.org.cn

[1] Department of Radiation Oncology, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, China

[2] Department of Thoracic Surgery, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, China

[3] Philips Healthcare, Shanghai, China

Published online: 13 July 2023

ESCC	Esophageal squamous cell carcinoma
GTV	Gross tumor volume
KPS	Karnofsky performance status
LOOCV	Leave-one-out cross-validation
MRI	Magnetic resonance imaging
nCRT	Neoadjuvant chemoradiotherapy
NPV	Negative predictive value
OAR	Organs at risk
OS	Overall survival
pCR	Pathological complete response
PD	Program death
PET	Positron emission tomography
PFS	Progression-free survival
ppCR	Primary tumor lesion pCR
PPV	Positive predictive value
PTV	Plan target volume
ROC	Receiver operating characteristics
TNR	True negative rate
tpCR	Total pCR
TPR	True positive rate
VOI	Volumn of interest

Introduction

Esophageal cancer is the sixth leading cause of cancer-related death and the eighth most common cancer worldwide [1]. Esophageal squamous cell carcinoma (ESCC) is the predominant histological subtype of esophageal carcinomas in China. Neoadjuvant chemoradiotherapy (nCRT) followed by surgery has proved to be effective for locally advanced EC by large-scale randomized clinical trials [2, 3].

Pathological complete response (pCR) is the ultimate goal of nCRT. Patients having a pCR appear to have superior overall survival (OS), while the progress may occur in the non-distant future for those who did not achieve a pCR [4]. Moreover, pCR after receiving NCT can help us guide treatment. In the CheckMate-577 study, patients with stage II or III esophageal or gastroesophageal junction cancer had received nCRT and had not reached pCR benefit from program death (PD)-1 immunotherapy irrespective of the PD-L1 status [5].

However, the pCR can be confirmed only after surgery. Those who responded to nCRT might have good prognoses regardless of whether they underwent surgery; thus, surgery could have been avoided in some patients. On the other hand, for those who do not respond to nCRT, earlier surgical intervention can be considered to avoid unnecessary chemoradiation-associated morbidity [6]. In a word, preselecting the potential responders prior to treatment could reduce the economic burden to patients and maximize their benefits. Therefore, the development of a preoperative, noninvasive, and accurate approach prior to treatment to predict pCR is meaningful [7].

Radiomics is an emerging field that uses algorithms to extract a large number of features from radiographic medical images, including CT, PET, and MRI. For instance, Van Rossum, P. S. and Jinrong Qu utilized diffusion-weighted MRI and dynamic contrast-enhanced MRI radiomics for the prediction of pCR to nCRT in esophageal cancer patients, respectively [8–10]. And they found that radiomics features of MRI are able to predict nCRT response.

However, neither FDG-PET nor MRI is commonly used as routine preoperative evaluations of EC in most countries. It would therefore be more efficient if one of the routine staging modalities (i.e., CT) was able to predict histopathological response. It is reported that CT radiomics–based model could facilitate noninvasive preselection for treatment of ESCC patients [11]. However, the accuracy and transferability of the prediction model still require improvement. An optimal approach, combined with more dimensions, is needed to improve the prediction model performance.

A recent study found that baseline clinic and pathological characteristics could predict pathological efficacy to anti-PD-1 combined with neoadjuvant chemotherapy in ESCC [12]. Patient's characteristics and clinicopathologic factors, such as age, poor differentiation grade, and tumor length, were identified as the predictors for pCR after nCRT [13]. Combining clinical factors and 18F-FDG PET–based radiomic features improved the prediction ability [14]. However, it is still not clear whether the combined model that included both clinical features and radiomics features could be more predictive relative to single-dimensional model. Here, we developed and validated the addition of clinical characteristic to a CT radiomics–based model, in order to improve the prediction performance for pCR in EC patients under nCRT. We believe this model is useful to help the doctors to make the best therapeutic management.

Methods

Patient data

The research protocol used in this study was approved by the Ethics Committee of Zhejiang Cancer Hospital (IRB-2020–189(Ke)). Requirement for informed consent was waived.

A total of 112 patients with histologically proven ESCC that underwent nCRT and surgical treatment from January 2008 to December 2018 in Zhejiang Cancer Hospital were recruited to this study. The inclusion criterion was as follows: (1) confirmed ESCC via histological examination; (2) age between 18 and 75 years old; (3) had completed enhanced CT examination before nCRT; (4) in stage T1N1 or T2-4aN0-1M0, and nonregional lymph node metastasis M1 patients (not including distant organ metastasis M1

patients) according to the 6th edition of the AJCC-TNM classification; (5) ECOG grade 0 to 1; (6) blood routine, liver function, and kidney function were normal in general. The exclusion criterion was as follows: (1) cervical esophageal squamous cell carcinoma patients; (2) distant organ metastasis; (3) did not receive surgical treatment after nCRT. The clinical features were collected, and patients underwent enhanced CT scan 1–2 weeks pretreatment (Philips Brilliance CT Big Bore Oncology Configuration) with the same protocol: CT tube voltage 120 kV, rack rotation time 0.75 s, field of view (FOV) 350 × 350 mm, pixel matrix 512 × 512, thickness 5 mm, layer spacing 5 mm. The images were applied with standard reconstruction without resampling.

Contrast medium (Omnipaque, Iohexol, GE Healthcare) was injected with a high-pressure syringe at a flow rate of 3.0 mL/s (1–1.5 mL/kg, ioproxamine injection 300), followed by 20 mL of normal saline for flushing (UlrichCT Injection System, Ulrich Medical). About 38 s after the administration of the contrast agent, single-phase-contrast-enhanced CT scanning was obtained.

Evaluation of tumor response

In this study, patients were categorized into tpCR group ($N = 48$) and non-tpCR group ($N = 64$) according to whether there was no residual invasive tumor in the surgical specimen removed following neoadjuvant therapy. For example, if no visible cancer cell was found neither in primary cancer lesion, nor in near lymph nodes, then that patient was categorized to the tpCR group; otherwise, that patient was categorized to the non-tpCR group. Such classification may more reasonable and usable in evaluating treatment efficacy.

Moreover, according to the primary cancer histopathological response (no visible cancer cells were found in primary cancer lesion), patients were also categorized into primary pathological clinical response (absence of any neoplastic cells in the primary site after receiving NCT, ppCR) group ($N = 65$) and non-ppCR group ($N = 47$).

Radiomic feature extraction and machine learning

Volume of interest (VOI) of CT images was delineated by a senior radiologist and checked by another senior radiologist; both were blinded to patient information. After that, 310 original radiomic features and 310 wavelet radiomic features were extracted by PyRadiomics package. Subsequently, we trained the radiomic models and clinic-based models for predicting ppCR as well as tpCR, respectively (Figure S1). The general procedure for machine learning model training was as follows: features were firstly standardized to the same scale (mean value equal to 0 and standard deviation equal to 1). Next step, correlation analysis was applied, and those highly correlated features (correlation coefficient ≥ 0.9) were deduplicated. In the remaining features, recursive feature elimination (RFE) using ridge regression as estimator was utilized for further feature selection. The process was as follows: (1) all the features were included to model training, then removed the least important feature according to the feature ranking; (2) the remaining features were included to second round model training, and removed the least important feature; (3) step 2 was repeated, until the feature number met the setting. After feature selection, the selected features were included to machine learning model training using logical regression classifier, and leave-one-out cross-validation (LOOCV) strategy was utilized. The machine learning model training process was as follows: (1) the first patient was selected as testing set, and the other 111 patients as training set, then we trained the first model with logical regression classifier and got a predictive result; (2) the second patient was selected as testing set and the other 111 patients as training set, then we trained another model and got second predict result; (3) step 2 was repeated until the last patient, then we got 112 predictive results; (4) the average value of the 112 predictive results was the final predictive result.

The combined model that included rad-score and clinical features was also trained for ppCR and tpCR prediction. Both the ppCR-rad-score and the tpCR-rad-score in this study were calculated via the following formula:

$$rad - score = \sum_{i=1}^{k} (coefficient_i * value_i)$$

The coefficient was a feature coefficient obtained from the model training, the value was a feature value, so the rad-score of a feature was obtained by multiplying its coefficient and value, and a patient rad-score could be calculated via adding up the feature rad-scores (for example, if the radiomic model was trained by four features, then the rad-score of a patient could be calculated by adding up the four feature rad-scores). Some previous studies also included intercept during rad-score calculation, while the values of intercept in our model were small, so we excluded them. Other processes for combined model training were similar to those of radiomic model and clinic-based model.

During the RFE feature selection, we set the feature number from 3 to 12 (the maximum feature number should not exceed 1/10 of sample size; otherwise, the model would overfit), respectively. Therefore, 10 models were trained for radiomic model, clinic-based model, and combined model for predicting ppCR and tpCR, respectively. Among the 10 models, the model that had the highest area under the curve (AUC) in receiver operating characteristics (ROC) curve in the testing set was regarded as the final model.

Statistical analysis

Python3.8 was used for feature extraction, feature selection, machine learning model training, model evaluation, and plotting. SPSS 21.0 was used for *t* test, chi-square test, and Fisher exact test. ROC curve was plotted for model performance evaluation. True positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV) were also calculated for model evaluation. AUC was calculated and bootstrap was utilized for calculating 95% confidence interval (95% CI). Coefficient of logical regression was applied for feature importance ranking. Continuous variables were displayed as mean ± standard deviation, and categorical variables were displayed as count (%). Student *t* test was used for comparison of continuous variables, and chi-square test or Fisher exact test was used for comparison of categorical variables. A two-sided $p < 0.05$ was regarded as statistical difference.

Results

Demographic characteristics

Demographic characteristics are shown in Table 1. From ppCR perspective, 65 patients achieved ppCR while the other 47 patients did not. In ppCR patients, 21.5%, 69.2%, and 9.2% patients were in T2, T3, and T4a stages, respectively; 1.5% and 98.5% patients were in N0 and N1 stages, respectively; 95.4% and 4.6% patients were in M0 and M1 stages, respectively. In non-ppCR patients, 21.3%, 76.6%, and 2.1% patients were in T2, T3, and T4a stages, respectively; 4.2% and 95.8% patients were in N0 and N1 stages, respectively; 87.2% and 12.8% patients were in M0 and M1 stages, respectively. There was no difference of demographic characteristics between ppCR patients and non-ppCR patients (all $p > 0.05$).

From tpCR perspective, 48 patients achieved tpCR whereas the other 64 patients did not. In tpCR patients, 25%, 66.7%, and 8.3% patients were in T2, T3, and T4a stages, respectively, and all patients were in N1 and M0 stages. In non-tpCR patients, 18.8%, 76.6%, and 4.7% patients were in T2, T3, and T4a stages, respectively; 4.7% and 95.3% patients were in N0 and N1 stages, respectively; 85.9% and 14.1% patients were in M0 and M1 stages, respectively. M stage ($p = 0.01$) and side effect rate ($p = 0.027$) were lower in tpCR patients compared to non-tpCR patients, whereas other characteristics were of no difference between the two groups (all $p > 0.05$). Figure S2 displayed a typical CT image in this study with delineation of the tumor.

Performance of radiomic models and clinic-based models in predicting ppCR

A total of 10 radiomics models with different feature numbers were trained (Table S1). Among them, radiomic model with 12 radiomic features had the best ppCR predicting performance, with an AUC of 0.883 (95% CI: 0.876–0.888) in the training set and an AUC of 0.817 (95% CI: 0.732–0.896) in the testing set (Fig. 1). Meanwhile, 10 clinic-based models with different feature numbers were trained to predict ppCR as well (Table S2), and the clinic-based model with 6 clinical features had a best performance, with an AUC of 0.709 (95% CI: 0.700–0.718) in the training set and 0.623 (95% CI: 0.510–0.729) in the testing set (Fig. 2). Radiomic model showed superior performance in predicting ppCR compared to that of clinic-based model.

Performance of combined models in predicting ppCR

To optimize the ppCR predicting performance, 10 combined models that included ppCR-rad-score and clinical features were trained at last (Table S3). Among those 10 combined models, the combined model with 3 features obtained a highest AUC of 0.905 (95% CI: 0.900–0.911) in the training set and 0.891 (95% CI: 0.823–0.950) in the testing set (Fig. 3). The ppCR-rad-score, drinking, and hypotension are three most important features in predicting ppCR. Comparison of those 3 features between ppCR patients and non-ppCR patients showed that ppCR-rad-score ($p < 0.001$) was significantly higher in ppCR patients compared to that of non-ppCR patients, and drinking history ($p > 0.05$) as well as hypotension ($p > 0.05$) were similar between the two groups.

Performance of radiomic models and clinic-based models in predicting tpCR

Besides models in predicting ppCR, another 9 radiomics models and 9 clinic-based models were trained to predict tpCR in parallel. Among the 9 radiomic models (Table S4), the radiomic model with 8 radiomic features showed best tpCR predicting performance, with an AUC of 0.787 (95% CI: 0.780–0.795) in the training set and 0.713 (95% CI: 0.613–0.808) in the testing set (Fig. 4). Among the 9 clinic-based models (Table S5), the clinic-based model with 7 clinical features exhibited best tpCR predicting performance, with an AUC of 0.713 (95% CI: 0.704–0.722) in the training set and 0.624 (95% CI: 0.522–0.725) in the testing set (Fig. 5). Similar to ppCR, radiomic model had a better performance compared to clinic-based model in predicting tpCR as well.

Performance of combined models in predicting tpCR

When combining tpCR-rad-score with clinical features, the tpCR predicting performance increased (Table S6), and the combined model with 4 features derived a best performance of 0.838 (95% CI: 0.832–0.845) in the training set and 0.814 (95% CI: 0.728–0.881) in the testing set (Fig. 6). The

Table 1 patient characteristics

Characteristics	ppCR ($n=112$)			tpCR ($n=112$)		
	ppCR ($n=65$)	Non-ppCR ($n=47$)	p value	tpCR ($n=48$)	Non-tpCR ($n=64$)	p value
Age	58.0 ± 6.8	58.5 ± 7.4	> 0.05	57.9 ± 6.8	58.5 ± 7.2	> 0.05
Gender			> 0.05			> 0.05
Male	61 (93.8)	40 (85.1)		45 (93.8)	56 (87.5)	
Female	4 (6.2)	7 (14.9)		3 (6.2)	8 (12.5)	
Height	167.9 ± 6.2	166.0 ± 6.4	> 0.05	168.2 ± 6.6	166.3 ± 6.0	> 0.05
Weight	61.0 ± 9.8	58.5 ± 7.7	> 0.05	61.3 ± 10.5	59.0 ± 7.7	> 0.05
Smoking			> 0.05			> 0.05
Yes	49 (75.4)	31 (66.0)		36 (66.7)	44 (68.8)	
No	16 (24.6)	16 (34.0)		12 (33.3)	20 (31.2)	
Drinking			> 0.05			> 0.05
Yes	50 (76.9)	29 (61.7)		36 (66.7)	43 (67.2)	
No	15 (23.1)	18 (38.3)		12 (33.3)	21 (32.8)	
Diabetes			> 0.05			> 0.05
Yes	4 (6.2)	3 (6.4)		2 (4.2)	5 (7.8)	
No	61 (93.8)	44 (93.6)		46 (95.8)	59 (92.2)	
Hypotension			> 0.05			> 0.05
Yes	18 (27.7)	9 (19.1)		13 (27.1)	14 (21.9)	
No	47 (72.3)	38 (80.9)		35 (72.9)	50 (78.1)	
Tumor location			> 0.05			> 0.05
Upper chest	27 (41.5)	16 (34.0)		19 (39.6)	24 (37.5)	
Middle chest	17 (26.2)	14 (29.8)		12 (25.0)	19 (29.7)	
Lower chest	21 (32.3)	17 (36.2)		17 (35.4)	21 (32.8)	
Lesion shape			> 0.05			> 0.05
Ulceration	10 (15.4)	9 (19.1)		9 (18.8)	10 (15.6)	
Mushroom	48 (73.8)	32 (68.1)		32 (66.7)	48 (75)	
Constrictive type	7 (10.8)	6 (12.8)		7 (14.6)	6 (9.4)	
T stage			> 0.05			> 0.05
T2	14 (21.5)	10 (21.3)		12 (25)	12 (18.8)	
T3	45 (69.2)	36 (76.6)		32 (66.7)	49 (76.6)	
T4a	6 (9.2)	1 (2.1)		4 (8.3)	3 (4.7)	
N stage			> 0.05			> 0.05
N0	1 (1.5)	2 (4.2)		0 (0.0)	3 (4.7)	
N1	64 (98.5)	45 (95.8)		48 (100.0)	61 (95.3)	
M stage			> 0.05			0.01
M0	62 (95.4)	41 (87.2)		48 (100.0)	55 (85.9)	
M1	3 (4.6)	6 (12.8)		0 (0.0)	9 (14.1)	
TNM staging			> 0.05			> 0.05
II	13 (20)	11 (23.4)		12 (25)	12 (18.8)	
III	47 (72.3)	30 (63.8)		35 (72.9)	42 (65.6)	
IV	5 (7.7)	6 (12.8)		1 (2.1)	10 (15.6)	
Total dose	46.0 ± 4.8	47.0 ± 5.1	> 0.05	46.1 ± 5.0	46.6 ± 4.9	> 0.05
Each dose	1.9 ± 0.2	1.8 ± 0.1	> 0.05	1.9 ± 0.2	1.9 ± 0.1	> 0.05
Side effects			> 0.05			0.027
Yes	23 (35.4)	23 (48.9)		14 (29.2)	32 (50.0)	
No	42 (64.6)	24 (51.1)		34 (70.8)	32 (50.0)	
Operation method			> 0.05			> 0.05
Thoracotomy	35 (53.8)	26 (55.3)		25 (52.1)	36 (56.3)	
Video-assisted thoracotomy surgery	30 (46.2)	21 (44.7)		23 (47.9)	28 (43.8)	

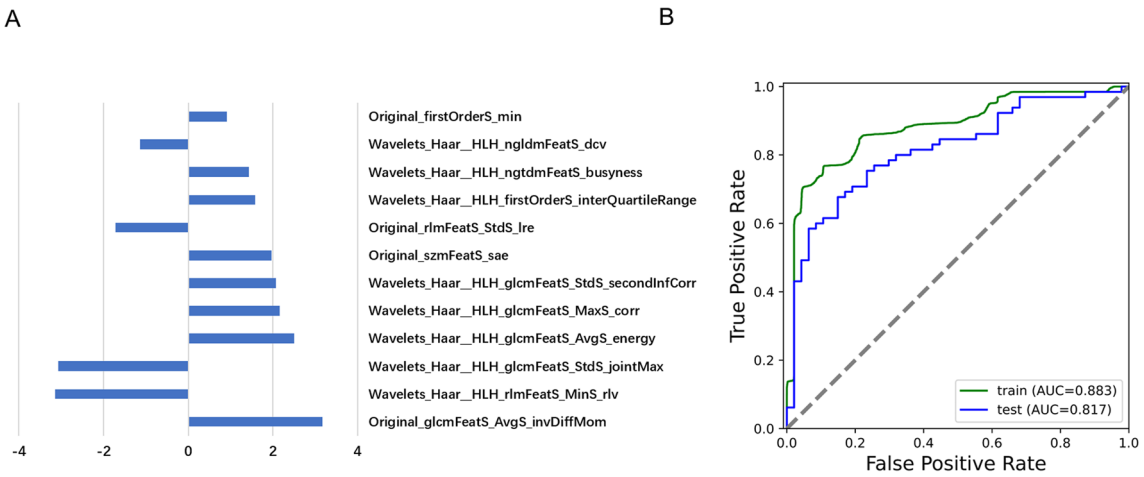

Fig. 1 Best-performance radiomic model for predicting ppCR. **A** Feature coefficients. **B** ROC curve of training set and testing set

tpCR-rad-score, M stage, TNM stage, and side effects were 4 most important features for predicting tpCR. Comparison of those 4 features between tpCR patients and non-tpCR patients revealed that tpCR-rad-score ($p=0.002$) was statistically higher while M stage ($p=0.007$) and side effects ($p=0.027$) were statistically lower in tpCR patients compared to those in non-tpCR patients. As for TNM stage ($p>0.05$), there was no difference between the two groups. Table 2 summarizes the TPR, TNR, PPV, and NPV of the models.

Discussion

Several studies utilized clinical characteristics, imaging features, or biological markers, to build pCR predicting models. For instance, Yihuai Hu et al trained radiomic model to predict the treatment response to chemoradiotherapy (accuracy of 77.1% in the testing set) [15]. In another study, Cuong Duong et al identified a 32-gene classifier that can be used to predict response to nCRT in ESCC patients [16]. However, most of those previous studies had models with AUCs less than 0.85 in the testing set, indicating that the predicting performance could be improved. What's more, features from different dimensions may be associated with pCR of ESCC (e.g., clinical characteristics, imaging features), so the two-dimensional features should be taken into consideration when model training. Last but not least, previous studies trained radiomic model to predict tpCR (including metastatic lymph node pCR). Enlarged lymph nodes are poorly distinguished from reactive inflammatory lymph nodes or normal lymph nodes by either CT or MRI, resulting in diagnostic uncertainty of metastatic lymph nodes. Thus, it is difficult to distinguish metastatic lymph nodes from benign lymph nodes when delineating VOI, leading to the decreased accuracy of tpCR predicting model. In this way, it would be better to train ppCR prediction models as well.

In this study, we trained machine learning models using both radiomic and clinical features, to predict both ppCR

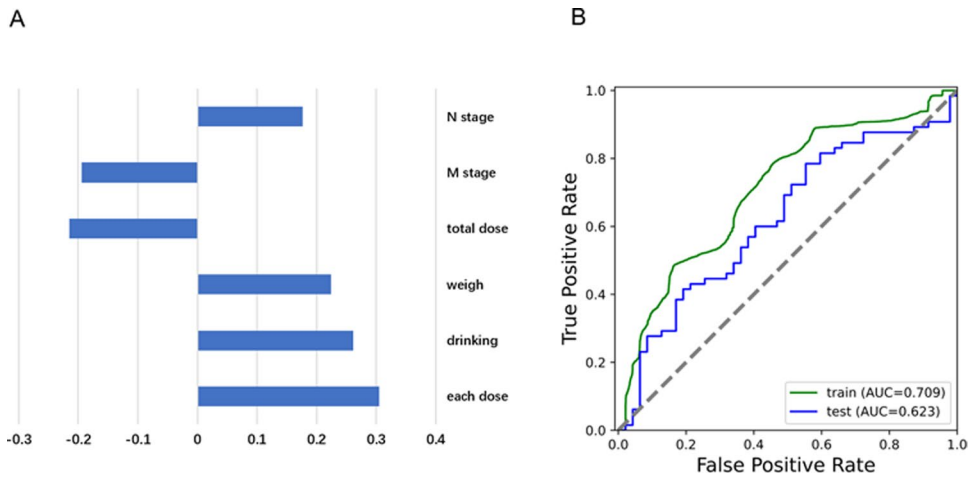

Fig. 2 Best-performance clinic-based model for predicting ppCR. **A** Feature coefficients. **B** ROC curve of training set and testing set

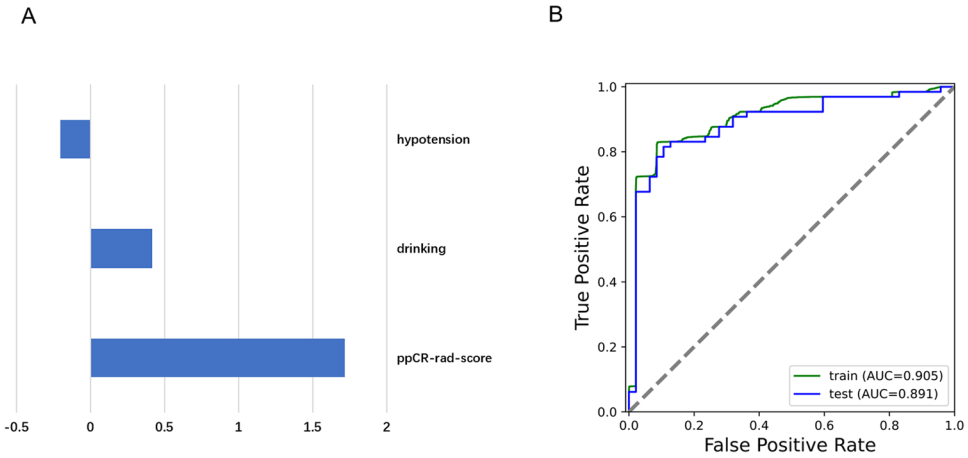

Fig. 3 Best-performance combined model for predicting ppCR. **A** Feature coefficients. **B** ROC curve of training set and testing set

and tpCR. The AUC for ppCR prediction model reached 0.891 in the testing set, and AUC for tpCR prediction model reached 0.814 in the testing set. During model training process, LOOCV strategy was applied to reduce the deviation. LOOCV needs extraordinarily more computing time during model training, whereas it is able to diminish deviation induced by training/testing set splitting, especially when the sample size is small [17, 18]. In the past few years, LOOCV has been widely used in artificial intelligence studies, suggesting the superiority of LOOCV in model training [19–21]. Therefore, we utilized LOOCV during model training, to strengthen the robustness of our model.

A number of studies utilized combination model that included imaging features and clinical features in model training, and obtained better predicting performance [22–24]. For example, Cui et al developed combined CT radiomics learning models (AUC: 0.833 for PFS and 0.768 for OS) with better performance than radiomics (AUC: 0.676 for PFS and 0.646 for OS) or clinical alone models (AUC: 0.823 for PFS and 0.695 for OS), in predicting 3 years progression-free survival (PFS) and OS of non-surgical ESCC patients [25]. However, limited studies used combined model to predict the pCR of EC. In our study, AUC for predicting ppCR was 0.623 in the clinic-based model, and 0.817 in the radiomic model, and AUC for predicting tpCR was 0.624 in the clinic-based model, and 0.713 in the radiomic model. Combination model that included rad-score and clinical features displayed with higher AUCs in predicting both ppCR (AUC: 0.891) and tpCR (AUC: 0.814), suggesting that predicting performance improved when combining imaging features with clinical features. For combination model of predicting ppCR, rad-score was the most important feature; other clinical features were drinking and hypotension. For combination model of predicting tpCR, rad-score remained the most important feature, other important features included M staging, side effects, and TNM staging. In this study, patients were classified as M1 because of supraclavicular lymph node metastases or abdominal lymph node metastases. Those patients were less likely to achieve tpCR. In addition, non-tpCR patients had higher adverse event rates. One possible explanation is that patients with adverse events may decrease their chemotherapy intensity or shorten their treatment cycle; thus, they were less likely to achieve tpCR. As for TNM staging, patients with higher TNM staging indicated increased tumor burdens, and elevated tumor burdens correlated with low probability of tpCR. Some other clinicopathologic factors had also been

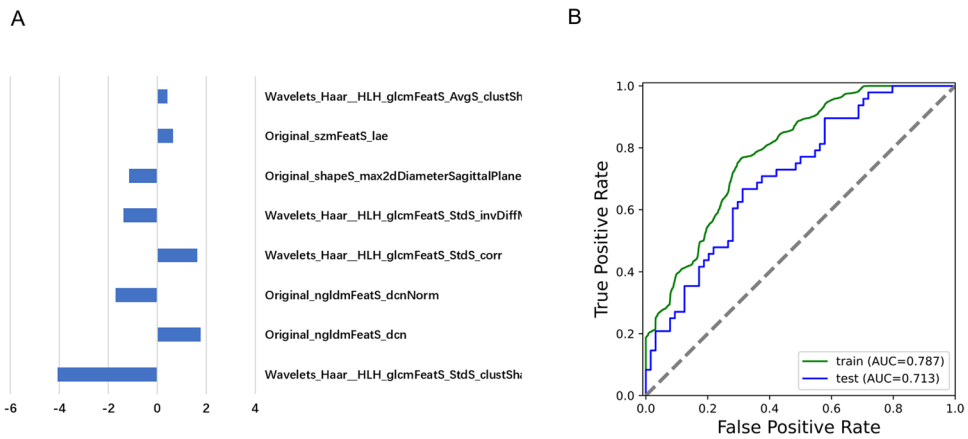

Fig. 4 Best-performance radiomic model for predicting total tpCR. **A** Feature coefficients. **B** ROC curve of training set and testing set

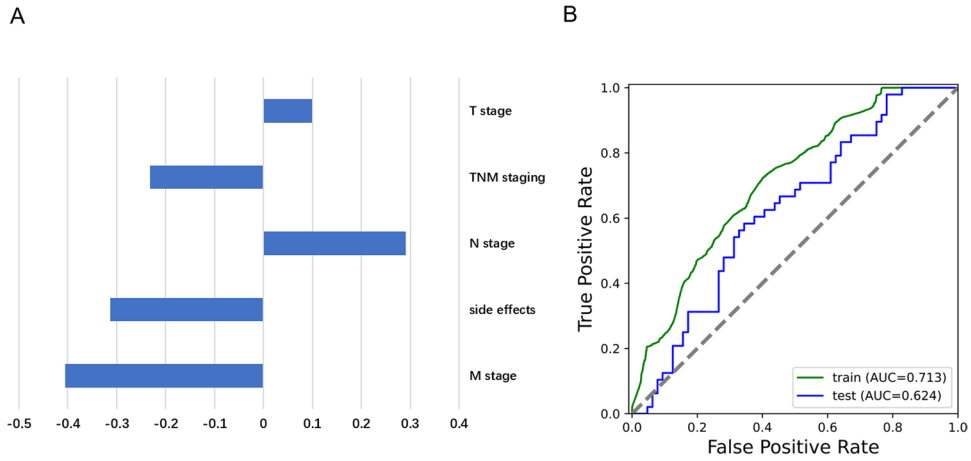

Fig. 5 Best-performance clinic-based model for predicting total tpCR. **A** Feature coefficients. **B** ROC curve of training set and testing set

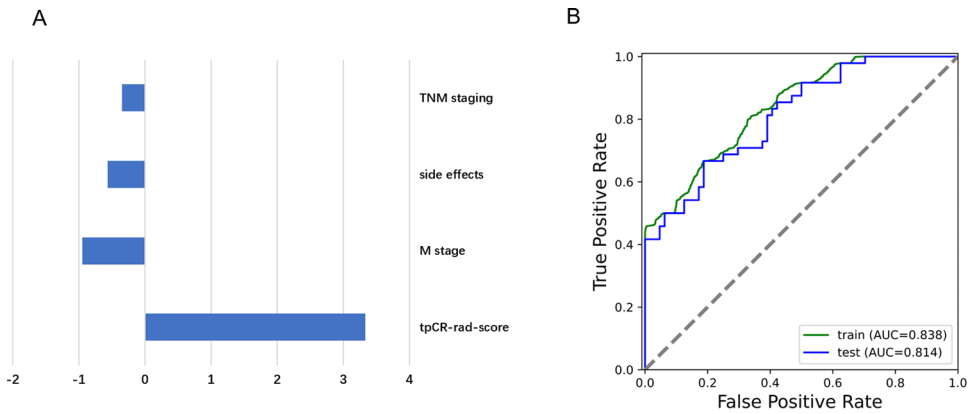

Fig. 6 Best-performance combined model for predicting total tpCR. **A** Feature coefficients. **B** ROC curve of training set and testing set

Table 2 Model evaluation

Y	Model type	Data set	Feature Number	AUC	TPR	TNR	PPV	NPV
ppCR	Imaging	Train	12	0.882524	0.846463	0.770617	0.727393	0.874076
ppCR	Imaging	Test	12	0.816694	0.765957	0.753846	0.692308	0.816667
ppCR	Clinical	Train	6	0.70902	0.433199	0.863617	0.696671	0.678167
ppCR	Clinical	Test	6	0.622586	0.361702	0.830769	0.607143	0.642857
ppCR	Combined	Train	3	0.90524	0.90531	0.82966	0.793515	0.923765
ppCR	Combined	Test	3	0.890998	0.87234	0.815385	0.773585	0.898305
tpCR	Imaging	Train	8	0.787436	0.779139	0.57226	0.708344	0.660243
tpCR	Imaging	Test	8	0.713216	0.734375	0.5000	0.661972	0.585366
tpCR	Clinical	Train	5	0.713273	0.693975	0.602853	0.699688	0.596361
tpCR	Clinical	Test	5	0.624023	0.671875	0.5625	0.671875	0.5625
tpCR	Combined	Train	4	0.838432	0.698198	0.741929	0.782952	0.648352
tpCR	Combined	Test	4	0.814128	0.671875	0.708333	0.754386	0.618182

identified as the predictors for pCR [13]. In this study, we also found that clinical features improve predicting performance. Although some clinical features did not exhibit statistical significance between two groups, they may influence treatment tolerance and toxicities, thereby improving predicting performance. Therefore, clinical features are important and should not be neglected during radiomic model training.

We also found that predicting performance was superior for ppCR (0.89 in the test set) compared with tpCR (0.81 in the test set); it was an interesting phenomenon that is worth exploring. As we mentioned before, ppCR was evaluated only on primary tumor lesion, while tpCR was evaluated on primary tumor lesion and lymph nodes around the tumor. Although VOI of lymph nodes was delineated in CT images in the tpCR model, indicating that imaging features also included lymph nodes characteristics, some confounding factors existed when delineating lymph nodes VOI, and may not be easy to be avoided. On one hand, some metastatic lymph nodes were too small and vague to be clearly delineated in CT image; thus, some lymph node features may be missing. On the other hand, some lymph nodes that were not metastatic lymph nodes but benign lymph nodes were also delineated; in this situation, cofounding bias was produced during feature extraction. Therefore, imaging features that reflected metastatic lymph node characteristics may be not accurate enough, and predicting performance of tpCR model was decreased consequently. This phenomenon reminds us that ppCR predicting models may be also taken into consideration as they exhibit better predicting performance compared to corresponding tpCR predicting models.

There were some limitations in this study. Firstly, we only used pretreatment single-phase CT images for machine learning, multi-phase enhanced CT, and multimodality such as MRI and PET/CT which may improve the model reliability; in addition, different timeline images such as posttreatment CT images may also enhance the results. Secondly, this study was one-centered study; further multicentric studies need to be conducted to obtain more reliable predicting models. In addition, some metastatic lymph nodes were too small to delineate, and some benign lymph nodes other than metastatic lymph nodes were also delineated, leading to the deceased accuracy of imaging features that related to metastatic lymph nodes.

Conclusion

In summary, this study built clinical and radiomic feature–based machine learning models for both ppCR and tpCR prediction with good predicting performance, which facilitates pCR prediction of chemoradiotherapy; and ppCR predicting performance was higher compared to that of tpCR. This study provides a new idea for predicting pCR after nCRT in ESCC patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-023-09884-7.

Author contribution WJ: data collection, statistical analysis, and writing and revising the manuscript. ZX, ZJ, LC, SW: statistical analysis and revising the manuscript. ZJ, SXJ: patient administration, and data collection. FJ, LQR: patient administration, and critical revision of the manuscript. JYL and CQX study design, statistical analysis, critical revision of the manuscript, and funds collection.

Funding This work was supported by the grants from the Zhejiang province public welfare funds (No. GF20H160009) and the Medical Science and Technology Project of Zhejiang Province (No. 2020382901, and 2021455891).

Declarations

Guarantor The scientific guarantor of this publication Yongling Ji.

Conflict of interest The authors declare that they have no competing interests.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Not applicable.

Ethical approval The study was approved by our hospital's institutional review board.

Study subjects or cohorts overlap No study subjects or cohorts have been previously reported.

Methodology
- retrospective
- observational
- performed at one institution

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A (2022) Cancer statistics, 2022. CA Cancer J Clin 72: https://doi.org/10.3322/caac.21708
2. Eyck BM, van Lanschot JJB, Hulshof MCCM et al (2021) Ten-year outcome of neoadjuvant chemoradiotherapy plus surgery for esophageal cancer: the randomized controlled CROSS trial. J Clin Oncol 39:1995–2004. https://doi.org/10.1200/JCO.20.03614
3. Yang H, Liu H, Chen Y, et al (2018) Neoadjuvant chemoradiotherapy followed by surgery versus surgery alone for locally advanced squamous cell carcinoma of the esophagus (NEOCRTEC5010): a phase III multicenter, randomized, open-label clinical trial. J Clin Oncol 36: https://doi.org/10.1200/JCO.2018.79.1483
4. Donahue JM, Nichols FC, Li Z, et al (2009) Complete pathologic response after neoadjuvant chemoradiotherapy for esophageal cancer is associated with enhanced survival. Ann Thorac Surg 87: https://doi.org/10.1016/j.athoracsur.2008.11.001
5. Kelly RJ, Ajani JA, Kuzdzal J, et al (2021) Adjuvant nivolumab in resected esophageal or gastroesophageal junction cancer. N Engl J Med 384: https://doi.org/10.1056/nejmoa2032125
6. Dittrick GW, Weber JM, Shridhar R, et al (2012) Pathologic non-responders after neoadjuvant chemoradiation for esophageal cancer demonstrate no survival benefit compared with patients treated with primary esophagectomy. Ann Surg Oncol 19: https://doi.org/10.1245/s10434-011-2078-4

7. Yang Z, He B, Zhuang X, et al (2019) CT-based radiomic signatures for prediction of pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant chemoradiotherapy. J Radiat Res 60: https://doi.org/10.1093/jrr/rrz027
8. Qu J, Ma L, Lu Y, et al (2022) DCE-MRI radiomics nomogram can predict response to neoadjuvant chemotherapy in esophageal cancer. Discover Oncol 13: https://doi.org/10.1007/s12672-022-00464-7
9. van Rossum PSN, van Lier ALHMW, van Vulpen M, et al (2015) Diffusion-weighted magnetic resonance imaging for the prediction of pathologic response to neoadjuvant chemoradiotherapy in esophageal cancer. Radiother Oncol 115: https://doi.org/10.1016/j.radonc.2015.04.027
10. Tan S, Kligerman S, Chen W, et al (2013) Spatial-temporal [18F] FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. Int J Radiat Oncol Biol Phys 85: https://doi.org/10.1016/j.ijrobp.2013.09.037
11. Zhu Y, Yao W, Xu BC, et al (2021) Predicting response to immunotherapy plus chemotherapy in patients with esophageal squamous cell carcinoma using non-invasive Radiomic biomarkers. BMC Cancer 21: https://doi.org/10.1186/s12885-021-08899-x
12. Zhang X, Gari A, Li M, et al (2022) Combining serum inflammation indexes at baseline and post treatment could predict pathological efficacy to anti PD 1 combined with neoadjuvant chemotherapy in esophageal squamous cell carcinoma. J Transl Med 20: https://doi.org/10.1186/s12967-022-03252-7
13. Toxopeus ELA, Nieboer D, Shapiro J, et al (2015) Nomogram for predicting pathologically complete response after neoadjuvant chemoradiotherapy for oesophageal cancer. Radiother Oncol 115: https://doi.org/10.1016/j.radonc.2015.04.028
14. van Rossum PSN, Fried DV, Zhang L, et al (2016) The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. J Nucl Med 57: https://doi.org/10.2967/jnumed.115.163766
15. Hu Y, Xie C, Yang H et al (2021) Computed tomography-based deep-learning prediction of neoadjuvant chemoradiotherapy treatment response in esophageal squamous cell carcinoma. Radiother Oncol 154:6–13. https://doi.org/10.1016/j.radonc.2020.09.014
16. Duong C, Greenawalt DM, Kowalczyk A et al (2007) Pretreatment gene expression profiles can be used to predict response to neoadjuvant chemoradiotherapy in esophageal cancer. Ann Surg Oncol 14:3602–3609. https://doi.org/10.1245/s10434-007-9550-1
17. Cheng H, Garrick DJ, Fernando RL (2017) Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. J Anim Sci Biotechnol 8:38. https://doi.org/10.1186/s40104-017-0164-6
18. Shao Z, Er MJ, Wang N (2016) An efficient leave-one-out cross-validation-based extreme learning machine (ELOO-ELM) with minimal user intervention. IEEE Trans Cybern 46:1939–1951. https://doi.org/10.1109/TCYB.2015.2458177
19. Jeong YS, Jeon M, Park JH et al (2021) Machine-learning-based approach to differential diagnosis in tuberculous and viral meningitis. Infect Chemother 53:53–62. https://doi.org/10.3947/ic.2020.0104
20. Cheng J, Dekkers JCM, Fernando RL (2021) Cross-validation of best linear unbiased predictions of breeding values using an efficient leave-one-out strategy. J Anim Breed Genet 138:519–527. https://doi.org/10.1111/jbg.12545
21. Nagata T, Noyori SS, Noguchi H et al (2021) Skin tear classification using machine learning from digital RGB image. J Tissue Viability 30:588–593. https://doi.org/10.1016/j.jtv.2021.01.004
22. Li M, Jin Y-M, Zhang Y-C et al (2021) Radiomics for predicting perineural invasion status in rectal cancer. World J Gastroenterol 27:5610–5621. https://doi.org/10.3748/wjg.v27.i33.5610
23. Xu W, Wu W, Zheng Y et al (2021) A computed tomography radiomics-based prediction model on interstitial lung disease in anti-MDA5-positive dermatomyositis. Front Med (Lausanne) 8:768052. https://doi.org/10.3389/fmed.2021.768052
24. Sun Y, Li C, Jin L et al (2020) Radiomics for lung adenocarcinoma manifesting as pure ground-glass nodules: invasive prediction. Eur Radiol 30:3650–3659. https://doi.org/10.1007/s00330-020-06776-y
25. Cui Y, Li Z, Xiang M et al (2022) Machine learning models predict overall survival and progression free survival of non-surgical esophageal cancer patients with chemoradiotherapy based on CT image radiomics signatures. Radiat Oncol 17:212. https://doi.org/10.1186/s13014-022-02186-0