Clinical and Translational Radiation Oncology 48 (2024) 100819

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

journal homepage: www.sciencedirect.com/journal/clinical-and-translational-radiation-oncology

Original Research Article

Integrative prediction model for radiation pneumonitis incorporating genetic and clinical-pathological factors using machine learning

Seo Hee Choi [a,1], Euidam Kim [b,1], Seok-Jae Heo [c], Mi Youn Seol [a], Yoonsun Chung [b,*], Hong In Yoon [a,**]

[a] Department of Radiation Oncology, Yonsei Cancer Center, Heavy Ion Therapy Research Institute, Yonsei University College of Medicine, Seoul, Republic of Korea
[b] Department of Nuclear Engineering, Hanyang University, Seoul, Republic of Korea
[c] Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Radiation pneumonitis
Single nucleotide polymorphism
Prediction model
Machine learning
Logistic model

ABSTRACT

Purpose: We aimed to develop a machine learning-based prediction model for severe radiation pneumonitis (RP) by integrating relevant clinicopathological and genetic factors, considering the associations of clinical, dosimetric parameters, and single nucleotide polymorphisms (SNPs) of genes in the TGF-β1 pathway with RP.

Methods: We prospectively enrolled 59 primary lung cancer patients undergoing radiotherapy and analyzed pretreatment blood samples, clinicopathological/dosimetric variables, and 11 functional SNPs in TGFβ pathway genes. Using the Synthetic Minority Over-sampling Technique (SMOTE) and nested cross-validation, we developed a machine learning-based prediction model for severe RP (grade $\geq$ 2). Feature selection was conducted using four methods (filtered-based, wrapper-based, embedded, and logistic regression), and performance was evaluated using three machine learning models.

Results: Severe RP occurred in 20.3 % of patients with a median follow-up of 39.7 months. In our final model, age (>66 years), smoking history, PTV volume (>300 cc), and AG/GG genotype in BMP2 rs1979855 were identified as the most significant predictors. Additionally, incorporating genomic variables for prediction alongside clinicopathological variables significantly improved the AUC compared to using clinicopathological variables alone (0.822 vs. 0.741, p = 0.029). The same feature set was selected using both the wrapper-based method and logistic model, demonstrating the best performance across all machine learning models (AUC: XGBoost 0.815, RF 0.805, SVM 0.712, respectively).

Conclusion: We successfully developed a machine learning-based prediction model for RP, demonstrating age, smoking history, PTV volume, and BMP2 rs1979855 genotype as significant predictors. Notably, incorporating SNP data significantly enhanced predictive performance compared to clinicopathological factors alone.

Introduction

Thoracic radiotherapy (RT) has long been the standard treatment for unresectable or locally advanced non-small cell lung cancer (NSCLC) and limited-stage small cell lung cancer (SCLC). However, the lungs are particularly sensitive to radiation, making radiation pneumonitis (RP) a significant side effect with an incidence ranging from 0 to 30 %. Therefore, predicting RP on an individual basis could optimize therapeutic strategies and minimize side effects. Both conventional statistical modeling and recent machine learning (ML) approaches have advanced the prediction of clinical outcomes like RP [1]. ML is effective in recognizing patterns within clinical data to predict disease progression and treatment side effects, such as RP, radiation-induced lymphopenia, and mucositis. This demonstrates its broad applicability in RT-related research [2–5].

Despite advances in RT techniques, the occurrence of RP remains a

* Corresponding author at: Department of Nuclear Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea.

** Corresponding author at: Department of Radiation Oncology, Yonsei Cancer Center, Heavy Ion Therapy Research Institute, Yonsei University College of Medicine, Yonsei University Health System, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea.

E-mail addresses: clickby_s@yuhs.ac (S.H. Choi), damdam1996@hanyang.ac.kr (E. Kim), sjheo@yuhs.ac (S.-J. Heo), myseol@yuhs.ac (M.Y. Seol), ychung@hanyang.ac.kr (Y. Chung), YHI0225@yuhs.ac (H.I. Yoon).

[1] These authors contributed equally to this paper.

https://doi.org/10.1016/j.ctro.2024.100819
Received 24 May 2024; Received in revised form 30 June 2024; Accepted 17 July 2024
Available online 26 July 2024

major clinical challenge. Severe RP not only impacts the quality of life but can also be life-threatening, emphasizing the need for accurate prediction models to mitigate risks. Traditional models often rely on clinical and dosimetric factors such as mean lung dose (MLD), lung volume exposed to specific dose thresholds, performance status, smoking history, and concurrent chemotherapy [6–10]. However, these models have shown limited predictive power, which underscores the necessity to explore additional, more predictive factors.

The interpatient variability in RP susceptibility, largely attributable to differences in radiosensitivity influenced by genetic variations, underscores this necessity. Studies suggest that transforming growth factor β1 (TGF-β1), a cytokine involved in inflammation, plays a crucial role in radiation-induced lung injury, including RP [11–18]. Genetic variables such as single nucleotide polymorphisms (SNPs) in TGF-β1 or BMP genes are significantly associated with changes in pulmonary function and treatment outcomes in patients with lung cancer [19–22]. We believe that recent ML approaches are optimal for developing the most accurate model by identifying the most significant genetic biomarkers within complex clinical data, considering their relationship with clinical factors, and enhancing predictive accuracy.

In this study, we aimed to develop an ML-based prediction model for severe RP by integrating relevant clinicopathological and genetic factors. Specifically, we sought to assess the utility of genomic variables in enhancing the predictive performance for severe RP, compare the predictive performance of dosimetric and clinical variables with and without the inclusion of genomic variables, and evaluate several ML predictive models and feature selection methods to identify the optimal combination for accurate RP prediction. This study has the potential to advance personalized medicine in RT. By incorporating genetic factors into predictive models, we can move towards more individualized treatment plans that minimize the risk of severe side effects, thereby improving the overall quality of care for lung cancer patients undergoing thoracic RT.

Patients and methods

Patient population

This prospective study was approved by the Institutional Review Board of our institute (4-2017-0382). Patient eligibility criteria included: histologically confirmed lung cancer; absence of distant metastasis; plan to receive definitive or preoperative RT; availability for pretreatment blood sampling and regular follow-up visits; and no previous history of thoracic RT. Patients who received stereotactic body radiation therapy were excluded. Patients with a history of previous thoracic RT or stereotactic body radiation therapy were excluded to form a homogeneous patient cohort aligned with our study objectives. This exclusion was due to the potential for different radiobiological impacts on the lungs and limitations of applying the same LQ model compared to conventional fractionation RT [23,24]. In total, 59 patients treated with thoracic RT between 2017 and 2019 fulfilled these criteria and were prospectively enrolled in this study. Informed consent obtained from patients at the time of enrollment. Pretreatment blood samples for genotyping were collected for each patient. Data on clinical variables, including age, sex, histology, smoking history, stage, results of pulmonary function tests (forced expiratory volume in one second (FEV1) to forced vital capacity (FVC) ratio (FEV1/FVC) and diffusing capacity of the lung for carbon monoxide (DLCO)), and dosimetric data (total dose, MLD, median lung volume receiving at least 20 Gy (V20), median lung volume receiving at least 5 Gy (V5), and volume of the planning target volume (PTV)), were also collected.

The primary endpoint was the development of 'severe RP', defined as grade 2 or higher RP. This was assessed and scored using the Common Terminology Criteria for Adverse Events (CTCAE) version 5.0. RP monitoring involved a combination of clinical examinations, symptom evaluations, chest X-rays, and follow-up chest CT scans conducted at intervals of 1, 3, and 6 months after RT, as required, and every 6–12 months thereafter.

Genotyping methods

For genotyping, we selected 11 functional SNPs across three genes (TGF-β1, BMP2, and BMP4) that are critical for the TGFβ pathways. The SNPs genotyped within the TGF-β1 gene included rs1800469 C>T, rs1800471 G>C, rs1982073 T>C, and rs11466345 A>G. The genotyped SNPs in the BMP2 gene were rs235768 A>T, rs3178250 T>C, rs1979855 A>G, and rs170986 C>A. The genotyped SNPs within the BMP4 gene were rs17563 T>C, rs4898820 T>G, and rs762642 T>G. Whole blood was collected from each patient before the start of RT, and genomic DNA was subsequently extracted from the fresh blood samples. A detailed description of the genomic DNA extraction method can be found in Supplementary text and Supplementary Table 1.

Genomic variables in severe radiation pneumonitis prediction

We constructed two logistic regression models, one using only clinical and dosimetric variables ("clinical model") and another incorporating genomic variables ("combined model") to investigate the importance of genomic variables in predicting severe RP. For both models, we selected variables highly correlated to severe RP, implementing the best subset selection algorithm with L0 and L2 regularizations [25–27]. We compared the predictive performance of these models for severe RP using a 5-fold cross-validation. The predictive performance of the models was assessed using the area under the receiver operating characteristic curve (ROC-AUC), and statistical differences between the two models for each fold were determined using paired t-tests. A value of $p < 0.05$ was considered statistically significant. All statistical analyses were performed using R software (version 4.2.0; https://www.r-project.org; R Foundation for Statistical Computing, Vienna) and SPSS 25.0 statistical software (SPSS Inc, Chicago, IL, USA).

Machine learning-based severe radiation pneumonitis prediction: SMOTE and nested cross-validation

Data regarding disease occurrence, like ours, often exhibit significant class imbalances, reflecting the nature that patients with the disease typically outnumber those without. To address this, we employed the Synthetic Minority Over-sampling Technique (SMOTE) [28] to augment the number of instances in the minority class, thereby balancing the dataset and creating a more equitable training environment for our models.

We employed Nested Cross-Validation (Nested CV) and SMOTE simultaneously to mitigate data leakage and reduce bias from the imbalanced dataset. Nested CV, a method for accurately evaluating a model's generalization performance with optimized hyperparameters, involves dividing the entire dataset into outer training and test sets. This division occurs in the outer loop, with further training/evaluation or cross-validation in the inner loop, treating the outer training set as the complete dataset. In our Nested CV approach, we ensured a consistent proportion of patients with non-severe RP to severe RP by applying stratified K-fold in the outer loop. SMOTE was then exclusively used on the training set of the outer loop, adjusting the number of patients with severe RP and non-severe RP to prevent data leakage caused by SMOTE [Fig. 1].

Machine learning-based severe radiation pneumonitis Prediction: Feature selection and performance evaluation

For the SMOTE-augmented outer training set, we identified relevant feature sets from the patient and SNP data using four feature selection methods: filter-based, wrapper-based, embedded, and multivariate logistic regression. In the filter-based approach, we iteratively removed

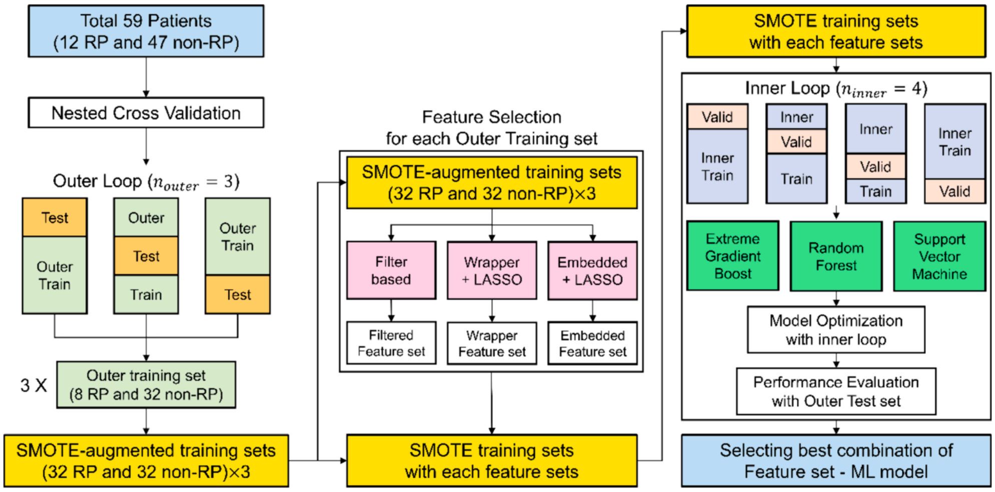


Fig. 1. Overall flow of machine learning-based severe radiation pneumonitis prediction.

the feature with the highest variance inflation factor (VIF) continuing until all features had a VIF below a predefined threshold (traditionally 10, however, we opted for 5 to achieve stricter criteria) to reduce multicollinearity among features. For both the wrapper-based and embedded methods, we utilized a LASSO-based logistic regression model [29], tuned on the outer loop training set via Grid Search with Cross-Validation (GridsearchCV, *sklearn.model_selection.GridSearchCV*). In the wrapper-based method [30], Recursive Feature Elimination with Cross-Validation (RFECV, *sklearn.feature_selection.RFECV*) was used to iteratively discard features, selecting the optimal set, while the embedded method employed the SelectFromModel (*sklearn.feature_selection.SelectFromModel*) function to choose features with above-average importance. This process was repeated across each fold of the Nested CV, selecting features consistently recognized as relevant more frequently than the average across all folds for the final feature sets for each method.

Subsequently, we assessed each feature set with three ML models: Extreme Gradient Boosting (XGBoost, *xgboost.sklearn.XGBClassifier*) [31], Random Forest (RF, *sklearn.ensemble.RandomForestClassifier*) [32], and Support Vector Machine (SVM, *sklearn.svm.SVC*) [33], utilizing another round of Nested CV for both hyperparameter optimization and performance evaluation. This evaluation was repeated five times, each with a different random seed to ensure a thorough assessment. Through this repetition, the feature set and model with the highest performance among nine combinations of three types of feature sets and models, were identified [Fig. 1]. Finally, we verified our identified feature set using Kaplan–Meier analysis and the log-rank test to assess the impact of different prognostic factors in our feature set on the cumulative probability of severe RP.

The Nested CV, applied for both feature selection and performance evaluation, was organized with four folds in the inner loop and three folds in the outer loop. All cross-validation and hyperparameter tuning were aimed at maximizing the AUC score. All ML-based analysis was performed using Scikit-learn package (version 1.3.0) [34] on Python 3.7.

Results

Patient characteristics and incidence of radiation pneumonitis

Table 1 presents the characteristics of the 59 patients; 43 men and 16 women, with a median age of 66 years (range, 29–82). Twenty-three patients received a median 63.0 (range, 55.5–70.9) Gy of definitive RT (with concurrent chemotherapy except one patient) and 36 patients received a median 46.1 (range, 40.0–66.0) Gy of preoperative RT (all with concurrent chemotherapy). Regarding the concurrent chemotherapy regimen, all patients with NSCLC received chemotherapy consisting of a paclitaxel and carboplatin regimen, while all with SCLC received chemotherapy consisting of a cisplatin and etoposide regimen. Supplementary Fig. 1 illustrates the genotype distribution within the SNPs of the TGF-β1, BMP2, and BMP4 genes. SNPs rs1800471 of TGF-β1 or rs170986 of BMP2 were not detected in any patient.

The overall incidence of RP was 47.5 % (28 patients) at a median follow-up of 39.7 (range, 2.3–77.2) months. Severe RP occurred in 12 (20.3 %) patients at a median of 2.8 months after RT. Any grade and severe RP occurred in 78.3 % and 34.8 % patients who received definitive RT, respectively. Any grade RP and severe RP occurred in 27.8 % and 11.1 % patients who received preoperative RT, respectively. No significant differences were observed in clinicopathological factors, dosimetric parameters, or frequency in the distribution of genotypes when patients were divided according to the occurrence of severe RP [Table 1]. Also, there were no significant differences based on whether concurrent chemotherapy was administered or based on the chemotherapy regimen ($p = 0.610$ and 0.668, respectively). The correlation plot examining clinical and dosimetric variables, and genomic variables in relation to severe RP is illustrated in Supplementary Fig. 2.

Genomic variables in severe radiation pneumonitis prediction

In our analysis, age (>66 years), smoking history, and PTV volume (≥300 cc) were selected to be the best variable subset in the clinical model, and age (>66 years), smoking history, PTV volume (≥300 cc), and BMP2 rs1979855 (AG/GG genotype) were selected to be the best variable subset in the combined model. As shown in Table 2, the combined model yielded a significantly ($p = 0.029$) higher performance in severe RP prediction (0.822 ± 0.158) than the clinical model (0.741 ± 0.121).

Comparison of feature selection models and performance using machine learning

The features selected by each method are summarized in Table 3. For filter-based, wrapper-based, and embedded feature selection methods, the average selection frequency of all features across three outer splits was 2.6, 1.3, and 2.2, respectively. Thus, variables selected in all three outer splits by the filter-based method, at least two times by the wrapper-based method, and in all three instances by the embedded method were selected to be a feature set of each method. Among these feature sets, the wrapper-based feature set (age (>66 years), smoking history, PTV volume (≥300 cc), and BMP2 rs1979855) was identical to

Table 1
Patients' characteristics.

Variables	Total	Occurrence of severe radiation pneumonitis	
	(n = 59)	Yes (n = 12)	No (n = 47)
Age (years)			
>66	29 (49.2 %)	9 (75.0 %)	20 (42.6 %)
≤66	30 (50.8 %)	3 (25.0 %)	27 (57.4 %)
Sex			
Male	43 (72.9 %)	11 (91.7 %)	32 (68.1 %)
Female	16 (27.1 %)	1 (8.3 %)	15 (31.9 %)
Pathology			
Adenoca	37 (62.7 %)	6 (50.0 %)	31 (66.0 %)
SCCa	19 (32.2 %)	6 (50.0 %)	13 (27.7 %)
SCLC	2 (3.4 %)	0 (0.0 %)	2 (4.3 %)
Others	1 (1.7 %)	0 (0.0 %)	1 (2.1 %)
T stage			
T1	9 (15.3 %)	1 (8.3 %)	8 (17.0 %)
T2	15 (25.4 %)	3 (25.0 %)	12 (25.5 %)
T3	24 (40.7 %)	7 (58.3 %)	17 (36.2 %)
T4	11 (18.6 %)	1 (8.3 %)	10 (21.3 %)
N stage			
N0	4 (6.8 %)	2 (16.7 %)	2 (4.3 %)
N1	8 (13.6 %)	0 (0.0 %)	8 (17.0 %)
N2	32 (54.2 %)	6 (50.0 %)	26 (55.3 %)
N3	15 (25.4 %)	4 (33.3 %)	11 (23.4 %)
RT total dose (cGy)*	5,440.34 ± 884.84	5,804.17 ± 965.05	5,347.45 ± 849.20
MLD (Gy)*	10.75 ± 4.14	11.98 ± 3.91	10.44 ± 4.18
V20			
<30 %	55 (93.2 %)	10 (83.3 %)	45 (95.7 %)
≥30 %	4 (6.8 %)	2 (16.7 %)	2 (4.3 %)
V5			
<60 %	48 (81.4 %)	8 (66.7 %)	40 (85.1 %)
≥60 %	11 (18.6 %)	4 (33.3 %)	7 (14.9 %)
FEV1/FVC			
<0.7	46 (78.0 %)	9 (75.0 %)	37 (78.7 %)
≥0.7	13 (22.0 %)	3 (25.0 %)	10 (21.3 %)
DLCO			
<0.8	48 (81.4 %)	11 (91.7 %)	37 (78.7 %)
≥0.8	11 (18.6 %)	1 (8.3 %)	10 (21.3 %)
Smoking history			
No	33 (55.9 %)	3 (25.0 %)	30 (63.8 %)
Yes	26 (44.1 %)	9 (75.0 %)	17 (36.2 %)
PTV volume (cc)			
<300	30 (50.8 %)	3 (25.0 %)	27 (57.4 %)
≥300	29 (49.2 %)	9 (75.0 %)	20 (42.6 %)
Concurrent chemotherapy			
Paclitaxel/ Carboplatin	56 (94.9 %)	12 (100.0 %)	44 (93.6 %)
Cisplatin/Etoposide	2 (3.4 %)	0 (0.0 %)	2 (4.3 %)
No	1 (1.7 %)	0 (0.0 %)	1 (2.1 %)
TGF- β1 rs1800469			
CC	21 (35.6 %)	3 (25.0 %)	18 (38.3 %)
CT/TT	38 (64.4 %)	9 (75.0 %)	29 (61.7 %)
TGF-β1 rs1982073			
TT	20 (33.9 %)	3 (25.0 %)	17 (36.2 %)
CT/CC	39 (66.1 %)	9 (75.0 %)	30 (63.8 %)
TGF-β1 rs11466345			
AA	33 (55.9 %)	5 (41.7 %)	28 (59.6 %)
AG/GG	26 (44.1 %)	7 (58.3 %)	19 (40.4 %)
BMP2 rs235768			
AA	1 (1.7 %)	0 (0.0 %)	1 (2.1 %)
AA/TT	58 (98.3 %)	12 (100.0 %)	46 (97.9 %)
BMP2 rs3178250			
TT	14 (23.7 %)	2 (16.7 %)	12 (25.5 %)
CT/CC	45 (76.3 %)	10 (83.3 %)	35 (74.5 %)
BMP2 rs1979855			
AA	37 (62.7 %)	4 (33.3 %)	33 (70.2 %)
AG/CC	22 (37.3 %)	8 (66.7 %)	14 (29.8 %)
BMP4 rs17563			
TT	22 (37.3 %)	5 (41.7 %)	17 (36.2 %)
TC/CC	37 (62.7 %)	7 (58.3 %)	30 (63.8 %)
BMP4 rs4898820			
TT	12 (20.3 %)	2 (16.7 %)	10 (21.3 %)
TG/GG	47 (79.7 %)	10 (83.3 %)	37 (78.7 %)
BMP4 rs762642			
TT	17 (28.8 %)	2 (16.7 %)	15 (31.9 %)
TG/GG	42 (71.2 %)	10 (83.3 %)	32 (68.1 %)

SCC, squamous cell carcinoma; SCLC, small cell lung cancer; RT, radiotherapy; MLD, mean lung dose; FEV1/EVC, the forced expiratory volume in 1 s (FEV1) divided by the forced vital capacity (FVC); DLCO, diffusing capacity of the Lung for CO; PTV, planning target volume; TGF-β1, Transforming Growth Factor-β1; BMP, bone morphogenetic protein.
*Expressed as mean ± standard deviation.

the subsets identified by logistic regression, and also included in the embedded-based feature set, as shown in Table 3.

In terms of performance evaluation, the results of five independent iterations of the entire Nested CV for each feature set and model are shown in Fig. 2. Overall, the wrapper-based feature set (identical to the logistic-based subset) exhibited the best performance across all feature sets (AUC 0.815, 0.805, 0.712 with XGB, RF, SVM model, respectively), followed by the embedded, and filtered feature sets. Among the models tested with the wrapper-based feature set, the XGBoost model demonstrated the highest performance (AUC 0.815 ± 0.078, 95 % confidence interval (CI) 0.772–0.858), with the RF model showing comparable results (AUC 0.805 ± 0.083, 95 % CI 0.758–0.851). However, the SVM model exhibited significant variability in performance, indicating less consistency compared to the other models (AUC 0.712 ± 0.244, 95 % CI 0.577–0.847).

The result of the Kaplan–Meier analysis and log-rank test for cumulative probability of severe RP according to presence of each of the four variables (age, smoking history, PTV volume, and BMP2 rs1979855) in the wrapper-based feature set and logistic-based subset are shown in Fig. 3. Age, smoking history, PTV volume, and BMP2 rs1979855 genotype showed p-values of 0.062, 0.021, 0.031, 0.036, respectively.

Discussion

In this study, we assessed whether genomic variables enhance the performance of RP prediction considering clinical, dosimetric and genomic variables. We found that genetic variants within the TGFβ signaling pathways significantly improve the predictive accuracy for radiation-induced pulmonary toxicities beyond conventional clinical and dosimetric factors. Specifically, our ML model identified that patients >66 years old, with a history of smoking, large PTV, and the BMP2 rs1979855 genetic variation, were notably more susceptible to severe RP following thoracic RT. To the best of our knowledge, this is the first study to confirm the crucial role of genomic variables in predicting severe RP using ML techniques. Notably, the wrapper-based method consistently yielded the highest performance across all ML algorithms, selecting features that closely aligned with those highlighted in the logistic regression analysis. Furthermore, these key features individually exhibited significant associations with severe RP in the log-rank tests for cumulative probability. This highlights not only the efficacy of the selection process but also the indispensable contribution of these variables to the model's predictive performance.

RP is characterized by inflammatory tissue damage, repair processes, and pulmonary fibrosis involving alveolar cells, endothelium, and fibroblasts, followed by inflammation, cell proliferation, and extracellular matrix remodeling. Studies have assessed plasma TGF-β1 levels as a potential predictor of radiation therapy toxicities, particularly focusing on RP [17,18,35]. In pursuit of establishing a more reliable correlation, some researchers investigated TGF-β1 genotypes, particularly TGF-β1 polymorphisms, and their association with normal tissue toxicities. Notably, polymorphisms such as T869C and G915C in TGF-β1 result in

Table 2
Selected variables in clinical model (using only clinical and dosimetric variables) and combined model (incorporating genomic variables), and comparison of performance using AUC obtained by the 5-folds cross-validation.

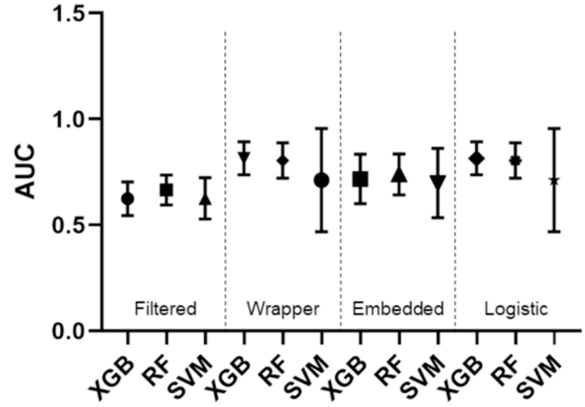
Models	Selected variables	Prediction AUC						P value*
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall	
Clinical Model	Age (>66 years) Smoking history PTV volume (≥300 cc)	0.800	0.806	0.778	0.525	0.796	0.741 ± 0.121	0.029
Combined Model	Age (>66 years) Smoking history PTV volume (≥300 cc) BMP2 rs197985	0.867	0.944	0.917	0.550	0.833	0.822 ± 0.158	

PTV, planning target volume; BMP, bone morphogenetic protein.
*Obtained by paired *t*-test.

Table 3
Selected features by each feature selection method.

	Filtered-based (n = 14)	Wrapper-based (n = 4)	Embedded (n = 14)
Age		○	○
Sex	○		
Pathology			○
T stage			
N stage			○
RT total dose	○		
RT fractional dose	○		
MLD	○		○
V20	○		
V5			
FEV1/FVC	○		
DLCO	○		
Smoking history		○	○
PTV volume		○	○
Concurrent chemotherapy			
Chemotherapy regimen			
TGF-β1 rs1800469			○
TGF-β1 rs1982073			
TGF-β1 rs11466345	○		○
BMP2 rs235768			
BMP2 rs3178250			
BMP2 rs1979855	○	○	○
BMP4 rs17563	○		○
BMP4 rs4898820			
BMP4 rs762642			

RT, radiotherapy; MLD, mean lung dose; FEV1/EVC, the forced expiratory volume in 1 s (FEV1) divided by the forced vital capacity (FVC); DLCO, diffusing capacity of the Lung for CO; PTV, planning target volume; TGF-β1, Transforming Growth Factor-β1; BMP, bone morphogenetic protein.



	Filtered			Wrapper			Embedded			Logistic		
	XGB	RF	SVM	XGB	RF	SVM	XGB	RF	SVM	XGB	RF	SVM
AUC (mean)	0.624	0.665	0.625	0.815	0.805	0.712	0.717	0.738	0.698	0.815	0.805	0.712

Fig. 2. The predictive performance (area under the curve (AUC): mean ± standard deviation) of the different combinations of feature selection methods and machine learning models, as evaluated through repeated nested cross-validation with SMOTE.

amino acid substitutions, potentially altering TGF-β1 function and repair kinetics, thereby affecting susceptibility to adverse effects [36]. Furthermore, TGF-β1 C509T and T869C SNPs have been linked to higher rates of severe fibrosis and RP [16,19,37]. Meta-analyses [38,39] also suggest that the T869C polymorphism is associated with an increased RP risk in Caucasians, with no similar association found for C509T and G915C polymorphisms.

BMP, a key member of the TGF-β superfamily, activates downstream signaling genes in TGF-β pathways, influencing inflammatory processes, cell proliferation, differentiation, apoptosis, and organ patterning [40,41]. Additionally, BMP genes can function as tumor suppressors or promoters, depending on the cell type, epigenetic background, or tumor stage [42]. Specifically, BMPs can antagonize TGF-β's effects on epithelial-to-mesenchymal transition (EMT) and induce the inverse process of mesenchymal-to-epithelial transition [43]. For example, it was shown that BMP-7 reverses EMT by counteracting TGF-β-induced Smad-dependent cell signaling [44]. The BMP2 and BMP4 genes have been investigated in relation to lung diseases. Although BMP-2 and BMP-4 share high sequence similarity and likely act on the same receptors, their biological roles may differ. BMP2 exerts pro-inflammatory effects in endothelial activation during airway inflammation, whereas BMP4 has anti-inflammatory effects in airway injury [45,46]. Previous studies have found that BMP2 levels are decreased and BMP4 levels are increased in idiopathic lung fibrosis, with the ratio between BMPs and TGF-β1 strongly correlating with the induction of EMT [47]. Although more research is needed to elucidate the exact mechanism, it is possible that the increase in BMP4 levels induces the inverse process of mesenchymal-to-epithelial transition, potentially exacerbating fibrosis.

In our study, a specific SNP in the BMP2 gene was significantly associated with an increased occurrence of severe RP. This may be because SNPs in the BMP2 gene can contribute to higher BMP2 expression [48], potentially exacerbating inflammation and impairing normal pulmonary tissue repair during radiation-induced lung injury. A previous study demonstrated the association between BMP2 SNPs and the incidence of RP for the first time after genotyping and tagging potentially functional SNPs of BMP2 and BMP4 genes [20]. BMP2 rs235768 and rs1980499 were associated with risk of grade ≥ 2 RP; and rs3178250 was associated with the risk of grade ≥ 3 RP in patients with NSCLC after definitive RT. However, no association was found between BMP4 SNPs and RP. It could be explained that aberrant BMP2 expression was caused by the location of BMP2 rs235768 and rs1980499 within the coding region or transcription factor binding sites. Similarly, we identified BMP2 rs1979855 as a significant risk factor for RP, instead of the BMP4. This specific BMP2 gene SNP is suspected to excessively induce

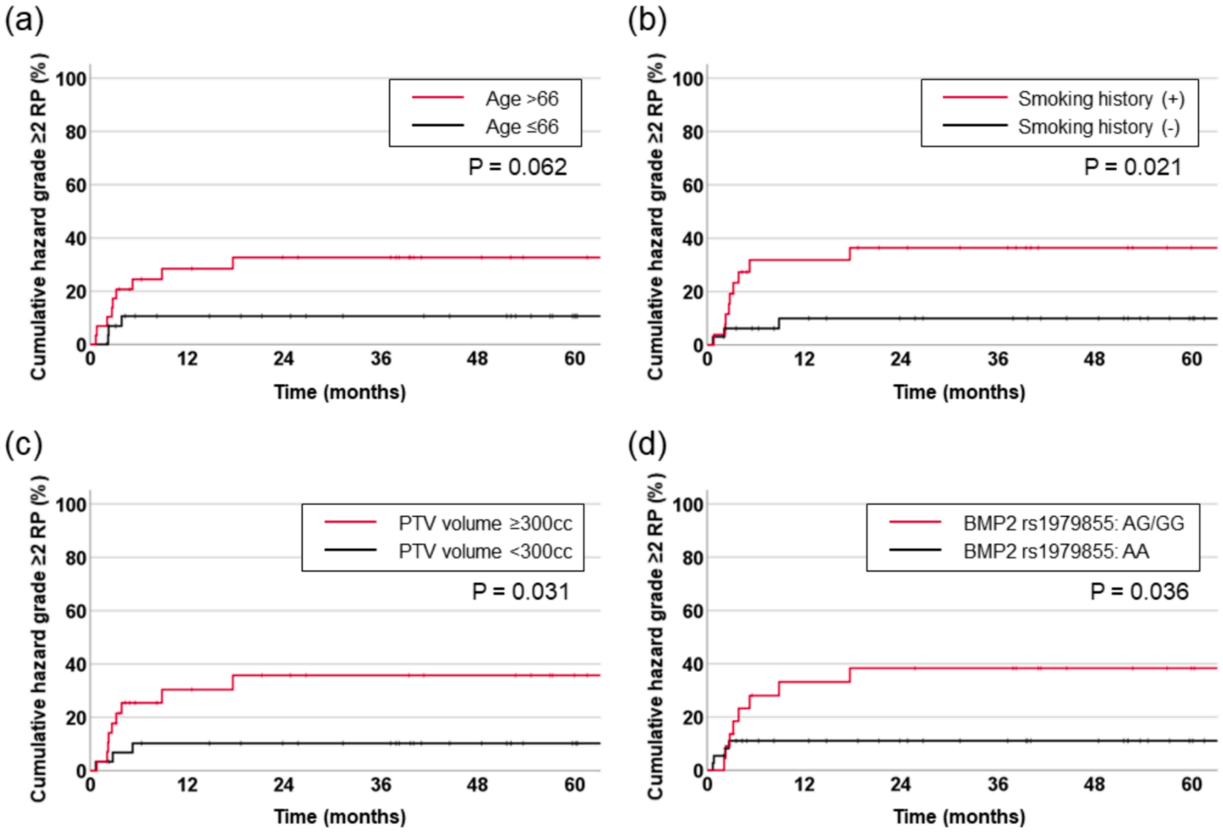


Fig. 3. Cumulative probability of severe (grade ≥ 2) radiation pneumonitis in all patients according to (a) patients' age (>66 year or ≤66 years), (b) smoking history, (c) planning target volume (PTV) volume (≥300 cc vs. <300 cc), and (d) BMP2 rs1979855 genotype (AG/GG vs. AA).

inflammation, thereby increasing the risk of RP. However, further investigation in subsequent studies is required to clarify this mechanism.

Several modeling studies incorporating TGF-β pathway genes as significant factors in predicting treatment outcomes or radiation-induced lung toxicities have been conducted [49]. Zhang et al. [22] identified four SNPs, including BMP2 rs235756, as significant predictors of overall survival, with the predictive power of their model significantly improving after incorporating these SNPs. Yang et al. reported significant associations between SNPs in BMP2 rs3178250 and BMP4 rs762642 with severe RP, developing a survival nomogram that integrated significant SNPs in the BMP/Smad4/Hamp hepcidin-regulating pathway [21,50]. Stenmark et al. [51] demonstrated that combining pretreatment levels of IL-8, TGF-β1, and MLD into a predictive model for RP improved predictive power compared to individual variables, a finding validated in their subsequent study [52]. Chinese researchers also highlighted the significance of BMP2 rs235768 for severe RP, with model performance improving upon adding rs235768 and rs1980499 SNPs to a clinical model comprising age, performance status, and MLD (C-index 0.6117–0.6235, p = 0.011) [20].

There are several limitations to our study that need to be addressed. A major limitation is the small sample size and the focus on a single ethnic group, which may limit the generalizability of our findings. The prognostic implications of some genes may have been underestimated due to unexpected selection bias. Previous studies indicate that allele frequencies of polymorphisms and their effects on RP susceptibility may vary by ethnicity [53]. However, since only one ethnic group was assessed in this study, differences in genetic distributions among ethnicities, including TGF-β1 and other genes, could not be considered. Additionally, to address the low event frequency and data imbalance in our small sample, we employed oversampling techniques such as SMOTE. However, the inherent limitations of building models from a small sample size remain. Finally, there is a lack of preclinical studies assessing the mechanism by which genetic variations in BMP2 rs1979855 impact radiosensitivity. Therefore, future in vivo/in vitro or clinical studies with larger patient cohorts are required to validate our findings.

Nonetheless, unlike existing studies, our study has the distinction of being the first to utilize various ML techniques to maximize the performance of predictive models. We utilized prospectively collected, well-curated homogeneous patient data, with blood samples uniformly collected at the same time, providing a significant advantage. Additionally, we genotyped 11 SNPs related to the TGF-β pathway and consistently identified a significant SNP across various statistical methods. Furthermore, we were able to enhance the reliability of our results by integrating specific clinical and dosimetric variables alongside genomic variables.

In conclusion, we confirmed that SNPs could serve as a reliable biomarker for predicting severe RPs while significantly improving predictive power compared to when only clinical factors were used. Using the SNPs in TGF-β signaling pathway genes in conjunction with age, smoking history, PTV volume, we successfully demonstrated decent performance in severe RP prediction in spite of the small and label-imbalanced patients in our study. We believe that our model will aid in the pretreatment prediction of radiation-related toxicities and enable personalized RT based on each patient's risk profile.

Funding

This work was supported by Ministry of Science, Korea through the R&D program of NRF2017M3A9E8029717. This study was supported by faculty research grants conferred by Yonsei University College of Medicine (6-2023-0070).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

None

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2024.100819.

References

[1] Shamout F, Zhu T, Clifton DA. Machine learning for clinical outcome prediction. IEEE Rev Biomed Eng 2021;14:116–26.

[2] Yu H, Chen F, Lam KO, Yang L, Wang Y, Jin JY, et al. Potential determinants for radiation-induced lymphopenia in patients with breast cancer using interpretable machine learning approach. Front Immunol 2022;13:768811.

[3] Hansen CR, Bertelsen A, Zukauskaite R, Johnsen L, Bernchou U, Thwaites DI, et al. Prediction of radiation-induced mucositis of H&N cancer patients based on a large patient cohort. Radiother Oncol 2020;147:15–21.

[4] Puttanawarut C, Sirirutbunkajorn N, Khachonkham S, Pattaranutaporn P, Wongsawat Y. Biological dosiomic features for the prediction of radiation pneumonitis in esophageal cancer patients. Radiat Oncol 2021;16:220.

[5] Teo PT, Rogacki K, Gopalakrishnan M, Das IJ, Abazeed ME, Mittal BB, et al. Determining risk and predictors of head and neck cancer treatment-related lymphedema: a clinicopathologic and dosimetric data mining approach using interpretable machine learning and ensemble feature selection. Clin Transl Radiat Oncol 2024;46:100747.

[6] Wang S, Liao Z, Wei X, Liu HH, Tucker SL, Hu CS, et al. Analysis of clinical and dosimetric factors associated with treatment-related pneumonitis (TRP) in patients with non-small-cell lung cancer (NSCLC) treated with concurrent chemotherapy and three-dimensional conformal radiotherapy (3D-CRT). Int J Radiat Oncol Biol Phys 2006;66:1399–407.

[7] Yom SS, Liao Z, Liu HH, Tucker SL, Hu CS, Wei X, et al. Initial evaluation of treatment-related pneumonitis in advanced-stage non-small-cell lung cancer patients treated with concurrent chemotherapy and intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys 2007;68:94–102.

[8] Kong FM, Hayman JA, Griffith KA, Kalemkerian GP, Arenberg D, Lyons S, et al. Final toxicity results of a radiation-dose escalation study in patients with non-small-cell lung cancer (NSCLC): predictors for radiation pneumonitis and fibrosis. Int J Radiat Oncol Biol Phys 2006;65:1075–86.

[9] Schallenkamp JM, Miller RC, Brinkmann DH, Foote T, Garces YI. Incidence of radiation pneumonitis after thoracic irradiation: dose-volume correlates. Int J Radiat Oncol Biol Phys 2007;67:410–6.

[10] Robnett TJ, Machtay M, Vines EF, McKenna MG, Algazy KM, McKenna WG. Factors predicting severe radiation pneumonitis in patients receiving definitive chemoradiation for lung cancer. Int J Radiat Oncol Biol Phys 2000;48:89–94.

[11] Anscher MS, Kong FM, Marks LB, Bentel GC, Jirtle RL. Changes in plasma transforming growth factor beta during radiotherapy and the risk of symptomatic radiation-induced pneumonitis. Int J Radiat Oncol Biol Phys 1997;37:253–8.

[12] Rabbani ZN, Anscher MS, Zhang X, Chen L, Samulski TV, Li CY, et al. Soluble TGFbeta type II receptor gene therapy ameliorates acute radiation-induced pulmonary injury in rats. Int J Radiat Oncol Biol Phys 2003;57:563–72.

[13] Zhao L, Sheldon K, Chen M, Yin MS, Hayman JA, Kalemkerian GP, et al. The predictive role of plasma TGF-beta1 during radiation therapy for radiation-induced lung toxicity deserves further study in patients with non-small cell lung cancer. Lung Cancer 2008;59:232–9.

[14] Andreassen CN, Alsner J, Overgaard J, Herskind C, Haviland J, Owen R, et al. TGFB1 polymorphisms are associated with risk of late normal tissue complications in the breast after radiotherapy for early breast cancer. Radiother Oncol 2005;75: 18–21.

[15] De Ruyck K, Van Eijkeren M, Claes K, Bacher K, Vral A, De Neve W, et al. TGFbeta1 polymorphisms and late clinical radiosensitivity in patients treated for gynecologic tumors. Int J Radiat Oncol Biol Phys 2006;65:1240–8.

[16] Quarmby S, Fakhoury H, Levine E, Barber J, Wylie J, Hajeer AH, et al. Association of transforming growth factor beta-1 single nucleotide polymorphisms with radiation-induced damage to normal tissues in breast cancer patients. Int J Radiat Biol 2003;79:137–43.

[17] Anscher MS, Kong FM, Andrews K, Clough R, Marks LB, Bentel G, et al. Plasma transforming growth factor beta1 as a predictor of radiation pneumonitis. Int J Radiat Oncol Biol Phys 1998;41:1029–35.

[18] Anscher MS, Garst J, Marks LB, Larrier N, Dunphy F, Herndon 2nd JE, et al. Assessing the ability of the antiangiogenic and anticytokine agent thalidomide to modulate radiation-induced lung injury. Int J Radiat Oncol Biol Phys 2006;66: 477–82.

[19] Yuan X, Liao Z, Liu Z, Wang LE, Tucker SL, Mao L, et al. Single nucleotide polymorphism at rs1982073:T869C of the TGFbeta 1 gene is associated with the risk of radiation pneumonitis in patients with non-small-cell lung cancer treated with definitive radiotherapy. J Clin Oncol 2009;27:3370–8.

[20] Yang J, Xu T, Gomez DR, Yuan X, Nguyen QN, Jeter M, et al. Polymorphisms in BMP2/BMP4, with estimates of mean lung dose, predict radiation pneumonitis among patients receiving definitive radiotherapy for non-small cell lung cancer. Oncotarget 2017;8:43080–90.

[21] Yang J, Xu T, Gomez DR, Yuan X, Nguyen QN, Jeter M, et al. Nomograms incorporating genetic variants in BMP/Smad4/Hamp pathway to predict disease outcomes after definitive radiotherapy for non-small cell lung cancer. Cancer Med 2018;7:2247–55.

[22] Zhang H, Wang W, Pi W, Bi N, DesRosiers C, Kong F, et al. Genetic variations in the transforming growth factor-beta1 pathway may improve predictive power for overall survival in non-small cell lung cancer. Front Oncol 2021;11:599719.

[23] Song CW, Glatstein E, Marks LB, Emami B, Grimm J, Sperduto PW, et al. Biological principles of stereotactic body radiation therapy (SBRT) and stereotactic radiation surgery (SRS): indirect cell death. Int J Radiat Oncol Biol Phys 2021;110:21–34.

[24] Brown JM. The biology of SBRT: LQ or something new? Int J Radiat Oncol Biol Phys 2018;101:964.

[25] Dedieu A, Hazimeh H, Mazumder R. Learning sparse classifiers: continuous and mixed integer optimization perspectives. J Mach Learn Res 2021;22.

[26] Hazimeh H, Mazumder R, Nonet T. L0Learn: A Scalable Package for Sparse Learning using L0 Regularization. arXiv preprint arXiv:2202.04820 2022.

[27] Hazimeh H, Mazumder R. Fast best subset selection: coordinate descent and local combinatorial optimization algorithms. Oper Res 2020;68:1517–37.

[28] Chawla NVBK, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[29] Tibshirani R. Regression shrinkage and selection via the Lasso. J Royal Stat Soc Ser B-Stat Methodol 1996;58:267–88.

[30] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97: 273–324.

[31] Chen TQ, Guestrin C. XGBoost: A Scalable Tree Boosting System. Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining 2016. https://doi.org/10.1145/2939672.2939785: 785-94.

[32] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[33] Hearst MA. Support vector machines. IEEE Intell Syst Appl 1998;13:18–21.

[34] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[35] Barthelemy-Brichant N, Bosquee L, Cataldo D, Corhay JL, Gustin M, Seidel L, et al. Increased IL-6 and TGF-beta1 concentrations in bronchoalveolar lavage fluid associated with thoracic radiotherapy. Int J Radiat Oncol Biol Phys 2004;58: 758–67.

[36] Elliott RL, Blobe GC. Role of transforming growth factor Beta in human cancer. J Clin Oncol 2005;23:2078–93.

[37] Giotopoulos G, Symonds RP, Foweraker K, Griffin M, Peat I, Osman A, et al. The late radiotherapy normal tissue injury phenotypes of telangiectasia, fibrosis and atrophy in breast cancer patients have distinct genotype-dependent causes. Br J Cancer 2007;96:1001–7.

[38] He J, Deng L, Na F, Xue J, Gao H, Lu Y. The association between TGF-beta1 polymorphisms and radiation pneumonia in lung cancer patients treated with definitive radiotherapy: a meta-analysis. PLoS One 2014;9:e91100.

[39] Shen ZT, Shen JS, Ji XQ, Li B, Zhu XX. TGF-beta1 rs1982073 polymorphism contributes to radiation pneumonitis in lung cancer patients: a meta-analysis. J Cell Mol Med 2016;20:2405–9.

[40] Nickel J, Mueller TD. Specification of BMP signaling. Cells 2019;8.

[41] Vukicevic S, Grgurevic L. Bone morphogenetic proteins in inflammation. Basel: Springer; 2016.

[42] Bach DH, Park HJ, Lee SK. The dual role of bone morphogenetic proteins in cancer. Mol Ther Oncolytics 2018;8:1–13.

[43] Thiery JP, Sleeman JP. Complex networks orchestrate epithelial-mesenchymal transitions. Nat Rev Mol Cell Biol 2006;7:131–42.

[44] Zeisberg M, Shah AA, Kalluri R. Bone morphogenic protein-7 induces mesenchymal to epithelial transition in adult renal fibroblasts and facilitates regeneration of injured kidney. J Biol Chem 2005;280:8094–100.

[45] Csiszar A, Ahmad M, Smith KE, Labinskyy N, Gao Q, Kaley G, et al. Bone morphogenetic protein-2 induces proinflammatory endothelial phenotype. Am J Pathol 2006;168:629–38.

[46] Li Z, Wang J, Wang Y, Jiang H, Xu X, Zhang C, et al. Bone morphogenetic protein 4 inhibits liposaccharide-induced inflammation in the airway. Eur J Immunol 2014; 44:3283–94.

[47] Selman M, Pardo A, Kaminski N. Idiopathic pulmonary fibrosis: aberrant recapitulation of developmental programs? PLoS Med 2008;5:e62.

[48] Li JM, Zhang Y, Ren Y, Liu BG, Lin X, Yang J, et al. Uniaxial cyclic stretch promotes osteogenic differentiation and synthesis of BMP2 in the C3H10T1/2 cells with BMP2 gene variant of rs2273073 (T/G). PLoS One 2014;9:e106598.

[49] Aguado-Barrera ME, Sosa-Fajardo P, Gomez-Caamano A, Taboada-Valladares B, Counago F, Lopez-Guerra JL, et al. Radiogenomics in lung cancer: where are we? Lung Cancer 2023;176:56–74.

[50] Yang J, Xu T, Gomez DR, Nguyen Q, Yuan X, Song Y, Levy LB, Komaki RU, Liao Z. Single nucleotide polymorphisms in BMP2/BMP4/Smad4 are associated with severe radiation pneumonitis in patients receiving definitive radiation therapy for non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2015;93:E440–1.

[51] Stenmark MH, Cai XW, Shedden K, Hayman JA, Yuan S, Ritter T, et al. Combining physical and biologic parameters to predict radiation-induced lung toxicity in

patients with non-small-cell lung cancer treated with definitive radiation therapy. Int J Radiat Oncol Biol Phys 2012;84:e217–22.

[52] Wang S, Campbell J, Stenmark MH, Zhao J, Stanton P, Matuszak MM, et al. Plasma levels of IL-8 and TGF-beta1 predict radiation-induced lung toxicity in non-small cell lung cancer: a validation study. Int J Radiat Oncol Biol Phys 2017;98:615–21.

[53] Wang L, Bi N. TGF-beta1 gene polymorphisms for anticipating radiation-induced pneumonitis in non-small-cell lung cancer: different ethnic association. J Clin Oncol 2010;28:e621–2.