INTERNATIONAL JOURNAL OF
RADIATION ONCOLOGY • BIOLOGY • PHYSICS
www.redjournal.org

CLINICAL INVESTIGATION

Prediction of Symptomatic Radiation Pneumonitis in Lung Cancer Patients: A Radiomics and Dosiomics Machine Learning Approach Using the Prospective Multicenter RTOG 0617 and REQUITE Trials

Lukas M. Reuter, MD, BS,[a,b] Kim M. Kraus, MD, PhD,[a,c,d,g] Stefan M. Fischer, MS,[a,b,e] Danai Pletzer, MD,[a,b] Denise Bernhardt, MD,[a] Stephanie E. Combs, MD,[a,c,d] on behalf of the REQUITE consortium, Julia A. Schnabel, PhD,[b,e,f] and Jan C. Peeken, MD, PhD[a,c,d]

[a]*Department of Radiation Oncology, School of Medicine, TUM Klinikum Rechts der Isar, Technical University of Munich (TUM), Munich, Germany;* [b]*Institute of Machine Learning in Biomedical Imaging (IML), Helmholtz Zentrum München (HMGU) GmbH, German Research Center for Environmental Health, Neuherberg, Germany;* [c]*Institute of Radiation Medicine (IRM), Helmholtz Zentrum München (HMGU) GmbH, German Research Center for Environmental Health, Neuherberg, Germany;* [d]*Partner Site Munich and German Cancer Research Center (DKFZ), Heidelberg, German Cancer Consortium (DKTK), Munich, Germany;* [e]*School of Computation, Information and Technology, Technical University of Munich (TUM), Munich, Germany;* [f]*School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom; and* [g]*Institute for Advanced Study, Technical University Munich, Garching, Germany*

Received Oct 13, 2025; Revised Dec 16, 2025; Accepted for publication Jan 27, 2026

Corresponding author: Lukas M. Reuter, MD, BS; E-mail: lukas.reuter@tum.de

Author Responsible for Statistical Analysis: Lukas M. Reuter was responsible for the statistical analyses.

Lukas M. Reuter and Kim M. Kraus made equal contributions to the study.

Disclosures: K.M.K. received funding for this project from the German Cancer Consortium (DKTK). This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 515279324 (J.C.P., J.A.S.). REQUITE received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant agreement no. 601826.

This article used data from the National Clinical Trials Network (NCTN)/National Cancer Institute (NCI) Community Oncology Research Program data archive of the NCTN of the NCI. The RTOG 0617 data were collected under clinical trial number NCT00533949. All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, or the NCI. Data from the REQUITE study (www.requite.eu) were originally collected under the study number ISRCTN98496463. The creators of the study are not involved in the analysis in this article.

Data Sharing Statement: The datasets analyzed in this study are available from the original study investigators upon reasonable reasonable request and with appropriate permissions.

Acknowledgments—The authors gratefully acknowledge all patients who participated in the Radiogenomics Consortium cohorts and the REQUITE study. We further thank all REQUITE investigators and research staff for their invaluable contributions at the participating institutions: Belgium: Ghent University Hospital, Ghent, and KU Leuven, Leuven; France: ICM Montpellier, and CHU Nîmes; Germany: Zentrum für Strahlentherapie Freiburg (P. Stegmaier); ViDia Christliche Kliniken Karlsruhe (J. Claßen); Klinikum der Stadt Ludwigshafen gGmbH (T. Schnabel); Universitätsmedizin Mannheim; DKFZ also thanks Anusha Müller and Irmgard Helmbold; Italy: Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, and Candiolo Cancer Institute − IRCCS, Candiolo; Spain: Complexo Hospitalario Universitario de Santiago, Santiago de Compostela; United Kingdom: University Hospitals Leicester, Leicester, and Manchester Biomedical Research Center, Manchester; United States: Mount Sinai Hospital, New York. The authors wish to acknowledge the valuable contribution of the REQUITE Publication Committee, including Catharine West, Jenny Chang-Claude, Tony Payton, Chris Talbot, Liv Veldeman, Dirk De Ruysscher, Barry Rosenstein, Tiziana Rancati, Ana Vega, David Azria, Ananya Choudhury, Petra Seibold, Adam Webb, Erik Briers, Hilary Stobart, and Tim Ward.

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ijrobp.2026.01.031.

https://doi.org/10.1016/j.ijrobp.2026.01.031

Purpose: Radiation-induced pneumonitis (RP) is a side effect after thoracic radiation therapy (RT). The ability to predict RP would facilitate treatment modifications. This study investigates the predictive capacity for symptomatic RP (Common Terminology Criteria for Adverse Events ≥ 2) employing Radiomics and Dosiomics models.

Methods and Materials: Computed tomography scans, along with physical and 2-Gy equivalent dose volumes (EQD2), dose-volume histograms, and clinical parameters, were evaluated for 708 multicenter lung cancer patients, among whom 89 developed RP ≥ 2. The training cohort consisted of 441 patients from the prospective RTOG 0617 trial. External validation was carried out on 267 patients from the prospective REQUITE (validating pREdictive models and biomarkers of radiotherapy toxicity to reduce side effects and improve QUalITy of lifE in cancer survivors) study. A Random Forest classifier was employed, with feature selection executed within the inner loop of a 10x5-fold nested cross-validation (nCV) utilizing the minimum-redundancy-maximum-relevance algorithm. To address class imbalances, synthetic oversampling and undersampling were implemented using SMOTE-Tomek. The QUANTEC Normal Tissue Complication Probability model served as a reference. Additionally, the experiments were stratified by subgroups (standard/high-dose and 3-dimensional conformal RT (3D-CRT)/intensity-modulated RT (IMRT).

Results: The best radiomics model identified in the nCV was trained on the standard-dose subgroup achieved a test ROC-AUC of 0.56. The baseline Normal Tissue Complication Probability model showed a predictive performance with a ROC-AUC of 0.56, which was largely dependent on radiation technique (ROC-AUCS: 3D-CRT: 0.75, IMRT: 0.50). The Dosiomics EQD2 model, trained on the full training cohort, attained the second-best performance in the nCV, demonstrating the same technique-dependence (ROC-AUC of 0.75 vs. 0.39). Using a Dosiomics EQD2 ensemble model trained separately on 3D-CRT and IMRT subgroups increased overall performance to a testing ROC-AUC of 0.61, outperforming other modeling strategies for IMRT, while being outperformed by clinical models for 3D-CRT.

Conclusions: This prospective trial-based study reveals an overall limited predictive capacity of radiomics and dosiomics models and a large influence of radiation technique. IMRT-specific models should be investigated further.

Introduction

Radiation therapy (RT) is a cornerstone treatment for locally advanced lung cancer.[1] Despite advancements aimed at protecting healthy lung tissue through increasingly conformal RT techniques, radiation pneumonitis (RP) remains one of the most prevalent complications associated with thoracic RT. The dose-limiting nature of RP not only exacerbates patient morbidity but may also adversely affect tumor control.[2] Clinical manifestations of RP typically encompass dry cough, shortness of breath, and fever, which may arise within the initial weeks and months after RI.[3] The incidence of RP varies, reported to be between 10% and 20%,[3] with its occurrence being significantly dose-dependent.[4]

Patients with dyspnea and respiratory dysfunction may benefit from early treatment with oral corticosteroids.[5] The ability to predict RP could facilitate individualized, risk-adapted adjustments to therapeutic protocols, thereby potentially enhancing clinical outcomes.[6]

Several dosimetric parameters derived from the dose-volume histogram (DVH), such as the volume receiving at least 5 Gy (V5), V10, V20, and V30, have been correlated with an elevated risk of RP.[7] Notably, the quantitative analysis of normal tissue effects in the clinic model leverages the mean lung dose in its prediction of symptomatic RP.[2] However, it is important to note that dosimetric parameters alone do not adequately capture the complexity of spatial dose distributions. This is particularly relevant, as distinct regions within the lungs, such as tumor localization in the mid-lower lung, are linked to an increased risk of RP.[8,9] Dosiomics analyses provide a means to extract these spatial features from 3-dimensional dose distributions, and a variety of studies have successfully employed dosiomics to predict RP.[6,10-12] For example, Liang et al[11] demonstrated that dosiomics outperformed traditional dosimetric and normal tissue complication probability (NTCP) models. Furthermore, integrating radiomics—where quantitative features are extracted from computed tomography (CT) images—has been shown to enhance predictive accuracy.[6,10,13,14]

Nevertheless, most existing studies rely on retrospective data, often from relatively small sample sizes. The meta-analysis conducted by Sheen et al[15] indicates that 15 of the 16 studies included were retrospective, with a mean cohort size of 155 patients. The retrospective assessment of RP presents challenges, particularly due to the nonspecific nature of its symptoms, which may overlap with other conditions such as anemia.[2] Additionally, reductions in tumor size after RT may enhance respiratory function, thereby inadvertently masking RP symptoms.[2] Another major limitation of current research is the absence of external validation cohorts, which severely constrains the generalizability of findings. For example, in the meta-analysis by Chen et al,[16] only 1 study utilized an external test cohort.

This study aimed to predict and externally validate the occurrence of symptomatic RP after fractionated thoracic irradiation, utilizing both radiomics and dosiomics methodologies on 2 large prospective trial cohorts. The models were benchmarked against NTCP and clinical baseline models. Additionally, we assessed the potential benefits of treatment fractionation and the integrative impact of the radiation technique.

Methods and Materials

Clinical cohort description

In total, our multicenter study included 708 lung cancer patients. The training cohort consisted of 441 patients from the NRG Oncology/Radiation Therapy Oncology Group (RTOG) 0617 dataset from the National Clinical Trials Network.[17,18] The external test set comprised 267 patients from the REQUITE study.[19] The clinical characteristics of both studies are shown in Table 1. An overview of the inclusion criteria are depicted in Figure E1. Both studies conducted a prospective evaluation of the adverse effects associated with thoracic RT. Data collection for the RTOG 0617 study was conducted at 185 institutions across the United States and Canada from November 2007 to November 2011.[20] Similarly, the REQUITE study was conducted between April 2014 and March 2017 at 10 sites in Europe, the United Kingdom, and the United States.[19] For detailed inclusion and exclusion criteria, please refer to the original publications.[19,20] The RTOG 0617 study focused exclusively on patients with stage III cancer (American Joint Committee on Cancer 6), whereas the REQUITE study included patients in stages I-III (American Joint Committee on Cancer 7). Additionally, there were differences in the treatment regimens between both trials. The RTOG 0617 study consisted of 2 study arms with patients receiving prescription doses of either 60 Gy (standard dose) or 74 Gy (high dose), administered in 2 Gy fractions. In contrast, the REQUITE study included total doses ranging from 45 to 70 Gy, delivered in fractions of 1.5-5 Gy. The clinical endpoint for this study was defined as the occurrence of acute or late-onset symptomatic RP classified at grade 2 or higher. The classification was based on the Common Terminology Criteria for Adverse Events,[21] utilizing version 3 for the RTOG 0617 cohort and version 4 for the REQUITE cohort. Follow-up examinations in the REQUITE study to check for RP occurred at 3, 6, 12, and 24 months post-RT.[19] The RTOG 0617 study scheduled follow-up assessments every 3 months during the first year, every 4 months in the second year, every 6 months from the third to fifth years, and annually thereafter.[20]

The analysis of event distribution reveals a notable disparity between the cohorts. The incidence of symptomatic RPs in the RTOG 0617 cohort was nearly twice that of the REQUITE cohort (15% vs 7.9%). Additionally, significant differences in the severity levels were observed. Although asymptomatic pneumonitis prevailed in the REQUITE study (16% vs 2.7%), severe forms (grade 3-5) occurred almost exclusively in the RTOG 0617 cohort (5.5% vs 0.7%). Notably, life-threatening or fatal cases (grades 4 and 5) were exclusive to the RTOG 0617 trial, occurring at a rate of 1.1%.

The clinical features examined were age, sex, smoking status (nonsmoker, former smoker, current smoker), histology (adenocarcinoma, squamous cell carcinoma, other),

Table 1 Comparison of the demographic, clinical and tumor biology characteristics of the training and test cohorts

Characteristic	Overall N = 708	REQUITE N = 267	RTOG 0617 N = 441
Age	65 (58-71)	67 (61-74)	64 (57-70)
Sex			
Female	259 (37%)	79 (30%)	180 (41%)
Male	449 (63%)	188 (70%)	261 (59%)
Smoker			
Current smoker	321 (48%)	119 (49%)	202 (48%)
Former smoker	303 (46%)	113 (46%)	190 (45%)
Nonsmoker	40 (6.0%)	12 (4.9%)	28 (6.7%)
Unknown	44	23	21
Histology			
Adenocarcinoma	281 (40%)	105 (40%)	176 (40%)
Large cell undifferentiated	25 (3.5%)	14 (5.3%)	11 (2.5%)
Other	110 (16%)	46 (17%)	64 (15%)
Squamous cell carcinoma	290 (41%)	100 (38%)	190 (43%)
Unknown	2	2	0
Stage			
Ia	15 (2.2%)	15 (6.6%)	0 (0%)
Ib	11 (1.6%)	11 (4.8%)	0 (0%)
IIa	14 (2.1%)	14 (6.2%)	0 (0%)
IIb	17 (2.5%)	17 (7.5%)	0 (0%)
IIIa	391 (59%)	100 (44%)	291 (66%)
IIIb	220 (33%)	70 (31%)	150 (34%)
Unknown	40	40	0
Technique			
3D-CRT	349 (49%)	115 (43%)	234 (53%)
IMRT	359 (51%)	152 (57%)	207 (47%)
Dose	60 (60, 74)	63 (56, 66)	60 (60, 74)
Chemotherapy schedule			
Concurrent	565 (80%)	127 (48%)	438 (99%)
No chemotherapy	77 (11%)	74 (28%)	3 (0.7%)
Sequential	64 (9.1%)	64 (24%)	0 (0%)
Unknown	2	2	0
Symptomatic pneumonitis (RP ≥2)			
No event	619 (87%)	246 (92%)	373 (85%)
Event	89 (13%)	21 (7.9%)	68 (15%)

Median (IQR) for age and absolute and relative frequencies (%) for categorical variables are shown.

Abbreviations: 3D-CRT = 3-dimensional conformal radiation therapy; IMRT = intensity modulated radiation therapy; RP = radiation pneumonitis.

radiation technique (intensity modulated radiation therapy [IMRT], 3-dimensional conformal RT [3D-CRT]), and chemotherapy schedule (none, concurrent, sequential).

Missing clinical features were handled using multiple imputation with 100 iterations in R.[22] Clinical differences between the study populations were evaluated using the Mann-Whitney U test for numerical variables that did not follow a normal distribution. For categorical characteristics with >2 classes, the χ^2 test was employed, whereas the Fisher exact test was utilized for binary variables. All statistical tests were performed using the SciPy module in Python.[23] The significance level was set at $\alpha = 0.05$.

Experiment overview

We conducted a total of 5 experiments. Initially, we trained on the entire RTOG 0617 dataset, as well as on the high dose and standard dose subgroups. For these instances, we evaluated our models on the complete REQUITE cohort, along with the radiation technique-specific subcohorts. To further investigate the impact of different radiation techniques, we conducted 2 experiments focusing solely on patients who underwent IMRT/3D-CRT, testing exclusively within this technique. Table E1 illustrates a comprehensive overview of the experiments conducted. Additionally, the clinical characteristics of the various subcohorts are detailed in Tables E2-E4.

CT and dose data

All patients received a noncontrast planning CT scan prior to RT. Initially, the digital imaging and communications in medicine CT images and dose volumes were converted into neuroimaging informatics technology initiative format using Plastimatch.[24] The regions of interest (ROIs) utilized included the gross tumor volume (GTV), lungs-GTV, planning target volume (PTV), and the PTV with an isotropic margin of 20 mm (PTV-2cm). Lung segmentation was performed automatically using TotalSegmentator[25] and subsequently refined manually via the open-source platform 3D Slicer.[26] To focus solely on lung tissue, intersections between the PTV and PTV+2cm with the lungs were created. To assess the impact of fractionation, the equivalent dose in 2 Gy fractions (EQD2) volume was calculated voxel-wise using Equation 1, where D represents the total dose administered across all fractions, d the individual dose per fraction, and the $\frac{\alpha}{\beta}$ ratio is set to 3 for lung tissue.

$$EQD2 = D\frac{d + \frac{\alpha}{\beta}}{2 + \frac{\alpha}{\beta}} \quad (1)$$

Feature extraction

For each ROI, a total of 104 features were extracted from the CT scan and dose volumes using the Python library PyRadiomics.[27] This includes 19 first-order statistics, 26 shape features, and 59 texture features. A list of all extracted features can be found in Table E5. In adherence to the Image Biomarker Standardization Initiative guidelines,[28] the kurtosis was adjusted by subtracting 3. For the dosiomics features, the dose values were interpreted as gray levels. DVH parameters were extracted using PlatiPy.[29] For the lung-GTV, this includes the V5, V10, V15, V20, V30, V40, and V50, both as absolute and relative volumes, in addition to the mean lung dose. For PTV and GTV, the mean dose and the doses encompassing 95% and 99% of the respective volumes were utilized.

We defined a total of 10 feature sets, comprising 5 single modality sets and 5 combined sets. All sets with regard to the modalities used are listed in Table E6.

Machine learning models

Scikit-learn[30] and imbalanced-learn[31] packages in Python were used as a basis for the model generation. Preliminary analyses comparing various classifiers (e.g., random forest, extreme gradient boosting, support vector machines) did not demonstrate significant differences in predictive performance. Consequently, random forest[32] was selected as the classifier based on evidence from previous studies demonstrating its superior performance in the prediction of RP.[6,10] For each feature set, a 5 -fold nested cross-validation (CV) was performed with 50 repetitions in the inner loop and 10 repetitions in the outer loop. Hyperparameter optimization was carried out through random search within the inner folds, with the specific ranges detailed in Table E7. To obtain a selection of the most relevant features while minimizing redundancy, feature reduction was implemented within the inner CV loop using the minimum redundancy-maximum relevance algorithm in Python.[33] The selected features were then ranked according to their selection frequency, with the highest-ranking features being utilized in the outer loop. The nested CV was conducted with varying numbers of features, and the optimal count for each feature set was determined based on the number that achieved the highest performance in the outer folds. To determine the maximum number of features, we adhered to the 1-in-10 rule, which specifies that there should be $\sim$10 RP $\geq$2 patients in the training set for every feature. Subsequently, the final models were trained again in a separate CV on the entire training dataset and evaluated on the independent test set. To address any distortions resulting from class imbalance and to enhance the classifier's discriminatory power, the training set was balanced to a 1:1 ratio between minority and majority classes using synthetic minority oversampling and random undersampling with SMOTE-Tomek.[34] Classical SMOTE-Tomek was applied for purely numerical features, whereas SMOTE N- or SMOTE NC-Tomek were employed for categorical or mixed datasets.[35] The top-performing models were subsequently evaluated on the external test set. The complete workflow is illustrated in Figure 1.

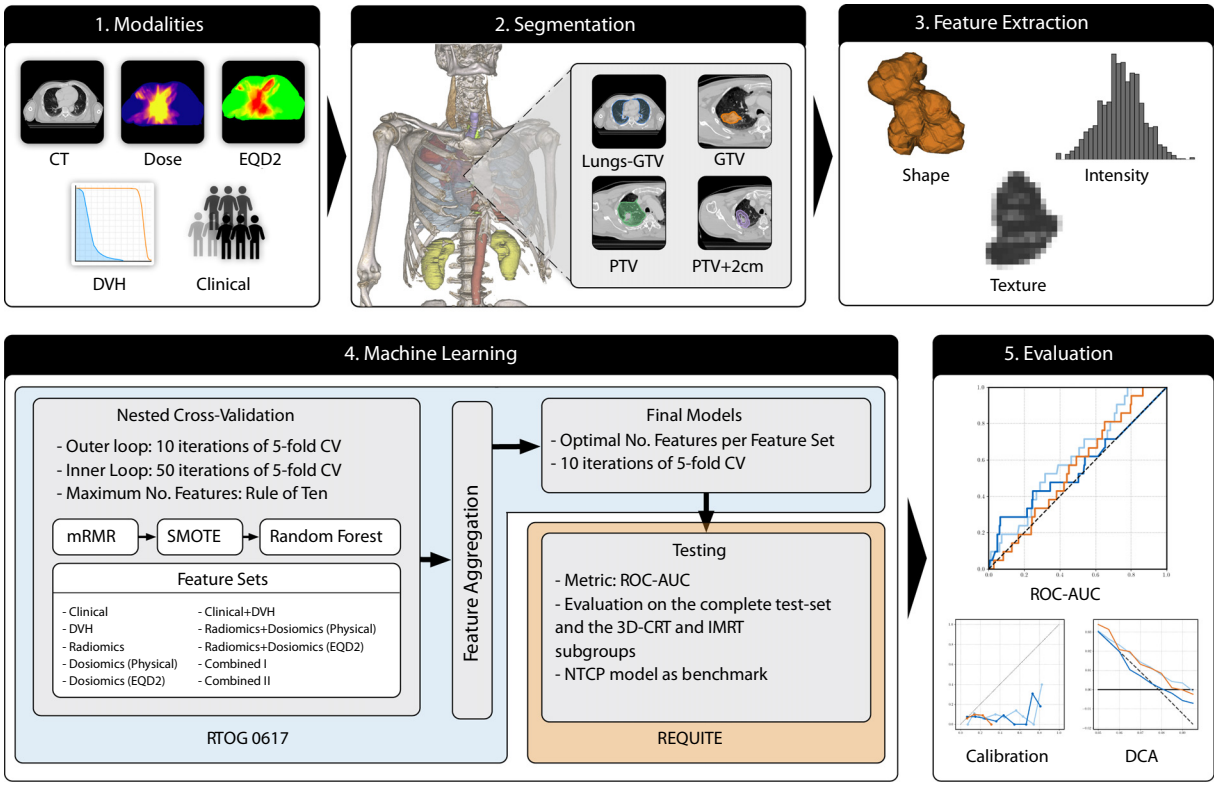

Fig. 1. An overview of the executed workflow. Along with clinical parameters and dose-volume histograms (DVHs), computed tomography (CT) and dose volumes (physical and biologically weighted) were utilized. Regions of interest (ROIs) were delineated based on the planning CT, including the gross tumor volume (GTV), the planning target volume (PTV), the PTV with a 2 cm margin (PTV+2 cm), and the lung volume excluding the tumor (lungs-GTV). Next, radiomics and dosiomics parameters were extracted for each ROI. A random forest classifier was trained to predict radiation pneumonitis (RP) in a nested cross-validation (CV). Features were selected using the minimum redundancy-maximum relevance (mRMR) algorithm, with varying feature counts. The maximum number was determined by the number of events in each respective experiment. Class imbalances were addressed using synthetic minority oversampling and random undersampling via SMOTE-Tomek. Based on the performance during the nested cross-validation, the optimal number of features was identified for each feature set, and final features were selected based on their frequency of selection. The hyperparameters of the final models were optimized in a separate CV process. Evaluation was carried out on the external test set, using a ROC-AUC curve, a calibration curve, and a decision curve analysis (DCA). *Abbreviations:* 3D-CRT = 3-dimensional conformal radiation therapy; CV = cross-validation; DVH = dose-volume histogram; EQD2 = equivalent dose in 2 Gy fractions; IMRT = intensity modulated radiation therapy; NTCP = normal tissue complication probability.

Normal tissue complication probability

As a benchmark for our models, we used the NTCP model based on logistic regression established by Marks et al.[2] It defines the relationship between the RP risk p and the *MLD* as follows:

$$p = \frac{\exp(b_0 + b_1 \cdot MLD)}{1 + \exp(b_0 + b_1 \cdot MLD)} \quad (2)$$

with $b_0 = -3.87$ and $b_1 = -0.126$.

Evaluation

We chose the area under the receiver operating characteristic curve (ROC-AUC) as our evaluation metric for comparing the models, as it effectively combines sensitivity and specificity, independent of the threshold. To calculate the 95% CIs, we conducted 500 bootstrap iterations.[36] We utilized calibration curves and the Hosmer-Lemeshow test to assess the alignment between predicted and observed event probabilities.[37] Additionally, we performed a decision curve analysis to evaluate the clinical implications of our prediction models.[38] Given the low incidence of RP, which may result in treatment discontinuation, predicting its occurrence is crucial. The lower limit was therefore set at a modest 5%. However, considering that a potential corticosteroid therapy can also result in adverse side effects, the upper limit was specified at 20%. To investigate the influence of patients with extreme values on model performance, a sensitivity analysis of the best-performing models was conducted based on 1000 iterations of 5-fold CV. Within

each iteration, 2 models were trained and their ROC-AUCs were compared: one including all patients and one excluding those whose feature values lay outside the 95% IQR of the respective training-fold feature space.

Reporting

Our work adheres to the guidelines for transparent reporting of a multivariable prediction model for individual prognosis or diagnosis. It corresponds to a type 3 prediction model. The ransparent reporting of a multivariable prediction model for individual prognosis or diagnosis statement can be found in Table E8.

Results

Clinical cohort

Apart from the age distribution, there were no significant differences in the clinical variables between the 2 studies. The patients in the REQUITE study had a median age of 67 years, which was significantly older than the median age of 64 years in the RTOG 0617 cohort (Mann-Whitney U test, $U = 47{,}238$, $P < .001$). Within the respective studies, there were no statistically significant differences in clinical characteristics between the 3D-CRT and IMRT subgroups.

Baseline models

In addition to the NTCP model, we used our clinical and DVH predictors as reference models. Table 2 provides an overview of the best models from the complete, standard, and high dose groups. The clinical models of the complete and standard dose experiments did not outperform random chance in the nested CV or on the test set (ROC-AUC, 0.45-0.50). Only the high dose experiment reached an ROC-AUC of 0.55 for the clinical features. Still, this model only achieved an ROC-AUC of 0.51 on the test set. DVH-based models performed better than random chance in both the nested CV and the test set, with respective ROC-AUC values of 0.52-0.55 and 0.55-0.59. With ROC-AUC values of 0.59 and 0.57, respectively, the complete and high dose models demonstrated superior performance compared with the NTCP model, which yielded an ROC-AUC of 0.56.

A subgroup analysis revealed that the predictive power of the DVH and NTCP models was stronger in patients treated with 3D-CRT than in those who received IMRT. In the 3D-CRT cohort, the NTCP model achieved an ROC-AUC of 0.69, while the DVH models reached ROC-AUCs of 0.72, 0.74, and 0.78 for the standard, complete, and high dose models, respectively. Conversely, neither the NTCP nor the DVH models could predict RP more effectively than random chance in the IMRT cohort, which showed ROC-AUC values ranging from 0.48 to 0.51.

Radiomics and dosiomics models

The top-performing model in the nested CV using standard dose radiomics achieved an ROC-AUC of 0.60 and an AUC of 0.56 on the independent test set. Notably, the difference between the 3D-CRT and IMRT cohorts was less pronounced here, with respective ROC-AUC values of 0.57 and 0.55. The second-best model, based on complete $\text{Dosiomics}_{\text{EQD2}}$, obtained an ROC-AUC of 0.56 in nested CV. However, there was a notable difference between the technique cohorts, showing an ROC-AUC of 0.75 for the 3D-CRT group compared with 0.39 for the IMRT group. In contrast, the standard dose $\text{Dosiomics}_{\text{EQD2}}$ model did not exhibit this effect, with ROC-AUC values of 0.60 and 0.59 for 3D-CRT and IMRT, respectively. The divergence of the $\text{Dosiomics}_{\text{EQD2}}$ model between the different irradiation techniques suggests that technique-specific dose-response patterns exist that were masked in the overall dataset. To account for this effect, an ensemble model was developed in which separate classifiers were trained for the 3D-CRT and IMRT cohorts based on biologically weighted $\text{Dosiomics}_{\text{EQD2}}$. This model achieved improved overall performance on the independent test dataset (ROC-AUC, 0.61; 95% CI, 0.50-0.73), as shown in Table 3. The ROC-AUC curves for the Ensemble-$\text{Dosiomics}_{\text{EQD2}}$ model, as well as the standard dose (StD)-Radiomics and NTCP models, are illustrated in Figure 2 for the entire test set (a) and for the irradiation technique subgroups (b-c). Additional representations of the ROC-AUC curves for all models can be found in Figure E2. The selected features of the best models are listed in Table E9. As illustrated in Figure E3a, b, the decision curve analysis indicated that the $\text{Dosiomics}_{\text{EQD2}}$ and NTCP models provided a positive net benefit within the 5% to 10% range. The calibration curve in Figure E3c indicates that all models tend to overestimate the actual risk of developing RP. The calibration curves of all models can be found in Figure E4. In the Hosmer-Lemeshow-Test, poor calibration was shown for the $\text{Dosiomics}_{\text{EQD2}}$ ($C(10) = 135.70$, $P < .001$) and StD-Radiomics ($C(10) = 123.73$, $P < .001$) models, whereas good calibration was observed for the NTCP model ($C(10) = 9.02$, $P = .530$). A sensitivity analysis was conducted to assess the impact of extreme values and patient characteristics, revealing only minimal differences across all feature sets and treatment subgroups. The detailed results of this analysis are included in Table E10. Furthermore, the testing revealed that the biologically weighted $\text{Dosiomics}_{\text{EQD2}}$ model performed comparably or even better than the physical $\text{Dosiomics}_{\text{Physical}}$ model on the test set for 3D-CRT dose volumes. Overall, integrating radiomics and dosiomics did not confer a tangible advantage over models based solely on individual modalities. Moreover, the inclusion of clinical features or DVH did not demonstrate any significant benefit. Furthermore, the models based on the high dose dataset tended to demonstrate poorer performance in the nested CV than those trained on the full and standard dose datasets. On average, the training ROC-AUC was 0.02 and 0.03 lower than when trained on the full and standard dose

Table 2 Results of the nested cross-validation and evaluation of the independent test set for the complete, standard dose, and high dose experiments

Experiment	Feature set	Training cohort (RTOG 0617) Nested CV	Test cohort (REQUITE)		
			Complete	3D-CRT	IMRT
Complete (COMP)	Clinical	0.50	0.48 (0.37-0.60)	0.41 (0.21-0.62)	0.53 (0.38-0.68)
	DVH	0.55	0.59 (0.47-0.70)	0.74 (0.55-0.93)	0.51 (0.38-0.63)
	$Dosiomics_{Physical}$	0.55	0.52 (0.40-0.65)	0.67 (0.53-0.81)	0.44 (0.28-0.61)
	$Dosiomics_{EQD2}$	**0.56**	**0.51 (0.37-0.64)**	**0.75 (0.57-0.93)**	**0.39 (0.24-0.54)**
	Radiomics	0.51	0.55 (0.44-0.66)	0.72 (0.58-0.87)	0.47 (0.33-0.61)
	DVH+Clinical	0.52	0.49 (0.35-0.64)	0.76 (0.59-0.93)	0.43 (0.30-0.57)
	Radiomics+$Dosiomics_{Physical}$	0.53	0.50 (0.39-0.61)	0.62 (0.45-0.79)	0.44 (0.30-0.57)
	Radiomics+$Dosiomics_{EQD2}$	0.54	0.53 (0.41-0.66)	0.67 (0.48-0.85)	0.46 (0.30-0.62)
	Combined I	0.53	0.49 (0.38-0.60)	0.61 (0.44-0.78)	0.42 (0.29-0.56)
	Combined II	0.53	0.55 (0.43-0.66)	0.54 (0.36-0.72)	0.54 (0.40-0.68)
High dose (HiD)	Clinical	**0.55**	**0.51 (0.39-0.62)**	**0.47 (0.30-0.65)**	**0.53 (0.38-0.68)**
	DVH	0.52	0.57 (0.46-0.68)	0.78 (0.64-0.91)	0.49 (0.36-0.62)
	$Dosiomics_{Physical}$	0.50	0.61 (0.48-0.74)	0.72 (0.52-0.91)	0.55 (0.38-0.71)
	$Dosiomics_{EQD2}$	0.51	0.57 (0.44-0.70)	0.72 (0.52-0.91	0.48 (0.32-0.64)
	Radiomics	0.50	0.49 (0.38-0.60)	0.41 (0.21-0.61)	0.53 (0.40-0.67)
	DVH+Clinical	0.50	0.53 (0.40-0.66)	0.60 (0.33-0.87)	0.48 (0.34-0.63)
	Radiomics+$Dosiomics_{Physical}$	0.51	0.52 (0.41-0.64)	0.58 (0.40-0.77)	0.48 (0.33-0.63)
	Radiomics+$Dosiomics_{EQD2}$	0.50	0.52 (0.40-0.63)	0.58 (0.39-0.77)	0.47 (0.32-0.63)
	Combined I	0.51	0.53 (0.41-0.64)	0.60 (0.41-0.78)	0.48 (0.34-0.63)
	Combined II	0.50	0.50 (0.38-0.62)	0.56 (0.36-0.75)	0.45 (0.30-0.61)
Standard dose (StD)	Clinical	0.49	0.45 (0.34-0.56)	0.50 (0.50-0.50)	0.50 (0.50-0.50)
	DVH	0.53	0.55 (0.43-0.67)	0.72 (0.48-0.96)	0.48 (0.36-0.61)
	$Dosiomics_{Physical}$	0.59	0.53 (0.41-0.66)	0.58 (0.36-0.79)	0.51 (0.37-0.66)
	$Dosiomics_{EQD2}$	0.56	0.60 (0.47-0.72)	0.60 (0.35-0.85)	0.59 (0.43-0.75)
	Radiomics	**0.60**	**0.56 (0.41-0.70)**	**0.57 (0.34-0.81)**	**0.55 (0.36-0.75)**
	DVH+Clinical	0.51	0.46 (0.32-0.59)	0.63 (0.42-0.85)	0.48 (0.34-0.62)
	Radiomics+$Dosiomics_{Physical}$	0.55	0.52 (0.38-0.65)	0.66 (0.45-0.88)	0.46 (0.29-0.62)
	Radiomics+$Dosiomics_{EQD2}$	0.52	0.59 (0.46-0.71)	0.66 (0.46-0.86)	0.55 (0.40-0.71)
	Combined I	0.55	0.56 (0.43-0.68)	0.72 (0.54-0.89)	0.48 (0.32-0.63)
	Combined II	0.51	0.55 (0.43-0.68)	0.73 (0.57-0.89)	0.46 (0.31-0.62)
NTCP	MLD	-	0.56 (0.44-0.67)	0.69 (0.47-0.91)	0.50 (0.37-0.62)

The area under the receiver operating characteristic curve (ROC-AUC) was utilized as the performance metric. It was calculated on the entire test set as well as for the 3D-CRT and IMRT subgroups. The 95% confidence intervals are based on 500 bootstrap iterations. The best models for each experiment are highlighted in bold.

Abbreviations: 3D-CRT = 3-dimensional conformal radiation therapy; CV = cross-validation; DVH = dose-volume histogram; EQD2 = equivalent dose in 2 Gy fractions; IMRT = intensity modulated radiation therapy; MLD = mean lung dose; NTCP = normal tissue complication probability.

subgroups, respectively. In contrast, the test ROC-AUC displayed only a modest average difference when compared to the others (complete, +0.01; standard dose, −0.002). Although a single model per experiment was selected based on the highest ROC-AUC in the nested CV, differences between top-performing models were small and often within a narrow range.

Technique stratification

To further investigate the influence of the irradiation technique, all models were trained and tested selectively on patients treated with either 3D-CRT or IMRT. The results showed that the 3D-CRT models generally outperformed

Table 3 Results of the nested cross-validation and evaluation of the independent test set for the IMRT and 3D-CRT experiments

	Training cohort (RTOG 0617)		Test cohort (REQUITE)		
Feature set	**3D-CRT**	**IMRT**	**3D-CRT**	**IMRT**	**Ensemble**
Clinical	0.53	**0.57**	0.40 (0.21-0.59)	**0.46 (0.31-0.62)**	0.43 (0.31-0.55)
DVH	**0.60**	0.53	**0.69 (0.48-0.90)**	0.49 (0.31-0.67)	**0.62 (0.49-0.74)**
$Dosiomics_{Physical}$	0.55	0.54	0.64 (0.44-0.84)	0.40 (0.25-0.55)	0.50 (0.38-0.62)
$Dosiomics_{EQD2}$	0.56	0.55	0.63 (0.46-0.80)	0.60 (0.43-0.76)	0.61 (0.50-0.73)
Radiomics	0.57	0.48	0.58 (0.42-0.73)	0.42 (0.25-0.58)	0.47 (0.35-0.58)
DVH+Clinical	0.59	0.55	0.73 (0.55-0.91)	0.48 (0.33-0.62)	0.61 (0.50-0.72)
Radiomics+$Dosiomics_{Physical}$	0.55	0.49	0.58 (0.42-0.74)	0.45 (0.26-0.63)	0.47 (0.34-0.61)
Radiomics+$Dosiomics_{EQD2}$	0.58	0.49	0.58 (0.42-0.74)	0.60 (0.43-0.76)	0.58 (0.45-0.71)
Combined I	0.55	0.49	0.58 (0.42-0.74)	0.42 (0.28-0.57)	0.45 (0.33-0.57)
Combined II	0.56	0.50	0.59 (0.34-0.83)	0.54 (0.37-0.71)	0.51 (0.37-0.65)

The performance of the ensemble models on the entire test set is presented. The area under the receiver operating characteristic curve (ROC-AUC) was used as the performance metric. It was calculated on the IMRT and 3D-CRT subgroup of the test set. The 95% CIs are based on 500 bootstrap iterations. The best models for each experiment are highlighted in bold.

Abbreviations: 3D-CRT = 3-dimensional conformal radiation therapy; DVH = dose-volume histogram; EQD2 = equivalent dose in 2 Gy fractions; IMRT = intensity modulated radiation therapy.

their IMRT counterparts in terms of both nested CV and testing performance. As illustrated in Table 3, the 3D-CRT models exceeded the IMRT models by an average ROC-AUC of 0.05 in the nested CV and 0.12 in the test set for each feature set. The top-performing 3D-CRT model achieved a ROC-AUC of 0.60 in nested CV and 0.69 in the test set. Among the IMRT models, the clinical model exhibited the best performance in the nested CV with an AUC of 0.57, although it yielded unsatisfactory results on the test set. In both experiments, combining radiomics and dosiomics, along with the inclusion of clinical and DVH parameters, did not enhance the test performance. Notably, unlike the 3D-CRT cohort, no predictive power can be achieved using DVH or DVH + clinical parameters for IMRT patients. On the other hand, the $Dosiomics_{EQD2}$ model demonstrated an ROC-AUC of 0.60 when evaluated on the test set.

Discussion

Our work investigated the potential of radiomics and dosiomics-based machine learning models in predicting symptomatic RP after thoracic RT. We validated the developed classifiers on a prospective, multicenter, external dataset. In contrast to extensive data published based on retrospective data, the models derived from StD-Radiomics and $Dosiomics_{EQD2}$ demonstrated predictive performance for RP that only slightly exceeded random chance in both the 3D-CRT and IMRT subgroups. Radiation technique has a large impact on predictive performance. Although for 3D-CRT, DVH and clinical parameters had the highest predictive value, for IMRT, $Dosiomics_{EQD2}$ achieved the best performance.

This analysis examined 10 different feature sets, with no definitive set proving broad superiority. Furthermore, several models demonstrated nearly equivalent predictive performance in nested CV. These marginal differences may not be sufficiently robust to reliably differentiate between competing models. The StD-Radiomics and $Dosiomics_{EQD2}$ models outperformed reference methods based on clinical parameters and DVH, yielding results comparable to, or marginally exceeding, those of the NTCP model. A subgroup analysis showed that the overall prediction quality among the patients who received 3D-CRT was considerably higher than that in those who received IMRT. The NTCP model had no predictive power among the latter. This is in line with considerations by Marks et al,[2] who questioned the applicability of the NTCP model to IMRT and stereotactic body RT protocols due to the distinct dose delivery patterns involved. Our stratified analysis also revealed that, within the IMRT subgroup, the DVH and DVH + clinical models, previously identified as the most robust predictive models for patients receiving 3D-CRT, demonstrated limited predictive efficacy. In contrast, models derived from $Dosiomics_{EQD2}$ exhibited superior performance. Given that IMRT has emerged as the predominant standard in modern RT, it suggests that a technique-specific development could improve prediction accuracy. Overall, none of the evaluated models achieved a consistent clinical net benefit within the relevant threshold range compared with the treatment of either all or no patients. A sensitivity analysis revealed that patients treated with IMRT exhibited slightly greater variations in performance. This suggests that extreme value distributions may have a modest effect on model performance within this

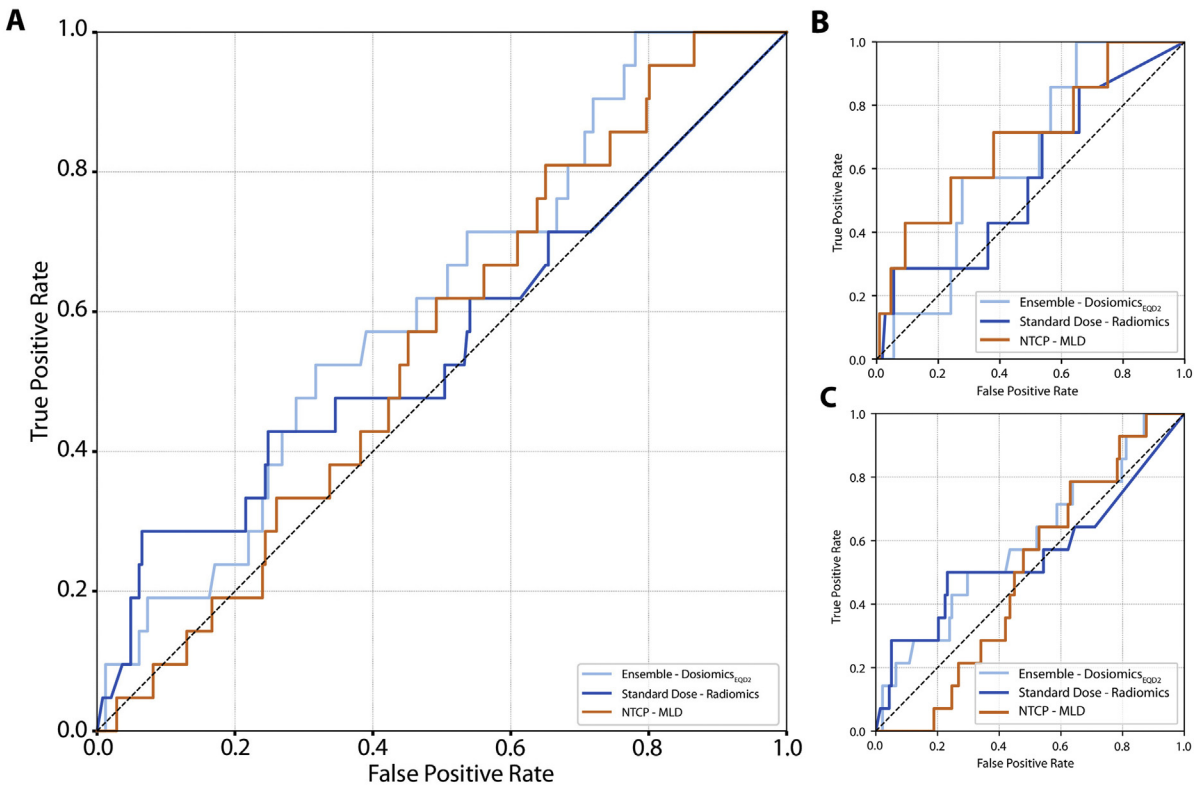

Fig. 2. Receiver operating characteristic (ROC) curves on the independent test set for (a) the entire test cohort (b) the 3D-CRT subgroup (c) the IMRT subgroup for the Ensemble-Dosiomics-EQD2 model, the Standard Dose Radiomics model, and the NTCP model. The Ensemble-Dosiomics-EQD2 model achieved the highest predictive performance on the entire test cohort with an ROC-AUC of 0.61, outperforming both the Standard Dose Radiomics and NTCP models (both ROC-AUC = 0.56). In the 3D-CRT subgroup, the NTCP model performed best. In contrast, in the IMRT subgroup, it showed no predictive ability and was outperformed by both the Ensemble-Dosiomics and Standard Dose Radiomics models. *Abbreviations:* 3D-CRT = 3-dimensional conformal radiation therapy; EQD2 = equivalent dose in 2 Gy fractions; IMRT = intensity modulated radiation therapy; NTCP = normal tissue complication probability.

subgroup. However, these differences were accompanied by significant variability, as indicated by large standard deviations. Notably, no similar effects were observed for the other treatment modalities, indicating that the overall findings are not influenced by sensitivity to extreme cases.

Various other studies have developed models for RP prediction. Hirose et al[39] achieved an ROC-AUC of 0.76 with a radiomics model on a test data set of 30 patients undergoing stereotactic body RT. Kawahara et al[40] reported an ROC-AUC of 0.86 for their radiomics-based prediction model in a study with a total of 77 patients. Liang et al[11] achieved an ROC-AUC of 0.78 in a retrospective analysis of 70 patients with a purely dosiomics model. Adachi et al[12] combined dosiomics features with DVH factors and reached an ROC-AUC of 0.85 in their study. Wang et al[41] were able to increase the model performance to an ROC-AUC of 0.80 by using deep learning-based radiomics and dosiomics features. In contrast, the prediction performance of the deep learning model by Zhang et al,[42] which combined CT and dose volumes, was lower when externally validated on the RTOG 0617 dataset (AUC 0.55 in the standard dose arm, 0.63 in the high dose arm, both without weight adjustment). The clinical heterogeneity between the training and testing cohorts could be a potential factor contributing to our comparatively lower performance. For example, in the RTOG 0617 study, patients were uniformly administered (predominantly concurrent) chemoradiation, whereas the REQUITE trial encompassed a diverse treatment strategy, with nearly half of the patients receiving sequential or no chemotherapy. Although previous publications have reported enhanced performance when integrating radiomics and dosiomics features, we did not observe such an improvement. In contrast, Li et al[13] reported an increase in AUC from 0.75 (radiomics-only) to 0.85 in a retrospective study involving 126 patients through a combination of radiomics with dosimetric features. Additionally, feature extraction from functionally segmented lung regions further enhanced the ROC-AUC to 0.93. Jiang et al[43] also showed in their study of 79 patients that augmenting radiomics data with dosimetric features improved performance. Instead, we did not find a clear benefit from the combination of different modalities compared with individual models. This disparity may stem from unaddressed inter-scanner variability, as radiomic features are known to exhibit sensitivity to discrepancies in

scanner type, pixel dimensions, slice thickness, and reconstruction algorithms.[44] Numerous preceding studies have relied on datasets from single institutions, potentially obscuring this variability. Notably, Zhou et al[45] also found no significant difference in test performance between DVH, Radiomics+DVH, Radiomics+Dosiomics$_{Physical}$, and Radiomics+Dosiomics$_{EQD2}$ models.

We also ensured a strict separation between the training and external test datasets. Many studies in the radiomics and dosiomics literature perform feature selection on the entire dataset prior to CV, which can introduce data leakage and lead to overly optimistic performance estimates. In this context, Demircioğlu estimated that such practices could inflate the ROC-AUC by up to 0.15 for radiomics analyses.[46] Another strength of this work is the utilization of 2 multicenter prospective data sets. This not only achieves a larger sample size, but the heterogeneity in treatment regimens and disease stages also reflects the diversity of daily clinical routine and thus allows insights into real-world generalizability. Diagnosing RP clinically is difficult. Kocak et al[47] showed that in 28% of the cases, RP diagnosis was confounded by other pulmonary diseases. The prospective data collection of our data sets may reduce this bias compared with the retrospective counterparts.

Nonetheless, this work has several limitations. First, the delineation of the ROIs is rater dependent. We tried to minimize this factor as much as possible by a semiautomatic segmentation approach of the lungs-GTV regions. Additionally, potential batch effects arising from differing imaging protocols or dose calculations were not accounted for due to the absence of information regarding individual treatment centers or scanner types in the RTOG 0617 study. Therefore, the models might have learned some dataset-specific characteristics, as the RTOG 0617 study was conducted in the mid-2000s whereas the REQUITE dataset spanned from 2014 to 2017.

Our datasets exhibited a pronounced class imbalance, with an event rate of 15% in the training set and just under 8% in the test set. To mitigate the effect of this imbalance in the RP rate, we employed synthetic minority oversampling in combination with random undersampling using SMOTE-Tomek. Finally, it should also be noted that 2 different versions of the Common Terminology Criteria for Adverse Events were employed in the RTOG 0617 and REQUITE studies. Although the differences are only minor, there are some instances in which both versions would define different RP grades for the same patient. For example, version 4 would classify patients needing therapy and limitation of instrumental activities of daily living as grade 2, whereas version 3 would often still assign grade 1.

Conclusion

This study, built upon 2 large-scale prospective cohorts, highlights the large relevance of dose prescription and radiation technique for radiomics and dosiomics model development. Technique-specific models outperformed the NTCP baseline model. For 3D-CRT, these models were outperformed by DVH+clinical machine learning models, whereas for IMRT, dosiomics achieved the highest, although still limited, predictive performance. However, given the relatively minor performance discrepancies observed among the top-performing models, it is challenging to outline a single superior solution.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors used Grammarly AI in order to improve the language and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

REFERENCES

1. Remon J, Soria JC, Peters S. ESMO Guidelines Committee. Early and locally advanced non-small-cell lung cancer: an update of the ESMO Clinical Practice Guidelines focusing on diagnosis, staging, systemic and local therapy. *Ann Oncol* 2021;32:637-1642.
2. Marks LB, Bentzen SM, Deasy JO, et al. Radiation dose-volume effects in the lung. *Int J Radiat Oncol Biol Phys* 2010;76(suppl 3):S70-S76.
3. Mehta V. Radiation pneumonitis and pulmonary fibrosis in non-small-cell lung cancer: Pulmonary function, prediction, and prevention. *Int J Radiat Oncol Biol Phys* 2005;63:5-24.
4. Palma DA, Senan S, Tsujino K, et al. Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis. *Int J Radiat Oncol Biol Phys* 2013;85(2):444-450.
5. Rahi MS, Parekh J, Pednekar P, et al. Radiation-induced lung injury-current perspectives and management. *Clin Pract* 2021;11:410-429.
6. Kraus KM, Oreshko M, Schnabel JA, Bernhardt D, Combs SE, Peeken JC. Dosiomics and radiomics-based prediction of pneumonitis after radiotherapy and immune checkpoint inhibition: the relevance of fractionation. *Lung Cancer* 2024;189:107507.
7. Zhang XJ, Sun JG, Sun J, et al. Prediction of radiation pneumonitis in lung cancer patients: a systematic review. *J Cancer Res Clin Oncol* 2012;138:2103-2116.
8. Vogelius IR, Bentzen SM. A literature-based meta-analysis of clinical risk factors for development of radiation induced pneumonitis. *Acta Oncol* 2012;51:975-983.
9. Kong FMS, Wang S. Nondosimetric risk factors for radiation-induced lung toxicity. *Semin Radiat Oncol* 2015;25:100-109.
10. Kraus KM, Oreshko M, Bernhardt D, Combs SE, Peeken JC. Dosiomics and radiomics to predict pneumonitis after thoracic stereotactic body radiotherapy and immune checkpoint inhibition. *Front Oncol* 2023;13:1124592.
11. Liang B, Yan H, Tian Y, et al. Dosiomics: extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis. *Front Oncol* 2019;9:269.
12. Adachi T, Nakamura M, Shintani T, et al. Multi-institutional dose-segmented dosiomic analysis for predicting radiation pneumonitis after lung stereotactic body radiation therapy. *Med Phys* 2021;48:1781-1791.
13. Li B, Ren G, Guo W, et al. Function-wise dual-omics analysis for radiation pneumonitis prediction in lung cancer patients. *Front Pharmacol* 2022;13:971849.

14. Huang Y, Feng A, Lin Y, et al. Radiation pneumonitis prediction after stereotactic body radiation therapy based on 3D dose distribution: dosiomics and/or deep learning-based radiomics features. *Radiat Oncol* 2022;17:188.
15. Sheen H, Cho W, Kim C, et al. Radiomics-based hybrid model for predicting radiation pneumonitis: a systematic review and meta-analysis. *Phys Med* 2024;123:103414.
16. Chen Z, Yi G, Li X, et al. Predicting radiation pneumonitis in lung cancer using machine learning and multimodal features: a systematic review and meta-analysis of diagnostic accuracy. *BMC Cancer* 2024;24:1355.
17. Bradley J, Forster K. Data from NSCLC-Cetuximab. RTOG-0617. The Cancer Imaging Archive. 2018. doi:10.7937/TCIA.2018.jze75u7v.
18. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013;26:1045-1057.
19. Seibold P, Webb A, Aguado-Barrera ME, et al. REQUITE: a prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer. *Radiother Oncol* 2019;138:59-67.
20. Bradley JD, Paulus R, Komaki R, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIa or IIIb non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. *Lancet Oncol* 2015;16:187-199.
21. US Department of Health and Human Services. Common Terminology Criteria for Adverse Events (CTCAE). Version 5.0. 2017. Accessed December 16, 2025. https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events/ctcae-v5-5x7.pdf.
22. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1-67.
23. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;17:261-272.
24. Sharp GC, Li R, Wolfgang J, et al. Plastimatch: an open source software suite for radiotherapy image processing. In: Proceedings of the XVI'th International Conference on the use of Computers in Radiotherapy (ICCR). *vol. 3*. Amsterdam, Netherlands.
25. Wasserthal J, Breit HC, Meyer MT, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell* 2023;5:e230024.
26. Kikinis R, Pieper SD, Vosburgh KG. 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. In: Jolesz FA, ed. *Intraoperative Imaging and Image-Guided Therapy*. Springer; 2013:277-289.
27. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017;77:e104-e107.
28. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020;295:328-338.
29. Chlap P, Finnegan RN. PlatiPy: processing library and analysis toolkit for medical imaging in python. *J Open Source Softw* 2023;8:5374.
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825-2830.
31. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017;18:1-5.
32. Breiman L. Random forests. *Mach Learn* 2001;45:5-32.
33. Galli S. Feature-engine: a python package for feature engineering for machine learning. *J Open Source Softw* 2021;6:3642.
34. Batista GEAPA, Bazzan ALC, Monard MC, et al. Balancing training data for automated annotation of keywords: a case study. *In:* Proceedings of the II Brazilian Workshop on Bioinformatics (WOB2003), December 3-5, 2003, Macaé, RJ, Brazil 2003;3:10-18.
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321-357.
36. Gildenblat J. *A python library for confidence intervals*; 2023. Accessed December 16, 2025; https://github.com/jacobgil/confidenceinterval.
37. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. John Wiley & Sons; 2013.
38. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.
39. Hirose TA, Arimura H, Ninomiya K, Yoshitake T, Fukunaga J, Shioyama Y. Radiomic prediction of radiation pneumonitis on pretreatment planning computed tomography images prior to lung cancer stereotactic body radiation therapy. *Sci Rep* 2020;10:20424.
40. Kawahara D, Imano N, Nishioka R, et al. Prediction of radiation pneumonitis after definitive radiotherapy for locally advanced non-small cell lung cancer using multi-region radiomics analysis. *Sci Rep* 2021;11:16232.
41. Wang X, Zhang A, Yang H, et al. Multicenter development of a deep learning radiomics and dosiomics nomogram to predict radiation pneumonia risk in non-small cell lung cancer. *Sci Rep* 2025;15:17106.
42. Zhang Z, Wang Z, Luo T, et al. Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy. *Radiother Oncol* 2023;182:109581.
43. Jiang W, Song Y, Sun Z, Qiu J, Shi L. Dosimetric factors and radiomics features within different regions of interest in planning ct images for improving the prediction of radiation pneumonitis. *Int J Radiat Oncol Biol Phys* 2021;110:1161-1170.
44. Ligero M, Jordi-Ollero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol* 2021;31:1460-1470.
45. Zhou L, Wen Y, Zhang G, Wang L, Wu S, Zhang S. Machine learning-based multiomics prediction model for radiation pneumonitis. *J Oncol* 2023;2023:5328927.
46. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights Imaging* 2021;12:172.
47. Kocak Z, Evans ES, Zhou SM, et al. Challenges in defining radiation pneumonitis in patients with lung cancer. *Int J Radiat Oncol Biol Phys* 2005;62:635-638.