Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy

Radiotherapy and Oncology 182 (2023) 109581

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy

Zhen Zhang [a,b,1], Zhixiang Wang [b,c,1], Tianchen Luo [d], Meng Yan [e], Andre Dekker [b], Dirk De Ruysscher [b], Alberto Traverso [b], Leonard Wee [b,2], Lujun Zhao [e,*]

[a] Zhejiang Cancer Hospital, Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China; [b] Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht 6229 ET, The Netherlands; [c] Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing, China; [d] Institute of System Science, National University of Singapore, 119260, Singapore; [e] Department of Radiation Oncology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin, Tianjin's Clinical Research Center for Cancer, 300060, China

ARTICLE INFO

Article history:
Received 23 August 2022
Received in revised form 17 February 2023
Accepted 17 February 2023
Available online 25 February 2023

Keywords:
Radiotherapy
Radiation pneumonitis
Deep learning
Artificial intelligence
Actuarial outcome models

ABSTRACT

Purpose: To develop a deep learning model that combines CT and radiation dose (RD) images to predict the occurrence of radiation pneumonitis (RP) in lung cancer patients who received radical (chemo)radiotherapy.

Methods: CT, RD images and clinical parameters were obtained from 314 retrospectively-collected patients (training set) and 35 prospectively-collected patients (test-set-1) who were diagnosed with lung cancer and received radical radiotherapy in the dose range of 50 Gy and 70 Gy. Another 194 (60 Gy group, test-set-2) and 158 (74 Gy group, test-set-3) patients from the clinical trial RTOG 0617 were used for external validation. A ResNet architecture was used to develop a prediction model that combines CT and RD features. Thereafter, the CT and RD weights were adjusted by using 40 patients from test-set-2 or 3 to accommodate cohorts with different clinical settings or dose delivery patterns. Visual interpretation was implemented using a gradient-weighted class activation map (grad-CAM) to observe the area of model attention during the prediction process. To improve the usability, ready-to-use online software was developed.

Results: The discriminative ability of a baseline trained model had an AUC of 0.83 for test-set-1, 0.55 for test-set-2, and 0.63 for test-set-3. After adjusting CT and RD weights of the model using a subset of the RTOG-0617 subjects, the discriminatory power of test-set-2 and 3 improved to AUC 0.65 and AUC 0.70, respectively. Grad-CAM showed the regions of interest to the model that contribute to the prediction of RP.

Conclusion: A novel deep learning approach combining CT and RD images can effectively and accurately predict the occurrence of RP, and this model can be adjusted easily to fit new cohorts.

Radiation pneumonitis (RP) is a relatively common radiotherapy(RT)-related side effect [1,2]; estimates of RP vary from 5%–58% [3] but it is challenging to forecast accurately on the individual patient level. The risk of RP constrains the tumoricidal dose that can be prescribed and, in serious instances, may directly threaten the life of the patient. Prediction models of a patient's RP risk are hence an active topic in current research work [4,5].

Dose-volume histogram (DVH) metrics, such as mean lung dose [6], V5 and V20 [7], are presently in broad clinical use as surrogates for RP risk. Normal tissue control probability (NTCP) can be computed from a DVH of total lungs [8]. These aforementioned DVH indicators do not explicitly account for the spatially heterogeneous distribution of dose in lungs, nor do they account for the functional state of lung parenchymal tissue prior to commencement of RT. Hand-crafted features that describe spatial dose non-uniformity (i.e. "dosiomics") have been recently investigated [9], as were characterization of non-tumour lung tissue via image-based analysis (i.e. "radiomics" and texture) [10–12]. To date, few RP studies have

* Corresponding author.
E-mail addresses: leonard.wee@maastro.nl (L. Wee), zhaolujun@tjmuch.com (L. Zhao).
[1] Z. Zhang and Z. Wang contributed as joint first authors.
[2] Senior author.

https://doi.org/10.1016/j.radonc.2023.109581
0167-8140/

been performed that combine both dosiomics from a clinical treatment plan and radiomics from its corresponding planning CT [13,14]. These studies have treated the two types of data as disjoint feature domains.

A promising direction for predicting RP is a deeper exploration of inter-related effects of dose and morphology. First, it is supposed that information about the underlying radio-sensitivity of lung tissue might be encoded into CT-based imaging features. Second, that variations in applying RT planning national guidelines leads to divergent spatial dose distributions that are not fully captured in traditional indices such as V20. For example, in China, the constraint V20 not exceed 25%–30% [15,16], however National Comprehensive Cancer Network (NCCN) guidelines recommend 35%–40%. Within a set of DVH constraints, there exists an unlimited number of feasible RT plans that would meet those constraints but result in non-comparable spatial dose distributions in normal lung. Third, it is not entirely clear how to explicitly define hand-crafted measures that combine both CT and dose information into a common feature domain. Last, it remains an open debate about the relative merits of hand-crafted features versus deep-learning features in regard to a given clinical question.

The objective of this study was to develop and evaluate a deep-learning (DL) model to predict RP on the basis of CT intensities and Radiotherapy Dose (RD) distributions, using a joint feature representation for CT attenuation (radiomics) and dose distribution (dosiomics), rather than making an ensemble of separated models. A design criterion was that any such DL-based predictions need to be ‘‘adjustable” in a relatively simple way to adapt to alternative prescribed dose and RT planning protocols.

This work describes the implementation a well-known 3D ResNet DL architecture as a generator of ‘‘deep features” in the joint CT-RD representation. A fully-connected (FC) network is appended to the end of the ResNet to estimate class probabilities of RP based on deep features. We assumed a linearly-weighted mixture of CT and RD, with tunable weights, as the input. In the event of different prescribed doses or dose planning procedures at different institutions, we assumed that a baseline model has to be subsequently adjusted only for a different mixing ratio of CT and RD, as well as to retrain the FC component to use the new deep features resulting from the alternative mixing. However, the ResNet part will be kept frozen after training an initial baseline model.

Method

Study design

The overall flow in this study has been illustrated in Fig. 1. This study utilizes private data from a single institution to train a baseline model. Subsequent model adjustments and model performance evaluations used a prospectively collected cohort from the same single institution, plus the RTOG-0617 randomized trial dataset [17–20] split into two sets according to the prescribed dose (60 Gy in control arm and 72 Gy in the experiment arm). Grad-CAM heatmaps were overlaid on the input CT and RD to support clinical interpretation. Model discrimination was reported as receiver-operator ‘‘area under the curve” (AUC) and model calibration was assessed as goodness-of-fit for binary classification. The details of each part of the study are as follows.

Study population

All patients included this study had confirmed diagnosis of lung cancer and were treated with radically-intended radiotherapy (IMRT or VMAT), either with or without concurrent chemotherapy. The primary endpoint was symptomatic RP grade 2 or higher according to Common Terminology Criteria for Adverse Events (CTCAE) v4.0. In the private institutional datasets, the presence (or absence) of RP was assessed by experienced radiation oncologists based on follow-up CT, blood test and symptoms. In the RTOG dataset, the status of RP was documented in individual case reports. In this work, we considered only RP events which occurred anytime from the last fraction of radiotherapy up to 6 months after the last fraction of radiotherapy, as specifically RT-treatment induced RP.

The *Training set* consisted of 314 routine care patients retrospectively extracted from archives at one medical university cancer hospital. These were primarily intended for treatment with 60 Gy, but a range of delivered doses between 50 Gy to 70 Gy was prescribed at the treating physicians discretion. *Test-set-1* comprised of 35 prospectively registered patients from the same institution, also predominantly 60 Gy total intended dose, with variations of delivered dose at treating physician’s discretion. Training set and Test-set-1 were obtained with approval from an internal review board (ref. IRBbc2021135). The discretionary deviations in delivered dose were based on each patient’s overall physical condition and best achievable normal tissue constraints. Specific details of Training set and Test-set-1 are provided in Supplementary Materials 1A.

Access for secondary re-use of data from the prospectively randomized controlled trial RTOG-0617 was obtained through the trial sponsor. From the control arm (60 Gy prescribed dose), 194 subjects were defined as *Test-set-2*, and from the intervention arm (74 Gy prescribed dose), 158 subjects were allocated as *Test-set-3*. Specific details for filtering the RTOG-0617 subjects are provided in Supplementary Materials 1B.

Data preparation

Planning CT and RD were originally extracted in DICOM format for all subjects. The voxel-wise values in the RD images were scaled to represent absolute physical dose in units of Gy. We used a deep-learning automatic lung contouring tool based on previous work [21] to automatically segment whole lungs. Experienced radiation oncologists (ZZ and MY) inspected and (where needed) manually corrected the auto-generated lung masks to ensure accuracy and segmentation consistency. Data preparation and preprocessing steps are described in Supplementary Materials 2A.

Development of deep learning RP models

A 3D ResNet architecture was implemented as the main backbone of the RP model (see technical schematic in Supplementary Materials Figure S1). In brief, the pre-processed CT and RD arrays of the same dimensions were passed to the ResNet via linear mixing (W layer) immediately followed by a 7×7 convolution layer. In the W layer, we defined the composite input source as CT_d, where $CT_d = (A \times CT) + (B \times RD)$. A and B were thus the mixing ratio of CT and RD, such that A was always fixed at unity. Values of A and B were tuned as part of the model training process and were determined by back-propagation of the error.

The ResNet was used as an image-based ‘‘deep feature” generator; its weights were determined by training an initial baseline model and thereafter the entire ResNet weights were frozen. The RP classification model consisted of average pooling and a fully-connected (FC) layer at the end, which uses the deep feature maps generated by the ResNet in order to compute a class probability of RP at a sigmoid function layer. A purely binary classification (RP or non-RP) was computed by applying a threshold of 0.5. The core of the ResNet comprised eight repeating residual blocks containing convolution (conv), batch normalization (BN) and Rectified Linear Unit (ReLU) activation. We used an Adam optimizer with a learning

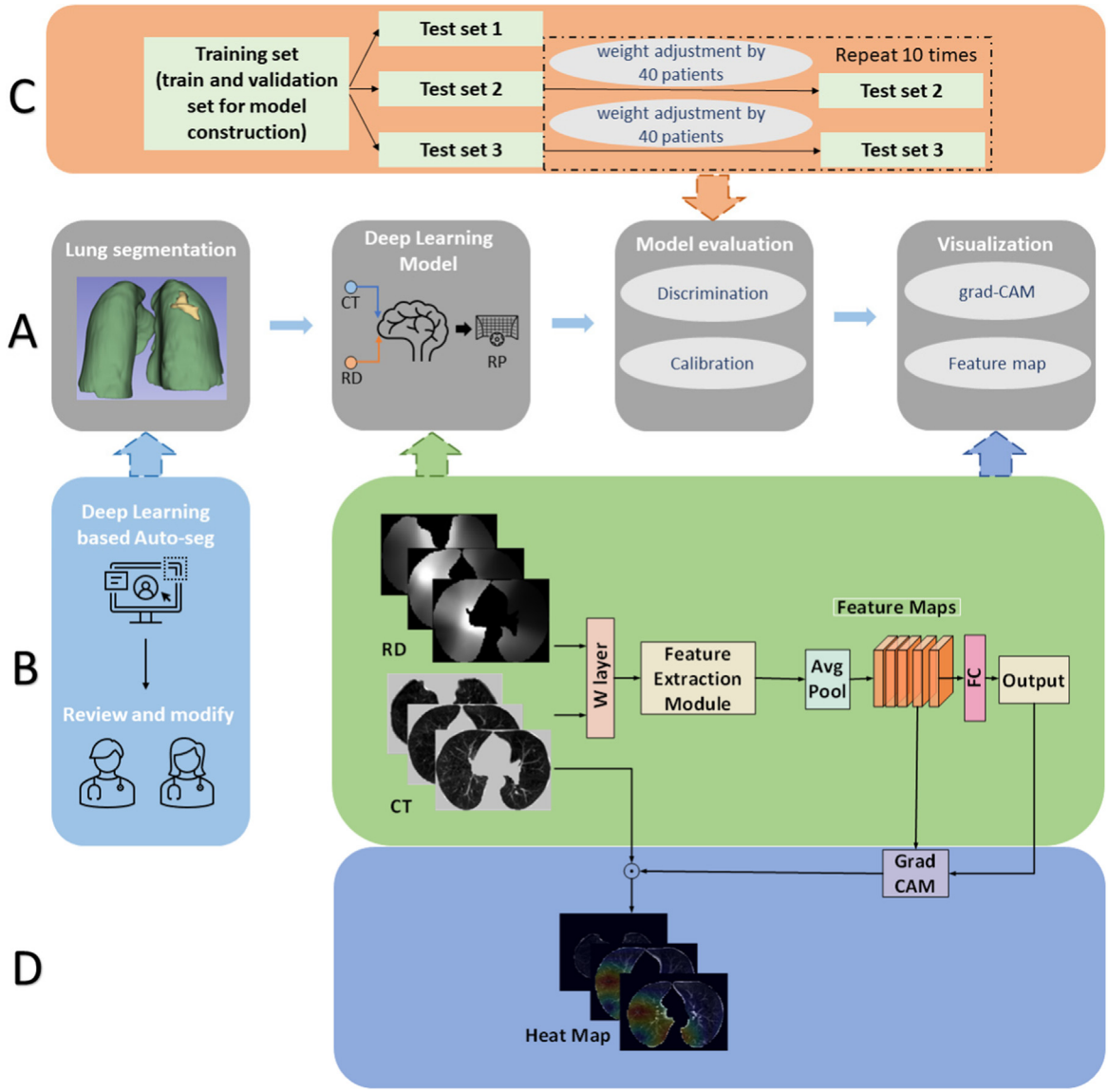


Fig. 1. A, The pipeline of this study: lung segmentation, model construction, model evaluation and visualization. B, Lung mask contouring using deep learning based automatic tool and reviewed and modified by two physicians. And model architecture. C, First, test-set-1, 2, and 3 were used to validate the base model built from training set. Second, forty patients from test-sets-2 and 3 were used to adjust the weight and fully connected layers, respectively, and validated by the test-sets-2 and 3 (without forty patients). This step was repeated ten times. D, Visualization of the model was achieved by the guided gradient weighted class activation mapping.

rate of 0.0001 and Binary Cross-Entropy as the loss function. The training strategy, loss function definition and model tuning hyper-parameters are shown in Supplementary Materials 2D.

After training using exclusively the Training set, the baseline model was evaluated in each of the three hitherto unseen cohorts i.e., Test-set-1 (medical university cancer hospital, 60 Gy prescribed), Test-set-2 (RTOG-0617 control arm, 60 Gy prescribed) and Test-set-3 (RTOG-0617 experiment arm, 74 Gy prescribed).

To examine the feasibility of "adjusting" the model for the same nominal prescribed dose but different planning protocol, we attempted two related experiments. First, we randomly chose 40 subjects from Test-set-2 *without replacement* and then proceeded to re-train only the CT-RD mixing ratio (i.e., the W layer) and the FC classifier – the ResNet was kept frozen as abovementioned.

As cross-validation, we evaluated the adjusted model using the remainder of Test-set-2 subjects (hereafter, Test-set-2* = the initial 194 subjects minus the 40 selected for adjustment = 154). To check for random vagaries of selecting 40 patients, we repeated the entire experiment 10 times, each time choosing different subsets of 40 patients. Secondly, to see if there was added value of using more patients, we adjusted the baseline model using all 194 subjects prescribed to 60 Gy in the RTOG control arm. However, it is no longer possible to check for over-optimism using repeated cross-validation, so 1000 times bootstrapping *with replacement* from Test-set-2 (hereafter Test-set-2$^{\#}$) was used to estimate a range of validation results.

To examine the feasibility of "adjusting" the model for simultaneously different prescribed dose and different planning protocol, we re-did the two related experiments above only utilizing Test-set-3.

To help visualize imaging and dose features that influence RP/non-RP prediction, and thus assist with clinical interpretation of the model attention area, activation heatmaps were generated by back-projecting Grad-CAM values as overlay on the planning CT and dose images (see detail in Supplementary Materials 2C).

Comparator RP models as alternatives to deep learning of mixed CT and RD models

A model employing only CT images and a model employing only RD images were constructed with the same process as the combined model described above for comparison.

We compared the aforementioned models against simple logistic regression based either on (i) dose-volume histograms (DVH) only, or (ii) clinical parameters only, or (iii) a combination of DVH and clinical parameters. Due to the high degree of correlation that is well-known in DVH metrics, we only considered V20 and mean lung dose (MLD) in the DVH-based model. For the clinical model, the patient age and presence of interstitial lung abnormalities were selected according to Supplementary Material Table S1. Further detailed information on the construction of the DVH model and clinical parameters model are provided in Supplementary Materials 5B.

Statistical analysis

The discrimination performance of the model was quantified using area under the receiver-operator curve (AUC), accuracy, sensitivity, and specificity of RP prediction. For all performance metrics reported, we estimated 95% confidence intervals by 1000 times bootstrapping. Goodness-of-fit was tested by calculating the model calibration error [22,23].

Patients' baseline characteristics for continuous variables are presented as mean ± standard deviation. For univariate analysis of clinical parameters, Pearson chi-squared tests and exact Fisher tests were used for categorical variables and logistic regression for continuous variables. For significance of clinical factors, a two-sided hypothesis test at the $\alpha = 0.05$ confidence level was assumed. Clinical and DVH data were analyzed in the Statistical Package for Social Science program (SPSS for Windows, version 27.0; SPSS Inc, Chicago, IL). All deep learning models were constructed and test set performance assessed using Python (version 3.8.5) and R software (version 4.0.5), respectively.

Code and data availability

Code packages and libraries for constructing our deep learning models are given in Supplementary Materials 2. The source code is made open access at https://gitlab.com/w654053334/rp_prediction.

The RTOG trial dataset may be obtained by contacting the sponsors for secondary re-use of data. Training set and Test-set-1 are private institutional collections, which may be made available to other researchers upon reasonable request and subject to data sharing agreements – please contact the corresponding author. To assist readers with using our RP model, we have prepared an open access online version with user interface (see Supplementary Materials 3).

Result

The characteristics of patients are shown in Table 1. Statistically significant heterogeneity between groups was observed across the majority of clinical factors, except for age and smoking. In Table 2, the clinical factors were grouped by RP versus non-RP. In univariate analysis, age, planning tumor volume (PTV), volume of the lung receiving 5 Gy (V5_lung) and 20 Gy (V20_lung), and mean lung dose (MLD) were each statistically significantly higher in patients with RP versus non-RP. Additional detailed clinical characteristics in the four datasets are given in Supplementary materials 1C (Table S1-3).

The predictive performance of models for RP is summarized in Table 3. The baseline model performed well on Test-set-1 (AUC 0.83) compared to Test-set-2 (AUC 0.55) and Test-set-3 (AUC 0.63). However, after adjustment, model discrimination was

Table 1
Patient characteristics in Training set, Test-set-1, 2, and 3.

Characteristics	Training set (n = 314) Mean ± SD	Test set 1 (n = 35) Mean ± SD	Test set 2 (n = 194) Mean ± SD	Test set 3 (n = 158) Mean ± SD	P-value
Age median, range	61 (30–85)	62 (34–75)	64 (37–82)	63 (41–82)	0.243
Gender					<0.001
Male	238 (75.8%)	23 (65.7%)	115 (59.3%)	89 (56.3%)	
Female	76 (24.2%)	12 (34.3%)	79 (40.7%)	69 (43.7%)	
Smoking					<0.001
Yes	71 (22.6%)	9 (25.7%)	14 (7.2%)	11 (7.0%)	
No	241 (76.8%)	26 (74.3%)	167 (86.1%)	144 (91.1%)	
Unknow	2 (0.6%)	0	13 (6.7%)	3 (1.9%)	
Histology					<0.001
LUSC	84 (26.8%)	8 (22.9%)	75 (38.7%)	70 (44.3%)	
LUAD	75 (23.9%)	10 (28.6%)	86 (44.3%)	63 (39.9%)	
LCU	–	–	4 (2.1%)	1 (0.6%)	
NOS	–	–	29 (14.9%)	24 (15.2%)	
SCLC	155 (49.4%)	17 (48.6%)	–	–	
Rt_technique					<0.001
3D-CRT	–	–	115 (59.3%)	81 (51.3%)	
IMRT	87 (27.7%)	5 (14.3%)	79 (40.7%)	77 (48.7%)	
VMAT	227 (72.3%)	30 (85.7%)	–	–	
Conso chemo					<0.001
Yes	179 (57.0%)	19 (54.3%)	173 (89.2%)	136 (86.1%)	
No	135 (43.0%)	16 (45.7%)	21 (10.8%)	22 (13.9%)	
PTV (cc)	446.82 ± 188.51	417.72 ± 179.70	507.93 ± 273.31	482.66 ± 261.40	0.014
V5_lung (%)	48.80 ± 10.15	48.82 ± 10.83	57.68 ± 15.29	57.11 ± 14.65	<0.001
V20_lung (%)	24.43 ± 5.24	24.06 ± 4.90	29.06 ± 7.47	31.22 ± 7.96	<0.001
MLD (Gy)	13.37 ± 2.62	13.06 ± 2.61	16.66 ± 4.15	19.16 ± 4.55	<0.001

Abbreviations: Pts = patients; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; LCU = Large cell undifferentiated; NOS = Non-small cell lung cancer; SCLC = small cell lung cancer; Rt_technique = radiotherapy technique used to treat patient; 3D-CRT = 3dimensional comformal radiation therapy; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; Conso chemo = consolidation chemotherapy; PTV = planning tumor volume; V5_lung = Lung V5 (%);V20_lung = Lung V20 (%); MLD = Mean lung dose (Gy).

improved in Test-set-2* (AUC 0.65) and Test-set-3* (AUC 0.70), respectively. The discrimination metrics of using only a subset of forty patients to adjust model were close to using the entire dataset, with small differences in AUC of 0.03 and 0.01, respectively. The accuracy, sensitivity and specificity largely followed the same pattern of findings as for AUC.

The mixing ratio, i.e. A and B weights, for CT and RD from each model were summarized in Supplementary Materials 4 Table S4. Across the baseline model and its subsequent adjustments, RD was overall more important than CT from an RP prediction perspective. Among nominally 60 Gy subjects, the post-training weight of RD relative to CT was reasonably stable around 1.5 (range 1.42–1.67). Among nominally 74 Gy subjects, the relative weight of RD to CT was suppressed to about 1.2 (range 1.20–1.25).

For comparison with the baseline model that included CT and RD, an alternative baseline model was constructed by either CT or RD alone and then tested on Test-set-1. The original baseline model (with both CT and RD, AUC 0.83) performed better than either CT-only or RD-only alternatives (AUC 0.63 for CT and 0.69 for RD, additionally accuracy, sensitivity and specificity were reported in Supplementary Materials 5A Table S5).

The discrimination of the DVH-based logistic model was poorer than that of the RD-only deep learning model (AUC 0.66 vs. 0.69) when evaluated in Test-set-1, and both were markedly poorer than the baseline model results. Discrimination of the logistic regression model based on clinical parameters (AUC 0.71 in Test-set-1) was poorer than the baseline model, but was slightly better than either of the RD-only deep learning and the DVH-only logistic regression.

The calibration error of the baseline model was 0.07 in Test-set-1, 0.22 in Test-set-2, and 0.18 in Test-set-3, indicating that there was no major calibration issue. However, after the model adjustment, the average expected calibration error was reduced to 0.14 for Test-set-2, and 0.13 for Test-set-3.

Some representative examples of 3D (Supplementary Materials Video) and 2D heatmaps (Fig. 2 and Supplementary Figure S11-13) generated by Grad-CAM may help to illustrate the global view for the whole lung and detailed view of each slice, respectively. In patients with pre-existing lung disease (the area indicated by the pointer in Fig. 2 and Supplementary Figure S11), such as interstitial lung abnormalities or emphysema, model attention appears more widely dispersed overall in the lungs. In contrast, for patients without pre-existing lung disease, relatively narrow distribution of model attention has been observed that follows the distribution of dose in the RD (Supplementary Figure S13). This clearly shows that, as far as the prediction of RP goes, a good model needs to be trained that can make use of (CT) features associated with pre-existing lung disease as well as (RD) features related to prominent dose distribution in the normal lungs.

Representative feature maps extracted from each residual block of the RP and non-RP cases are shown in Supplementary Materials 2E (Figure S2-9). As the level deepens in the model, the extracted features become more complex and abstract. While these features maps are very important since the FC layer uses these ResNet-generated feature maps to estimate the probability of RP, it nonetheless remains challenging to interpret the feature maps and thus visually associate them with clinically meaningful features. Thus, in this respect, the grad-CAM heatmaps overlaid onto the CT and RD might be potentially more useful by way of clinical interpretation.

Discussion

In this study, we used pre-treatment radiotherapy planning CT and planned radiation dose distribution to build a ResNet-based deep learning model to predict RP. The baseline model is trained using a joint representation of features from CT and RD, which we implemented using a linear mixing method of the intensity/-dose magnitudes. We then showed that such a baseline model can be subsequently adjusted by only re-training the mixing ratio (i.e., the W layer) and the FC classifier for RP, at the start and at the end of the ResNet, respectively, without changing any other weights in the ResNet feature extractor itself.

The combination of CT and RD predicted RP reasonably well in Test-set-1, which was expected since the test set most closely resembled the Training set in terms of prescribed dose, RT planning procedure and race cohort. Model performance and model calibration on the RTOG-0617 datasets, i.e., Test-set-2 and Test-set-3 were overall improved after adjusting the baseline model with either some or all of the each dataset.

However, the adjusted model did not perform as well on either of the RTOG-0617 subsets as it performed on Test-set-1. We hypothesize this is because RTOG-0617 data was contributed unevenly across 185 institutions [17], which may leave a large amount of heterogeneity among patients as well as residual differences between scanners, physicians delineations and RT planners that the trial protocol could not reconcile, as one can see in Table 1. It was interesting that the baseline model initially performed better in Test-set-3 (74 Gy) with higher AUC and sensitivity compared to Test-set-2 (60 Gy), which should have been closer to the prescription setting of the training dataset. However, we cannot rule out random chance since the baseline model initially performed sub-optimally for both Test-set-2 and Test-set-3. This may also suggest that treatment delivery modality may not be the critical factor for the model, at least relative to lung tissue and dose hotspots, and other sources of clinical heterogeneity may be more important. We are unable to resolve this question at present, and resolution of such questions needs more detailed study.

Grad-CAM heatmaps overlaid onto CT and RD suggested synergistic information for the prediction of RP, that is, the influential features point towards pre-existing lung injury in CT and regions of high dose in normal lung. Moreover, we proposed a computationally simplified way to adjust the model to fit different clinical settings. We suggest this a feasible method to adapt to different dose groups and planning protocols. However, it must be noted that even this limited adjustment-based retraining is still more computationally intensive than retraining a conventional machine learning model from scratch; as such, it is presently computationally unfeasible to perform more than a dozen repetitions of cross-validation or bootstraps during training.

This study included a retrospective single-institutional dataset as training set, and three other cohorts to evaluate the performance of our model. All test sets were prospectively collected to ensure the best available accuracy of registering the primary outcome of RP. In clinical practice, an RP event needs to be diagnosed by following up patients' symptoms and examinations. To distinguish RP from other types of pneumonia, follow-up CT examinations, routine blood tests, and C-reactive protein may be used. The endpoint of this study is grade 2 or higher RP, because patients with grade 2 RP require medical intervention and their activities of daily living are affected.

In this study, As mentioned, the relative importance of RD relative to CT was about 1.5 in most cases, except for the 74 Gy Test-set-3 where it appeared suppressed to about 1.2. A possible reason for this is that the standard dose (60 Gy) can induce RP in patients with intrinsic lung susceptibility to RP, but increasing prescribed dose to 74 Gy seems *not to be additionally effective* at inducing RP. Although the method proposed in this study is potentially an efficient way to update the baseline model for a new clinical setting, it is still possible to obtain a biased dataset with randomly sampling 40 patients [24], therefore if a larger dataset may be used for adjustment, we expect the model will be more robust.

Table 2
Patient characteristics group according to outcome of RP or without RP.

Characteristics	Without RP (n = 565) Mean ± SD	With RP (n = 136) Mean ± SD	P-value
Age median, range (years)	62 (30–86)	65 (38–80)	0.004
Gender			0.170
Male	368 (65.1%)	97 (71.3%)	
Female	197 (34.9%)	39 (28.7%)	
Smoking			0.229
Yes	464 (82.1%)	114 (83.8%)	
No	84 (14.9%)	21 (15.4%)	
Unknow	17 (3.0%)	1 (0.7%)	
Histology			0.926
LUSC	189 (33.5%)	48 (35.3%)	
LUAD	191 (33.8%)	43 (31.6%)	
LCU	5 (0.9%)	0	
NOS	43 (7.6%)	10 (7.4%)	
SCLC	137 (24.2%)	35 (25.7%)	
Histology			0.778
LUSC	189 (33.5%)	48 (35.3%)	
NSC-NSCLC	239 (42.3%)	53 (39.0%)	
SCLC	137 (24.2%)	35 (25.7%)	
Rt_technique			0.620
3D-CRT	162 (28.7%)	34 (25.0%)	
IMRT	200 (35.4%)	48 (35.3%)	
VMAT	203 (35.9%)	54 (39.7%)	
Conso chemo			0.046
Yes	418 (74.0%)	89 (65.4%)	
No	147 (26.0%)	47 (34.6%)	
PTV (cc)	459.54 ± 226.35	515.30 ± 253.67	0.013
V5_lung (%)	52.50 ± 13.66	55.72 ± 12.70	0.013
V20_lung (%)	26.91 ± 7.22	28.52 ± 6.92	0.020
MLD (Gy)	15.42 ± 4.37	16.21 ± 3.99	0.056

Abbreviations: Pts = patients; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; LCU = Large cell undifferentiated; NOS = Non-small cell lung cancer; SCLC = small cell lung cancer; Rt_technique = radiotherapy technique used to treat patient; 3D-CRT = 3dimensional comformal radiation therapy; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; Conso chemo = consolidation chemotherapy; PTV = planning tumor volume; V5_lung = Lung V5 (%);V20_lung = Lung V20 (%); MLD = Mean lung dose (Gy).

Some interesting points were found based on the attention maps (Fig. 2 and Supplementary Materials 6 Figure S11-13), where we tried to understand the "diagnostic" logic of the model. The results of this study are consistent with previous insights [25–30] and based on our data set, interstitial lung abnormality is an influential factor for the occurrence of RP. In the future, as the sample size expands, the model based only on patients with interstitial lung abnormalities can be developed and compared with the model developed in this study by Grad-CAM approach. Another feature is that the attentional areas tend to be located more in the central area of the lungs than in the peripheral areas. We speculate that there are two reasons for this phenomenon. First, RD is denser in the central part because of the irradiation of metastatic lymph nodes [28,29]. Secondly, the dose received by the heart may be another factor in the development of RP [30,31]. Krafft et al. found that cardiac DVH metrics improved the predictive power of radiomics models for RP prediction [12]. In our previous study, cardiac comorbidity was also found to be an independent predictor of RP [32].

Based on these observations, we speculate that the predictive logic of the model may be as follows: for patients with pre-existing lung disease, which was determined in collaboration with radiologists, the model pays attention to lung tissue with disease and analyzes these areas in conjunction with RD distribution. For patients with overall good (no lung disease) status, the model preferentially pays attention to regions of high dose and predicts RP mainly using the RD features. For most patients, the central part of the lung and the regions adjacent to the heart are more important than the peripheral lung. We also compared an RD-only deep learning model with the DVH-based model, which is another commonly used model in clinical practice. From the results, the predictive power of the DVH-based model is not better than that of the RD-based deep learning model.

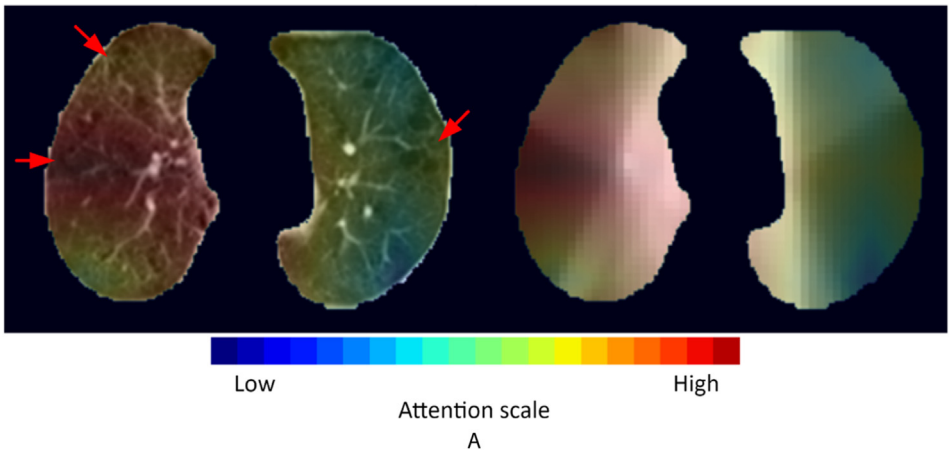
The result of this study has a few real-world clinical implications. In this study, we did not iteratively tune the decision threshold of the model. In practice, we may select the thresholds that prioritize either higher sensitivity or higher specificity, but we could not do both. Patients with very low probability of RP could receive standard or adequate doses if adjusted models with high specificity were used for these hospitals, which might improve their prognosis [1,2]. For patients with a very high probability of RP, physicians can give these patients more frequent examinations or preventive medications to lower the grade of RP or prevent it from occurring [33]. Alternatively, this clinical tool may be of assistance during the doctor-patient consultation about risks and expectations of treatment.

There were several limitations in this study. The deep learning model with complex neural networks needs a large dataset to avoid overfitting. We included 701 patients in this study and although, to best of our knowledge, this is the largest dataset on the topic of artificial intelligence model to predict RP, model development will benefit further from even larger datasets including

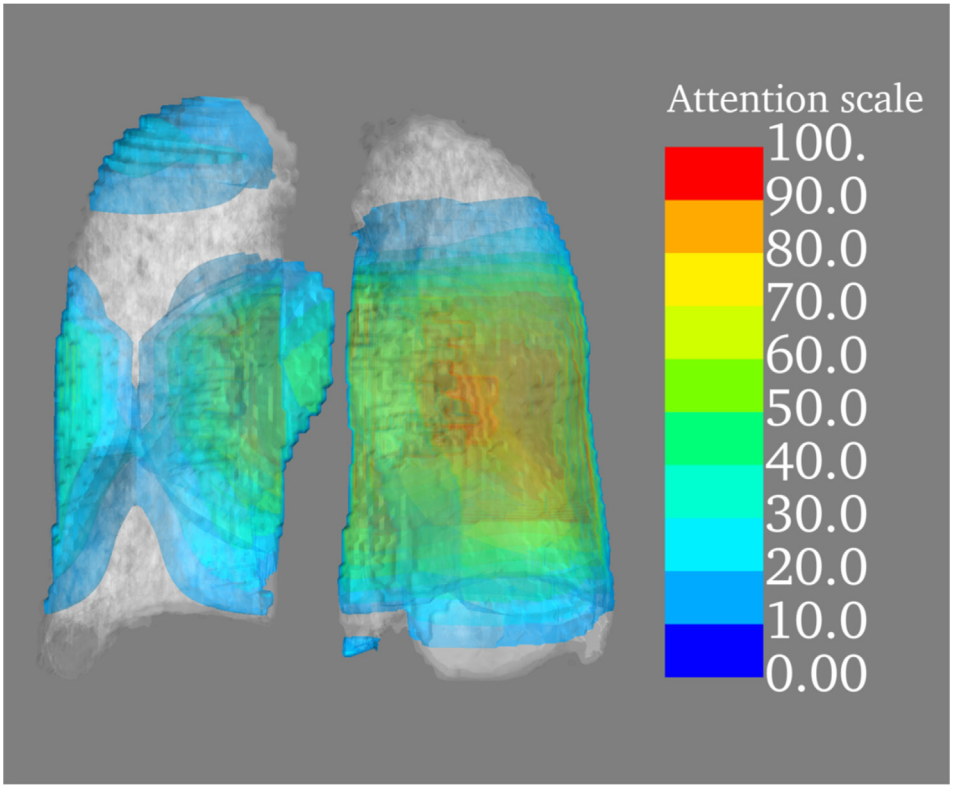
Table 3
Performance of baseline model and adjustments (using 60 Gy and 74 Gy arms of RTOG-0617 trial).

Model	Adjustment	Evaluation	AUC (95%CI)	Accuracy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
Baseline model	No adjustment	**Test-set-1** (35)	**0.83** (0.82–0.91)	**0.82** (0.75–0.81)	**0.70** (0.67–0.77)	**0.88** (0.83–0.89)
		Test-set-2 (194)	**0.55** (0.47–0.69)	**0.70** (0.57–0.82)	**0.41** (0.39–0.52)	**0.69** (0.61–0.84)
		Test-set-3 (158)	**0.63** (0.53–0.72)	**0.66** (0.60–0.73)	**0.60** (0.46–0.74)	**0.68** (0.61–0.75)
Adjusted for RTOG 60 Gy arm	40 randomly selected from Test-set-2	**Test-set-2*** (154)	**0.65** (0.54–0.77)	**0.76** (0.63–0.91)	**0.58** (0.48–0.83)	**0.70** (0.66–0.98)
Adjusted for RTOG 60 Gy arm	All subjects from Test-set-2	**Test-set-2#** (194 bootstrap samples)	**0.68** (0.58–0.83)	**0.78** (0.80–0.89)	**0.77** (0.62–0.97)	**0.65** (0.60–0.74)
Adjusted for RTOG 74 Gy arm	40 randomly selected from Test-set-3	**Test-set-3*** (118)	**0.70** (0.63–0.76)	**0.71** (0.63–0.83)	**0.62** (0.56–0.86)	**0.73** (0.67–0.95)
Adjusted for RTOG 74 Gy arm	All subjects from Test-set-3	**Test-set-3#** (158 bootstrap samples)	**0.71** (0.62–0.81)	**0.78** (0.73–0.84)	**0.68** (0.54–0.83)	**0.77** (0.71–0.82)

Abbreviations: AUC = area under receiver operating characteristic curve; 95% CI = 95% confidence interval; * the asterisk indicates that the coefficients of CT and RD for this model are adjusted with 40 patients for each set; # The pound symbol indicates that the coefficients of CT and RD for this model are adjusted using the entire data set. The number in parentheses are the sample size for the evaluations.



A



B

Fig. 2. Illustration of attention (heat) map of a 60-year-old male with non-small cell lung cancer. A, Two-dimensional attention map. The left image is an overlay of the CT image and the attention map, with blue to red representing increasing levels of importance (attention scale). The area of interstitial lung abnormalities indicated by pointers. The image on the right is an overlay of the radiation dose (RD) image and the attention map. From dark to light represents low to high dose, and from blue to red represents increasing importance. B, Three-dimensional attention map. The different colors represent different levels of importance (attention scale).

heterogeneity of CT scanners, dose planning systems, etc. different institutions with improvements expected both in terms of performance and in terms of generalizability across backgrounds, scanners, treatment strategies and patients. Second, this model did not include combinations of clinical parameters including cytokines. Our previous studies and others have demonstrated that it has predictive value for RP [34,35]. The combination of cytokines could improve the performance of the model [36], however, the present aim of our study was to focus on a non-invasive approach to modeling and therefore cytokines were not included. A potential benefit is that this model can be directly embedded into RT planning systems, as it only needs CT and RD information and can export its predictions directly to other systems for clinical decision support. In addition, patients included in this study did not receive concurrent chemotherapy with the same regimen, and we think that the predictive power of the model could be improved if clinical factors were harmonized. On the reverse side, it is difficult to maintain the same treatment regimen everywhere in the world, and the generalizability of the model would be affected if only patients receiving the same chemotherapy regimen were included.

Finally, we did not include patients who received immunotherapy, which is already a standard therapy for local advanced lung cancer patients now. And the incidence of pneumonitis is higher with the addition of durvalumab after concurrent chemoradiotherapy [37,38]. There are still challenges to be addressed before including patients receiving immunotherapy in the analysis,

such as differential diagnosis of immune checkpoint inhibitor therapy-related pneumonitis and RP and datasets containing large sample sizes of patients receiving immunotherapy. The model we developed in this study can served as a base (pre-trained) model for future studies that include patients receiving immunotherapy [39].

In summary, we successfully developed a deep learning model to predict RP, and this model can be adjusted easily to fit new cohorts. We tried to uncover the model prediction logic by a visualization approach. In addition, a ready-to-use online software was developed to assist clinical practice. Despite several limitations, we believe that deep learning algorithm possesses great potential to sever as a clinical assistant tool.

Conflict of 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.

Acknowledgments

This manuscript was prepared using data from datasets (RTOG-0617; NCT00533949-D1, D2, D3) from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data were originally collected from a clinical trial (identifier NCT00533949; "A Randomized Phase III Comparison of Standard-Dose (60 Gy) Versus High-Dose (74 Gy) Conformal Radiotherapy with Concurrent and Consolidation Carboplatin/Paclitaxel +/− Cetuximab (IND #103444) in Patients With Stage IIIA/IIIB Non-Small Cell Lung Cancer"). 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, the NCORP or the NCI.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109581.

References

[1] Luo H-S, Huang H-C, Lin L-X. Effect of modern high-dose versus standard-dose radiation in definitive concurrent chemo-radiotherapy on outcome of esophageal squamous cell cancer: a meta-analysis. Radiat Oncol 2019;14:178. https://doi.org/10.1186/s13014-019-1386-x.

[2] Ladbury CJ, Rusthoven CG, Camidge DR, Kavanagh BD, Nath SK. Impact of radiation dose to the host immune system on tumor control and survival for Stage III non-small cell lung cancer treated with definitive radiation therapy. Int J Radiat Oncol*Biol*Phys 2019;105:346–55. https://doi.org/10.1016/j.ijrobp.2019.05.064.

[3] Arroyo-Hernández M, Maldonado F, Lozano-Ruiz F, Muñoz-Montaño W, Nuñez-Baez M, Arrieta O. Radiation-induced lung injury: current evidence. BMC Pulm Med 2021;21:9. https://doi.org/10.1186/s12890-020-01376-4.

[4] Ullah T, Patel H, Pena GM, Shah R, Fein AM. A contemporary review of radiation pneumonitis. Curr Opin Pulm Med 2020;26:321–5. https://doi.org/10.1097/MCP.0000000000000682.

[5] Käsmann L, Dietrich A, Staab-Weijnitz CA, Manapov F, Behr J, Rimner A, et al. Radiation-induced lung toxicity - cellular and molecular mechanisms of pathogenesis, management, and literature review. Radiat Oncol (London, England) 2020;15:214. https://doi.org/10.1186/s13014-020-01654-9.

[6] Liu Y, Wang W, Shiue K, Yao H, Cerra-Franco A, Shapiro RH, et al. Risk factors for symptomatic radiation pneumonitis after stereotactic body radiation therapy (SBRT) in patients with non-small cell lung cancer. Radiother Oncol: J Eur Soc Therapeutic Radiol Oncol 2021;156:231–8. https://doi.org/10.1016/j.radonc.2020.10.015.

[7] Saha A, Beasley M, Hatton N, Dickinson P, Franks K, Clarke K, et al. Clinical and dosimetric predictors of radiation pneumonitis in early-stage lung cancer treated with Stereotactic Ablative radiotherapy (SABR) - An analysis of UK's largest cohort of lung SABR patients. Radiother Oncol: J Eur Soc Therapeutic Radiol Oncol 2021;156:153–9. https://doi.org/10.1016/j.radonc.2020.12.015.

[8] Prasanna PG, Rawojc K, Guha C, Buchsbaum JC, Miszczyk JU, Coleman CN. Normal tissue injury induced by photon and proton therapies: gaps and opportunities. Int J Radiat Oncol Biol Phys 2021;110:1325–40. https://doi.org/10.1016/j.ijrobp.2021.02.043.

[9] Bourbonne V, Da-Ano R, Jaouen V, Lucia F, Dissaux G, Bert J, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. Radiother Oncol: J Eur Soc Therapeutic Radiol Oncol 2021;155:144–50. https://doi.org/10.1016/j.radonc.2020.10.040.

[10] Wang L, Gao Z, Li C, Sun L, Li J, Yu J, et al. Computed tomography-based delta-radiomics analysis for discriminating radiation pneumonitis in patients with esophageal cancer after radiation therapy. Int J Radiat Oncol Biol Phys 2021. https://doi.org/10.1016/j.ijrobp.2021.04.047.

[11] Du F, Tang N, Cui Y, Wang W, Zhang Y, Li Z, et al. A novel nomogram model based on cone-beam CT radiomics analysis technology for predicting radiation pneumonitis in esophageal cancer patients undergoing radiotherapy. Front Oncol 2020;10:. https://doi.org/10.3389/fonc.2020.596013596013.

[12] Krafft SP, Rao A, Stingo F, Briere TM, Court LE, Liao Z, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. Med Phys 2018;45:5317–24. https://doi.org/10.1002/mp.13150.

[13] Puttanawarut C, Sirirutbunkajorn N, Tawong N, Jiarpinitnun C, Khachonkham S, Pattaranutaporn P, et al. Radiomic and dosiomic features for the prediction of radiation pneumonitis across esophageal cancer and lung cancer. Front Oncol 2022;12.

[14] Zhang Z, Wang Z, Yan M, Yu J, Dekker A, Zhao L, et al. Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: a model development study with prospective external validation and decision-curve analysis. Int J Radiat Oncol Biol Phys 2022;S0360-3016:03189-3. https://doi.org/10.1016/j.ijrobp.2022.08.047.

[15] Wei J, Zhang Z, Yu J, Jia H, Tian J, Meng C, et al. Meta-analysis of the incidence of radiation pneumonitis between European, American and Asian populations. Chinese J Radiat Oncol 2021;30:556–62. https://doi.org/10.3760/cma.j.cn113030-20201114-00554.

[16] Liu Z, Liu W, Ji K, Wang P, Wang X, Zhao L. Simultaneous integrated dose reduction intensity-modulated radiotherapy applied to an elective nodal area of limited-stage small-cell lung cancer. Exp Ther Med 2015;10:2083–7. https://doi.org/10.3892/etm.2015.2835.

[17] Bradley JD, Paulus R, Komaki R, Masters G, Blumenschein G, Schild S, 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;16:187–99. https://doi.org/10.1016/S1470-2045(14)71207-0.

[18] Bradley JD, Hu C, Komaki RR, Masters GA, Blumenschein GR, Schild SE, et al. Long-term results of NRG oncology RTOG 0617: standard- versus high-dose chemoradiotherapy with or without cetuximab for unresectable stage III non-small-cell lung cancer. J Clin Oncol 2020;38:706–14. https://doi.org/10.1200/JCO.19.01162.

[19] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57. https://doi.org/10.1007/s10278-013-9622-7.

[20] Bradley J, Forster K. Data from NSCLC-Cetuximab. Cancer Imag Archive 2018. https://doi.org/10.7937/TCIA.2018.jze75u7v.

[21] Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 2020;4:50. https://doi.org/10.1186/s41747-020-00173-2.

[22] Shi Z, Zhang Z, Liu Z, Zhao L, Ye Z, Dekker A, et al. Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal cancer: a systematic review of clinical outcome prediction after concurrent chemoradiotherapy. Eur J Nucl Med Mol Imaging 2021. https://doi.org/10.1007/s00259-021-05658-9.

[23] Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. Proc Conf AAAI Artif Intell 2015;2015:2901–7.

[24] Shahinfar S, Meek P, Falzon G. "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Eco Inform 2020;57:101085. https://doi.org/10.1016/j.ecoinf.2020.101085.

[25] Kocak Z, Evans ES, Zhou S-M, Miller KL, Folz RJ, Shafman TD, et al. Challenges in defining radiation pneumonitis in patients with lung cancer. Int J Radiat Oncol Biol Phys 2005;62:635–8. https://doi.org/10.1016/j.ijrobp.2004.12.023.

[26] Doi H, Nakamatsu K, Nishimura Y. Stereotactic body radiotherapy in patients with chronic obstructive pulmonary disease and interstitial pneumonia: a review. Int J Clin Oncol 2019;24:899–909. https://doi.org/10.1007/s10147-019-01432-y.

[27] Okumura M, Hojo H, Nakamura M, Hiyama T, Nakamura N, Zenda S, et al. Radiation pneumonitis after palliative radiotherapy in cancer patients with interstitial lung disease. Radiother Oncol: J Eur Soc Therapeutic Radiol Oncol 2021;161:47–54. https://doi.org/10.1016/j.radonc.2021.05.026.

[28] Jiang X, Li T, Liu Y, Zhou L, Xu Y, Zhou X, et al. Planning analysis for locally advanced lung cancer: dosimetric and efficiency comparisons between intensity-modulated radiotherapy (IMRT), single-arc/partial-arc volumetric modulated arc therapy (SA/PA-VMAT). Radiat Oncol 2011;6:140. https://doi.org/10.1186/1748-717X-6-140.

[29] Chang JY. Intensity-modulated radiotherapy, Not 3 dimensional conformal, is the preferred technique for treating locally advanced lung cancer. Semin Radiat Oncol 2015;25:110–6. https://doi.org/10.1016/j.semradonc.2014.11.002.

[30] Shepherd AF, Iocolano M, Leeman J, Imber BS, Wild AT, Offin M, et al. Clinical and dosimetric predictors of radiation pneumonitis in patients with non-small cell lung cancer undergoing postoperative radiation therapy. Pract Radiat Oncol 2021;11:e52–62. https://doi.org/10.1016/j.prro.2020.09.014.
[31] Keffer S, Guy CL, Weiss E. Fatal radiation pneumonitis: literature review and case series. Adv Radiat Oncol 2020;5:238–49. https://doi.org/10.1016/j.adro.2019.08.010.
[32] Nalbantov G, Kietselaer B, Vandecasteele K, Oberije C, Berbee M, Troost E, et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiother Oncol: J Eur Soc Therapeutic Radiol Oncol 2013;109:100–6. https://doi.org/10.1016/j.radonc.2013.08.035.
[33] Konkol M, Śniatała P, Milecki P. Radiation-induced lung injury — what do we know in the era of modern radiotherapy? Reports Practical Oncol Radiother 2022. https://doi.org/10.5603/RPOR.a2022.0046.
[34] Niu X, Li H, Chen Z, Liu Y, Kan M, Zhou D, et al. A study of ethnic differences in TGFβ1 gene polymorphisms and effects on the risk of radiation pneumonitis in non-small-cell lung cancer. J Thorac Oncol 2012;7:1668–75. https://doi.org/10.1097/JTO.0b013e318267cf5b.
[35] Zhao L, Wang L, Ji W, Wang X, Zhu X, Hayman JA, et al. Elevation of plasma TGF-beta1 during radiation therapy predicts radiation-induced lung toxicity in patients with non-small-cell lung cancer: a combined analysis from Beijing and Michigan. Int J Radiat Oncol Biol Phys 2009;74:1385–90. https://doi.org/10.1016/j.ijrobp.2008.10.065.
[36] Wang L, Liang S, Li C, Sun X, Pang L, Meng X, et al. A novel nomogram and risk classification system predicting radiation pneumonitis in patients with esophageal cancer receiving radiation therapy. Int J Radiat Oncol Biol Phys 2019;105:1074–85. https://doi.org/10.1016/j.ijrobp.2019.08.024.
[37] Bi J, Qian J, Yang D, Sun L, Lin S, Li Y, et al. Dosimetric risk factors for acute radiation pneumonitis in patients with prior receipt of immune checkpoint inhibitors. Front Immunol 2021;12:828858. https://doi.org/10.3389/fimmu.2021.828858.
[38] Vansteenkiste J, Naidoo J, Faivre-Finn C, Özgüroğlu M, Villegas A, Daniel D, et al. MA05.02 PACIFIC subgroup analysis: pneumonitis in stage III, unresectable NSCLC patients treated with durvalumab vs. placebo after CRT. J Thorac Oncol 2018;13:S370–1. https://doi.org/10.1016/j.jtho.2018.08.350.
[39] Greenspan H, Bv G, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35:1153–9. https://doi.org/10.1109/TMI.2016.2553401.