# Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer

V. Bourbonne, R. Da-Ano, V. Jaouen, F. Lucia, G. Dissaux, J. Bert, O. Pradier, D. Visvikis, M. Hatt, U. Schick

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167814020308872
Manuscript_78071bdc2d2e63585b57bc9767cd1747

# Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer

V. Bourbonne[1,2], MD. MSc, R. Da-ano[2], MSc, V. Jaouen[2], PhD, F. Lucia[1,2], MD. MSc, G. Dissaux[1,2], MD. MSc, J. Bert, PhD[2], O. Pradier[1,2], MD. PhD, D. Visvikis[2], PhD, M. Hatt[2*], PhD, U. Schick[1,2*], MD. PhD

1. Radiation Oncology Department, University Hospital, Brest, France
2. LaTIM, UMR 1101, INSERM, Univ Brest, Brest, France

*equally contributed

**Corresponding author :**

Dr Vincent Bourbonne, MD, MSc

Radiation Oncology Department

CHRU Brest

2 avenue Foch, 29200 Brest, France

**Mail :** vincent.bourbonne@chu-brest.fr

**Tel :** +33298223398

**Short running title:** radiomics prediction of lung RT's toxicity

**Conflict of interest:** none

**Funding source:** none

**Highlights:**

- The miscalibration of pulmonary and esophageal toxicities in patients with lung cancer treated by (chemo)-radiotherapy is frequent.
- Usual dose-volume histograms do not account for dose spatial distribution.
- Radiomics based models contribute to a significant improvement in acute and late pulmonary toxicities prediction. For acute esophageal toxicity, dosimetric and radiomics based models achieve similar results.

**Keywords :**

- Toxicities prediction
- Lung Cancer
- Radiomics
- Dose spatial distribution

**Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer**

**Purpose:** (Chemo)–radiotherapy (RT) is the gold standard treatment for patients with locally advanced lung cancer non accessible for surgery. However, current toxicity prediction models rely on clinical and dose volume histograms (DVHs) and remain unsufficient. The goal of this work is to investigate the added predictive value of the radiomics approach applied to dose maps regarding acute and late toxicities in both the lungs and esophagus.

**Methods:** Acute and late toxicities scored using the CTCAE v4.0 were retrospectively collected on patients treated with RT in our institution. Radiomic features were extracted from 3D dose maps considering Gy values as grey-levels in images. DVH and usual clinical factors were also considered. Three toxicity prediction models (clinical only, clinical + DVH and combined*, i.e.,* including clinical + DVH + radiomics) were incrementally trained using a neural network on 70% of the patients for prediction of grade ≥ 2 acute and late pulmonary toxicities (APT/LPT) and grade ≥ 2 acute esophageal toxicities (AET). After bootstrapping (n = 1000), optimal cut-off values were determined based on the Youden Index. The trained models were then evaluated in the remaining 30% of patients using balanced accuracy (BAcc).

**Results:** 167 patients were treated from 2015 to 2018: 78% non small-cell lung cancers, 14% small-cell lung cancers and 8% other histology with a median age at treatment of 66 years. Respectively, 22.2%, 16.8% and 30.0% experienced APT, LPT and AET. In the training set (*n=117*), the corresponding BAcc for clinical only/clinical + DVH/combined were 0.68/0.79/0.92, 0.66/0.77/0.87 and 0.68/0.73/0.84. In the testing evaluation (*n=50*), these trained models obtained a corresponding BAcc of 0.69/0.69/0.92, 0.76/0.80/0.89 and 0.58/0.73/0.72.

**Conclusion:** In patients with a lung cancer treated with RT, radiomic features extracted from 3D dose maps seem to surpass usual models based on clinical factors and DVHs for the prediction of APT and LPT.

## INTRODUCTION

With an incidence of over 385.000 cases/year and a crude rate of 52.2 deaths/100.000, lung cancer is the first cause of death by cancer in Europe [1]. (Chemo)–Radiotherapy is the gold standard therapeutic option for patients with locally advanced lung cancer non accessible or ineligible for surgery [2]. In the definitive setting or as a neoadjuvant treatment, it has led to good clinical outcomes at the cost of high treatment-related toxicities. Modern techniques such as intensity-modulated radiotherapy (IMRT) or Volumetric Modulated Arctherapy (VMAT) [3,4] offer high target conformation but remain poorly used in this setting because of the unknown effects of the "low dose" bath, especially with VMAT.

A better assessment of the of toxicities' risk could result in substantial treatment modifications such as dose-escalation for patients at low risk of toxicity or treatment optimization for patients at high risk of toxicity.

Most existing toxicity prediction models rely only on clinical factors and dose-volume histograms (DVH). However, the strict application of the corresponding current dose constraints does not prevent serious toxicities in some patients. This may be explained by the fact that DVH do not efficiently account for spatial dose distribution or organ architecture. Radiomic features are statistical, geometrical or textural metrics designed to provide quantitative measurements of intensity, shape or heterogeneity of a given volume of interest (VOI) in medical images [5]. These features could relate to dose distribution heterogeneity when applied to dose maps [6].

Current normal tissue complication probability models (NTCP) mainly use logistic regression for model building. Machine learning methods, especially artificial neural networks (NN) [7], have the potential to efficiently model the synergistic interaction between variables using a flexible nonlinear relationship and could optimize prediction capacity.

The goal of the present work was to investigate the added predictive value of radiomic features extracted from dose maps for acute and late pulmonary and esophageal toxicities using a NN's approach.

## MATERIAL AND METHODS

### Population

All patients treated with a curative intent from 2015 to 2018 at our institution were considered. Patients with an age > 18 yo, a histologically proven lung cancer (non-small cell or small cell lung cancer), a treatment by (chemo)-radiotherapy with VMAT in a curative setting and a minimum follow-up of 1 year after radiotherapy (RT) completion were included. When performed, chemotherapy could be delivered as a sequential or a concomitant treatment. The study was approved by the hospital ethical committee (B2020CE.34).

### Toxicities

Acute and late toxicities, defined and scored using the CTCAE v4.0 [8], were retrospectively collected for all patients. Acute pulmonary or esophageal toxicity (APT or AET) was defined as a grade ≥ 2 toxicity event occurring during the 6 months following the start of RT treatment. Late pulmonary or esophageal toxicity (LPT or LET) was defined as a grade ≥ 2 toxicity event occurring later than 6 months after the start of RT. Actuarial incidences were used for both the acute and late toxicities.

### Clinical and dosimetric factors

Usual clinical such as age, mean expiratory volume/second (MEVS), clinical (CTV) and planning target volume (PTV) and dosimetric factors [9,10] were included, according to current recommendations [11,12]. Vx (Gray: Gy) will further be defined as the percentage of the volume

of interest (VOI) receiving x dose (Gy). The full list of included features is available in the Supplementary Materials.

**Dose map conversion**

The full protocol for the dose-map conversion is available in the Supplementary Materials. Three-dimensional dose maps corresponding to the delivered treatment plans were extracted and converted, using the Platismatch® extension in Slicer®, as a grey-level volume where the voxel value equals the corresponding dose. VOIs delineation were recovered from the treatment planning (target volumes, contralateral and homolateral lungs and the esophagus).
Radiomic features extraction on the dose-maps was performed using an in-house developed software (MIRAS® [13,14]) compliant with the most up-to-date Image Biomarker Standardisation Initiative (IBSI) guidelines and benchmark values [15,16], resulting in 352 features per patient. From here onwards, each feature will be denoted as $x\_y_z$, with x being the organ, y the feature name and z the matrix.

**Features selection and model building**

Features were processed for classification and toxicity modeling via a supervised NN approach relying on the NN library in SPSS Statistics® (Perceptron Multilayer Network). The detailed protocol is presented in the Supplementary Materials.
Three models were incrementally built: using only clinical factors as inputs, adding DVH metrics to clinical factors, and finally considering all available parameters i.e., clinical, DVH and dose map radiomic features as inputs to the network.
The cohort was partitioned into a training set (~70%, *n=117*) and a testing set (~30%, *n=50*) using the stratified random sampling method [17]. In the training set, an evaluation of the importance of independent features in the NN determination was performed inside each model. The maximum number of features accepted in a model was set to be lower than 10% of the sample size [18]. If more features than this cut-off value ended up being selected, the features having the lowest importance were deleted and the model retrained.
A final exploratory step was added defining the number of retained features on the number of cases, and not the overall sample size. Given the expected imbalance between the positive and negative cases, each positive case was associated with 2 negatives cases; together making the number of cases. If more features than 10% of the cases were selected, the feature having the lowest importance was deleted and the model retrained. This last step stopped as soon as a drop $\geq 5\%$ in performance (Area Under the Curve: AUC) was observed.
In the training set, the performance of all models was evaluated and compared using the AUC and the $R^2$ (R-Squared : square of the Pearson correlation coefficient). Optimal cut-off values for each feature/model were determined according to the Youden Index. Quantitative performance evaluation was carried out using balanced accuracy (BAcc) defined as the average of sensitivity (Se) and specificity (Sp), regarding the prediction of each aforementioned toxicity event.
To further enhance the robustness of the models and their selected features, the bootstrap resampling method with $n = 1000$ replications was used relying upon the bagging library implemented in SPSS Modeler®. For each sub-sample, a new NN was trained using the previously selected features resulting in an individual performance (assessed using BAcc). Results over all 1000 replicates were then reported as a mean BAcc.
Finally, the best trained models for each endpoint prediction were evaluated in the testing set using the BAcc, the AUC and the $R^2$.

To prevent the development of miscalibrated models, toxicity prediction models were built only if the toxicity rate was $\geq 10\%$.

**RESULTS**

Between 2015 and 2018, 167 patients were treated in our institution. Main patients' characteristics are summarized in table 1. No significant difference was observed between the training and the testing sets.

After a median follow-up of 14.0 months, the grade ≥2 APT, LPT, AET and LET rates for the entire cohort were 22.2%, 16.8%, 30.0% and 5.4%, respectively (table 2). The LET rate being < 10%, no prediction model was developed for this specific toxicity.

With a training cohort of 117 patients, the maximum number of features per model was set to 11 features (10% of the training sample size). Regarding the exploratory rule (10% of the cases in the training set), the number of features was defined as 7 for APT, 5 for LPT and 10 for AET. Description and abbreviation of each further described radiomics feature combined in the NTCP models can be found in supplemental Table 1.

After pre-selection, the clinical model for APT was based on the combination of 5 features (age, MEVS, CTV and PTV volumes and RT duration) resulting in the training set in an AUC of 0.67 ($p$ = 0.005) and a BAcc of 0.68. In the testing set, the model resulted in a BAcc of 0.69. The clinical + DVH model for APT was also based on the combination of 5 features (age, MEVS, CTV and PTV volumes and V20Gy to both lungs) with an AUC of 0.86 ($p < 0.0001$) and a BAcc of 0.79, in the training set and a BAcc of 0.69 in the testing set. The combined model for APT was based on 5 radiomic features extracted from the Grey-Level Co-Occurrence Matrix (GLCM) which are $LungH\_IC1_{GLCM}$, $LungH\_Entropy_{GLCM}$, $LungH\_Contrast_{GLCM}$, $LungH\_DVAR_{GLCM}$, $LungH\_Var_{GLCM}$ and 1 feature extracted from the Histogram (Hist) $Lungs\_Energy_{Hist}$, the normalized weight of each feature being available as Supplementary Figure 1. This model was strongly associated with APT in the training set: AUC 0.92 ($p < 0.0001$), Bacc of 0.92. In the testing set, the model remained highly predictive, reaching a BAcc of 0.92. ROC curves for each APT prediction model are available as figure 1a (training set) and figure 2a (testing set). Example of the NN for the APT combined prediction model can be found in Supplementary figure 2. Concordance between observed and predicted APT events are presented in Supplemental figure 3.

After pre-selection, the clinical model for LPT was based on the combination of 2 features (CTV and PTV volumes) resulting in an AUC of 0.58 ($p$ = 0.40) and a BAcc of 0.66 in the training set. In the testing set, this model achieved a BAcc of 0.76. The clinical + DVH model for LPT was based on the combination of 6 features (CTV and PTV volumes, V5Gy, V10Gy and V13Gy to the homolateral lung and V13Gy to both lungs) with an AUC of 0.79 ($p < 0.0001$) and a BAcc of 0.77 in the training set. It remained highly predictive in the testing set, reaching a BAcc of 0.80. The combined model for LPT was based on 9 radiomic features extracted from the Grey-Level Size Zone Matrix (GLSZM): $LungC\_LZSE_{GLSZM}$, $LungC\_LZLGE_{GLSZM}$ $LungC\_LZHGE_{GLSZM}$, $LungH\_LZSE_{GLSZM}$, $LungH\_LZLGE_{GLSZM}$, $LungH\_LZHGE_{GLSZM}$, $LungH\_ZSVAR_{GLSZM}$, $Lungs\_Prominence_{GLCM}$ and $Lungs\_LZHGE_{GLSZM}$, the normalized weight of each feature being available as Supplementary Figure 4. This model led to an AUC of 0.89 ($p < 0.0001$) and a BAcc of 0.87. In the testing set, the model remained highly predictive with a BAcc of 0.89. ROC curves for each LPT prediction model are available as figure 1b (training set) and figure 2b (testing set). Concordance between observed and predicted LPT events are presented in Supplemental figure 5.

Regarding AET, two robust prediction models were built with respective AUCs of 0.78 and 0.85 in the training cohort (Clinical + DVH and Combined). On the testing cohort, the 2 models performed similarly with respective Baccs of 0.73 and 0.72. None of the clinical or DVH variables were retained in the Combined model, consisting of 5 radiomic features which were Oesophagus_Average$_{GLCM}$, Oesophagus_DiffAverage$_{GLCM}$, Oesophagus_DVAR$_{GLCM}$, Oesophagus_Contrast$_{GLCM}$, Oesophagus_IC2$_{GLCM}$ and Oesophagus_Contrast$_{GLCM}$, the normalized weight of each feature being available as Supplementary Figure 6. ROC curves for each AET prediction model are available as figure 1c (training set) and figure 2c (testing set). Concordance between observed and predicted AET events are presented in Supplemental figure 7.

Detailed prediction results regarding each studied toxicity endpoint can be found in Table 3.

Regarding the exploratory step in the selection's process, the number of retained features exceeded the limit pre-defined in the exploratory rule for LPT (9 vs 5 features). However, the exploratory step did not allow to reduce the number of features with a decrease in performance exceeding the 5% cut-off (Supplementary Table 2).

After bootstrap aggregation ($n$ = 1000 replications), Bacc's drops of 0.12 and 0.15 were respectively observed for the APT and AET-combined models in the training set. For the LPT model, bootstrap aggregation increased the performance of the clinical + dosimetric model with a gain of 0.09, levelling the performance of the combined model which slightly increased of 0.02. Performances for the clinical model were enhanced with a respective Bacc's increase of 0.12, 0.15 and 0.03 for the APT, LPT and AET prediction models. Detailed results are available in Table 4.

**DISCUSSION**

To our knowledge, our study is the first to combine the use of a NN approach and dose maps radiomics-extracted features for lung RT-induced toxicity prediction.

With 22.2% APT, 16.8% LPT, 30.0% AET and 5.4% LET (grades ≥ 2), the toxicity rates observed in our cohort are in line with previous published reports focusing on IMRT [19-21]. The LET rate being < 10%, no prediction model could be developed.

Eight previously published NTCP models focusing on APT and AET were recently compared [22]. Globally, NTCP models for AET were superior than those for APT (AUC 0.63-0.65 *vs*. 0.51-0.65). The best AET model achieved an AUC of 0.65 ($p < 0.0001$) combining DMean to the esophagus and concurrent chemotherapy, and the best APT model resulted in an AUC of 0.73 ($p < 0.0001$) combining age, DMean to the lung and pulmonary comorbidities

To the best of our knowledge, models for prediction of late pulmonary toxicity have never been reported. Here, we developed two efficient prediction models (clinical + DVH and combined) for LPT. For this toxicity, the number of features exceeded the limit set in the exploratory step (respectively, 9 vs 5). Nevertheless, this step has shown the indispensability of additional parameters while complying with the selection rule, supporting the robustness of our selection's workflow.

The LPT prediction model is based on 9 predictors extracted from the GLSZM and the GLCM. The LZSE (Large Zone Small Emphasis) is the distribution of the large homogeneous zones in an image. Similarly, LZLGE (Large Zone Low Gray-level Emphasis) is the distribution of the low-grey level zones whereas LZHGE (Large Zone High Gray-level

Emphasis) is the distribution of the high-grey level zones. In our case, the low-grey level zones could be interpreted as the "low-dose bath". ZSVAR (Zone Size Variance) measures the variance in zone size volumes for the zones. The last radiomic feature, the Prominence, is extracted from the "Lungs" VOI on the GLCM. The GLCM examines the spatial relationship among pixels and defines how frequently a combination of pixels is present in an image. Prominence is a measure of the GLCM asymmetry. By definition, these radiomic features appreciate the heterogeneity in dose distribution [23]. A similar physiological reasoning for the 2 other models can be applied.

The analysis of a tridimensional (3D) spatial dose distribution through texture analysis applied to dose maps remains scarce. Rossi previously reported the development of 2 prediction models (gastro-intestinal and genito-urinary) based on 3D dose maps and logistic regression [24]. On a limited population (70 patients) and with a bootstrap internal validation, Liang *et al* developed a logistical regression based APT prediction with an AUC of 0.78 [25]. The main drawback of this study is the lack of a testing cohort. Nevertheless, it demonstrated the benefit of the addition of 3D spatial features from dose distributions for APT prediction.

3D voxel-wise based analysis was recently developed as an alternative approach to toxicity modelling [26]: this technique evaluates the significance of dose differences between groups of patients (patient with and without the studied outcome), trying to identify dose-sensitive sub-regions of normal tissues. Thus, on a cohort of 178 patients treated by RT, Palma *et al* found the lower lungs (especially the right lung) and the heart to be significantly correlated with grade $\geq$ 2 radiation pneumonitis [27,28]. The main drawback from such a technique is the complexity of the impact of the elastic registration applied to the dose maps [29]. Furthermore, despite the commonly accepted pathophysiological picture of the lungs, the physiological explanation of such highly-sensitive anatomical sub-regions warrants further research, especially on the heart's implication with conflicting results to this date[30]. Analysis of the value of the dose heterogeneity to the heart and comparison with a 3D voxel-wise based analysis could also be of interest in this context.

NN have been previously used in the RT area, especially in toxicity prediction. Gulliford *et al* described one of the first use of NN for toxicity prediction (nocturia and rectal bleeding) after prostate RT [31]. In the same setting, Carrara *et al* developed an efficient prediction model (accuracy 80.8%) for late rectal incontinence, based on clinical and dosimetric features only [32]. Combining a NN's approach with 2D spatial dose distribution, Buettner *et al* showed the importance of spatial dose distribution (dose-surface map) over dose-surface histograms to predict late rectal toxicity prediction [33]. However, although these previous studies exploited NN methods for toxicity prediction, only Buettner's was based on the analysis of the dose map.

Our combined models can be easily implemented in the clinical workflow. Within minutes after delineation, the radiomics features are extracted from the dose-maps and a toxicity probability is given based on the previously developed models. Interestingly, the probability does not rely on any clinical or dosimetric features as only radiomics features remain in the combined models.

Apart from the retrospective setting and the limited size of the cohort, a few limitations of our study have to be acknowledged. Regarding the generalizability of our results, with a training cohort of 117 patients, we believe the model reflects the diversity of clinical situations (tumor localization, cancer stage…). However, all patients having been treated by VMAT-RT,

generalizability to other technique (especially 3D-conformal RT) has yet to be studied. For practical issues (high number of features compared to a relatively small cohort), we chose to perform bootstrapping after selection of the predictors. Such a technique, while it is often used, is commonly known to induce a possible underestimation of the optimism and thus narrow standard deviations [34]. The bootstrap aggregation showed a decrease in the combined-models' performances for both the APT and AET and an increase in the clinical models' performances for all toxicities. This could be due to an overly optimistic fit of the combined model in the training set. Nevertheless, results in the testing set indicate a higher robustness of the combined model compared to the other models for APT and LPT. NNs are often criticized as being "black boxes" [35]. Our approach offers classification by normalized importance of the features, thus partly addressing this issue and providing models with some explainability for the users. Finally, the gold-standard validation of a diagnostic model remains comparisons with clinically used NTPC models and an external prospective validation.

The NN approach appears as a feasible statistical approach for toxicity prediction. In patients with a lung cancer treated by chemoradiotherapy, radiomic features extracted from 3D dose maps seem to surpass usual models in the prediction of APT and LPT. For AET, the addition of radiomic features to clinical + DVH features did not improve toxicities modelling. Prospective validation is currently under investigation in our institution.

**TABLES**



**FIGURES**

**Table 1: Patients and disease characteristics**

| Characteristics | Overall | | Training (117) | | Testing (50) | | p |
|---|---|---|---|---|---|---|---|
| **Age median** (year, range) | 66 (39-88) | | 66 (42.0 - 88.0) | | 66 (39.0 - 87.0) | | *1* |
| **Gender** | | | | | | | *0.34* |
| Male (nb, %) | 113 | 67.7 | 76 | 65 | 37 | 74 | |
| Female (nb, %) | 54 | 32.3 | 41 | 35 | 13 | 26 | |
| **PS** (median) | 1 | | 1 | | 1 | | *1* |
| **Smoking** | | | | | | | *0.53* |
| Activ (nb, %) | 65 | 39 | 43 | 37.4 | 22 | 44 | |
| Former/never (nb, %) | 102 | 61 | 74 | 62.6 | 28 | 56 | |
| **Known COPD** | 63 | 37.7 | 43 | 38.0 | 20 | 40 | *0.94* |
| **Mean MEVS (%, range)** | 74 (23-122) | | 73.9 (23.0-122.0) | | 74.2 (40.0 - 113.0) | | *0.92* |
| **Histology** | | | | | | | |
| SCC (nb, %) | 63 | 37.7 | 46 | 39.3 | 17 | 34 | *0.21* |
| ADC (nb, %) | 67 | 40.1 | 46 | 39.3 | 21 | 42 | *0.88* |
| SCLC (nb, %) | 24 | 14.4 | 14 | 12.0 | 10 | 20 | *0.27* |
| Others (nb, %) | 13 | 8 | 11 | 9.4 | 2 | 4 | *0.38* |
| **AJCC stage (Median)** | 3 | | 3 | | 4 | | |
| **Total RT Dose** | | | | | | | |
| Median (Gy, range) | 66 (60-66) | | 66 (60-66) | | 66 (60-66) | | *1* |
| **Chemotherapy sequence** | | | | | | | |
| Concomitant (nb, %) | 54 | 32.3 | 37 | 31.6 | 17 | 34.0 | *0.90* |
| Induction (nb, %) | 49 | 29.3 | 36 | 30.8 | 13 | 26.0 | *0.66* |
| Induction + concomitant (nb, %) | 35 | 21.0 | 23 | 19.7 | 12 | 24.0 | *0.68* |
| None (nb, %) | 29 | 17.4 | 21 | 17.9 | 8 | 16.0 | *0.94* |
| **Adjuvant durvalumab** (nb, %) | 16 | 9.6 | 9 | 7.7 | 7 | 14 | *0.32* |

*Abbreviations: nb: number, % : percentage, PS : Performance Status, COPD: Chronic Obstructive Pulmonary Disease; MEVS: Maximum Expiratory Volume per Second, PTV: Planning Target Volume, CDDP: cisplatine; RT: radiotherapy; AJCC: American Joint Committee on Cancer; SCC: Squamous Cell Carcinoma; ADC: Adenocarcinoma; SCLC: Small Cell Lung Cancer*

**Table 2: prevalence of each toxicity endpoint**

| **Toxicity** | **Overall (*n = 167*)** | | **Training (*n = 117*)** | | **Testing (*n = 50*)** | | **p** |
|---|---|---|---|---|---|---|---|
| | *n* | *%* | *n* | *%* | *n* | *%* | |
| **APT** | 37 | 22.2 | 26 | 22.2 | 11 | 22.0 | *0.86* |
| **LPT** | 28 | 16.8 | 18 | 15.3 | 10 | 20.0 | *0.60* |
| **AET** | 50 | 30.0 | 34 | 29.1 | 16 | 32.0 | *0.85* |
| **LET** | 9 | 5.4 | 6 | 5.1 | 3 | 6.0 | *0.89* |

*Abbreviations: n: number, %: percentage, APT: acute pulmonary toxicity ≥ G2, LPT: late pulmonary toxicity ≥ G2, AET: acute oesophageal toxicity ≥ G, LET: late oesophageal toxicity*

**Table 3: results of each toxicity prediction model**

| | | Training | | | | | | | | Testing | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUC | p | $R^2$ | Best cut-off value | Bacc | Se (%) | Sp (%) | $R^2$ | Bacc | Se (%) | Sp (%) |
| APT | Clinical | 0.67 | 0.005 | 0.04 | ≤ 0.93 | 0.68 | 68.0 | 68.5 | 0.05 | 0.69 | 63.6 | 75.0 |
| | Clinical + DVH | 0.86 | < 0.0001 | 0.31 | ≤ 0.63 | 0.79 | 92.3 | 65.9 | 0.11 | 0.69 | 81.8 | 56.8 |
| | Radiomics | 0.92 | < 0.0001 | 0.57 | ≤ 0.99 | 0.92 | 92.3 | 91.2 | 0.87 | 0.92 | 90.9 | 92.1 |
| LPT | Clinical | 0.58 | 0.40 | 0.04 | ≤ 0.90 | 0.66 | 55.6 | 75.8 | 0.16 | 0.76 | 70.0 | 82.1 |
| | Clinical + DVH | 0.79 | < 0.0001 | 0.15 | ≤ 0.83 | 0.77 | 72.2 | 80.8 | 0.47 | 0.80 | 70.0 | 89.5 |
| | Radiomics | 0.89 | < 0.0001 | 0.45 | ≤ 0.05 | 0.87 | 77.8 | 97.0 | 0.69 | 0.89 | 80.0 | 97.4 |
| AET | Clinical | 0.72 | < 0.0001 | 0.13 | ≤ 0.71 | 0.68 | 61.8 | 74.7 | 0.14 | 0.58 | 43.8 | 72.7 |
| | Clinical + DVH | 0.78 | < 0.0001 | 0.21 | ≤ 0.71 | 0.73 | 76.5 | 68.7 | 0.27 | 0.73 | 81.3 | 64.7 |
| | Radiomics | 0.85 | < 0.0001 | 0.44 | ≤ 0.46 | 0.83 | 76.5 | 90.4 | 0.32 | 0.72 | 56.3 | 88.2 |

*Abbreviations: APT: acute pulmonary toxicity ≥ G2, LPT: late pulmonary toxicity ≥ G2 and AET: acute oesophageal toxicity ≥ G2, AUC: Area Under the Curve, $R^2$: square of the Pearson correlation coefficient, DVH: Dose Volume Histogram, Bacc: balanced accuracy, Se : sensitivity, Sp : specificity*

**Table 4: results of each toxicity prediction model in the Bootstrap validation**

| | | Before Bootstrap | Bootstrap : *n = 1000* | |
|---|---|---|---|---|
| | | Bacc | Bacc | Standard deviation |
| APT | Clinical | 0.68 | 0.80 | 0.03 |
| | Dosimetric | 0.79 | 0.78 | 0.03 |
| | Combined | 0.92 | 0.84 | 0.02 |
| LPT | Clinical | 0.68 | 0.83 | 0.01 |
| | Dosimetric | 0.73 | 0.82 | 0.02 |
| | Combined | 0.83 | 0.85 | 0.03 |
| AET | Clinical | 0.66 | 0.69 | 0.02 |
| | Dosimetric | 0.77 | 0.70 | 0.02 |
| | Combined | 0.87 | 0.72 | 0.03 |

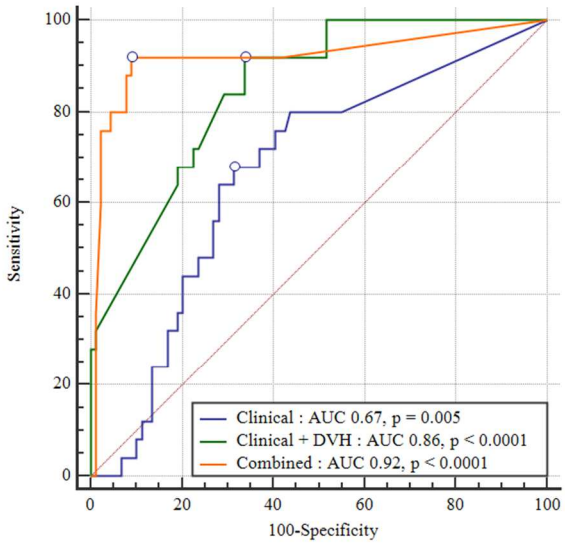
*Abbreviations: APT: acute pulmonary toxicity ≥ G2, LPT: late pulmonary toxicity ≥ G2 and AET: acute oesophageal toxicity ≥ G2, Bacc: balanced accuracy*

## Figures:

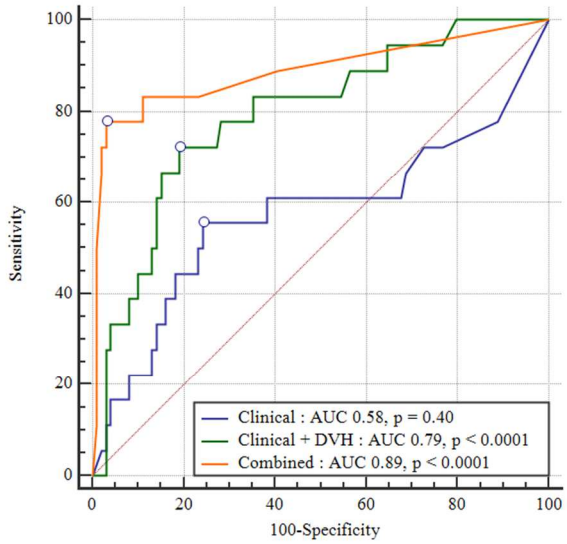
**Figure 1:** ROC curves for each toxicity prediction model in the training set

- 1a: APT prediction models in the training set
- 1b: LPT prediction models in the training set
- 1c: AET prediction models in the training set

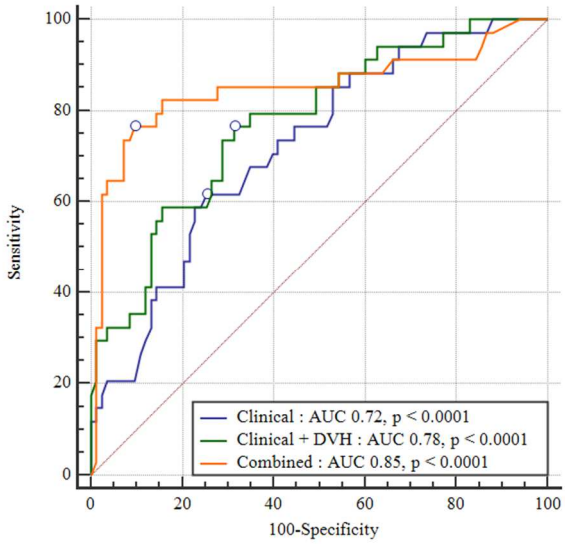
1a




1b




1c

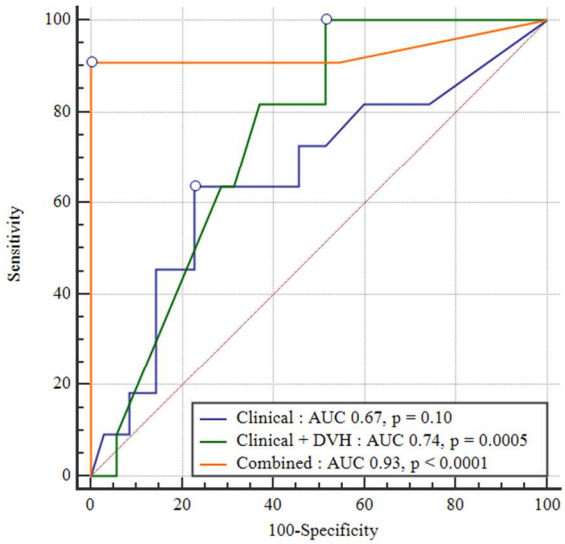



*Abbreviations: AUC: Area Under the Curve, DVH: Dose-Volume Histogram, Combined: Clinical + DVH + Radiomics*

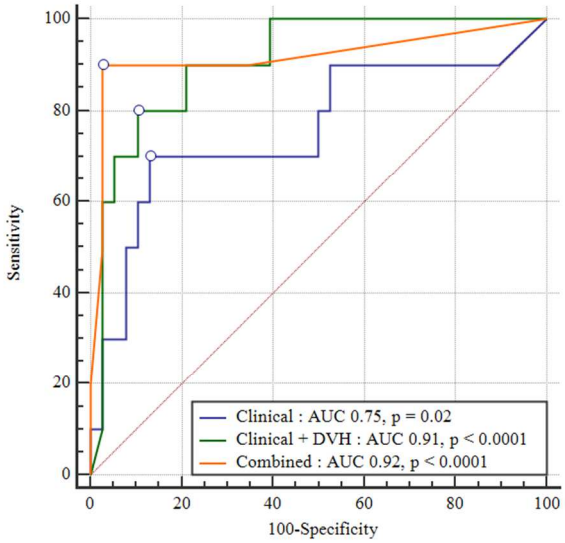
**Figure 2:** ROC curves for each toxicity prediction model in the testing set

- 2a: APT prediction models in the testing set
- 2b: LPT prediction models in the testing set
- 2c: AET prediction models in the testing set

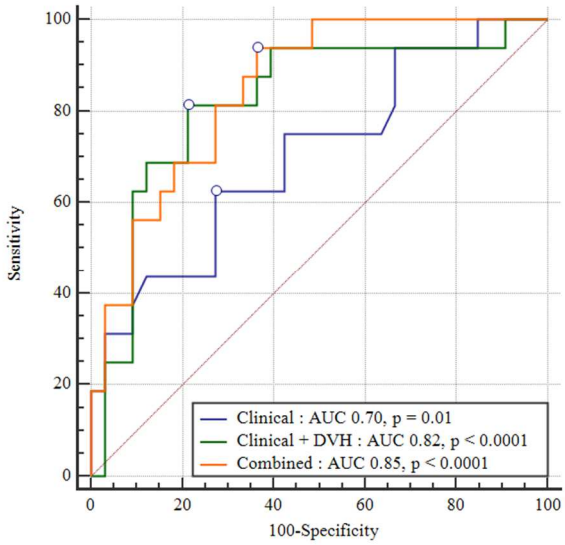
2a




2b




2c




*Abbreviations: AUC: Area Under the Curve, DVH: Dose-Volume Histogram, Combined: Clinical + DVH + Radiomics*

**References**


1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians.* 2018;68(6):394-424.
2. Ettinger DS, Wood DE, Akerley W, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 4.2016. *Journal of the National Comprehensive Cancer Network : JNCCN.* 2016;14(3):255-264.
3. Grills IS, Yan D, Martinez AA, Vicini FA, Wong JW, Kestin LL. Potential for reduced toxicity and dose escalation in the treatment of inoperable non-small-cell lung cancer: a comparison of intensity-modulated radiation therapy (IMRT), 3D conformal radiation, and elective nodal irradiation. *International journal of radiation oncology, biology, physics.* 2003;57(3):875-890.
4. Christian JA, Bedford JL, Webb S, Brada M. Comparison of inverse-planned three-dimensional conformal radiotherapy and intensity-modulated radiotherapy for non-small-cell lung cancer. *International journal of radiation oncology, biology, physics.* 2007;67(3):735-741.
5. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer.* 2012;48(4):441-446.
6. Mylona E, Acosta O, Lizee T, et al. Voxel-Based Analysis for Identification of Urethrovesical Subregions Predicting Urinary Toxicity After Prostate Cancer Radiation Therapy. *International journal of radiation oncology, biology, physics.* 2019;104(2):343-354.
7. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. *Japanese journal of radiology.* 2018;36(4):257-272.
8. National Cancer Institute PROCSG. Common Terminology Criteria for Adverse Events v4.0. Available.
9. Marks LB, Bentzen SM, Deasy JO, et al. Radiation dose-volume effects in the lung. *International journal of radiation oncology, biology, physics.* 2010;76(3 Suppl):S70-76.
10. Rudra S, Al-Hallaq HA, Feng C, Chmura SJ, Hasan Y. Effect of RTOG breast/chest wall guidelines on dose-volume histogram parameters. *Journal of applied clinical medical physics.* 2014;15(2):4547.
11. Kim TH, Cho KH, Pyo HR, et al. Dose-volumetric parameters of acute esophageal toxicity in patients with lung cancer treated with three-dimensional conformal radiotherapy. *International journal of radiation oncology, biology, physics.* 2005;62(4):995-1002.
12. Werner-Wasik M, Yorke E, Deasy J, Nam J, Marks LB. Radiation dose-volume effects in the esophagus. *International journal of radiation oncology, biology, physics.* 2010;76(3 Suppl):S86-93.
13. Lucia F, Visvikis D, Vallieres M, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *European journal of nuclear medicine and molecular imaging.* 2019;46(4):864-877.
14. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE transactions on medical imaging.* 2009;28(6):881-893.
15. Zwanenburg A, Stefan Leger, Martin Vallières, Steffen Löck. Image Biomarker Standardisation Initiative *arXiv.* 2016;preprint arWiv:1612.07003.

16. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Löck S, Initiative ftIBS. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high throughput image-based phenotyping. *Radiology.* 2020.
17. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *Journal of clinical epidemiology.* 1999;52(1):19-26.
18. Abu-Mostafa YS. Hints. *Neural computation.* 1995;7(4):639-671.
19. Khalil AA, Hoffmann L, Moeller DS, Farr KP, Knap MM. New dose constraint reduces radiation-induced fatal pneumonitis in locally advanced non-small cell lung cancer patients treated with intensity-modulated radiotherapy. *Acta oncologica.* 2015;54(9):1343-1349.
20. Wijsman R, Dankers F, Troost EGC, et al. Comparison of toxicity and outcome in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-)radiotherapy using IMRT or VMAT. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2017;122(2):295-299.
21. Ling DC, Hess CB, Chen AM, Daly ME. Comparison of Toxicity Between Intensity-Modulated Radiotherapy and 3-Dimensional Conformal Radiotherapy for Locally Advanced Non-small-cell Lung Cancer. *Clinical lung cancer.* 2016;17(1):18-23.
22. Thor M, Deasy J, Iyer A, et al. Toward personalized dose-prescription in locally advanced non-small cell lung cancer: Validation of published normal tissue complication probability models. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2019;138:45-51.
23. Zwanenburg A, Vallieres M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology.* 2020:191145.
24. Rossi L, Bijman R, Schillemans W, et al. Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2018;129(3):548-553.
25. Liang B, Yan H, Tian Y, et al. Dosiomics: Extracting 3D Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis. *Frontiers in oncology.* 2019;9:269.
26. Drean G, Acosta O, Ospina JD, et al. Identification of a rectal subregion highly predictive of rectal bleeding in prostate cancer IMRT. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2016;119(3):388-397.
27. Palma G, Monti S, Xu T, et al. Spatial Dose Patterns Associated With Radiation Pneumonitis in a Randomized Trial Comparing Intensity-Modulated Photon Therapy With Passive Scattering Proton Therapy for Locally Advanced Non-Small Cell Lung Cancer. *International journal of radiation oncology, biology, physics.* 2019;104(5):1124-1132.
28. Samavati N, Velec M, Brock KK. Effect of deformable registration uncertainty on lung SBRT dose accumulation. *Medical physics.* 2016;43(1):233.
29. Cunliffe AR, Contee C, Armato SG, 3rd, et al. Effect of deformable registration on the dose calculated in radiation therapy planning CT scans of lung cancer patients. *Medical physics.* 2015;42(1):391-399.
30. Tucker SL, Liao Z, Dinh J, et al. Is there an impact of heart exposure on the incidence of radiation pneumonitis? Analysis of data from a large clinical cohort. *Acta oncologica.* 2014;53(5):590-596.
31. Gulliford SL, Webb S Fau - Rowbottom CG, Rowbottom Cg Fau - Corne DW, Corne Dw Fau - Dearnaley DP, Dearnaley DP. Use of artificial neural networks to predict

biological outcomes for patients receiving radical radiotherapy of the prostate. (0167-8140 (Print)).
32. Carrara M, Massari E, Cicchetti A, et al. Development of a Ready-to-Use Graphical Tool Based on Artificial Neural Network Classification: Application for the Prediction of Late Fecal Incontinence After Prostate Cancer Radiation Therapy. *International journal of radiation oncology, biology, physics.* 2018;102(5):1533-1542.
33. Buettner F, Gulliford Sl Fau - Webb S, Webb S Fau - Partridge M, Partridge M. Using dose-surface maps to predict radiation-induced rectal bleeding: a neural network approach. (0031-9155 (Print)).
34. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of clinical epidemiology.* 2003;56(5):441-447.
35. Clark T, Nyberg E. Creating the Black Box: A Primer on Convolutional Neural Network Use in Image Interpretation. *Current problems in diagnostic radiology.* 2019.