Design and implementation of predictive models based on radiomics to assess response to immunotherapy in non-small-cell lung cancer

M. Corral Bolaños[1,2], B. Farina[2], A.D. Ramos Guerra[2], C. Palacios Miras[3], G. Gallardo Madueño[4], A. Muñoz-Barrutia[5], G. R. Peces-Barba[3], L. M. Seijo[4], J. Corral[4], I. Gil-Bazo[4,6], M. Dómine Gómez[3], M. J. Ledesma-Carbayo[2]

[1] ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain, m.corral@alumnos.upm.es

[2] Biomedical Image Technologies, Universidad Politécnica de Madrid & CIBER-BBN, Madrid, Spain

[3]Hospital Universitario Fundación Jiménez Díaz & CIBERES, Madrid, Spain

[4]Clínica Universidad de Navarra & CIBERES & CIBERONC, Madrid, Spain

[5]Dpto. Bioingeniería e Ingeniería Aeroespacial, Universidad Carlos III de Madrid & IiSGM, Spain

[6]Instituto de Investigación Sanitaria de Navarra (IDISNA), Pamplona, Spain

Summary

Lung cancer is the leading cause of cancer-related deaths in Europe. Immunotherapy treatments have been proved as the new standard of care for stage III-IV non-small cell lung cancer patients. However, the treatments vary in success, and there is not a reliable biomarker. This retrospective project aimed to develop a predictive model based on radiomics through machine learning or deep learning techniques to assess the response to the treatment, understood as the progression (or not) of the disease. Then, the study was complemented with an analysis of the progression-free survival time and an attempt of association with biological data.
We used the basal computed tomography images of the primary tumour lesions from a cohort with 84 patients with IV stage non-small-cell lung cancer. The best performance model reached an AUC of 0.80 – 90 % CI [0.62, 0.99]. Our results suggest that the radiomics models may be useful for patient classification.

1. Introduction

Lung cancer is the leading cause of cancer-related deaths in Europe. In 2016, 239 000 people died from lung cancer, more than one fifth (20.5 %) of all deaths from cancer [1]. In most patients, lung cancer is diagnosed once it has reached an advanced stage, with more than 60 % of patients presenting advanced or metastatic stage, with a typical survival time of less than a year [2]. At this point, new treatments, such as immunotherapy and targeted therapies have emerged as new standard of care treatments for non-small-cell lung cancer (NSCLC) patients, the most common subtype.

Despite their success, clinical benefits are only observed in a subset of patients. Hence, there is an important need of development of predictive markers to assess which treatment is more likely to be appropriate for each patient and to allow the tumour surveillance during the treatment, since immunotherapy may present important side effects and is expensive. Different biomarkers have been studied with limited success for patient stratification according to their potential benefit before the treatment, such as gene alterations or the expression of immune checkpoints proteins, like the PD-L1 [3]. The current standard to determine the treatment response is based mainly on tumour size evolution according to the RECIST and iRECIST criteria. Unfortunately, targeted treatments can cause morphological changes without changing size, so it seems insufficient.

The solution to this challenge could be found in radiomics. Radiomics is the high-throughput extraction and analysis of quantitative features from medical imaging to transform images into mineable high-dimensional data. This process is motivated by the idea that images reflect the underlying pathophysiology, and the quantitative features offer information on the tumour phenotype and its microenvironment.

In this study, we hypothesize that the basal computed tomography images and the radiomic features of the main tumour can be used to develop predictive biomarkers with prognostic value about the progression of the disease.

2. Methods

2.1. Patients cohort

Patients with confirmed stage IV NSCLC receiving immunotherapy treatment from January 2013 to December 2019 at Hospital Universitario Fundación Jiménez Diaz (FJD) and Clínica Universidad de Navarra (CUN) were analysed retrospectively after the approval of the corresponding institutional review boards. Immunotherapy treatment could be monotherapy, a combination of immune-based treatments or immunotherapy combined with chemotherapy or radiotherapy. A patient was excluded from the study if the clinical data were not available, primary tumour boundaries were not clear or there was not basal CT. Finally, 84 patients were included in the study, 60 with progression and 24 with not progression. Our models aimed to predict the progression of the disease (labelled as 1), or the non-progression (labelled as 0). The progression was defined by an oncologist based on radiological or clinical evidence on patient status. The whole dataset was divided into a training set with 58 patients (70 %) and a validation set with

26 patients (30 %), they were balanced to get a similar proportion of progression/non-progression response to the treatment. From the whole dataset we had information of the PD-L1 expression levels in 34 patients.

2.2. Imaging and tumour segmentation process

In this project, the basal computed tomography lung images were employed. The acquisition and reconstruction parameters are summarized in Table 1.

The segmentation of the main lesions was performed using syngo®.via software. This tool provides a semiautomatic seed-based method for tumour segmentation, although manual correction was commonly required. After segmentation, images were exported to DICOM format and transformed into Nearly Raw Raster Data (NRRD) [4] format through a MATLAB script to work with them in other environments.

	Acquisition	
Tube voltage	80 KVP	1 image
	100 KVP	22 images
	120 KVP	58 images
	140 KVP	3 images
Tube current	< 300 mA	34 images
	300 – 500 mA	21 images
	> 500 mA	28 images
Exposure time	285 ms	1 image
	426 – 493 ms	12 images
	500 – 562 ms	71 images
	Image reconstruction	
Slice Thickness	1 mm	38 images
	1.5 mm	22 images
	2 mm	21 images
	3 mm	2 images
Convolutional kernel	B, B26f, B30f, B31f	59 images
	C	16 images
	FC01	9 images

Table 1: *Image acquisition and* reconstruction *parameters.*

2.3. Radiomic features extraction and test-retest

The feature extraction was performed through *PyRadiomics*, an open-source Python package [5]. The radiomic features were extracted from both intranodular and perinodular regions, as well as from both regions merged. The perinodular region mask was generated through a morphological dilation of the tumour mask. It was used a 4 mm, 5 mm, or 6 mm radius 3-D spherical structural element for tumours with major axis length (2D) <25 mm, ≥ 25 mm and 50 mm, respectively. After dilation, the tumour mask was subtracted from the dilated area which produced the border mask.

In this project, first order features, shape-based features and texture features (obtained from the Gray Level Cooccurrence Matrix, GLCM; Gray Level Run Length Matrix, GLRLM; Gray Level Size Zone Matrix, GLSZM; Neighbouring Gray Tone Difference Matrix, NGTDM; and Gray Level Dependence Matrix, GLDM) were extracted from original images as well as from different filtered images: the Laplacian of gaussian, wavelet decomposition, square root, and local 3D binary pattern. Shape features were not extracted for the perinodular mask. Moreover, we performed the feature extraction with four different pairs of hyperparameters associated with the resampling resolution and bin width discretization to choose the best performance radiomic features subset.

Then, non-reproducible features were discarded using test/re-test scans from the Reference Image Database to Evaluate Therapy Response (RIDER) dataset [6]. This data set is composed of 31 NSCLC patients who underwent two chest CT scanners by using the same imaging protocol and the same scanner with a difference of 15 minutes. All those features whose Lin’s concordance correlation coefficient value was lower than 0.9 were discarded.

2.4. Machine learning model building

We developed four types of models attending to the different feature extraction regions of interest: a model for the intranodular region, the perinodular region, the model which merged the regions and a model which uses both intranodular and perinodular features (Figure 1). For the feature selection and training strategy, we used the Python open-source package *Scikit-learn* [7].

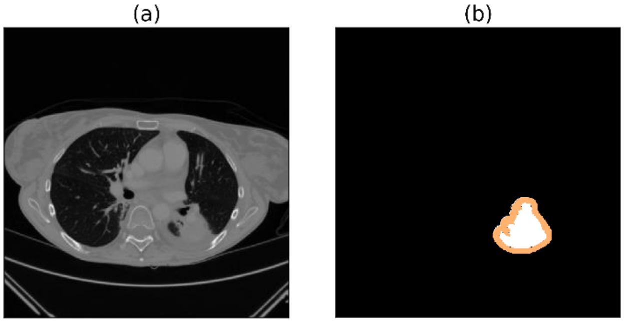

Figure 1: *a) Original CT slide, b) In orange, perinodular mask. In white, intranodular mask. The union of both masks composes the merged mask.*

To avoid the *curse of dimensionality* several steps for dimensionality reduction were performed. Firstly, Pearson’s correlation coefficient was used to remove one feature from each pair of highly correlated features. Afterwards, we used two additional methods for feature selection based on a logistic regression algorithm: recursive feature selection (a wrapper method) and LASSO feature selection (an embedded method). Then, each subset of selected features was used to train three different types of machine learning classical algorithms: logistic regression, support vector machines (SVM) and random forest, and their performance was evaluated on the validation test.

For both feature selection and training stages, each classifier was trained on the training set using 3-fold cross-validation. Moreover, we used Bayesian sequential model-based optimization (SMBO) for hyperparameter tuning with the Python package *Scikit-optimize* [8]. In all the steps, the classifier which achieved the highest AUC score was selected as the candidate solution.

2.5. Deep learning model building

Due to the limited number of available samples, we opted to develop a convolutional neural network (CNN) model through transfer learning on an already developed CNN model for prediction of lung nodule malignancy [9], whose architecture followed the one published by Causey et al. [10], *NoduleX*. This task was done with *Keras* software package (a Python deep learning API) [11].

The input of the CNN was a small 3D volume of dimensions 5x47x47 pixels. These volumes have been extracted selecting the centroid of the tumour and getting the above and below slices through a Python implementation. The pixel values were rescaled by dividing by the maximum pixel intensity in the cropped volume. To deal with overfitting, real-time data augmentation has been performed. Since the model is developed through transfer learning, the weights of the first layers are exported from the previous model, while only the three top layers are trained to adjust the output to the new task during 200 epochs. Moreover, we tested the effect of *early stopping* on the models. In this case, we use the 20 % of samples of the training set for validation (i.e., 46 samples for training, 12 for validation, 26 for testing). After, we also tested to *unfreeze* all layers and adjust their weights with a small learning rate and during just 30 epochs.

2.6. Model selection and statistical analysis

We considered various metrics to choose the candidate solutions for each region of interest. Firstly, all classifiers whose AUC differed more than 0.05 between training and validation were discarded, and when there were two similar performance models, the one with the lowest standard deviation during training was preferred. We computed the 90 % confidence interval (CI) of the validation AUC using the Delong method. To check whether the classifier achieved a significant stratification for each kind of patient (i.e., if the classifier probability scores are significantly different between the patients in which the disease progresses or not), we calculated the p-value using the Mann-Whitney U test. We used the Youden Index to select the optimal cut-off point of the classifiers. Then, we compute other metrics such as the accuracy, balanced accuracy, sensitivity, and specificity to get a better insight into the behaviour of the models. All metrics were compared with a baseline classifier which always classified the most common class.

To correlate the classifier behaviour with biological meaning, we used the spearman rank-order correlation to analyse the relevance between the PD-L1 expression levels and the selected features of the best performance classifier.

Finally, we evaluated the progression-free survival (PFS) (i.e., the time between the first immunotherapy cycle and the progression/death date) with the Kaplan-Meier method for each one of the stratified groups by the selected classifier. Endpoints were the progression of the disease, death from any cause or any recurrence. The log-rank test was used to compare the survival distributions of the two groups. Moreover, we used Cox's proportional hazard model to evaluate how the different features (covariates) of each sample affect the survival time of the subject. The Kaplan-Meier curves and Cox's models have been developed through the Python open-source package *lifelines* [12].

3. Results

The main issue of most of the developed models was their high variance between the training and validation performance. Indeed, no perinodular model achieved the requirements to be considered. Anyway, some models of the other regions of interest reached acceptable stability (Figure 2). In the intranodular region of the tumour, an SVM model achieved an AUC of 0.70 – 90% CI [0.52, 0.88], however, its accuracy was worse than the baseline classifier, so it had no practical purpose. For the features in the merged region, the best classifier reached an AUC of 0.76 – 90% CI [0.58, 0.93]. Anyway, the best performance was achieved by a logistic regression method which uses 20 selected features by recursive feature elimination among the intra- and perinodular features, AUC of 0.80 – 90% CI [0.62, 0.99].

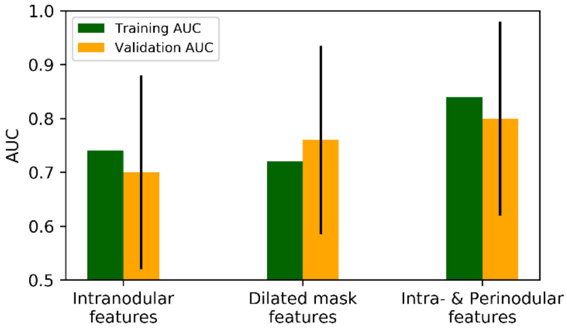

Figure 2: *Performance of the best classifiers in the training and validation set. The validation set shows the 90 % CI.*

The 20 selected features were extracted with a voxel resampling of 3 mm isotropic resolution and 25 bin width discretization. Five of them correspond to the intranodular region, and 15 to the perinodular region. Moreover, just one was computed from the original image, and the other 19 were computed from filtered images. The spearman rank-order correlation revealed three significant correlations between the features and the PD-L1 expression: the intranodular wavelet-HHH GLSZM Small Area Emphasis, perinodular wavelet-LHH first order Root Mean Squared and the perinodular square root NGTDM Busyness (p-value < 0.05). Nonetheless, the correlation coefficient value of these features is below 0.5, which means soft correlation.

Regarding the CNN models, all of them obtained similar results with AUCs between 0.643 and 0.703 in the validation phase. The model with slightly better performance was the one with early stopping after 153 epochs and only the training of the three last layers. This model reached an AUC of 0.703 - 90% CI [0.48, 0.92]. To compare the machine learning and the deep learning model, we used additional metrics which are presented in Table 2 and the ROC curves (Figure 3)

	p-value	AUC [90% CI]	Acc.	Sens.	Spec.
ML	**0.01**	**0.80 [0.62, 0.99]**	**0.85**	0.84	**0.86**
DL	0.06	0.70 [0.48, 0.92]	0.81	0.84	0.71

Table 2: *Different metrics for the ML and DL performance. Acc: Accuracy, Sens: Sensitivity, Spec: Specificity*

We have used the selected machine learning method for the PFS analysis. There is not a significant difference between the Kaplan-Meier curves for each of the classified groups by the machine learning algorithm, log-rank test p-value = 0.55.

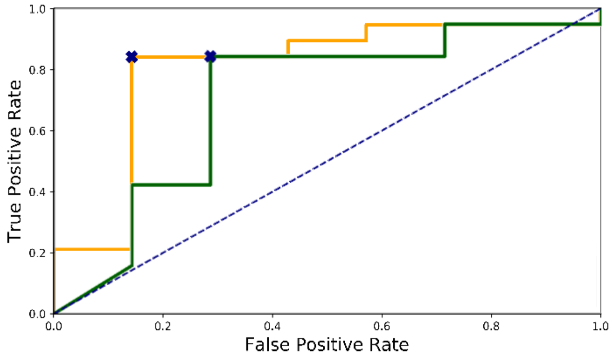

Figure 3: *ROC curve of the ML model (in yellow) and the DL model (green). The blue cross represents the optimum threshold.*

We used the Cox's models to analyse the effect of the 20 features in the PFS time of the patients. Two of them obtained significant results (p-value < 0.05): the perinodular wavelet-HHL NGTDM Strength (the higher the value, the shorter PFS) and perinodular wavelet-LLL NGTDM Contrast (the higher the value, the longer PFS). However, there was no significant association between the classifier scores and the PFS time of the validation patients (p-value = 0.26). But if we only consider the group of the progression patients, results were borderline significant (p-value = 0.05).

4. Discussion

This project aimed to analyse the suitability of basal CT images of advanced NSCLC to build radiomic biomarkers.

The development of different models considering different regions of interest allowed us to confirm the relevance of the surrounding peritumoral area. By its own, the perinodular regions did not obtain stable results but combined with the intranodular features the model achieved good classifications results with high sensitivity and specificity despite the imbalance of our dataset. These results support the current studies which assert the biological relevance of the surrounding regions due to different reasons such as the tumour vascularization or the lymphocyte infiltrations. We attempted to integrate biological evidence into our model with the PD-L1 correlation since it is a key step to be able to develop clinically applicable models. Three features showed significant results, although the limited number of patients with this information does not allow us to consider them definitive findings.

One of the main limitations of our work is the limited number of patients in the study. Hence, we were not able to develop a complete deep learning model with our data. Nonetheless, this first transfer learning approach obtained comparable results with respect to the machine learning model in most of the metrics (despite that the CNN classification was not significant, p-value = 0.06).

The PFS time analysis also suffered the lack of non-progression patients, not showing significant difference in the Kaplan-Meier curves. This analysis was deepened through the Cox's models, in which we observed the same effect since we can associate the classifier scoring to the PFS in only progression patients, but not in non-progression patients. Besides, two radiomic features revealed some association between our radiomic biomarker and the PFS.

5. Conclusion

Despite the limitations of this work, mainly associated to the limited number of patients, this analysis succeeds to demonstrate the predictive value of the intranodular and perinodular radiomic features as a feasible way to develop cheap, non-invasive and easy-obtainable biomarkers for immunotherapy effectiveness prediction at a patient-level. The current use of CT images in the follow-up of NSCLC patients motivates to continue in this direction, increasing the number of available patients and associate the model's behaviour to biological tumour biomarkers to be able to develop a clinically applicable radiomics model to support the oncological decision-making process.

6. Acknowledgements

This work was partially funded by projects RTI2018-098682-B-I00 and PID2019-109820RB-I00 (MCIU/ AEI/ FEDER, UE) by the Spanish Ministry of Science and Innovation, and the Leonardo grant 2019 from Fundación BBVA. MCB, BF and ADRG were supported by the Spanish Ministry of Education and Comunidad de Madrid.

References

[1] "Cancer statistics - specific cancers," *Eurostat*, Aug. 2020. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cancer_statistics_-_specific_cancers#Lung_cancer (Oct. 18, 2020).

[2] C. Gérard and C. Debruyne, "Immunotherapy in the landscape of new targeted treatments for non-small cell lung cancer," *Molecular Oncology*, vol. 3, no. 5–6. John Wiley and Sons Ltd, pp. 409–424, Dec. 01, 2009.

[3] P. Villalobos and I. I. Wistuba, "Lung Cancer Biomarkers," *Hematology/Oncology Clinics of North America*, vol. 31, no. 1. W.B. Saunders, pp. 13–29, Feb. 01, 2017.

[4] "nrrd: Definition of NRRD File Format." http://teem.sourceforge.net/nrrd/format.html (Jun. 18, 20).

[5] J. J. M. van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, Nov. 2017

[6] B. Zhao, L. H. Schwartz, and M. G. Kris, "Data From RIDER_Lung CT." The Cancer Imaging Archive., 2015.

[7] F. Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] T. Head, "scikit-optimize/scikit-optimize: v0.5.2." [Sw package] 2018.

[9] B. Farina, "Deep Learning-Based Models for Prediction of Lung Nodule Malignancy with CT scans," Master Thesis, University Federico II, 2018.

[10] J. L. Causey *et al.*, "Highly accurate model for prediction of lung nodule malignancy with CT scans," *Scientific Reports*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-27569-w.

[11] F. Chollet and others, "Keras." [Sw package] 2015.

[12] C. Davidson-Pilon *et al.*, "CamDavidsonPilon/lifelines: v0.24.9.." [Sw package] 2020,