research article

[^{18}F]FDG PET immunotherapy radiomics signature (iRADIOMICS) predicts response of non-small-cell lung cancer patients treated with pembrolizumab

Damijan Valentinuzzi[1,2], Martina Vrankar[3,4], Nina Boc[3], Valentina Ahac[3], Ziga Zupancic[3], Mojca Unk[3], Katja Skalic[3], Ivana Zagar[3], Andrej Studen[1,2], Urban Simoncic[1,2], Jens Eickhoff[5], Robert Jeraj[1,2,6]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
[3] Institute of Oncology Ljubljana, Ljubljana, Slovenia
[4] Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[5] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA
[6] Department of Medical Physics, University of Wisconsin, Madison, WI, USA

Radiol Oncol 2020; 54(3): 285-294.

Received 30 April 2020
Accepted 5 June 2020

Correspondence to: Prof. Robert Jeraj, Ph.D., Department of Medical Physics, University of Wisconsin, 1111 Highland Avenue, Madison, WI 53705, USA. Phone: +1 608 263 8619; E-mail: rjeraj@wisc.edu

Disclosure: No potential conflicts of interest were disclosed.

Background. Immune checkpoint inhibitors have changed the paradigm of cancer treatment; however, non-invasive biomarkers of response are still needed to identify candidates for non-responders. We aimed to investigate whether immunotherapy [^{18}F]FDG PET radiomics signature (iRADIOMICS) predicts response of metastatic non-small-cell lung cancer (NSCLC) patients to pembrolizumab better than the current clinical standards.

Patients and methods. Thirty patients receiving pembrolizumab were scanned with [^{18}F]FDG PET/CT at baseline, month 1 and 4. Associations of six robust primary tumour radiomics features with overall survival were analysed with Mann-Whitney U-test (MWU), Cox proportional hazards regression analysis, and ROC curve analysis. iRADIOMICS was constructed using univariate and multivariate logistic models of the most promising feature(s). Its predictive power was compared to PD-L1 tumour proportion score (TPS) and iRECIST using ROC curve analysis. Prediction accuracies were assessed with 5-fold cross validation.

Results. The most predictive were baseline radiomics features, e.g. Small Run Emphasis (MWU, $p = 0.001$; hazard ratio $= 0.46$, $p = 0.007$; AUC = 0.85 (95% CI 0.69–1.00)). Multivariate iRADIOMICS was found superior to the current standards in terms of predictive power and timewise with the following AUC (95% CI) and accuracy (standard deviation): iRADIOMICS (baseline), 0.90 (0.78–1.00), 78% (18%); PD-L1 TPS (baseline), 0.60 (0.37–0.83), 53% (18%); iRECIST (month 1), 0.79 (0.62–0.95), 76% (16%); iRECIST (month 4), 0.86 (0.72–1.00), 76% (17%).

Conclusions. Multivariate iRADIOMICS was identified as a promising imaging biomarker, which could improve management of metastatic NSCLC patients treated with pembrolizumab. The predicted non-responders could be offered other treatment options to improve their overall survival.

Key words: anti-PD-1; [^{18}F]FDG PET/CT; non-small-cell lung cancer; radiomics analysis; iRADIOMICS

Introduction

In spite of the advances in lung cancer treatment, prognosis for patients has been poor with a 5-year survival rate around 15%.[1] A new hope has come with renaissance of immunotherapy, such as programmed death-1 antibodies (anti-PD-1), which invigorate a patient's immune system to fight

doi: 10.2478/raon-2020-0042

against malignant cells.[2] In non-small-cell lung cancer (NSCLC), which represents 85% of all lung cancer cases, treatment outcomes of anti-PD-1 immunotherapy are significantly better compared to conventional cytotoxic therapies. In selected patient population, response rates can be over 40%.[3] The responding patients usually achieve durable benefit and prolonged survival. Occasionally, even complete remissions of metastatic disease are observed, but such complete responses are still in minority.

Due to possible unusual response patterns (e.g. pseudoprogression), treatment response assessment in immunotherapy is challenging.[4] The most routinely used methods are Response Evaluation Criteria in Solid Tumours (RECIST) and its modification for use in immunotherapy (iRECIST), among others.[5] Although iRECIST was found superior to RECIST in identifying pseudoprogression, iRECIST is a late response assessment method, because anatomical changes observed on computed tomography are usually delayed, and the suspicion of progressive disease needs to be confirmed with an additional scan 1–2 months after the first assessment.[6] Importantly, studies have shown that none of the RECIST-based endpoints could be used as valid surrogates for overall survival (OS) in anti-PD-1 trials, while the correlation of iRECIST-based endpoints with OS is yet to be explored.[7,8] Since the molecular and functional tumour changes are known to appear faster compared to anatomical changes, several immunotherapy response assessment methods, based on 2-deoxy-2-[fluorine-18] fluoro-D-glucose positron emission tomography/ computed tomography ([^{18}F]FDG PET/CT), have been proposed.[9-12] However, there is still a lack of sufficient evidence to infer, which method, if any, might be the most appropriate for the routine clinical use.[13-15]

Recently, research into the identification of new biomarkers for use in immunotherapy has also increased. Various predictive and prognostic biomarkers of response have been identified, including tumour PD-1 ligand (PD-L1) expression, tumour mutation burden, tumour infiltrating lymphocytes density, mismatch repair deficiency, microsatellite instability, and gut microbiota.[16,17] However, the reports from different studies sometimes oppose each other, therefore the current biomarkers need further validation.[18] Moreover, most of them require invasive biopsies, and are impractical or too expensive for a routine clinical use. On the other hand, few immunotherapy clinical studies examined possible non-invasive imaging biomarkers, but there is still a lack of research performed in NSCLC patients.[14] Three retrospective anti-PD-1 studies showed associations of pre-treatment sum of maximum standardized uptake values (SUV_{max}) of all lesions (SUV_{maxwb})[19], SUV_{max} of the most avid lesion[20], and volumetric parameters (metabolic tumour volume [MTV], and total lesion glycolysis [TLG])[21], with NSCLC patient response as defined by RECIST. However, significant correlations of these features with OS were not observed. There is also a lack of clinical studies in immunotherapy investigating more sophisticated image analysis methods such as radiomics analysis. Radiomics analysis harnesses the full power of medical imaging by extracting numerous quantitative features, hypothesized to reflect more deeply the tumour phenotype, as well as the genotype.[22,23] Recent anti-PD-(L)1 radiomics studies have shown associations of CT radiomics signatures with tumour immune phenotype[24], hyperprogression[25], and progression-free survival (PFS)[26]. Moreover, two studies also examined the predictive value of PET radiomics features. Polverari *et al*. observed significant differences in tumour heterogeneity (as defined by kurtosis and skewness) between patients with progressive disease (PD) and non-PD[21], while the study by Mu *et al*. proposed a combined PET and CT radiomics signature for predicting patient PFS and OS.[27] In these studies (except Polverari *et al*.), data mining using vast number of features (up to 1160) was performed in order to build multivariate radiomics signatures containing up to eight features. Although on one hand, such approach might allow for a more precise quantification of tumour characteristics, on the other hand, the so obtained predictive models could be prone to overfitting, and probably too complex and non-intuitive for a successful clinical translation. Moreover, it is also well known that a lot of radiomics features are not suitable candidates for biomarkers, for example due to an excessive test-retest variability.[28]

The primary aim of our prospective study was to determine whether immunotherapy [^{18}F]FDG PET radiomics signature (iRADIOMICS) predicts response of stage IV NSCLC patients to pembrolizumab better than the current routinely used clinical standards (PD-L1 immunohistochemistry, and iRECIST). To overcome the aforementioned pitfalls, we deliberately analysed only a small subset of radiomics features, which were previously proven to be robust and reliable according to test-retest variability[28], and built iRADIOMICS with minimum number of features.

Patients and methods

Patients

Thirty consecutive patients who met the following inclusion criteria were enrolled from January 2017 – March 2019 at the Institute of Oncology Ljubljana (Slovenia): ≥ 18 years old, cytologically or histologically confirmed stage IV NSCLC (8th TNM classification of the International Association for the Study of Lung Cancer), no history of other malignancies, PD-L1 tumour proportion score (TPS) > 1% (assessed by a validated immunohistochemistry assay), Eastern Cooperative Oncology Group criteria (ECOG) performance status 0–2. Enrolment required approval of the multidisciplinary tumour board that the patient was a candidate for treatment with pembrolizumab. The study (NCT04007068) was approved by the institutional review board committee and the National Ethics Committee (KME 117/02/17). All patients gave informed consent to participate.

Study protocol

All patients underwent standard diagnostic procedures including clinical examination and blood tests. Baseline [^{18}F]FDG PET/CT was performed ≤ 4 weeks before treatment, and follow-up [^{18}F]FDG PET/CTs were performed 1 month (± 5 days) and 4 months (± 14 days) after treatment initiation. Patients were treated with pembrolizumab until progression, clinical benefit, or unaccepted toxicities. Pembrolizumab dosage was 2 mg/kg or 200 mg/patient (depending on the guidelines at the time of treatment), intravenously, every three weeks (q3w). Patients could also receive palliative radiotherapy in case of symptomatic lesions. Such treatment intervention required approval of the multidisciplinary tumour board.

Imaging acquisition and analysis

Patients fasted for at least 6 hours before intravenous application of 3.7 MBq/kg [^{18}F]FDG and remained seated or recumbent for 60 minutes. Data acquisition was performed on a Biograph 40 mCT (Siemens Healthcare, Erlangen, Germany) with the following parameters: CT (tube current 100 kV, tube voltage 80 mAs, Care dose 4D and Care kV dose modulation, collimation 16×1.2 mm, pitch 1.2, reconstruction using 3 mm slice thickness in 2 mm increment, abdominal window, B40f kernel), [^{18}F]FDG PET (acquired from skull base to mid-thigh, 2 minutes per bed position, reconstruction using TruX+TOF (UltraHD-PET) algorithm, 2 iterations per 21 subsets, matrix size 200×200, 3 mm slice thickness, 2.5 mm pixel size). Two physicians segmented the lesions semi-automatically in 3D Slicer using SUV > 4.0 g/ml as the threshold. The segmentations were then examined by an experienced radiologist and, if necessary, manually edited. The radiologist also performed iRECIST assessment. All researchers involved in tumour segmentations were blinded to the outcome of the study.

Feature extraction

At first, eight [^{18}F]FDG radiomics features were extracted from primary tumours, including three volume-based features (volume, maximum standardized uptake value (SUV$_{max}$), total SUV (SUV$_{total}$)) and five texture-based heterogeneity features, derived from Grey-Level Co-occurrence Matrix (GLCM) (Sum Entropy, Entropy-GLCM, Difference Entropy) and Grey-Level Run Length Matrix (GLRLM) (Small Run Emphasis (SRE), Run Percentage).[29,30] Importantly, these five texture-based features were deliberately chosen, because they were identified as very robust and reliable, based on test-retest variability in a prospective multicentre study of NSCLC tumours imaged with [^{18}F]FDG PET/CT.[28] Feature definitions and their intuitive explanations are summarized in Table S1. Feature extraction was performed using an in-house software, see references.[31-33] Briefly, features were extracted using a voxel-based method. The image was discretized into 256 grey levels. For each voxel, the feature was calculated over a 5 × 5 voxel patch in axial, coronal, and sagittal planes, and averaged over the three planes for each voxel. The final feature was calculated by averaging over all voxels. After examining the correlation between features using Pearson correlation coefficient, we excluded SUV$_{total}$ and Run Percentage from further analysis, because they were too closely correlated with other features (Figure S1).

Statistical analysis

Response was defined based on overall survival (OS), the gold standard end-point in immunotherapy[8], therefore OS was the primary outcome measure in our study. OS was defined as the time from initiation of pembrolizumab until death from any cause. Patients with OS > 14.9 months were defined as responders. The selected threshold was median OS in the multicentre KEYNOTE-10 study (subgroup of NSCLC patients with PD-L1 TPS > 50%, treated with pembrolizumab dose 2 mg/kg)).[34] Although the inclusion criteria in our study was PD-L1 TPS > 1%, the majority of patients (26/30, 87%) had PD-L1 TPS > 50%, resulting in comparable median OS (15.95 months).

Mann-Whitney U-test and Fisher exact test were used to investigate the differences in radiomics features and demographic data between the responders and non-responders. Receiver operating characteristic (ROC) curve analysis was used to assess the predictive power of each radiomics feature. Univariate and multivariate Cox proportional hazards (Cox PH) regression analyses were used to study the relationship between the radiomics features and OS. A multivariate Cox PH model was constructed utilizing forward selection, considering univariate predictors of level $p < 0.05$. The results of the variable selection procedure were confirmed using backward selection based on the Akaike Information Criterion (AIC). Since the hazard ratio depends on the unit of the measurement, all radiomics features were normalized into z-scores.[35] Probability of OS as a function of time was analysed with Kaplan-Meier diagrams, and the difference between survival curves was tested with the log-rank test.

iRADIOMICS, iRECIST, and PD-L1 signatures were constructed using univariate or multivariate logistic regression analyses. The iRADIOMICS signatures consisted of the most promising radiomics features. The iRECIST signature consisted of one categorical variable with five ordered iRECIST response categories.[5] The predictive power of each model was assessed by calculating the area under the curve (AUC) of the corresponding ROC analysis. The accuracy of each model (percentage of correctly classified patients) was assessed with repeated (10×) 5-fold cross validation, so that the patients were randomly split into five groups: at each validation step, four unique groups were chosen to train the model and the remaining group was used to validate accuracy of model predictions.

A planned sample size of 30 evaluable patients was deemed to be sufficient for evaluating the predictive power of each model. Specifically, assuming an anticipated response rate of 50%, a sample size of 30 evaluable patients provided >85% power to detect an AUC of at least 0.80 (high predictive power) at the two-sided 0.05 significance level under the null hypothesis that the AUC is at most 0.5. All analyses were performed in R (3.5.3.) and were considered statistically significant if $p < 0.05$.

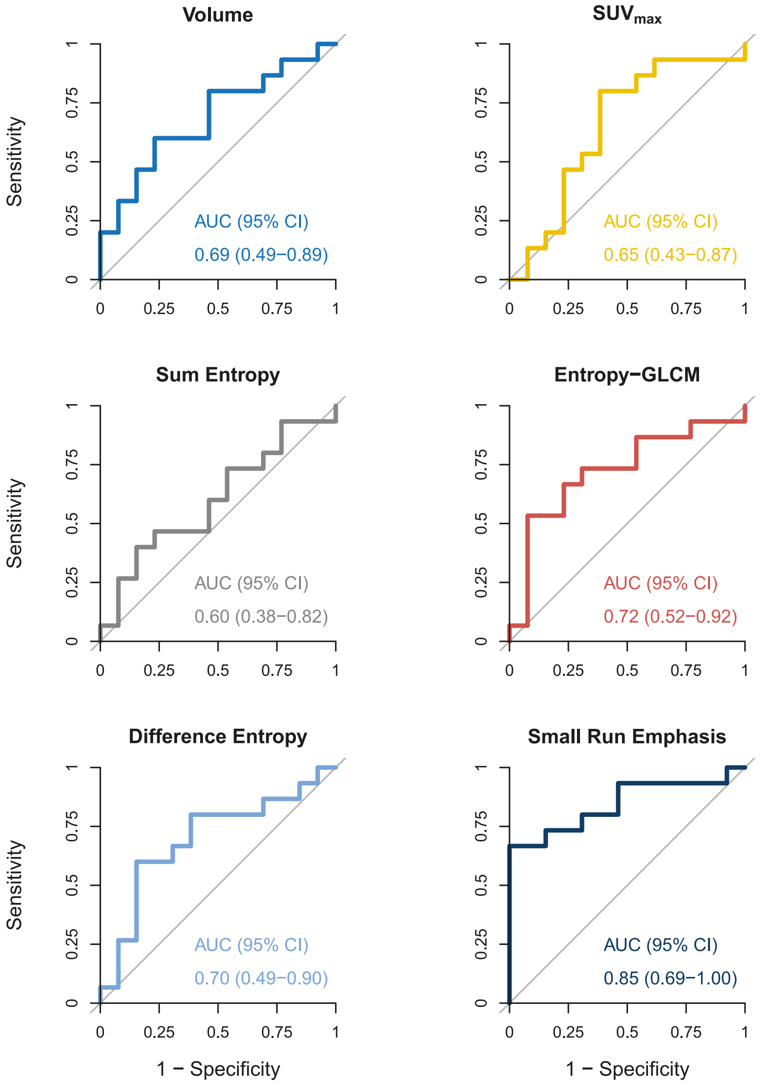

FIGURE 1. Baseline radiomics features of primary tumours – Receiver operating characteristic curve (ROC) analysis. For each radiomics feature, the area under the ROC curve (AUC) with the corresponding 95% confidence interval (CI) is reported. AUC of 0.8 or above indicates a high level of predictive power, while an AUC of 0.6 or less indicates poor level of predictive power.

Results

Patient demographic and clinical data

Thirty patients were enrolled in the study. Median follow-up time (time to censoring) was 21.4 months. A full list of demographic characteristics is presented in Table 1. The examination of demographic data did not reveal any significant differences between the responders and non-responders.

Individual radiomics features as predictors of overall survival (OS)

We analysed radiomics features extracted from primary tumours at baseline, month 1, and month

TABLE 1. Patient demographic and clinical data. The data is presented for all patients, responders (overall survival [OS] > 14.9 months), and non-responders (OS < 14.9 months). The reported p-value is the result of Mann-Whitney U-test (MWU) (continuous variables) and Fisher exact test (categorical variables) comparing differences between responders and non-responders

Characteristic	All patients median (range)	Responders (OS > 14.9 months) median (range)	Non-responders (OS < 14.9 months) median (range)	p-value
Number of patients	30	16	14	
Age [years]	65 (46–77)	67 (48–76)	61 (46–77)	0.298
PD-L1 TPS [%]	75 (3–100)	77.5 (3–100)	75 (10–100)	0.933
Sex				**0.715**
Female	15	9	6	
Male	15	7	8	
Histology				**0.532**
Adenocarcinoma	17	8	9	
Squamous cell carcinoma	8	4	4	
Other	5	4	1	
Smoking status				**0.672**
Never	1	0	1	
Former > 3 years ago	12	7	5	
Former < 3 years ago	5	3	2	
Until current disease	8	3	5	
Current smoker	4	3	1	
ECOG PS				**0.162**
0	8	2	6	
1	18	12	6	
2	4	2	2	
Line of treatment (immunotherapy)				**0.096**
1st	15	10	5	
2nd	13	4	9	
3rd	2	2	0	
Palliative RT during treatment				**0.657**
No	24	12	12	
Yes	6	4	2	

ECOG PS = Eastern Cooperative Oncology Group performance status; RT = radiotherapy; TPS = tumour proportion score (TPS)

4. Two patients did not have primary tumours, excluding them from this analysis (N = 28). The analysis of the features extracted at baseline is presented in Table 2 and Figure 1. Neither standard volume-based features (volume, SUV_{max}) were able to discriminate responders from non-responders. Among the texture-based features, Entropy-GLCM (p = 0.046) and Small Run Emphasis (SRE) (p = 0.001) were found to be significantly different between the two groups. ROC curve analysis revealed SRE having high level of predictive power (AUC = 0.85 (95% CI 0.69–1.00)), while the predictive power of other features was moderate (0.6 < AUC < 0.8).

At month 1, only volume was significantly different between the responders and non-responders (p = 0.035, AUC = 0.75 (0.55-0.95)), while none of the radiomics features reached high level of predictive power (AUC < 0.8). At month 4, none of the features were significantly different between responders and non-responders, and all radiomics features had AUC < 0.7.

To further explore the impact of baseline radiomics features on OS, we performed Cox proportional

TABLE 2. Baseline radiomics features of primary tumours – Mann-Whitney U-test (MWU) and receiver operating characteristic (ROC) curve analysis. Patients were dichotomized into 2 groups: responders (OS > 14.9 months) and non-responders (OS < 14.9 months). For each radiomics feature median value, range, p-value of MWU, and the area under the ROC curve (AUC) with the corresponding 95% confidence interval (CI), are reported. See also Figure 1

Feature	Responders (OS > 14.9 months) median (range)	Non-responders (OS < 14.9 months) median (range)	p-value	AUC (95% CI)
Volume [cm^3]	27.9 (2.64–351)	44.4 (7.81–792)	0.098	0.69 (0.49–0.89)
SUV$_{max}$ [g/ml]	20.6 (5.21–32.1)	15.6 (9.54–37.0)	0.185	0.65 (0.43–0.87)
Sum entropy	3.69 (3.53–3.77)	3.7 (3.54–3.76)	0.387	0.60 (0.38–0.82)
Entropy-GLCM	**4.07 (3.99–4.15)**	**4.11 (4.03–4.14)**	**0.046**	**0.72 (0.52–0.92)**
Difference entropy	2.98 (2.74–3.07)	2.89 (2.74–3.06)	0.080	0.70 (0.49–0.90)
Small Run Emphasis (SRE)	**0.0382 (0.00962–0.0615)**	**0.0163 (0.00854–0.0303)**	**0.001**	**0.85 (0.69–1.00)**

GLCM = Grey-Level Co-occurrence Matrix; SUV$_{max}$ = maximum standardized uptake value

TABLE 3. Baseline radiomics features of primary tumours – univariate and multivariate Cox proportional hazards regression analysis (Cox PH). For each radiomics feature, the hazard ratio (HR), corresponding 95% confidence interval (CI), and p-value of univariate analysis are reported. The 2-variable multivariate regression model was chosen based on the Akaike information criterion (AIC). In order to achieve comparable HRs, all radiomics features were normalized into z-scores

Feature	Univariate HR (95% CI)	Univariate p-value	Multivariate HR (95% CI)	Multivariate p-value
Volume	**1.6 (1.1–2.4)**	**0.015**		
SUV$_{max}$	0.77 (0.46–1.3)	0.320		
Sum Entropy	0.96 (0.60–1.5)	0.860		
Entropy-GLCM	1.4 (0.82–2.3)	0.230		
Difference entropy	**0.62 (0.40–0.97)**	**0.037**	**0.54 (0.31–0.93)**	**0.026**
Small Run Emphasis (SRE)	**0.46 (0.26–0.81)**	**0.007**	**0.39 (0.20–0.76)**	**0.006**

GLCM = Grey-Level Co-occurrence Matrix; SUV$_{max}$ = maximum standardized uptake value

hazards (Cox PH) regression analysis (Table 3). In univariate analysis, volume (hazard ratio (HR) = 1.6, p = 0.015), Difference Entropy (HR = 0.62, p = 0.037), and SRE (HR = 0.46, p = 0.007) showed statistically significant relationship with patient OS. Multivariate Cox PH regression model with the lowest AIC consisted of Difference Entropy (HR = 0.54, p = 0.026) and SRE (HR = 0.39, p = 0.006). As shown in Figure S1, SRE and Difference Entropy also exhibited low correlation (ϱ = 0.20), confirming that these two features were independent predictors of survival.

For the feature SRE, which was found to be the most informative in all statistical tests, we performed Kaplan-Meier survival analysis for baseline SRE where patients were dichotomized by the median (Figure 2). Survival probability was significantly different between groups (p = 0.015). Median OS of the patients with SRE < SRE$_{median}$ was 10.4 months (95% CI 6.0 months–not reached), while median OS of the patients with SRE ≥ SRE$_{median}$ was not reached (95% CI 15.9 months–not reached).

Ability of iRADIOMICS, iRECIST, and PD-L1 signatures to predict patient overall survival

Finally, we examined the predictive power of iRADIOMICS (baseline), iRECIST (month 1 and 4), and PD-L1 (baseline) signatures. 25 patients, which had both baseline and month 1 scans available, were suitable for this analysis. Two patients were excluded because they had no primary tumours (impossible to extract iRADIOMICS), and three other patients had no month 1 scans (impossible to assess iRECIST). For the three additional patients, who died before the scheduled month 4 scanning, we used month 1 iRECIST assessment for the construction of month 4 iRECIST signature. Otherwise, the statistics of month 4 iRECIST signature could

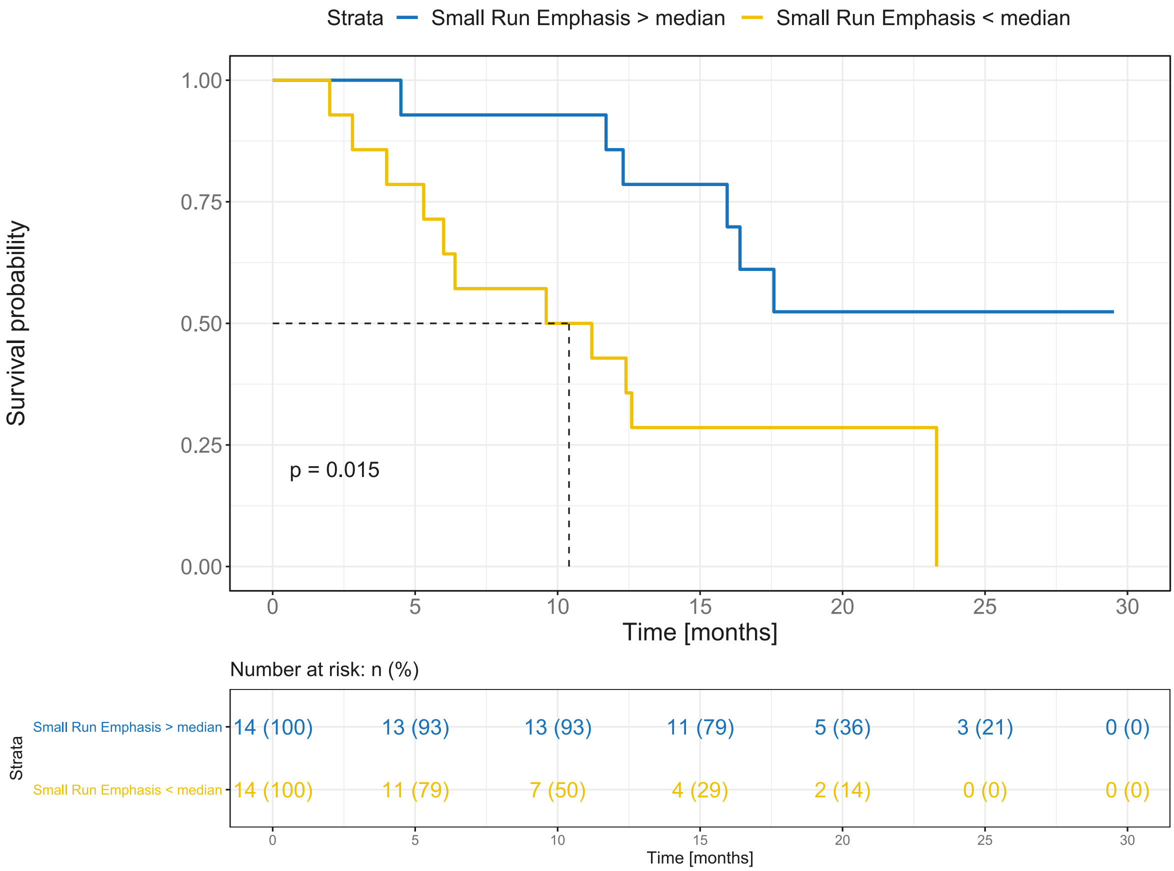

FIGURE 2. Kaplan-Meier diagram – Small Run Emphasis (SRE). Blue: patients with SRE ≥ SRE_{median}, yellow: patients with SRE < SRE_{median}. The reported p-value is the result of log-rank test.

be biased due to the exclusion of hyperprogressive patients. The results are presented in Figure 3. PD-L1 TPS showed poor predictive power (AUC = 0.60 (0.37-0.83)). The AUC of iRECIST signatures were 0.79 (0.62–0.95) and 0.86 (0.72–1.00) for month 1 and month 4, respectively. On the other hand, the AUC of the univariate iRADIOMICS at baseline was 0.81 (0.62–0.99), which was comparable to iRECIST at month 1. The highest predictive power was achieved by the multivariate baseline iRADIOMICS (consisting of SRE and Difference Entropy) with AUC = 0.90 (0.78–1.00). Model coefficients of iRADIOMICS are summarized in Table S2.

To further validate the predictive ability of all models, the accuracy of predictions was calculated using 5-fold cross validation. PD-L1 TPS achieved poor accuracy of only 53% (standard deviation SD = 18%). iRECIST signatures at month 1 and month 4 correctly classified 76% (16%) and 76% (17%) of patients, respectively. The accuracy of univariate iRADIOMICS at baseline was slightly lower, 73% (18%). The highest accuracy was achieved by multivariate baseline iRADIOMICS, which correctly classified 78% (18%) of patients.

Additionally, we performed a sensitivity study by repeating the same analyses either with a subset of 22 patients, who were scanned at all three time-points (excluding hyperprogressive patients who died before month 4), or by using all available data at each specific time-point (resulting in different number of analysed patients at baseline, month 1 and month 4), but the change of the results was negligible. In each scenario, multivariate iRADIOMICS reached AUC around 0.90 with accuracy up to 80%, and always performed better than the other models.

Discussion

New biomarkers of response to immunotherapy are urgently needed. In NSCLC, PD-L1 TPS is still the only predictive biomarker routinely used in clinics, in spite of the growing evidence sug-

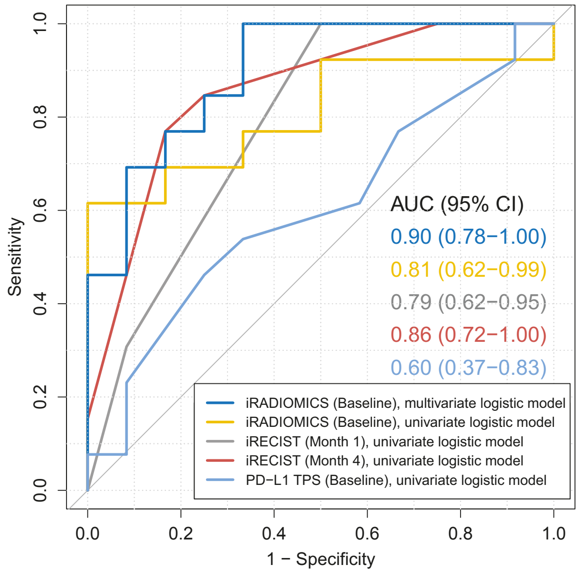

FIGURE 3. Receiver operating characteristic (ROC) curve analysis. Blue: baseline iRADIOMICS multivariate logistic model (independent variables: Small Run Emphasis [SRE], Difference Entropy), yellow: baseline iRADIOMICS univariate logistic model (independent variable: SRE), grey: month 1 iRECIST univariate logistic model (independent variable: iRECIST response category), red: month 4 iRECIST univariate logistic model (independent variable: iRECIST response category). For each model, area under curve (AUC) and 95% confidence interval (CI) are reported.

gesting that it is far from optimal.[36] Among the reasons for its questionable predictive power are inconsistent measurement methodologies, intratumour PD-L1 expression heterogeneity, and the fact that immune cells infiltrating the tumour can express PD-L1.[37] Even in our study, the survival predictions based on PD-L1 TPS performed poorly. Additionally, because it is not clear to what extent the current standards for treatment response assessment (RECIST, iRECIST) correlate with overall survival (OS), the duration of treatment, as well as the decision about cessation of anti-PD-1 immunotherapy, rely on the subjective judgment of the treating physician, which is mainly based on the observed immune-related adverse events and achieved clinical benefit.

We aimed to address these issues with the use of [^{18}F]FDG PET/CT imaging, since it is widely used, affordable, and non-invasive. When we examined the predictive ability of individual radiomics features, we found that some of the features showed high predictive power at baseline, while at month 1 and month 4 their informative value decreased significantly. This is consistent with a number of studies suggesting that intrinsic tumour characteristics, such as tumour histopathology, tumour microenvironment, and immune contexture, most likely have a major impact on response to immunotherapy.[16,17,38] The most dominant feature was Small Run Emphasis (SRE), which was able to discriminate responders from non-responders to anti-PD-1 therapy, it had a significant relationship with patient OS, and high predictive power. In patients with SRE > SRE$_{median}$, the probability of survival by Kaplan-Meier analysis was also significantly higher. Although studies have shown that texture-based features might reflect tumour heterogeneity on macroscopic, cellular, or even molecular or genomic level[39], their clear relationship with the underlying biology still needs to be elucidated. However, from the definitions of texture features used in our study we can infer that at baseline, primary tumours of responders have finer and more homogeneous metabolic structure, as reflected by higher SRE and lower Entropy-GLCM, respectively. See Table S1 for formal mathematical definitions, as well as intuitive descriptions of the studied texture features. In terms of underlying biology we could speculate that these findings might reflect tumours with spatially more homogeneous clonal structure, more homogeneous intrinsic infiltration of immune cells, more homogeneous tumour microenvironment, or fewer hypoxic or necrotic regions. Interestingly, this finding is in agreement with the study by Polverari *et al*., where patients with progressive disease (PD) exhibited higher tumour heterogeneity at baseline (reflected by higher kurtosis and skewness), compared to non-PD patients. On the other hand, the finding is at odds with the study by Mu *et al*., where heterogeneous tumours presumably had a higher chance to achieve durable clinical benefit.[27] However, heterogeneous tumour phenotype in this study was inferred from two components of eight-variable radiomics signature, making intuitive conclusions about the underlying tumour biology even more difficult compared to our study. In agreement with the study by Takada *et al*., we also observed the trend of higher SUV$_{max}$ among the responding patients, although it was not statistically significant.[20] A similar lack of statistical significance of SUV$_{max}$, or even the opposite trend, was observed by other groups, therefore the predicitve value of SUV$_{max}$ should be considered highly questionable.[13,19,21]

We analysed only primary tumours, yet neglected lymph nodes (LN) and distant metastases (DM). The main reason for this approach is that radiomics analyses might not accurately quantify intra-tumour heterogeneity of small lesions due to the partial volume effects, which could be even more pronounced in PET imaging with limited spatial resolution.[40] However, inclusion of LN and DM in future predictive models could additionally improve their predictive power and accuracy. Especially an [^{18}F]FDG PET signal of LN might be connected with the cancer immunity cycle, possibly capturing the processes that occur in LN after the initiation of anti-PD-1 therapy, including T cell priming and activation.[41]

The analysis of the predictive ability of iRECIST, PD-L1, and iRADIOMICS signatures revealed some interesting aspects. First, the response to anti-PD-1 therapy seems to occur fast, as iRECIST signature was able to predict the response of 76% of patients already at month 1, while the predictive ability at month 4 had not improved. These results suggest that treatment response assessment could be performed as soon as 1 month after treatment initiation. Moreover, its satisfactory ability to predict OS indicates that clinical decisions about (dis) continuation of anti-PD-1 therapy could (at least in part) rely on iRECIST assessment rather than purely on the observed clinical benefit. However, the correlation of other iRECIST-based endpoints with patient survival should be further explored.

Lastly, the iRADIOMICS was found superior to PD-L1 and iRECIST both in terms of predictive power and, importantly, timing. From the clinical point of view, each additional month (or day) of an ineffective therapy can be crucial for metastatic NSCLC patients. The fact that the iRADIOMICS was able to correctly predict the response of almost 80% of patients before therapy, could have an important clinical impact. The predicted non-responders to pembrolizumab could be offered other treatment options to improve their OS. However, the predictive ability of iRADIOMICS needs to be confirmed in future independent studies with a higher number of patients.

Our study compared the predictive power of baseline biomarkers (iRADIOMICS and PD-L1) to the early treatment response assessment method (iRECIST) – single point vs. multiple point assessment. However, from the practical standpoint, the baseline prediction is desirable to the treatment response assessment as it is earlier and allows more time for favourable clinical decision making. Potentially the two approaches could be combined, but such study would require higher number of patients to secure clinical significance because of more degrees of freedom (variables).

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding P1-0389), and the University of Wisconsin Carbone Cancer Center (support grant P30 CA014520).

Trial registration: ClinicalTrials.gov
NCT04007068. Registered 5 July 2019
(retrospectively registered).

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019; **69:** 7-34. doi: 10.3322/caac.21551

2. Hoos A. Development of immuno-oncology drugs – from CTLA4 to PD1 to the next generations. *Nat Rev Drug Discov* 2016; **15:** 235-47. doi: 10.1038/nrd.2015.35

3. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A; KEYNOTE-024 investigators, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N Engl J Med* 2016; **375:** 1823-33. doi: 10.1056/NEJMoa1606774

4. Vrankar M, Unk M. Immune RECIST criteria and symptomatic pseudoprogression in non-small cell lung cancer patients treated with immunotherapy. *Radiol Oncol* 2018; **52:** 365-9. doi:10.2478/raon-2018-0037

5. Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol* 2017; **18:** e143-52. doi: 10.1016/S1470-2045(17)30074-8

6. Tazdait M, Mezquita L, Lahmar J, Ferrara R, Bidault F, Ammari S, et al. Patterns of responses in metastatic NSCLC during PD-1 or PDL-1 inhibitor therapy: comparison of RECIST 1.1, irRECIST and iRECIST criteria. *Eur J Cancer* 2018; **88:** 38-47. doi: 10.1016/j.ejca.2017.10.017

7. Mushti SL, Mulkey F, Sridhara R. Evaluation of overall response rate and progression-free survival as potential surrogate endpoints for overall survival in immunotherapy trials. *Clin Cancer Res* 2018; **24:** 2268-2275. doi: 10.1158/1078-0432.CCR-17-1902

8. Nie RC, Chen FP, Yuan SQ, Luo YS, Chen S, Chen YM, et al. Evaluation of objective response, disease control and progression-free survival as surrogate end-points for overall survival in anti-programmed death-1 and anti-programmed death ligand 1 trials. *Eur J Cancer* 2019; **106:** 1-11. doi: 10.1016/j.ejca.2018.10.011

9. Cho SY, Lipson EJ, Im HJ, Rowe SP, Gonzalez EM, Blackford A, et al. Prediction of response to immune checkpoint inhibitor therapy using early-time-point 18F-FDG PET/CT imaging in patients with advanced melanoma. *J Nucl Med* 2017; **58:** 1421-8. doi: 10.2967/jnumed.116.188839

10. Anwar H, Sachpekidis C, Winkler J, Kopp-Schneider A, Haberkorn U, Hassel JC, et al. Absolute number of new lesions on 18F-FDG PET/CT is more predictive of clinical response than SUV changes in metastatic melanoma patients receiving ipilimumab. *Eur J Nucl Med Mol Imaging* 2018; **45:** 376-83. doi: 10.1007/s00259-017-3870-6

11. Goldfarb L, Duchemann B, Chouahnia K, Zelek L, Soussan M. Monitoring anti-PD-1-based immunotherapy in non-small cell lung cancer with FDG PET: introduction of iPERCIST. *EJNMMI Res* 2019; **9:** 8. doi: 10.1186/s13550-019-0473-1

12. Ito K, Teng R, Schöder H, Humm JL, Ni A, Michaud L, et al. 18 F-FDG PET/CT for monitoring of ipilimumab therapy in patients with metastatic melanoma. *J Nucl Med* 2019; **60:** 335-41. doi: 10.2967/jnumed.118.213652

13. Kaira K, Higuchi T, Naruse I, Arisaka Y, Tokue A, Altan B, et al. Metabolic activity by 18F–FDG-PET/CT is predictive of early response after nivolumab in previously treated NSCLC. *Eur J Nucl Med Mol Imaging* 2018; **45:** 56-66. doi: 10.1007/s00259-017-3806-1

14. Aide N, Hicks RJ, Le Tourneau C, Lheureux S, Fanti S, Lopci E. FDG PET/CT for assessing tumour response to immunotherapy. *Eur J Nucl Med Mol Imaging* 2019; **46:** 238-50. doi: 10.1007/s00259-018-4171-4

15. Rossi G, Bauckneht M, Genova C, Rijavec E, Biello F, Mennella S, et al. Comparison between 18F-FDG-PET- and CT-based criteria in non-small cell lung cancer (NSCLC) patients treated with Nivolumab. *J Nucl Med* 2019; [Ahead of print]. doi: 10.2967/jnumed.119.233056

16. Yi M, Jiao D, Xu H, Liu Q, Zhao W, Xinwei Han H, et al. Biomarkers for predicting efficacy of PD-1/PD-L1 inhibitors. *Mol Cancer* 2018; **17:** 129. doi: 10.1186/s12943-018-0864-3

17. Zou W, Wolchok JD, Chen L. PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: Mechanisms, response biomarkers, and combinations. *Sci Transl Med* 2016; **8:** 328rv4. doi: 10.1126/scitranslmed.aad7118

18. Shukuya T, Carbone DP. Predictive markers for the efficacy of anti–PD-1/PD-L1 antibodies in lung cancer. *J Thorac Oncol* 2016; **11:** 976-88. doi: 10.1016/j.jtho.2016.02.015

19. Evangelista L, Cuppari L, Menis J, Bonanno L, Reccia P, Frega S, et al. 18F-FDG PET/CT in non-small-cell lung cancer patients: a potential predictive biomarker of response to immunotherapy. *Nucl Med Commun* 2019; **40:** 802-7. doi: 10.1097/MNM.0000000000001025

20. Takada K, Toyokawa G, Yoneshima Y, Tanaka K, Okamoto I, Shimokawa M, et al. 18F-FDG uptake in PET/CT is a potential predictive biomarker of response to anti-PD-1 antibody therapy in non-small cell lung cancer. *Sci Rep* 2019; **9:** 1-7. doi: 10.1038/s41598-019-50079-2

21. Polverari, G. Ceci F, Bertaglia V, Reale MC, Rampado O, Gallio E, et al. 18F-FDG PET parameters and radiomics features analysis in advanced NSCLC treated with immunotherapy as predictors of therapy response and survival. *Cancers* 2020;. **12:** 1163. doi: 10.3390/cancers12051163

22. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012; **48:** 441-6. doi: 10.1016/j.ejca.2011.11.036

23. Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, et al. Radiomics and radiogenomics in lung cancer: a review for the clinician. *Lung Cancer* 2018; **115:** 34-41. doi: 10.1016/j.lungcan.2017.10.015

24. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol* 2018; **19:** 1180-91. doi: 10.1016/S1470-2045(18)30413-3

25. Tunali I, Gray JE, Qi J, Abdalah M, Jeong DK, Guvenis A, et al. Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: an early report. *Lung Cancer* 2019; **129:** 75-9. doi: 10.1016/j.lungcan.2019.01.010

26. Dercle L, Fronheiser M, Lu L, Du S, Hayes W, Leung DK, et al. Identification of non-small cell lung cancer sensitive to systemic cancer therapies using radiomics. *Clin Cancer Res* 2020. [Aheqad of print]. doi: 10.1158/1078-0432.CCR-19-2942

27. Mu W, Tunali I, Gray JE, Qi J, Schabath MB, Gillies RJ. Radiomics of 18F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy. *Eur J Nucl Med Mol Imaging* 2020; **47:** 1168-82. doi: 10.1007/s00259-019-04625-9

28. Desseroit MC, Tixier F, Weber WA, Siegel BA, Le Rest CC, Visvikis D, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. *J Nucl Med* 2017; **58:** 406-11. doi: 10.2967/jnumed.116.180919

29. Tang X. Texture information in run-length matrices. *IEEE Trans image Process* 1998; **7:** 1602-9. doi: 10.1109/83.725367

30. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern* 1973; **3:** 610-21. doi: 10.1109/TSMC.1973.4309314

31. Lin C, Harmon S, Bradshaw T, Eickhoff J, Perlman S, Liu G, et al. Response-to-repeatability of quantitative imaging features for longitudinal response assessment. *Phys Med Biol* 2019; **64:** 025019. doi: 10.1088/1361-6560/aafa0a

32. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol* 2010; **49:** 1012-6. doi: 10.3109/0284186X.2010.498437

33. Chen S, Harmon S, Perk T, et al. Diagnostic classification of solitary pulmonary nodules using dual time 18F-FDG PET/CT image texture features in granuloma-endemic regions. *Sci Rep*. 2017; **7:** 9370. doi: 10.1038/s41598-017-08764-7

34. Herbst RS, Baas P, Kim D-W, Felip E, Pérez-Gracia JL, Han JY, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* 2016; **387:** 1540-50. doi: 10.1016/S0140-6736(15)01281-7

35. Kickingereder P, Burth S, Wick A, Götz M, Eidel O, Schlemmer HP, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*. 2016; **280:** 880-9. doi: 10.1148/radiol.2016160845

36. Gubens MA, Davies M. NCCN guidelines updates: new immunotherapy strategies for improving outcomes in non-small cell lung cancer. *J Natl Compr Canc Netw* 2019; **17:** 574-8. doi: 10.6004/jnccn.2019.5005

37. McLaughlin J, Han G, Schalper KA, Carvajal-Hausdorf D, Pelekanou V, Rehman J, et al. Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer. *JAMA Oncol* 2016; **2:** 46. doi: 10.1001/jamaoncol.2015.3638

38. Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C, et al. Towards the introduction of the 'immunoscore' in the classification of malignant tumours. *J Pathol* 2014; **232:** 199-209. doi: 10.1002/path.4287

39. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014; **5:** 4006. doi: 10.1038/ncomms5006

40. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol* 2016; **61:** R150-66. doi: 10.1088/0031-9155/61/13/R150

41. Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* 2013; **39:** 1-10. doi: 10.1016/j.immuni.2013.07.012

42. Santos TA, Maistro CEB, Silva CB, Oliveira MS, Franca MC, Castellano G. MRI texture analysis reveals bulbar abnormalities in Friedreich ataxia. *Am J Neuroradiol* 2015; **36:** 2214-8. doi: 10.3174/ajnr.A4455

43. Galloway MM. Texture analysis using gray level run lengths. *Comput Graph Image Process* 1975; **4:** 172-9. doi: 10.1016/s0146-664x(75)80008-6