ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO REAL WORLD DATA AND DIGITAL ONCOLOGY

ORIGINAL ARTICLE

Integrating radiomics and real-world data to predict immune checkpoint inhibitor efficacy in advanced non-small-cell lung cancer☆

L. Provenzano[1,2†*], M. Favali[1,2†], L. Mazzeo[1,2], A. Spagnoletti[2], M. Ruggirello[3], G. Calareso[3], F. G. Greco[3], R. Vigorito[3], A. Quarta[4,5], F. Calimeri[4], M. Monteleone[1], G. Baselli[1], E. De Momi[1], B. Guirges[1], A. Di Lello[2], A. Zec[1,2], A. Ferrarin[1], C. Giani[2], C. Silvestri[2], M. Occhipinti[2], M. Brambilla[2], R. Leporati[2], S. Manglaviti[2], C. Cavalli[2], G. Mazzoli[2], D. Miliziano[2], G. Di Liberti[2], B. M. Marino[3], S. Frasca[3], R. Di Mauro[2], A. D. Dumitrascu[2], T. Beninato[2], C. Proto[2], M. Ganzinelli[2], A. Vingiani[6], D. Lorenzini[6], C. Agosta[2], M. Borraccino[2], C. Bonalume[2], V. Bartolomeo[7], R. Romanò[2], P. Solli[8], A. R. Filippi[9], S. Sangaletti[10], M. Restelli[1], A. Marchianò[3], M. C. Garassino[11], F. de Braud[2,12], F. Trovò[1], A. L. G. Pedrocchi[1], G. Lo Russo[2‡], V. Miskovic[1,2‡] & A. Prelaj[2‡]

[1]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan; Departments of [2]Medical Oncology; [3]Radiology, Fondazione IRCCS Istituto Nazionale Tumori, Milan; [4]Department of Mathematics and Computer Science, University of Calabria, Calabria; [5]Department of Computer, Control and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Rome; Departments of [6]Pathology; [7]Radiation Oncology; [8]Thoracic Surgery; [9]Radiotherapy, Fondazione IRCCS Istituto Nazionale Tumori, Milan; [10]Molecular Immunology Unit, Department of Experimental Oncology, Fondazione IRCCS Istituto Nazionale Tumori, Milan, Italy; [11]Department of Medicine, Section of Hematology/Oncology, University of Chicago Medical Center, Chicago, USA; [12]Oncology Department, University of Milan, Milan, Italy

Available online 22 September 2025

Background: Immunotherapy (IO) revolutionized the prognosis of patients with non-small-cell lung cancer (NSCLC). However, identifying optimal candidates for this treatment remains challenging. Based on previous studies suggesting the potential power of radiomics in predicting clinical outcomes in different clinical settings, we aimed to assess its capability in predicting IO efficacy in advanced NSCLC patients.

Materials and methods: A total of 375 advanced NSCLC patients treated with IO-based regimens from April 2013 to May 2022 were enrolled. Primary lung lesions were segmented and radiomic features extracted. Using clinical benefit rate and overall survival status at 6 and 24 months (OS6 and OS24) as endpoints, machine learning classifiers were trained and then evaluated on a test set.

Results: Model achieving the highest performance predicting long-term survival (OS24) reached an accuracy of 0.71 and area under the curve of 0.79 on the test set, using the combination of radiomic features and real-world data (RWD) as input. Combining radiomics with RWD consistently allowed to outperform predictions obtained using the current standard predictive biomarker, i.e. programmed death-ligand 1 expression, for most of the outcomes.

Conclusions: We explored a radiomics-based signature with potential utility in predicting the prognosis of NSCLC patients undergoing IO. Further validation is required to confirm its clinical applicability and to support oncologists in making prognostic assessments.

Key words: radiomics, non-small-cell lung cancer, immunotherapy, machine learning, explainable artificial intelligence

INTRODUCTION

Non-small-cell lung cancer (NSCLC) is a significant health burden globally, accounting for 11% of newly diagnosed cancer cases and 18% of all cancer-related deaths.[1] Immunotherapy (IO), in particular immune checkpoint inhibitors (ICIs), revolutionized the treatment approach of NSCLC patients, significantly impacting survival outcomes. More specifically, pivotal trials demonstrated long-term responses in a subgroup of these patients, with 5-year survival rates reaching 15% for pre-treated and 20% for treatment-naive NSCLC patients.[2-5] On the other hand, an equally significant percentage of NSCLC patients derive poor benefit from ICI, including patients who hyperprogress after IO treatment initiation.[6] The only recognized and commonly used biomarker to support treatment decisions in this scenario is programmed death-ligand 1 (PD-L1) expression, as evaluated with immunohistochemistry assay on tumor tissue.

*Correspondence to: Dr Leonardo Provenzano, Fondazione IRCCS Istituto Nazionale dei Tumori, Medical Oncology 1, Fondazione IRCCS Istituto Nazionale dei Tumori, Via G. Venezian, 1, Milan 20133, Italy. +39 02 2390 3066
E-mail: leonardoprovenzano51@gmail.com (L. Provenzano).

†These authors contributed equally to this work.
‡These authors contributed equally to this work.

☆Note: This study was previously presented at ESMO Congress, 20 - 24 October 2025.

However, its accuracy in predicting IO benefit is limited, with long-lasting responses occurring even in patients with intermediate or low PD-L1 expression.[7] Several efforts have focused on identifying novel biomarkers beyond PD-L1, including blood-count markers like the neutrophil-to-lymphocyte ratio (NLR), genomic indicators such as tumor mutational burden, and specific gene mutations like STK11 and KEAP1, as well as transcriptomic profiles, microRNA (miRNA) signatures, and features derived from the tumor microenvironment.[7,8] However, none of these biomarkers are currently used to support treatment decisions in clinical practice, due to the low accuracy when used alone (e.g. NLR) or the high cost and poor accessibility (e.g. transcriptomic, miRNA). Artificial intelligence (AI) offers a promising approach for both biomarker discovery, by extracting complex textural information from medical images that is imperceptible to the human eye, and biomarker aggregation, enabling the development of valuable tools to support clinical decision making.[9,10] Radiomics is a computational approach that involves the quantitative extraction of data from radiological images, providing insight into diagnostic, predictive, and prognostic factors in cancer patients. Although numerous studies have applied radiomics to predict outcomes in NSCLC patients undergoing IO treatment, none of these AI tools have been approved for clinical utilization, primarily due to small sample size and poor generalizability.[11-14]

This study aims to develop machine learning (ML)-based radiomic models to predict the effectiveness of IO in a large-scale cohort of advanced NSCLC patients receiving any-line ICIs. The study expands beyond unimodal feature models, combining radiomic features and real-world data (RWD), enhancing the exploration of multimodal integration in oncological data analysis.

MATERIALS AND METHODS

Study population and objectives

This is an observational, retrospective, monocentric, real-world study including consecutive advanced NSCLC patients who received ICIs as any line of treatment for advanced disease at a single Italian institution. Eligibility criteria were as follows: (i) cytologically or histologically confirmed diagnosis of stage IV or recurrent NSCLC; (ii) age $\geq$18 years; (iii) at least one administration of first- or further-line ICI either alone or in combination with chemotherapy; and (iv) availability of computed tomography (CT) images before ICI initiation, carried out within 90 days before treatment start.

The aim of this study was to evaluate the predictive capability of radiomic features, either alone or in combination with RWD, for predicting the response to ICIs in patients with advanced NSCLC.

Outcomes and endpoints

Binary classification was used to identify patients who benefit or not from ICIs, based on radiological response and survival status at pre-specified timepoints. In particular, three clinical endpoints were selected. Regarding clinical benefit rate (CBR) composite endpoint, patients were attributed to the positive class if they achieved one of the following: complete response, partial response, stable disease lasting for at least 4 months, or progressive disease at the first radiological evaluation but treatment duration $\geq$9 months. The latter group of patients was included based on clinical decision making suggesting perceived clinical benefit from treatment despite radiological progression, a scenario increasingly recognized in clinical practice as 'pseudoprogression'. Overall survival status at 6 months (OS6) and 24 months (OS24) were also used as endpoints; for survival endpoints, patients without an OS event and censored before the timepoint (6 and 24 months for OS6 and OS24, respectively) were excluded from the analysis.

RWD curation

RWD, i.e. clinical data retrieved from patients' electronic health records (EHRs), included 16 baseline clinically relevant variables collected before treatment initiation (Supplementary Table S1, available at https://doi.org/10.1016/j.esmorw.2025.100182). Data were manually retrieved from both paper-based and digital EHRs, depending on availability. Data curation process included removing duplicate entries, fixing inconsistencies (e.g. wrong data format and out of range values), and converting textual data into structured binary or categorical values. Missing data were imputed using the iterative imputer method, which creates models for each missing feature as a function of the available features.

CT scan acquisition and volume of interest identification

Baseline non-contrast-enhanced and contrast-enhanced CT scans were analyzed by four experienced radiologists (with a minimum of 8 years of experience). Each CT scan was segmented by a single radiologist, who identified the lung primary lesion as volume of interest. The three-dimensional (3D) segmentation was carried out semi-automatically using the Syngo.via software.[15] The follow-up radiological assessment was carried out for each patient with a total body CT scan, conducted every 9-12 weeks, or before if clinical disease progression was suspected, as per local practice. The tumor response evaluation was conducted based on RECIST 1.1.[16]

CT scan preprocessing

CT scans and the corresponding segmentations were converted to nearly raw raster data (.nrrd) format with 3D Slicer software.[17] Image preprocessing was used to homogenize images with respect to pixel spacing and gray-level intensities, using the z-score normalization technique.[18,19] Both the CT scan and the segmentation mask were resampled to isometric voxels of $1 \times 1 \times 1$ m^3 to ensure complete matching resolutions and voxel sizes.

Radiomic feature extraction

The extraction of radiomic features was then carried out with the Pyradiomics package.[20] A total of 107 features

were extracted, categorized into seven different classes: 18 features of the first-order class, 14 shape descriptors, 75 texture features of gray-level co-occurrence matrix, gray-level size zone matrix, gray-level run length matrix, neighboring gray tone difference matrix, and gray-level dependence matrix.[20]

Feature selection

Three feature sets were provided as input to the ML classifiers: radiomic features (R), RWD, and a combination of radiomics and RWD (R + RWD). To address multicollinearity among radiomic features, pairwise Spearman correlation coefficients were computed. The analysis was conducted separately for each of the three endpoints. When the correlation coefficient between two features exceeded 0.8, only the feature with the highest correlation with the outcome was retained. After removing highly correlated radiomic features, the maximum relevance minimum redundancy (MRMR) technique was applied to each of the three feature sets across the three analyzed endpoints.[21] To determine the optimal number of features for each ML classifier across the three feature sets and the three endpoints, we tested intervals of 5 feature increments (such as 5, 10, and 15) using the MRMR feature selector and subsequently evaluated the performance of the ML model.

Classification analysis with ML models

Since survival classes were unbalanced [OS6: class 1 (OS $\geq$ 6) = 252 (72%), class 0 (OS $<$ 6) = 100 (28%); OS24: class 1 (OS $\geq$ 24) = 84 (28%), class 0 (OS $<$ 24) = 214 (72%)], in order to optimize model accuracy and balance performance metrics between the two classes, undersampling with the NearMiss method was carried out for the two survival endpoints.[22]

Four ML classifiers were used: logistic regression (LR), random forest, AdaBoost, and K-nearest neighbors.[23-26] ML models were trained using the three feature sets (R, RWD, and R + RWD) after applying the feature selection process. For each endpoint and each feature set, the best model was selected through 10-fold cross-validation, and the model was then evaluated on a test set. Comparisons of area under the curve (AUC) performance on PD-L1, RWD, and R versus the combined R + RWD model were conducted using statistical methods appropriate to each comparison. Specifically, the bootstrap method was applied to compare PD-L1 against R + RWD models, while DeLong's test was used to compare RWD and R models against R + RWD. All classifiers were implemented using scikit-learn in Python 3.7.0.[27] The probability of survival for patients in the test set, categorized into two prediction classes by the best-performing ML classifier, was visualized using Kaplan–Meier curves.[28]

Explainability analysis

Given the complexity of understanding the inner workings of ML models, we applied an explainable AI technique, namely SHapley Additive exPlanation (SHAP) values, to gain insight into the underlying mechanisms governing ML model predictions.[29,30] SHAP analyses are provided for both global and local explainability to assess which features most influenced the model in patient classification in the whole study population and at the single-patient level, respectively.[31] SHAP values were calculated and plotted.

In the global explainability plots, each data point corresponds to a patient observation. Features are arranged on the *y*-axis based on their importance for the model, with the most important feature positioned on the top of the graph. Each data point on the plot corresponds to a patient observation, and its position with respect to the *x*-axis indicates whether the effect of the feature value is associated with a higher (class 1) or lower (class 0) target value. Additionally, a color map applied on data points is used to represent the specific values of features for each patient, where red and blue colors indicate higher and lower feature values, respectively.

In local explainability graphs, features are ordered from the top to the bottom based on their importance for the patient prediction, and the contribution of each feature to the individual prediction is displayed by SHAP values. Features that move the prediction toward class 1 are represented by red bars, whereas those shifting the prediction toward class 0 are represented by blue bars. Figure 1 summarizes the workflow from image preprocessing to model development and explainability.

Ethical considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the ethical committee (INT 128/22).[32] Patients who were alive at the time of study initiation signed informed consent for the treatment of personal data for research purposes during the first clinical visit. For patients not alive at study initiation and data collection, the study adhered to the rules for the protection of personal data.

RESULTS

Study population

Eligible consecutive patients treated with ICI-based regimens from April 2013 to May 2022 were enrolled in the study. Of the 664 assessable cases, 375 patients were included for the final analysis and randomly split into training (300, 80%) and test (75, 20%) sets for evaluating the CBR endpoint. For survival outcomes, the dataset was undersampled to balance the outcome classes by reducing the majority class, with a final sample size of 242 and 204 for OS6 and OS24, respectively. The OS6 dataset was split into training (193, 80%) and test (49, 20%) sets, as well as the OS24 dataset (163, 80%; 41, 20%, respectively). Supplementary Figure S1, available at https://doi.org/10.1016/j.esmorw.2025.100182, displays the Consolidated Standards of Reporting Trials (CONSORT) diagram of the study.

Patients' clinical characteristics

The most relevant clinical characteristics of patients enrolled in the study are shown in Table 1. Most of the

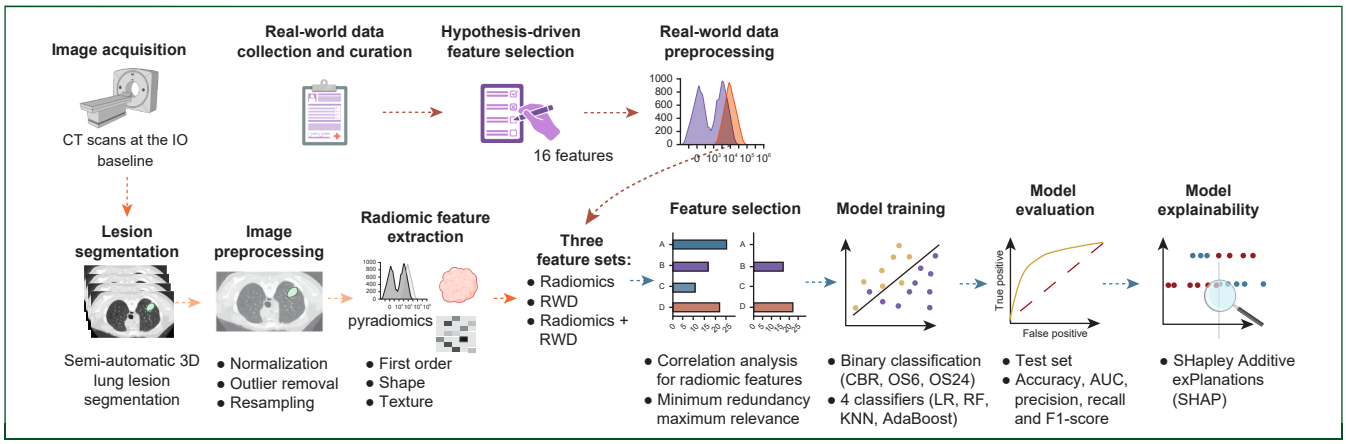

Figure 1. Study workflow, including data acquisition, preprocessing, ML modelling, and explainability.
3D, three-dimensional; AUC, area under the curve; CBR, clinical benefit rate; CT, computed tomography; IO, immuno therapy treatment; KNN, K-nearest neighborhood; LR, logistic regression; ML, machine learning; OS24, overall survival at 24 months; OS6, overall survival at 6 months; RF, random forest; RWD, real-world data.

patients received IO-based treatment as first-line therapy for metastatic disease (58%) and were treated with ICIs as monotherapy (77%).

Feature selection

The initial set of 107 radiomics underwent correlation analysis. Specifically, 75 radiomic features were identified as highly correlated and excluded from the dataset. The residual 32 radiomics and 16 RWD underwent MRMR feature selection, separately for each endpoint.

Comparison between R, RWD, and R + RWD for CBR, OS6, and OS24 on cross-validation

The accuracy, AUC, and F1-score of all ML classifiers in predicting CBR, OS6, and OS24 outcomes on training and cross-validation by using the three feature modalities are reported in Supplementary Table S2, available at https://doi.org/10.1016/j.esmorw.2025.100182.

For the CBR endpoint, the selected best-performing unimodal radiomic model, using 15 R and AdaBoost as classifier, achieved an accuracy of 0.61 and an AUC of 0.64 on cross-validation. When using RWD feature set, the LR model reached an accuracy of 0.69 and an AUC of 0.73 on cross-validation. Finally, with the combination of 20 R + RWD features, LR reached an accuracy of 0.70 and an AUC of 0.75 on cross-validation.

For the OS6 endpoint, LR achieved an accuracy of 0.76 and an AUC of 0.82 using five R features, while AdaBoost achieved an accuracy of 0.75 and an AUC of 0.84 on cross-validation when using five RWD features. An accuracy of 0.80 and an AUC of 0.90 were obtained when training LR with five R + RWD features.

Regarding the OS24 endpoint, LR was the best classifier for the three feature sets. Specifically, it reached an accuracy of 0.64 and an AUC of 0.68 using 10 R features, an accuracy of 0.66 and an AUC of 0.74 with 10 RWD, and an accuracy of 0.68 and an AUC of 0.78 with 10 R + RWD features on cross-validation.

Test set performances of R, RWD, and R + RWD for CBR, OS6, and OS24

The performance of the best ML classifier for each endpoint (CBR, OS6, and OS24) and feature set (R, RWD, and R + RWD) was evaluated on the test set (Supplementary Table S3, available at https://doi.org/10.1016/j.esmorw.2025.100182). Figure 2 shows the comparison of the models' predictive capability across the three feature set modalities and PD-L1 expression and across the three endpoints, based on test AUC. The best predictive performance was achieved for OS24, with an accuracy of 0.71 and an AUC of 0.79 on the test set, using LR with 10 R + RWD features. Although statistically significant differences in test AUCs were observed only for the CBR endpoint R versus R + RWD ($P = 0.002$) and PD-L1 versus R + RWD ($P = 0.002$), a trend toward better AUC was consistently observed when comparing the R + RWD model with unimodal models and with PD-L1 (R + RWD versus PD-L1 $P = 0.13$, R + RWD versus RWD $P = 0.91$, R + RWD versus R $P = 0.31$, for the OS6 endpoint; R + RWD versus PD-L1 $P = 0.08$, R + RWD versus RWD $P = 0.14$, R versus RWD $P = 0.24$, for the OS24 endpoint).

Figure 3 displays the Kaplan–Meier curves showing OS probabilities of patients predicted by the best models across the three endpoints (LR with 20 R + RWD for CBR, LR with 5 R + RWD for OS6, and LR with 10 R + RWD for OS24). All the three models were able to identify two classes of patients with distinct prognosis (median OS 30.87 among patients predicted as CBR responders versus 5.77 non-responders, log-rank $P < 0.001$; median OS 13.67 among patients predicted as $\geq$6 months' survivors versus 4.73 in the other group, log-rank $P = 0.003$; median OS 27.03 among patients predicted as $\geq$24 months' survivors versus 9.27 in the other group, log-rank $P = 0.003$).

SHAP: global explainability

Supplementary Table S4, available at https://doi.org/10.1016/j.esmorw.2025.100182, shows the list of the 10 selected R + RWD features used by the best-performing model for predicting OS24 endpoint. To understand which features most influenced model decisions, we applied the

Table 1. Clinical characteristics (RWD) of the study population

Characteristics	Dataset (n = 375)
Sex, *n* (%)	
Male	227 (61)
Female	148 (39)
Age, years	
Median (IQR)	68 (60-74)
BMI, kg/m^2	
Median (IQR)	24.2 (21.6-27.5)
ECOG PS, *n* (%)	
0	124 (33)
1	203 (54)
2	46 (12)
Tobacco Smoking habit, *n* (%)	
Smoker	314 (83.7)
Nonsmoker	56 (15)
NA	5 (1.3)
Histology, *n* (%)	
Squamous	72 (19.2)
Non-squamous	300 (80)
NA	3 (0.8)
PD-L1, *n* (%)	
<1%	97 (25.9)
1%-49%	103 (27.5)
≥50%	97 (25.9)
NA	78 (20.7)
Presence of distant metastases, *n* (%)	
Yes	321 (85.6)
No	35 (9.4)
NA	19 (5)
Previous surgery for early disease, *n* (%)	
Carried out	61 (16)
Not carried out	273 (73)
NA	41 (11)
ICI administration, *n* (%)	
Alone	289 (77)
With chemotherapy	86 (23)
Therapy line for metastatic disease, *n* (%)	
1	219 (58.4)
2	94 (25)
≥3	56 (15)
NA	6 (1.6)

BMI, body mass index; ECOG PS, Eastern Cooperative Oncology Group performance status; ICI, immune checkpoint inhibitor; IQR, interquartile range; NA, not available; PD-L1, programmed death-ligand 1; RWD, real-world data.

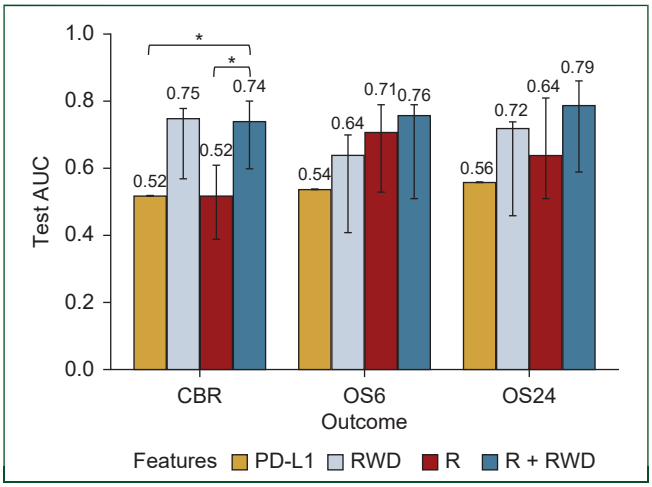

Figure 2. Test AUC comparison between the three feature sets (R, RWD, and R + RWD) for CBR, OS6, and OS24 endpoints. Statistical significance between PD-L1 versus R + RWD was assessed using the bootstrap method, while comparisons between RWD and R versus R + RWD were evaluated using DeLong's test, with * indicating statistically significant differences ($P < 0.05$). AUC, area under the curve; CBR, clinical benefit rate; OS24, overall survival status at 24 months after immunotherapy start; OS6, overall survival status at 6 months after immunotherapy start; PD-L1, programmed death-ligand 1; R, radiomics; R + RWD, radiomics plus real-world data.

SHAP technique to the best-performing model (Figure 4). Of note, Eastern Cooperative Oncology Group performance status (ECOG PS), metastatic burden, expressed as the number of metastatic sites, as well as the presence of brain metastases resulted among the most important RWD characteristics, which are in line with established clinical knowledge and available literature. In detail, a low value of ECOG PS, a low number of metastatic sites, and the absence of brain metastases are associated with a longer survival (class 1). Large area emphasis (LAE), which is a measure of the distribution of areas that share the same gray-level intensities, resulted as the most important radiomic feature. SHAP plot revealed that higher values of LAE are correlated with class 1, suggesting that a coarser texture of tumor lesion could be associated with better prognosis.

SHAP: local explainability

To better understand how features impact model predictions at a single-patient level, waterfall plots for local explainability were generated. Figure 5 displays the local explainability plot in four representative cases: (i) true positive (TP), i.e. a patient who is correctly classified as a longer survivor by the model; (ii) true negative, a patient who is correctly classified as a shorter survivor; (iii) false positive, a shorter survivor patient who is incorrectly classified as a longer survivor; (iv) false negative, a longer survivor patient who is incorrectly classified as a shorter survivor. Taking the TP case as an example, the prediction was strongly influenced by ECOG PS. In particular, a low ECOG PS (0 in this case) moved the prediction toward longer survival (class 1). On the contrary, a low value of LAE radiomic feature pushed the prediction toward shorter survival (class 0).

DISCUSSION

The present study evaluates the predictive power of radiomic features, either alone or combined with RWD, in predicting benefit and survival outcomes from ICIs in a monocentric cohort of advanced NSCLC patients. Two different endpoints were assessed: CBR, a mixed clinical—radiological endpoint, and survival status at 6 and 24 months (OS6 and OS24, respectively). Survival status endpoints are of particular clinical relevance for stratifying patients for treatment escalation or de-escalation. Indeed, OS6 aimed at identifying early progressors, i.e. patients who do not benefit from ICI with a life expectancy of <6 months. In contrast, OS24 endpoint identifies long-term survivors, i.e. patients particularly sensitive to ICI and potentially 'cured' with the use of ICI. Three sets of features, R, RWD, and R + RWD, were analyzed using four different ML classifiers for each of the three endpoints.

Overall, survival prediction models (OS6 and OS24) outperformed those for CBR, reflecting the importance of selecting appropriate endpoints in evaluating treatment efficacy, especially when evaluating IO efficacy. Indeed, IO-based

https://doi.org/10.1016/j.esmorw.2025.100182

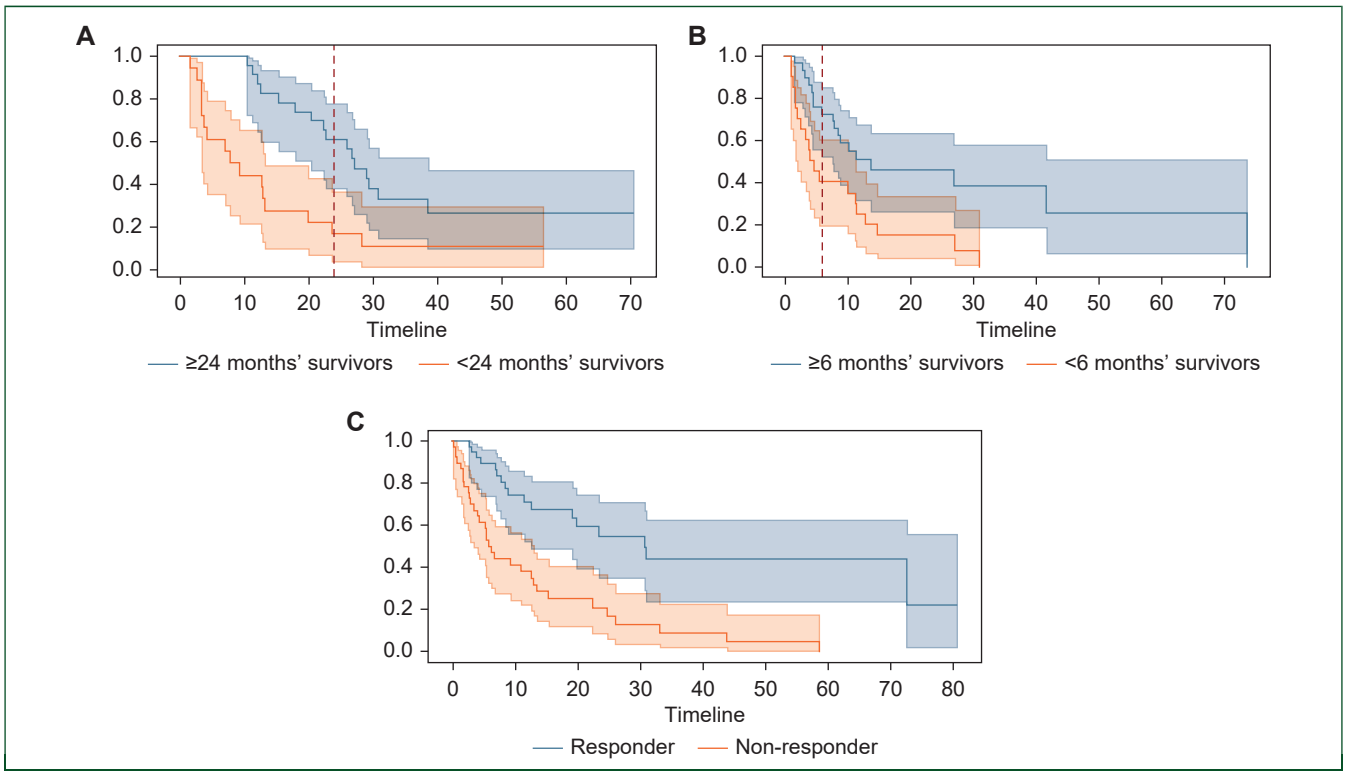

Figure 3. Kaplan–Meier curves showing probabilities of overall survival of predicted patients based on the three endpoints in the test set using the best classifiers (LR with 10 R + RWD for OS24, LR with 5 R + RWD for OS6, and LR with 20 R + RWD for CBR). (A) OS24. (B) OS6. (C) CBR.
CBR, clinical benefit rate; LR, logistic regression; OS24, overall survival at 24 months; OS6, overall survival at 6 months; R + RWD, radiomics plus real-world data.

treatments can exert its antitumor effects even in the absence of significant radiological changes.

Leveraging multimodal data (R + RWD) as input for ML models results in better prediction performances when compared with the use of unimodal features alone across multiple endpoints. Importantly, both the radiomic and the multimodal models outperformed the prediction performances of the categorized PD-L1 expression (<1%, 1%-49%, ≥50%), which is the current biomarker used in clinical practice to support treatment selection and to predict long-term benefit from ICI. The best predictive performance in the independent test set was achieved using OS24 as endpoint, with an accuracy of 0.71 and an AUC of 0.79, using LR as classifier with the combination of 10 radiomic features and RWD. Although the AUC difference when comparing the multimodal model with RWD-

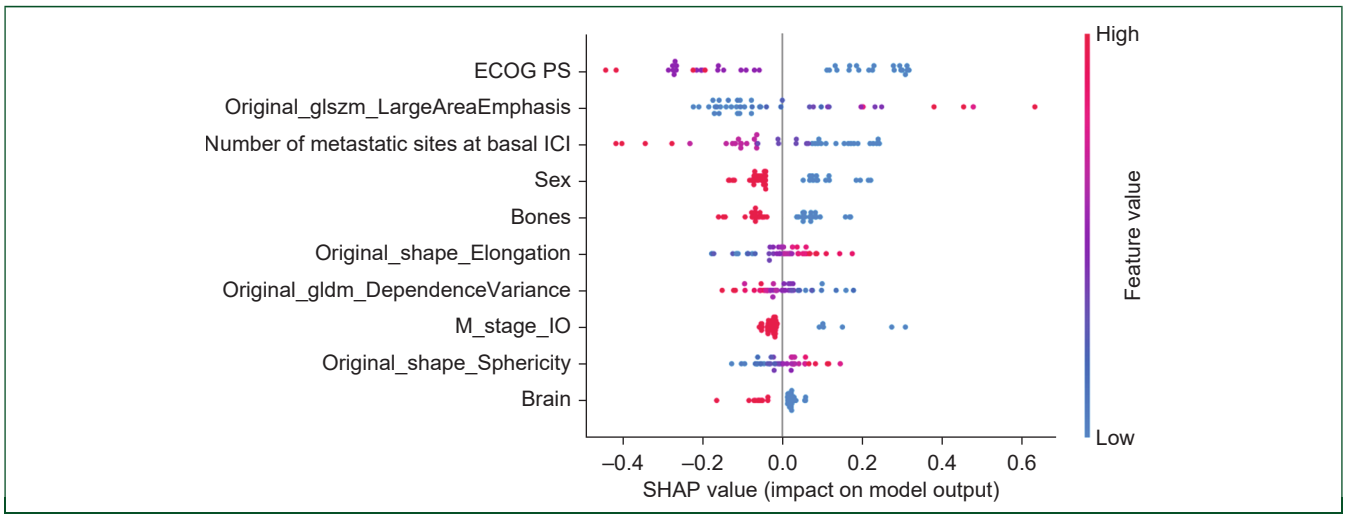

Figure 4. SHAP global explainability for the best-performing model LR utilizing the combination feature set for OS24 on the test set.
ECOG PS, Eastern Cooperative Oncology Group performance status; ICI, immune checkpoint inhibitor; LR, logistic regression; M_stage_IO, metastatic stage at the start of immunotherapy; OS24, overall survival at 24 months; SHAP, SHapley Additive explanation.

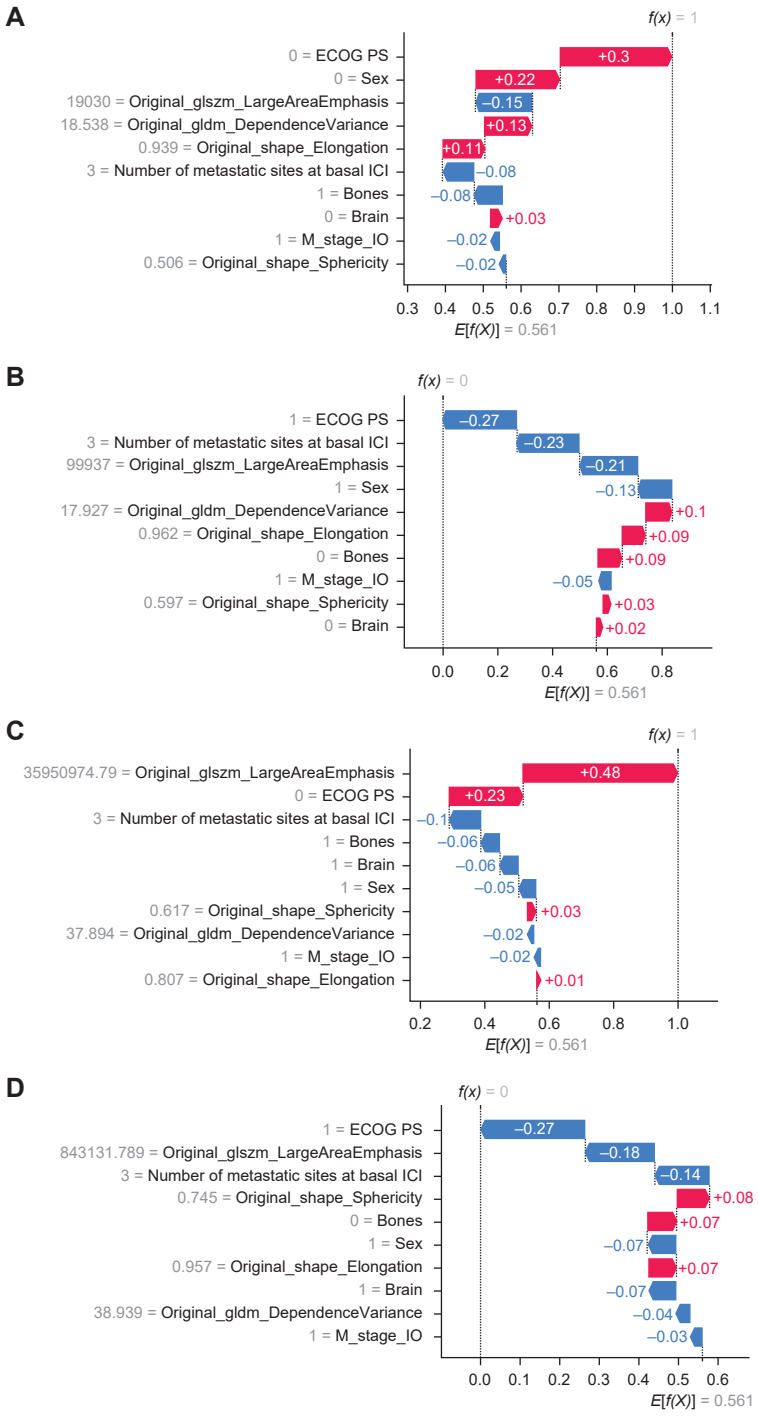

Figure 5. SHAP local explainability for the best-performing model LR utilizing 10 R + RWD for OS24 on the test set. (A) True positive. (B) True negative. (C) False positive. (D) False negative.
ECOG PS, Eastern Cooperative Oncology Group performance status; ICI, immune checkpoint inhibitor; LR, logistic regression; M_stage_IO, metastatic stage at the start of immunotherapy; OS24, overall survival at 24 months; R + RWD, radiomics plus real-world data.

only and PD-L1 was not statistically significant except for CBR endpoint, a consistent trend favoring the combined feature set emerged for both the survival endpoints (AUC 0.76 versus 0.64 versus 0.54 with R + RWD, RWD, and PD-L1 feature sets for the OS6 endpoint; AUC 0.79 versus 0.72 versus 0.56 for the OS24 endpoint). This observation supports the added value of radiomic features in improving the predictive performance of classification models predicting survival in NSCLC patients.

Our results are consistent with findings from recent literature and contribute to support radiomics in predicting ICI effectiveness in advanced NSCLC patients. Lesion heterogeneity and non-uniform density patterns were associated with better response to ICIs, while integrating

radiomics with RWD like albumin and lymphocyte levels further improved prediction performance over using a single data source.[33,34] Finally, the multimodal integration of radiomics, histopathology, and genomics enabled early stratification of responders versus non-responders.[34,35]

If validated in a larger prospective dataset, our results, which are consistent with the available literature, may support the utility of radiomics in clinical decision making. Specifically, early identification of high-risk patients may guide treatment intensification via the addition of chemotherapy or the use of emerging strategies like antibody—drug conjugates. Conversely, predicting long-term responders could benefit from de-intensified, chemotherapy-free regimens, potentially optimizing toxicity without compromising efficacy.

To enhance interpretability, we used the SHAP technique to identify which features most influenced the models' predictions. The top 10 selected features for the best-performing model included both RWD (e.g. ECOG PS, sex, presence of brain and bone metastases) and radiomic features (e.g. LAE, elongation, sphericity, and dependence variance), reinforcing the usefulness of integrating radiomic features when predicting IO efficacy in NSCLC patients (see Supplementary Table S4, available at https://doi.org/10.1016/j.esmorw.2025.100182, for details). The key RWD in our study aligned with the ones associated with prognostic impact on cancer patients' survival (i.e. patient's ECOG PS, sex, presence of brain and bone metastases), based on clinical experience and previous studies.[36-38] The observation that some well-established features (e.g. PD-L1 expression, tumor histology, or timing of IO administration) were not considered important by the model may suggest that these variables had a limited impact on treatment prediction within the specific test set analyzed. This phenomenon may be attributed by the underrepresentation of certain subgroups in the test data, as in the case of the line of therapy at which IO was administered. While some radiomic features are intuitively associated with prognostic outcomes [e.g. major axis length, which approximates to 'T' dimension in TNM (tumor—node—metastasis), classification], others cannot be captured by the human eye and are associated with clinical and/or biological surrogates. Our results highlighted the LAE radiomic feature as an important factor; as a high value of LAE is typically associated with a coarser texture, we could speculate that higher intratumor heterogeneity may be linked to enhanced immune infiltration and improved prognosis.[39]

This study has some limitations. Firstly, its retrospective nature, which encompasses the use of CT scans acquired during a wide time frame (from 2013 to 2022), introduces variability in image acquisition protocols, potentially affecting feature extraction. However, this potential pitfall may enhance the generalizability of the models across external datasets. In addition, the heterogeneity of the patient cohort, which included patients treated with ICIs in different treatment lines and with different strategies (with or without chemotherapy), could limit the applicability of the model, especially as ICIs are now predominantly used as first-line therapy in advanced NSCLC. The small sample size may have also impacted model performances and statistical power, highlighting the need for a larger patient cohort to strengthen the observed results.

Our model can be further improved. Firstly, predicting continuous survival outcomes could result in better prognosis prediction. However, this approach may limit clinical applicability, as classification task, by identifying a specific subgroup of patients, can more effectively guide treatment escalation or de-escalation strategies. Furthermore, deep learning approaches could be adopted to build an end-to-end model, potentially improving model performance. However, these architectures need a larger dataset for training, and their increased complexity may reduce interpretability, which is a crucial aspect when aiming for clinical translation. In addition, integrating RWD and radiomics with additional data sources, like genomics and digital pathology, could provide better insights into tumor biology and, therefore, even increase model performances, as already demonstrated in literature.[34,35] Moreover, incorporating longitudinal imaging data for the computation of delta-radiomic features (difference of features extracted from CT/positron emission tomography scan carried out at baseline and at first radiological evaluation) could enrich temporal tumor characterization.[40] To pursue these aims, we are conducting the international, retrospective—prospective I3LUNG study (NCT05537922), which will use data from multiple sources to build a comprehensive predictive model for advanced NSCLC patients treated with ICIs.[41]

Conclusions

We demonstrated that predicting clinical and survival outcomes of advanced NSCLC patients treated with ICIs using radiomics and RWD is feasible and effective. The explainability of the model assured consistency with available literature and clinical knowledge. Further efforts, including population expansion, external validation, and prospective cohort enrollment, are needed to finally apply this data in clinical practice.

ACKNOWLEDGEMENTS

We acknowledge donors for the Excalibur project in memory of Giorgiana Marchesi Bianchini. We especially thank all patients who took part in this clinical trial and their families.

FUNDING

This work was supported by 5xMille funding from the Italian Ministry of Health and the Fondazione IRCCS Istituto Nazionale dei Tumori, through its Call for the Valorisation of Institutional Research program (Institutional Grant: BRI 2021).

DISCLOSURE

CP reports personal fees from BMS and MSD, outside the submitted work. MCG reports personal financial interests with the following organizations: AstraZeneca, MSD International GmbH, BMS, Boehringer Ingelheim Italia S.p.A, Celgene, Eli Lilly, Ignyta, Incyte, Inivata, MedImmune, Novartis, Pfizer, Roche, Takeda, outside the submitted work. FdB reports a patent for PCT/IB2020/055956 pending

and a patent for IT201900009954 pending; honoraria from or consultant role for Roche, EMD Serono, NMS Nerviano Medical Science, Sanofi, MSD, Novartis, Incyte, BMS, Menarini Healthcare Research & Pharmacoepidemiology, Merck Group, Pfizer, Servier, AMGEN, Incyte, outside the submitted work. ALGP holds shares of Agade srl, outside the submitted work. GLR reports consultant role for Roche, Novartis, BMS, MSD, AstraZeneca, Takeda, Amgen, Sanofi, Italfarmaco, Pfizer; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events from Roche, Novartis, BMS, MSD, AstraZeneca, Takeda, Amgen, Sanofi; support for attending meetings and/or travel from Roche, BMS, MSD; data safety monitoring board or advisory board for Roche, Novartis, BMS, MSD, AstraZeneca, Sanofi; has acted as principal investigator in sponsored clinical trials for Roche, Novartis, BMS, MSD, AstraZeneca, GSK, Amgen, Sanofi, outside the submitted work. AP reports consulting or advisory role for BMS, AstraZeneca; travel, accommodations, or other expenses paid or reimbursed by Roche, Italfarmaco; principal investigator of Spectrum Pharmaceuticals; personal fees from Roche, AstraZeneca, BMS, outside the submitted work. All other authors have declared no conflicts of interest.

DATA SHARING

Data analyzed during the current study are available upon a reasonable request.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249.
2. Garon EB, Hellmann MD, Rizvi NA, et al. Five-year overall survival for patients with advanced non—small-cell lung cancer treated with pembrolizumab: results from the phase I KEYNOTE-001 study. *J Clin Oncol*. 2019;37(28):2518-2527.
3. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Updated analysis of KEYNOTE-024: pembrolizumab versus platinum-based chemotherapy for advanced non—small-cell lung cancer with PD-L1 tumor proportion score of 50% or greater. *J Clin Oncol*. 2019;37(7):537-546.
4. Gadgeel S, Rodríguez-Abreu D, Speranza G, et al. Updated analysis from KEYNOTE-189: pembrolizumab or placebo plus pemetrexed and platinum for previously untreated metastatic nonsquamous non—small-cell lung cancer. *J Clin Oncol*. 2020;38(14):1505-1517.
5. Paz-Ares L, Vicente D, Tafreshi A, et al. A randomized, placebo-controlled trial of pembrolizumab plus chemotherapy in patients with metastatic squamous NSCLC: protocol-specified final analysis of KEYNOTE-407. *J Thorac Oncol*. 2020;15(10):1657-1669.
6. Ferrara R, Mezquita L, Texier M, et al. Hyperprogressive disease in patients with advanced non—small cell lung cancer treated with PD-1/PD-L1 inhibitors or with single-agent chemotherapy. *JAMA Oncol*. 2018;4(11):1543-1552.
7. Prelaj A, Tay R, Ferrara R, Chaput N, Besse B, Califano R. Predictive biomarkers of response for immune checkpoint inhibitors in non—small-cell lung cancer. *Eur J Cancer*. 2019;106:144-159.
8. Petitprez F, Meylan M, de Reyniès A, Sautès-Fridman C, Fridman WH. The tumor microenvironment in the response to immune checkpoint blockade therapies. *Front Immunol*. 2020;11:784.
9. Wu G, Jochems A, Refaee T, et al. Structural and functional radiomics for lung cancer. *Eur J Nucl Med Mol Imag*. 2021;48:3961-3974.
10. Prelaj A, Miskovic V, Zanitti M, et al. Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. *Ann Oncol*. 2024;35(1):29-65.
11. Saad MB, Hong L, Aminu M, et al. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: a retrospective study. *Lancet Digit Health*. 2023;5(7):e404-e420.
12. Mu W, Tunali I, Gray JE, Qi J, Schabath MB, Gillies RJ. Radiomics of ^{18}F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy. *Eur J Nucl Med Mol Imaging*. 2020;47:1168-1182.
13. He S, Feng Y, Lin Q, et al. CT-based peritumoral and intratumoral radiomics as pretreatment predictors of atypical responses to immune checkpoint inhibitor across tumor types: a preliminary multicenter study. *Front Oncol*. 2021;11:729371.
14. Vaidya P, Bera K, Patil PD, et al. Novel, non-invasive imaging approach to identify patients with advanced non-small cell lung cancer at risk of hyperprogressive disease with immune checkpoint blockade. *J Immunother Cancer*. 2020;8(2):e001343.
15. Siemens Healthineers. syngo.via. Available at https://www.siemens-healthineers.com/digital-health-solutions/syngovia. Accessed January 10, 2025.
16. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247.
17. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;30(9):1323-1341.
18. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights Imaging*. 2020;11(1):91.
19. Patro SGK, Sahu KK. Normalization: a preprocessing stage. *arXiv preprint*. 2015. arXiv:1503.06462.
20. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107.
21. Laajili R, Said M, Tagina M. Application of radiomics features selection and classification algorithms for medical imaging decision: MRI radiomics breast cancer cases study. *Inform Med Unlocked*. 2021;27:100801.
22. Imbalanced-Learn. NearMiss. Available at https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.NearMiss.html. Accessed January 10, 2025.
23. LaValley MP. Logistic regression. *Circulation*. 2008;117(18):2395-2399.
24. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*.1. Montreal, QC, Canada: IEEE; 1995. p. 278-282.
25. Schapire RE. Explaining adaboost. In: *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Berlin, Heidelberg: Springer; 2013. p. 37-52.
26. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. *OTM Confederated International Conferences*. In: On the Move to Meaningful Internet Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003.
27. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
28. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
29. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
30. Kundu S. AI in medicine must be explainable. *Nat Med*. 2021;27(8):1328.
31. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.
32. Prelaj A, Ganzinelli M, Provenzano L, et al. APOLLO 11 project, consortium in advanced lung cancer patients treated with innovative therapies: integration of real-world data and translational research. *Clin Lung Cancer*. 2024;25(2):190-195.

33. Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol*. 2019;30(6):998-1004.
34. Vanguri RS, Luo J, Aukerman AT, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(l)1 blockade in patients with non-small cell lung cancer. *Nat Cancer*. 2022;3(10):1151-1164.
35. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. 2022;3(6):723-733.
36. van Laar M, van Amsterdam WAC, van Lindert ASR, de Jong PA, Verhoeff JJC. Prognostic factors for overall survival of stage iii non-small cell lung cancer patients on computed tomography: a systematic review and meta-analysis. *Radiother Oncol*. 2020;151:152-175.
37. Meira DD, de Castro E Caetano MC, Casotti MC, et al. Prognostic factors and markers in non-small cell lung cancer: recent progress and future challenges. *Genes*. 2023;14(10):1906.
38. Garinet S, Wang P, Mansuet-Lupo A, Fournel L, Wislez M, Blons H. Updated prognostic factors in localized NSCLC. *Cancers*. 2022;14(6):1400.
39. Wu W, Liu Y, Zeng S, Han Y, Shen H. Intratumor heterogeneity: the hidden barrier to immunotherapy against MSI tumors from the perspective of IFN-γ signaling and tumor-infiltrating lymphocytes. *J Hematol Oncol*. 2021;14:1-28.
40. Dercle L, Fronheiser M, Lu L, et al. Identification of non-small cell lung cancer sensitive to systemic cancer therapies using radiomics. *Clin Cancer Res*. 2020;26(9):2151-2162.
41. Prelaj A, Ganzinelli M, Trovo F, et al. The EU-funded I^3LUNG project: integrative science, intelligent data platform for individualized lung cancer care with immunotherapy. *Clin Lung Cancer*. 2023;24(4):381-387.