Radiomics and gene expression profile to characterize the disease and predict outcome in patients with lung cancer

Margarita Kirienko
Fondazione IRCCS Istituto Nazionale dei Tumori
Martina Sollini (martinasollini@msn.com)
Humanitas University https://orcid.org/0000-0003-2214-6492
Marinella Corbetta
Humanitas University
Emanuele Voulaz
Istituto Clinico Humanitas - Ospedale a Milano: Humanitas Research Hospital
Noemi Gozzi
Istituto Clinico Humanitas - Ospedale a Milano: Humanitas Research Hospital
Matteo Interlenghi
Consiglio Nazionale delle Ricerche
Francesca Gallivanone
Consiglio Nazionale delle Ricerche
Isabella Castiglioni
Consiglio Nazionale delle Ricerche
Rosanna Asselta
Humanitas University
Stefano Duga
Humanitas University
Giulia Soldà
Humanitas University
Arturo Chiti
Humanitas University

Research Article

Keywords: Artificial intelligence, Image analysis, Lung cancer, Radiogenomics, PET/CT, mutation, gene expression.

Posted Date: February 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-157951/v1

Abstract

Objectives The objectives of our study were to assess the association of radiomic and genomic data with histology and patient outcome in non-small cell lung cancer (NSCLC).

Methods In this retrospective single-centre observational study, we selected 151 surgically treated patients with adenocarcinoma or squamous cell carcinoma who performed baseline [18F]-FDG PET/CT. A subgroup of patients with cancer tissue samples at the Institutional Biobank (n=74/151) was included in the genomic analysis. Features were extracted from both PET and CT images using an in-house tool. The genomic analysis included detection of genetic variants, fusion transcripts, and gene expression. Generalized linear model (GLM) and machine learning (ML) algorithms were used to predict histology and tumour recurrence.

Results Standardized Uptake Value (SUV) and kurtosis (among the PET and CT radiomic features, respectively), and the expression of *TP63*, *EPHA10*, *FBN2*, and *IL1RAP* were associated with the histotype. No correlation was found between radiomic features/genomic data and relapse using GLM. The ML approach identified several radiomic/genomic rules to predict the histotype successfully. The ML approach showed a modest ability of PET radiomic features to predict relapse, while it identified a robust gene expression signature able to predict patient relapse correctly. The best-performing ML radiogenomic rule in predicting the outcome resulted in an Area Under the Curve (AUC) of 0.87.

Conclusions: Radiogenomic data provided clinically relevant information in NCSCL patients, regarding the histotype, aggressiveness, and progression. Gene expression may provide additional valuable information to guide patient management. The application of ML allows to increase the efficacy of radiogenomic analysis and provide novel insights into cancer biology.

Introduction

Lung cancer is a leading global cause of cancer-related deaths, with more than 2.2 million people diagnosed and 1.9 million deaths documented worldwide in 2017 [1]. The five-year survival rate is less than 25% when diagnosis occurs at a locally advanced or metastatic disease stage. The survival rate rises above 50% if the disease is diagnosed early when local treatment is feasible [2]. In unresectable disease, systemic therapy with cytotoxic and targeted drugs are the main treatment options [3, 4]. Targeted therapy requires the molecular profiling of tumour tissue to identify specific biomarkers to tailor treatments [5]. Cancer cells are characterized by high genomic instability, responsible for accumulating somatic mutations in crucial oncogenic/oncosuppressor genes, driving uncontrolled cancer cell proliferation [6]. Genetic information may be used to predict survival, as a prognostic biomarker, or response to treatment, as a predictive biomarker to support clinical decisions [7]. In particular, in lung cancer genetic mutations in *ALK*, *BRAF*, *EGFR*, and *ROS1* guide treatment decisions in patients affected by advanced disease and recurrence [3, 4]. Beyond these alterations, other oncogenic driver mutations - even if currently not targetable - include *RET*, *HER2*, *KRAS*, and *MET* [8].

In cancer patients, features of tumours identified from imaging data (e.g., CT and PET) can be used as biomarkers to reveal diagnostic, predictive, and prognostic associations, based on the identification of correlations with pathological or molecular reference, response to treatment, or survival outcomes: this process is defined radiomics. Within this framework, image features extracted and used as predictors include lesion volume, shape, and texture descriptors [9]. Radiogenomics refers to the integration of imaging-derived parameters and genomic data to find clinically relevant associations. Imaging-based typing has the advantage that it can capture information from the whole tumour lesion, can be performed at multiple time points for treatment monitoring, and can be carried out when a biopsy is not feasible. It is cost-effective, relying on routinely acquired clinical imaging.

On the other hand, imaging-genomic maps may be valuable for predicting targeted therapies' efficacy [10]. Additionally, the integration of data generated from complementary "omics" sources can improve the performance of single-domain predictive models [11]. Current data supporting the efficacy of radiomics in lung cancer predicting diagnosis, prognosis, and optimal therapy are ample and promising and supports a future role for computer-assisted diagnosis and management in clinical oncology [12]. Nonetheless, radiogenomics in lung cancer patients is still in its early stages, and extensive data studies are needed to validate the concept [5].

The objectives of our study were to assess: 1) the association of [18F]-FDG PET/CT radiomic features with histology and patient outcome; 2) the association of mutations and gene expression data with histology and patient outcome, and 3) the association of radiogenomics with histology and patient outcome.

Methods

Study design

In this retrospective single-centre observational study, we applied the following criteria to select patients from the Institutional database. Inclusion criteria were: i) age > 18 years, ii) pathologic diagnosis of non-small cell lung cancer (NSCLC), iii) enrolment from November 2011 to April 2018; iv) availability of baseline [18F]-FDG PET/CT, v) surgical treatment, and vi) availability of cancer tissue sample at the Institutional Biobank for those patients to be included in a subpopulation of the study cohort. Exclusion criteria were: i) diagnosis of other malignancies, but non-melanoma skin cancer, in the previous three years; ii) interval time between PET/CT and surgery > 3 months; iii) neoadjuvant treatment; iv)NSCLC other than squamous cell carcinoma (SQC) and adenocarcinoma (AC) to avoid inhomogeneity within the patient cohort. The selection workflow is reported in Fig. 1. Demographic parameters such as age and sex were collected for all patients. Smoking habits were recorded. Performance status was not considered in this analysis.

The institutional ethics committee approved the study (study number 1751). All the patients who donated their tissue samples to the biobank signed informed consent to use their data, imaging and samples for research purposes; for the remaining patients, because of the observational and retrospective study design, a specific, informed consent was waived.

Image acquisition and processing

[18F]-FDG PET/CT image acquisition was performed according to versions 1.0 [13] and 2.0 [14] of the European Association of Nuclear Medicine (EANM) guidelines until and from February 2015, respectively. Briefly, patient preparation with fasting at least 4 hours before [18F]-FDG injection and blood glucose levels below 200 mg/dl were requested. Images were acquired 60 ± 5 min after injection of [18F]-FDG, using either a Siemens Biograph 6 LSO (Siemens, Erlangen, Germany) or a General Electric Discovery 690 (General Electric Healthcare, Waukesha, WI, USA) PET/CT scanner. All PET images were corrected for attenuation using the acquired CT data. Image acquisition parameters are reported in Supplementary Table 1.

The primary lung cancer lesions were delineated on PET images applying a fully automatic segmentation method, combining an automatic threshold-based algorithm to define the tumour volume and a k-means clustering algorithm for estimation of the background [15]. CT target lesion was delineated using 3-D Slicer application FastGrowCut implementing a competitive region growing algorithm using cellular automata [16]. The images were resampled to have isometric voxels with 2 mm length. Calculation of PET parameters (n = 60) was performed using an in-house image processing tool, running on MATLAB [17, 18]. The partial volume effect correction was used for Standardized Uptake Value (SUV) calculation. CT radiomic features (n = 57) extraction was performed using HeterogeneityCAD tool implemented in the 3-D Slicer, according to Aerts et al. [19]. Image processing and calculation of image-derived parameters are reported in Supplementary Table 1, according to the imaging biomarker standardization initiative (IBSI) reporting guidelines [20].

Pathology

Histological type and staging classification were assessed according to good clinical practice on pathology samples obtained at surgery (AJCC manual). Fresh-frozen samples were collected and stored according to the Biological Biobank's Institutional procedures for those patients who donated their tissue.

Molecular analyses (mutations and gene expression)

Molecular analysis of 74 tumour samples (21 SQC, 53 AC) was performed using a targeted RNAseq approach. Besides, six normal tissue samples were evaluated as reference. Molecular analyses by targeted RNAseq included: 1) detection of genetic variants (both single nucleotide variants, SNVs, and small insertions/deletions, indels); 2) detection of fusion transcripts; 3) gene expression analysis. RNA extraction from fresh-frozen lung tissues, preserved in RNALater-ICE (Thermo Fisher Scientific, Waltham, MA, USA), was performed using either an automated procedure with the Maxwell RSC miRNA Tissue kit (Promega, Madison, WI, USA) or a standard protocol using the Eurogold TriFast reagent (Euroclone, Wetherby, UK). RNA quality was assessed on an Agilent 4200 TapeStation (Agilent Technologies; Santa Clara, CA, USA), obtaining a mean RNA integrity number (RIN) of 6.7 (max: 9, min: 4). Libraries were prepared starting from 55 ng of total RNA with the TruSight RNA Pan-Cancer Panel (Illumina, San Diego, CA, USA), following the manufacturer's instructions. This panel allows the simultaneous detection of fusion transcripts, point mutations, and gene expression changes, and it is characterized by a broad dynamic range, which can robustly detect RNAs of low abundance; it covers a total of 1,385 genes, including all major ones found mutated in lung cancer. Sequencing (76-bp paired-end reads) was performed on a NextSeq500 platform (Illumina). Data were analyzed using the RNAseq alignment v. 2.0.10 pipeline on BaseSpace (Illumina). Briefly, input reads were filtered against abundant sequences, such as mitochondrial or ribosomal sequences, using Bowtie 0.12.9, and then aligned to the reference human genome (UCSC hg19) and the RefSeq annotation of transcripts with the Spliced Transcripts Alignment to a Reference (STAR) program (v. 2.6.1a). SNVs were identified with the Strelka Variant Caller v.2.9.9, and the presence of fusion transcripts was detected with the Manta Structural Variant Caller v.1.4.0. Gene/transcript expression was quantified by Salmon v.0.11.2. Differential expression analysis among histotypes (SQC, AC, and normal tissue) was evaluated with a likelihood ratio test (LRT) for significance using the DeSeq2 Bioconductor package [21]. We looked for known and possibly recurrent oncogenic variants to extract meaningful information to correlate with imaging and clinical data. Therefore, we selected variants with the following features: SNV, nonsynonymous (missense, nonsense, splice variants mapping at +/-2 position of the splice sites), and annotated in the Clinvar database (https://www.ncbi.nlm.nih.gov/clinvar/) as pathogenic or likely pathogenic. Concerning indels, we selected rare frameshift variants occurring only in cancer samples, covered by at least 10 reads, and with a Combined Annotation Dependent Depletion, CADD, score > 15 [22].

Follow-up and outcome assessment

After surgery, further treatment and follow-up were performed according to standard procedures and guidelines after discussion at the multidisciplinary lung tumour board. As for outcome prediction, the endpoints of this study were disease recurrence, disease-free survival (DFS), and overall survival (OS). Disease recurrence was defined as relapse occurrence during follow-up. DFS was defined as the time between the date of surgery and either the date of recurrence or tumour-related death (event) or the date of last patient access (censored). OS was defined as the time between surgery and death (event) or the date of last patient access (censored).

Statistical analyses

All statistical procedures were carried out using specific R program packages, release 3.6.1 (http://www.r-project.org/). All P values were two-sided. P values of < 0.05 were considered statistically significant.

Conventional statistics

Patient characteristics were summarized in frequency tables, and descriptive statistics were provided. Principal component analysis (PCA) was used to explore data.

The Kaplan–Meier method was used to generate survival curves for the subgroups in each dataset, and the log-rank test was used to determine the statistical significance of differences (survminer R package).

For *KRAS/TP53/EGFR* mutations, analyses of carrier frequency data were performed using the Fisher exact test.

For data analysis, we first performed an unsupervised clustering, using the pheatmap R package on log-transformed data, evaluating whether patients are distinguishable in the feature space (either by histotype or the tumour recurrence). After that, we used variables to fit a Generalized Linear Model (GLM), after

having checked for the correct normal distribution of residuals and the homogeneity of variance across the fitted values of the model. We used either the histotype (SQC or AC) or the tumour recurrence (yes or no) as the output variable. More specifically, for gene expression analysis, differentially expressed genes were selected using a false discovery rate (FDR) ≤ 0.001, a fold change ≥ 2, and a minimum average expression (baseMean) of at least 50 counts. Differential expression analyses identified genes specifically altered in cancer status compared to normal tissues and genes specifically altered in each histotype (SQC and AC). Differential expression analysis to identify transcriptional signatures associated with tumour recurrence (yes or not) was evaluated with a Wald test for significance and the DeSeq2 package. In this case, differentially expressed genes were selected using a FDR ≤ 0.05 and no threshold for fold change.

For radiogenomic analysis, we focused on the top differentially expressed genes, setting a stringent threshold for significance of FDR ≤ 0.001, fold change ≥ 2, and average expression (baseMean) among samples of at least 50 counts.

Machine learning analysis

The significance of classical statistical analysis (e.g. univariate analysis and GLM previously introduced) decreases for high dimensional datasets (i.e. large number of features), especially when the biological/clinical significance of features is unknown [23]. In recent years, the quantitative nature and the multitude of omics data have driven research towards the application of ML for the analysis [24]. ML is one of the major subfields in artificial intelligence and may be defined as a set of algorithms, developed on input data, able to infer a prediction of labels on unseen observations (e.g. a diagnosis or outcome prediction). The advantage of ML lies in learning directly from data and improving the prediction process [25]. Therefore, we decided to apply the Forecast environment of the Rulex (RULe eXtractor) 4.0 suite (www.rulex-inc.com), an integrated suite for the analysis of data through statistical/ML approaches. Rulex can manage all data types, including categorical/continuous variables, variables showing a high degree of correlation, or characterized by any type of data distribution.

Three datasets were used for analyses: 1) a dataset including only complete radiomic data (149 patients; two additional patients were not included because they were missing either PET or CT data); 2) a dataset using only molecular data (gene expression and *KRAS/TP53/EGFR* mutational status), comprising 74 patients; and 3) a dataset including only those 73 patients for which both radiomic and molecular data were available.

We applied the Logic Learning Machine (LLM) algorithm to our datasets using as output variable either the histotype (SQC or AC) or the patient outcome (tumour recurrence yes or no). Each dataset was split into a training cohort and a test set (70% and 30%, respectively). Rulex LLM takes as input the features and returns a series of "rules" characterized by *n* conditions. For each rule, the algorithm provides the percentage of exact output predictions (i.e. covering) and the error percentage.

We used the radiogenomic rule (integrating gene expression and radiomic data) which best predicted tumour recurrence to build a score, and we calculated the corresponding receiving operating characteristic (ROC) curve and the area under the curve (AUC). For each condition, the score corresponded to the sum of 1 or 0 points.

Results

Overall, 151 patients were included. Patient characteristics are reported in Table 1. 70% of patients developed an AC, 30% an SQC; the ratio between those that relapsed versus those that did not was exactly 1:1.

Table 1
Patient characteristics.

Characteristics		*Whole dataset (N = 151)**	*Genetics (n = 74)***
Age - median (range)		70 (41–84) years	70 (41–80) years
Sex (M: F)		95:56	47:27
Histology (AC: SQC)		106:45	53:21
Smoking status (Yes:No: Ex-smokers)***		42:31:77	23:14:36
Outcome	Lost at follow-up	7/151	6/74
	Relapse Yes: No	72:72	31:37
	Follow-up/OS - median (range)	39 (1-102) months	24 (3–79) months
	DFS	44(1-102) months	40 (4–81) months

* The whole dataset consisted of 151 patients, for 2 of them either PET or CT data were missing and hence not included in the ML analysis. ** This category indicates the subset of patients submitted to mutational and differential gene-expression analyses (for 1 of them, we did not have radiomics data, and hence not included in the ML analysis of combined radiomics and transcriptomics data); *** For one person, we do not have data on smoking status. AC, adenocarcinoma; DFS, disease-free survival; F, females; M, males; OS, overall survival; SQC, squamous cell carcinoma.

No correlations were observed between patients' age and the histotype of the relapse status (P = 0.078 and P = 0.538, respectively). As for sex, a weak correlation was observed with histotype (being the male sex more common among SQC, with a male: female ratio = 3.5, P = 0.016). Instead, no correlation was evident between sex and relapse. Finally, we observed a weak correlation between the histotype and the relapse status, with AC cases more prone to relapse than SQC patients (P = 0.032; the significance was also retained after correction for age, sex, and smoking status, P = 0.026) (Supplementary Fig. 1).

The two histotypes did not show any significant difference in the overall survival rate of patients, nor on their tendency to relapse (Supplementary Fig. 2).

Mutation screening by targeted RNAsequencing identified more than 130,500 variants in the 80 analyzed samples (74 tumours + 6 normal), with a mean of 1,632 variants per sample (1,577 SNVs, 27 deletions and 28 insertions). Each variant was covered on average by 162 reads (min: 3, max: 12,592). We obtained a list of the 142 topmost pathogenic variants, although very few of them were present in more than one sample. We found a missense *KRAS* mutation at codon 12 in 28% (21/74) of tumour samples, and a *TP53* mutation in 55% (41/74) of cases (Supplementary Fig. 3).

No known pathogenic variants or hotspot mutations were detected in *EGFR*, although eight samples carried rare nonsynonymous SNVs of uncertain significance (Fig. 2). Based on the variant type, location within the gene/protein, and predicted deleteriousness (evaluated by CADD, score > 15), at least 6 (75%) of them might represent a likely pathogenic mutation. No recurrent gene fusions in the *ALK*, *ROS*, and *RET1* were detected.

Gene expression analysis on RNAseq data on 80 samples (21 SQC, 53 AC, and 6 non-tumoral tissues) showed a prevalent clusterization based on histotype at PCA, with normal tissues separating from tumour samples (Supplementary Fig. 4). Samples were not separated based on relapse status, at least when considering the first two principal components (data not shown).

1. Association of PET/CT radiomic data with histology and outcome

A total of 60 PET and 57 CT radiomic features were extracted in 149 patients (see Table 1).

Unsupervised hierarchical clustering, based on log-transformed data of extracted radiomic features, showed a good clusterization based on histotype (Supplementary Fig. 5). Conversely, the same analysis did not perform well when discriminating patients based on relapse (data not shown).

PET and CT features were further independently analyzed using GLM. GLM showed that among the 117 analyzed features, two outperformed in discriminating AC vs SQC patients. SUV and kurtosis resulted in the best PET and CT features, respectively, in predicting histology (Fig. 3A). SUV resulted higher in SQC than in AC (16.91 ± 7.92 and 10.13 ± 5.77, respectively; $P = 2.89*10^{-6}$). Similarly, CT-derived kurtosis was greater in SQC than in AC (11.97 ± 11.17 and 3.81 ± 5.85, respectively; $P = 5.49*10^{-6}$). Notably, both variables survived Bonferroni corrections for multiple testing (threshold corresponding to P = 0.00083 and P = 0.00088 for the SUV and the kurtosis, respectively). Conversely, PET and CT features poorly correlated with the relapse status. Indeed, the best performing PET and CT features in discriminating tumour recurrence were Kurtosis (P = 0.035) and LRE (P = 0.0096), respectively, but neither kurtosis nor LRE survived correction for multiple testing (Fig. 3B).

The Rulex LLM software, comprehensively analyzing all the 117 radiomic features, identified four and two rules to predict histology and tumour recurrence, respectively (Table 2). The rule number 4 - composed of 8 different conditions, three based on PET and five based on CT features - was the most interesting, reaching coverage of 85.7% with an error rate of 3.6%.

Table 2
Histotype and relapse predictions based on radiomics features - Rulex analysis results.

Rule	*Output*	*Covering (%)**	*Error (%)***	*Cond 1*	*Cond 2*	*Cond 3*	*Cond 4*	*Cond 5*	*Cond 6*	*Cond*
A - HISTOTYPE										
1	Histotype = AC	48.7	3.6	Max_Intensity CT ≤ 73						
2	Histotype = AC	43.4	3.6	PVC_SUV PET ≤ 16.946515	Spherical_disp PET ≤ 2.354150	GLV PET ≤ 0.065135	Voxel_Count CT > 799	Max_Intensity CT ≤ 140		
3	Histotype = AC	42.1	3.6	Kurtosis PET ≤ 2.654415	SumAverage PET ≤ 0.009197	Stand_Deviation CT > 43.485000				
4	Histotype = SQC	85.7	3.6	Min PET > 1.312135	Uniformity PET > 0.005588	GLN PET > 0.019649	Compactness_2 CT ≤ 0.179247	Min_Intensity CT ≤ -309	Max_Intensity CT > 69	Clust CT > 1241
B - RELAPSE										
1	Relapse = NO	66. 7	3.7	Energy PET ≤ 17375	Energy PET > 0.000853	SRLGE PET ≤ 0.039183	SZLGE PET > 0.002677	0.009156 < LZLGE PET ≤ 0.092522	Complexity PET ≤ 11892	
2	Relapse = YES	61. 1	4.2	Mean PET ≤ 15.503765	Mean abs dev PET ≤ 4.515045	Energy PET > 17010	Correlation PET > 0.316205	SZE PET > 0.601404	HGZE PET ≤ 1262	LRLG CT ≤

* Covering: the percentage of the training patterns that satisfy the rule whose output value is equal to the rule's output; ** the percentage of the training patter whose output value is different from the output of the rule. The best-performing rule in each section is highlighted in grey.

Table 3
Association between KRAS/TP53/EGFR mutational status and histotype/relapse.

A - HISTOTYPE			
KRAS	**Non-carriers (N)**	**Carriers (N)**	**P-value ***
AC	33	20	**0.0040**
SQC	20	1	
TP53			
AC	24	29	0.80
SQC	9	12 (1 case with 2 mutations)	
EGFR			
AC	48	5	0.43
SQC	18	3	
B - RELAPSE **			
KRAS			
YES	18	13	**0.029**
NO	31	6	(0.086)
TP53			
YES	15	16	0.80
NO	16	21 (1 case with 2 mutations)	(1)
EGFR			
YES	31	0	**0.028**
NO	31	6	(0.17)

* Fisher exact test. ** Analysis performed on a total of 68 cases; in this analysis, the P-value presented in parenthesis are corrected for the histotype. Significant P values are indicated in bold.

2. Association of genomic data with histology and outcome

Mutation and gene expression data on the 74 cases were used to search for possible correlations with histotype and the relapse status. Genes specifically altered by relapse were investigated on the 68 samples with available data (31 relapses, 37 no relapse). Only four genes resulted differentially expressed with an FDR ≤ 0.05. Mutation analysis focused on the two genes most commonly mutated in our cohort (*KRAS*, *TP53*), as well as on the *EGFR* gene, i.e. the sole - among genes guiding treatment decision in lung cancer patients [3, 4] - being mutated in the analyzed cases. Concerning histotype, we evidenced a profound difference in the frequency of *KRAS* mutation carriers between AC and SQC (37.7% and 4.8%, respectively; P = 0.0040) (Table 3A). We also observed a significant association between *KRAS*/ *EGFR* mutations and relapse status, not retained after correction for histotype (values between parentheses in Table 3B).

Gene expression analysis focused on the 238 genes that resulted differentially expressed (at FDR ≤ 0.001 and a fold change ≥ 2) according to different tissue samples (normal tissue, AC, SQC; 187 genes), tumour histotypes (47 genes), and recurrent status (4 genes). Among these 238 genes, *TP63*, *FBN2*, *EPHA10*, and *IL1RAP* emerged as strongly associated with the histotype ($P < 1.5*10^{-4}$; all genes survive the Bonferroni corrections for multiple tests) (Fig. 4).

None among the analyzed genes was associated below the Bonferroni threshold (P = 0.00021) with the relapse status. The comprehensive analysis performed by the Rulex LLM approach focused on all the 238 genes, and the data on *KRAS*/ *TP53*/ *EGFR* mutational status are summarized in Table 5. Rulex LLM proved to be very powerful both in predicting the histotype (rule number 1 reached a coverage > 94%) and, above all, in predicting the outcome (rule number 4 reached a coverage = 92%). In all cases, no conditions related to the mutational status emerged. The best-performing rule predicting the histotype was based only on two conditions (one related to expression levels of *TP63*), whereas the best rule predicting relapse stem on the expression levels of 5 different genes (*AURKA*, *HIST1H2AM*, *IL12Rb2*, *CXXC4*, and *RYR3*).

Table 5
Histotype and relapse predictions based on mutation and gene expression data - Rulex analysis results.

Rule	*Output*	*Covering (%)**	*Error (%)***	*Cond 1*	*Cond 2*	*Cond 3*	*Cond 4*	*Cond 5*
A - HISTOTYPE								
1	Histotype = AC	94.3	0	*HIF1A* ≤ 12.662616	*TP63* ≤ 7.748115			
2	Histotype = AC	74.3	0	*EPCAM* > 10.788665				
3	Histotype = SQC	88.2	0	*TP63* > 7.142827	*EPCAM* ≤ 10.788665			
4	Histotype = SQC	29.4	0	*HIF1A* > 12.627919				
B - RELAPSE								
1	Relapse = NO	58.3	2.6	*HIF1A* > 11.268115	*RYR3* > 5.052014			
2	Relapse = NO	54.2	2.6	*CDK4* ≤ 9.801157	*GNAQ* ≤ 8.387779			
3	Relapse = NO	37.5	2.6	*CXXC4* ≤ 2.124897				
4	Relapse = YES	91.7	0	*AURKA* > 6.211039	*HIST1H2AM* ≤ 9.677214	*IL12RB2* ≤ 7.370931	*CXXC4* > 1.892740	*RYR3* ≤ 5.160035
5	Relapse = YES	20.8	4.6	*ECT2L* > 4.866101				

* Covering: the percentage of the training patterns that satisfy the rule whose output value is equal to the rule's output; ** the percentage of the training patterns that satisfy the rule whose output value is different from the output of the rule. The best-performing rule in each section is highlighted in grey.

3. Association of radiogenomic with histology and outcome

Finally, we used the Rulex LLM approach for analyzing the 73 cases having the entire set of variables available (i.e. all PET and CT features) and data on *KRAS/ TP53/ EGFR* mutational status and on 238 differentially expressed genes).

Results of radiogenomics analysis are summarized in Table 6. Interestingly, using hystotype as an output variable results almost entirely overlapping with those already obtained for the predictions based on genomic data (Table 5A). The slightly different covering values depended on the missing sample.

Table 6
Histotype and relapse predictions based on radiomic features, mutation, and gene expression data - Rulex analysis results.

Rule	*Output*	*Covering (%)**	*Error (%)***	*Cond 1*	*Cond 2*	*Cond 3*	*Cond 4*
A - HISTOTYPE							
1	Histotype = AC	92.3	4.8	*HIF1A* ≤ 12.717780	*TP63* ≤ 7.853678		
2	Histotype = AC	73.1	0	*EPHA10* > 5.143141			
3	Histotype = SQC	90.5	3.8	*FCGBP* ≤ 10.437165	*TP63* > 7.125095		
4	Histotype = SQC	23.8	0	*HIF1A* > 12.657739			
B - RELAPSE							
1	Relapse = NO	67.6	3.3	*SMO* > 5.381590	*ATRNL1* > 1.689360	*GHR* > 4.673396	Busyness_PET ≤ 0.222300
2	Relapse = NO	35.1	3.3	*FEN1* ≤ 8.477941	*GRID1* ≤ 5.241994	*HIST1H2BO* > 8.055055	
3	Relapse = NO	24.3	3.3	*ACKR3* ≤ 6.493873	*LHX2* > 0.672730		
4	Relapse = YES	73.3	0	*CXXC4* > 1.871573	*GHR* ≤ 4.706566	2.614300 < *PAK3* ≤ 9.111535	LRHGE_PET > 853
5	Relapse = YES	23.3	2.7	*DLL3* > 0.655998	6.049279 < *LGR5* ≤ 8.796951		
6	Relapse = YES	13.3	0	*IBSP* ≤ 0.255760			

* Covering: the percentage of the training patterns that satisfy the rule whose output value is equal to the rule's output; ** the percentage of the training patterns that satisfy the rule whose output value is different from the output of the rule. The best-performing rule in each section is highlighted in grey.

The best-performing rule in predicting the outcome (covering rate > 73%) combined conditions based on gene expression data and a PET-derived feature (i.e. LRHGE). This rule's ROC curve resulted in an area under the curve of 0.87 (Fig. 5).

Discussion

Our study showed the ability of radiomics, genomics, and radiogenomics to provide clinically relevant information in lung cancer patients. We found that image-derived features were able to discriminate between NSCLC histotype (Fig. 3). Several studies have successfully demonstrated an association between radiomic features and NSCLC tumour histology based on both CT and PET radiomic features. In the study by Wu et al. [26] 53 CT radiomic features from lung tumours of 350 patients were significantly associated with tumour histological subtype. Applying multivariate classifiers using radiomic features as input tumour histological subtype could be reliably predicted (AUC = 0.72) [26].

Similarly, Aerts et al. [19] reported a radiomic analysis of 440 features extracted from CT data of 1,019 patients affected by NSCLC and head-and-neck cancer. They found a significant association with histology (P = 0.019, Chi-square test). While in the study by Koyasu et al., the authors found PET-based models to identify histological lung cancer subtype with an AUC up to 0.84 [27]. In a previous cohort of 534 patients with lung nodules, we have demonstrated radiomic features' ability to potentially classify primary lung cancer subtypes (AUC = 0.59–0.70 for CT and = 0.61–0.88 for PET)[28]. SUV appeared to be the best predictor in the present work, following literature data [29], regardless of the statistical approach (i.e. conventional statistics and ML). Indeed, SUVmax was recognized by GLM as significant, and minimum SUV was included in the best radiomic rule to predict histology. Randomness due to high correlation and the inherent redundancy among SUVs parameters was probably at the basis of the model's selection of SUVmax instead of the minimum.

Even though PET and CT features failed in predicting tumour recurrence when analyzed with classical statistics, we found a couple of rules through the Rulex LLM approach capable of correctly prognosticating the outcome in a good percentage of cases (61–67% of covering). These findings underlined that one radiomic feature might efficiently differentiate tumour subtypes, but it is not sufficient to explain the complexity of the disease (i.e. biological phenotyping). Conversely, combining more radiomic features, enclosing several - complementary - information, may summarize all those biological properties beyond the histotype that contributes to disease aggressiveness. Notably, the best-performing rule to predict tumour relapse comprised a total of 6 conditions all related to PET features. The inclusion of conditions related to CT-derived parameters determined a drop of performances, confirming literature data [30].

We identified the levels of expression of several genes (e.g. *TP63* and *EPHA10*) to be associated with AC vs SQC. In particular, *TP63* overexpression, often due to gene amplification, is frequently found in SQC and has been associated with prolonged survival [31]. *EPHA10* (the ephrin receptor A10 belonging to the subfamily of receptor tyrosine kinases and involved in cell-cell communication, regulating cell attachment, shape, and mobility in neuronal and epithelial cells) is expressed many breast cancers [32]. Conversely, it has never been described to be associated with lung cancer subtypes. This finding needs further investigation to better characterize this protein's role in lung cancer as a prognostic biomarker or potential therapeutic target.

Gene expression data - analyzed by Rulex LLM - proved to be very powerful both in predicting the histotype (rule number 1 reached a coverage > 94%) and, above all, in predicting the outcome (rule number 4 reached a coverage = 92%). The best-performing genomic rule in predicting relapse stem on the expression levels of *AURKA*, *HIST1H2AM*, *IL12Rb2*, *CXXC4*, and *RYR3*. The expression of *AURKA* - a gene that contributes to the regulation of cell cycle progression [33] - has been reported to be associated with poor prognosis in smoking-related lung AC [34]. Histones variants *HIST1H2*, *HIST1H3*, and *HIST1H4*, acting as transcriptional promoters or repressors of cancer-related genes, have been reported to be involved in tumour progression and metastasis [35]. In particular, in lung AC, *HIST1H2AM* has been counted among genes differentially expressed between smoking and non-smoking [36]. Our data suggested that the smoking habits induced changes in *AURKA* and *HIST1H2AM* genes, and their upregulation or downregulation might play a pivotal role in determining the outcome. Literature data supported the involvement of IL-12Rβ2 in tumour cell proliferation, apoptosis, and metastasis. The down-regulation of IL-12Rβ2 in lung AC seems to be a tumour escape mechanism [37], and the IL-12Rβ2 expression has been negatively associated with tumour progression [38]. The Disheveled (Dvl) inhibitor Idax, coded by the *CXXC4* gene [39], seems to be involved in tumour cell invasiveness and proliferation [40, 41], and its expression is associated with poor prognosis [42]. Furthermore, *CXXC4*, being capable of inhibiting the mitogen-activated protein kinases (MAPK) signalling pathway [43], is emerging as a novel tumour suppressor [44]. Moreover, the downregulation of the MAPK signalling pathway seems to reduce the expression of programmed cell death 1 ligand 1 (PD-L1) in lung AC cells [45]. The *RYR3* gene encodes for a ryanodine receptor which mediates the calcium release for many cellular processes. Ryanodine receptor has been reported to be involved in epithelial-mesenchymal transition, in tumour cell apoptosis and treatment resistance in some cancers, including lung AC [46–48]. These data may provide the rationale for post-operative risk stratification with a differential follow-up scheme. The patients operated on a more aggressive tumour may be closely investigated during follow-up to identify recurrence at an earlier time point. However, molecular testing to identify molecular biomarkers are currently performed on tumour samples collected from biopsies or cytological specimens; these are invasive procedures which are not always feasible, may result in inadequate sampling, and cannot characterize intra- and inter-tumour heterogeneity. Moreover, in the case of recurrence, repetition of a biopsy is not mandatory. Indeed, targeted therapies may be administered based on the molecular testing on the specimens obtained at diagnosis, assuming that no molecular modification arises between disease onset and recurrence [49]. Consequently, other methods to identify actionable biomarkers in NSCLC are emerging to address the need for complementing or replacing traditional testing on tissue and cytological samples.

From the Rulex LLM analysis on radiogenomics, our study emerged that gene expression data alone prevail on those coming from [18F]-FDG PET/CT analyses in predicting histology. Interestingly, the best-performing rule in predicting the outcome (covering rate > 73%) combined conditions based on gene expression data and PET-derived feature (i.e. LRHGE). Though this rule - less performing than the one found when analyzing only gene expression data (Table 5B) - it was noteworthy to underline that: i) the number of conditions for this rule was lower, ii) we "forced" the software to give priority to radiomic features (which indeed only came up for relapse predictions), and iii) the corresponding ROC curve gave an overall significant AUC of 0.87.

The study is limited by the retrospective design that determined using routinely acquired PET/CT images on two different scanners. However, in our previous study, we demonstrated radiomics analysis's reliability since the predictive models performed in the same way when considering and not considering significantly different features among scanners [28]. Secondly, the outcome has been evaluated in terms of occurrence or not of disease. This evaluation may have determined a modest prognostic ability. Future analyses will take into account time information related to recurrence and overall survival.

In conclusion, the radiogenomic approach promises the extraction of relevant information regarding lung cancer histotype, aggressiveness, and progression. Gene expression may provide additional valuable information to guide patient management and follow-up. ML algorithms' application allows to increase the efficacy of transcriptomic and radiogenomic analysis and provides novel insights into cancer biology.

Declarations

Funding: The present project was supported by the grant won by AC (IG-2016-18585) funded by the AIRC (Italian Association for Cancer Research).

Authors' contributions: AC, MK, and MS conceptualized the study; AC, SD, GS, RA, MK, and MS designed the study; MK screened patients; EV enrolled patients; MS and MK collected the clinical data; MK, FG, IC, and MI, performed image analysis; MC and GS performed the genomic analysis; RA performed data analysis; AC, SD, MS, MK, RA, GS and NG critically interpreted results; AC, RA, MK, MS, GS, and NG drafted the paper. AC coordinated and supervised the project's activities. All the authors critically revised the paper and approved the submitted version of the manuscript.

Availability of data and material: The manuscript represents valid work, and neither this manuscript nor one with substantially similar content under the same authorship has been published or is being considered for publication elsewhere. Arturo Chiti had full access to all the data in the study and takes responsibility for the data integrity and the accuracy of the data analysis. Raw data are available on specific request to the corresponding author.

Code availability: not applicable.

Compliance with ethical standards:

Consent to participate and consent for publication: All the patients who donated their tissue samples to the biobank signed informed consent for the usage of their data, imaging and samples for a research purpose; for the remaining patients, because of the observational and retrospective study design, a specific, informed consent was waived.

Ethics approval: The Ethics Committee of the Humanitas Clinical and Research Centre IRCCS approved the study on 18 April 2017, with the authorization number 1751.

Conflict of interest: Chiti reports a fellowship grant from Sanofi, personal fees from AAA, Blue Earth Diagnostics and General Electric Healthcare, outside the submitted work. The other authors do not report any conflict of interest.

References

1. Fitzmaurice C, Abate D, Abbasi N, Abbastabar H, Abd-Allah F, Abdel-Rahman O, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017. JAMA Oncol. American Medical Association; 2019;5:1749.
2. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 2016;11:39–51.
3. Postmus PE, Kerr KM, Oudkerk M, Senan S, Waller DA, Vansteenkiste J, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2017;28:iv1–21.
4. Planchard D, Popat S, Kerr K, Novello S, Smit EF, Faivre-Finn C, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29:iv192–237.
5. Ninatti G, Kirienko M, Neri E, Sollini M, Chiti A. Imaging-Based Prediction of Molecular Therapy Targets in NSCLC by Radiogenomics and AI Approaches: A Systematic Review. Diagnostics. 2020;10:359.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
7. Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol. 2018;15:353–65.
8. Kalemkerian GP, Narula N, Kennedy EB, Biermann WA, Donington J, Leighl NB, et al. Molecular testing guideline for the selection of patients with lung cancer for treatment with targeted tyrosine kinase inhibitors: American society of clinical oncology endorsement of the college of American pathologists/ international association for the. J Clin Oncol. 2018;36:911–9.
9. Sollini M, Cozzi L, Antunovic L, Chiti A, Kirienko M. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. Sci Rep. Springer US; 2017;7:358.
10. Jansen RW, van Amstel P, Martens RM, Kooi IE, Wesseling P, de Langen AJ, et al. Non-invasive tumor genotyping using radiogenomic biomarkers, a systematic review and oncology-wide pathway analysis. Oncotarget. Oncotarget; 2018;9:20134–55.
11. Bodalal Z, Trebeschi S, Nguyen-Kim TDL, Schats W, Beets-Tan R. Radiogenomics: bridging imaging and genomics. Abdom Radiol. 2019;44:1960–84.
12. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019;46:2656–72.
13. Boellaard R, O'Doherty MJ, Weber WA, Mottaghy FM, Lonsdale MN, Stroobants SG, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. Eur J Nucl Med Mol Imaging. Springer; 2010;37:181–200.
14. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2014;42:328–54.
15. Gallivanone F, Interlenghi M, Canervari C, Castiglioni I. A fully automatic, threshold-based segmentation method for the estimation of the Metabolic Tumor Volume from PET images: validation on 3D printed anthropomorphic oncological lesions. J Instrum. 2016;11:C01022–C01022.
16. Vezhnevets V, Konushin V. "GrowCut" - Interactive Multi-Label ND Image Segmentation By Cellular Automata. Proc Graph. 2004;1:150–6.
17. Stefano A, Gallivanone F, Grosso E, Messa C, Gianolli L, Gilardi C, et al. TOUCH-SUV: a Touchscreen-Assisted Tool for Quantitative, Reproducible, Clinically Feasible and Collaborative Diagnostic Reporting of Whole-Body PET-CT. Softw Eng. 2011;1:1–8.
18. Stefano A, Gallivanone F, Grosso E, Messa C, Gianolli L, Gilardi MC, et al. TOUCH-SUV: a Touchscreen-Assisted Tool for Quantitative, Reproducible, Clinically Feasible and Collaborative Diagnostic Reporting of Whole-Body PET-CT Studies. Int J Softw Eng. 2012;1:24–31.
19. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
20. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. BioMed Central; 2014;15:550.
22. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
23. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019;46:2656–72.
24. Sollini M, Bartoli F, Marciano A, Zanca R, Slart RHJAJA, Erba PA. Artificial intelligence and hybrid imaging: the best match for personalized medicine in oncology. Eur J Hybrid Imaging. 2020;4:24.
25. Kirienko M, Biroli M, Gelardi F, Seregni E, Chiti A, Sollini M. CNN-based approaches for PET/CT and PET/MR. Where do we stand?". Clin Transl Imaging. 2021;
26. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. Front Oncol. 2016;6:1–11.
27. Koyasu S, Nishio M, Isoda H, Nakamoto Y, Togashi K. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT. Ann Nucl Med. 2020;34:49–57.
28. Kirienko M, Cozzi L, Rossi A, Voulaz E, Antunovic L, Fogliata A, et al. Ability of FDG PET and CT radiomics features to differentiate between primary and metastatic lung lesions. Eur J Nucl Med Mol Imaging. 2018;45:1649–60.

29. Groheux D, Quere G, Blanc E, Lemarignier C, Vercellino L, de Margerie-Mellon C, et al. FDG PET-CT for solitary pulmonary nodule and lung cancer: Literature review. Diagn Interv Imaging. Elsevier Masson; 2016;97:1003–17.
30. Kirienko M, Cozzi L, Antunovic L, Lozza L, Fogliata A, Voulaz E, et al. Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. Eur J Nucl Med Mol Imaging. 2018;45:207–17.
31. Nagano K, Maeda Y, Kanasaki S, Watanabe T, Yamashita T, Inoue M, et al. Ephrin receptor A10 is a promising drug target potentially useful for breast cancers including triple negative breast cancers. J Control Release. 2014;189:72–9.
32. Wang Y-T, Pan S-H, Tsai C-F, Kuo T-C, Hsu Y-L, Yen H-Y, et al. Phosphoproteomics Reveals HMGA1, a CK2 Substrate, as a Drug-Resistant Target in Non-Small Cell Lung Cancer. Sci Rep. 2017;7:44021.
33. Carvalhal S, Ribeiro SA, Arocena M, Kasciukovic T, Temme A, Koehler K, et al. The nucleoporin ALADIN regulates Aurora A localization to ensure robust mitotic spindle formation. Zheng Y, editor. Mol Biol Cell. 2015;26:3424–38.
34. Zhang M-Y, Liu X-X, Li H, Li R, Liu X, Qu Y-Q. Elevated mRNA Levels of AURKA, CDC20 and TPX2 are associated with poor prognosis of smoking related lung adenocarcinoma using bioinformatics analysis. Int J Med Sci. 2018;15:1676–85.
35. Mehta A, Dobersch S, Romero-Olmedo AJ, Barreto G. Epigenetics in lung cancer diagnosis and therapy. Cancer Metastasis Rev. 2015;34:229–41.
36. Sui Q, Liang J, Hu Z, Chen Z, Bi G, Huang Y, et al. Genetic and microenvironmental differences in non-smoking lung adenocarcinoma patients compared with smoking patients. Transl Lung Cancer Res. Transl Lung Cancer Res; 2020;9:1407–21.
37. Airoldi I, Di Carlo E, Cocco C, Caci E, Cilli M, Sorrentino C, et al. IL-12 Can Target Human Lung Adenocarcinoma Cells and Normal Bronchial Epithelial Cells Surrounding Tumor Lesions. Ng IO-L, editor. PLoS One. 2009;4:e6119.
38. Liu Z, Yang W, Yang S, Cai K. The close association between IL-12Rβ2 and p38MAPK, and higher expression in the early stages of NSCLC, indicates a good prognosis for survival. Mol Med Rep. 2018;
39. Kojima T, Shimazui T, Hinotsu S, Joraku A, Oikawa T, Kawai K, et al. Decreased expression of CXXC4 promotes a malignant phenotype in renal cell carcinoma by activating Wnt signaling. Oncogene. 2009;28:297–305.
40. Wissmann C, Wild PJ, Kaiser S, Roepcke S, Stoehr R, Woenckhaus M, et al. WIF1, a component of the Wnt pathway, is down-regulated in prostate, breast, lung, and bladder cancer. J Pathol. 2003;201:204–12.
41. Uematsu K, He B, You L, Xu Z, McCormick F, Jablons DM. Activation of the Wnt pathway in non small cell lung cancer: evidence of dishevelled overexpression. Oncogene. 2003;22:7218–21.
42. Zhao Y, Yang Z-Q, Wang Y, Miao Y, Liu Y, Dai S-D, et al. Dishevelled-1 and dishevelled-3 affect cell invasion mainly through canonical and noncanonical Wnt pathway, respectively, and associate with poor prognosis in nonsmall cell lung cancer. Mol Carcinog. 2010;n/a-n/a.
43. Lu H, Jin W, Sun J, Feng L, Lan H, Shen Q, et al. New tumor suppressor CXXC finger protein 4 inactivates mitogen activated protein kinase signaling. FEBS Lett. 2014;588:3322–6.
44. Li P, Luo X, Xie Y, Li P, Hu F, Chu J, et al. GC-Derived EVs Enriched with MicroRNA-675-3p Contribute to the MAPK/PD-L1-Mediated Tumor Immune Escape by Targeting CXXC4. Mol Ther Nucleic Acids. Mol Ther Nucleic Acids; 2020;22:615–26.
45. Stutvoet TS, Kol A, Vries EG, Bruyn M, Fehrmann RSN, Terwisscha van Scheltinga AGT, et al. MAPK pathway activity plays a key role in PD-L1 expression of lung adenocarcinoma cells. J Pathol. 2019;249:52–64.
46. Zhang L, Chen J, Cheng T, Yang H, Li H, Pan C. Identification of the key genes and characterizations of Tumor Immune Microenvironment in Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). J Cancer. 2020;11:4965–79.
47. Law BYK, Michelangeli F, Qu YQ, Xu S-W, Han Y, Mok SWF, et al. Neferine induces autophagy-dependent cell death in apoptosis-resistant cancers via ryanodine receptor and Ca2+-dependent mechanism. Sci Rep. Sci Rep; 2019;9:20034.
48. Davis FM, Parsonage MT, Cabot PJ, Parat M-O, Thompson EW, Roberts-Thomson SJ, et al. Assessment of gene expression of intracellular calcium channels, pumps and exchangers with epidermal growth factor-induced epithelial-mesenchymal transition in a breast cancer cell line. Cancer Cell Int. 2013;13:76.
49. Vansteenkiste J, Crino L, Dooms C, Douillard JY, Faivre-Finn C, Lim E, et al. 2nd ESMO Consensus Conference on Lung Cancer: early-stage non-small-cell lung cancer consensus on diagnosis, treatment and follow-up. Ann Oncol. 2014;25:1462–74.

Figures

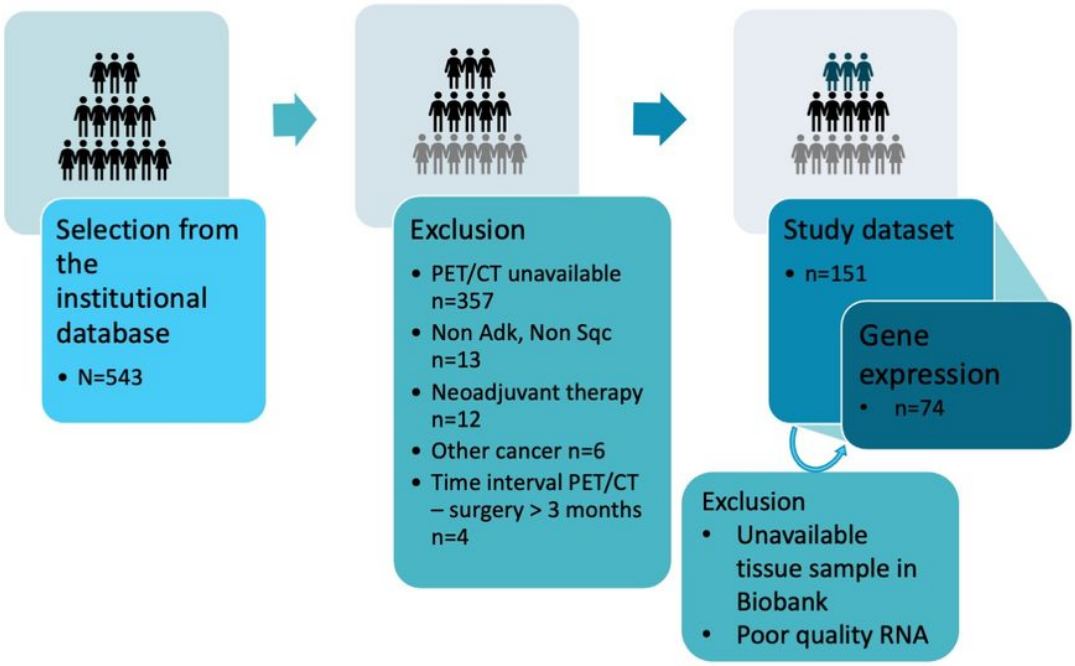

Figure 1

Patients' selection workflow (title). Schematic representation of the patients' selection workflow.

A.

Patient (Cancer histotype)	gene	location (hg19)	SNV type	Variant name* (cDNA:protein)	CADD score	Coverage	Alt Allele fraction	Existing variant
LC4_1T (AC)	EGFR	chr7: 55210120	MS	c.230C>T: p.Ser77Phe	**26.2**	43	0.09	novel
LC3_13T (AC)	EGFR	chr7: 55214383	MS	c.509G>C: p.Ser170Thr	0.004	69	0.12	novel
LC1_27T (SQC)	EGFR	chr7: 55224527	SD	c.1207+2T>G	**23**	3	0.67	novel
LC3_4T (AC)	EGFR	chr7: 55231493	MS	c.1699A>G: p.Met567Val	**16.2**	542	0.45	rs771278492
LC2_9T (SQC)	EGFR	chr7: 55270271	MS	c.3224G>T: p.Gly1075Val	**15.24**	1225	0.82	COSV51826987
LC3_20T (SQC)	EGFR	chr7: 55273012	MS	c.3335A>G: p.Asn1112Ser	**22.6**	101	0.95	rs75533058, COSV51837290
LC3_18T (AC)	EGFR	chr7: 55273292	MS	c.3615T>A: p.Ser1205Arg	11.29	19	0.26	novel
LC4_T (AC)	EGFR	chr7: 55273306	MS	c.3629C>T: p.Ala1210Val	**24.1**	853	0.41	rs35918369

B.

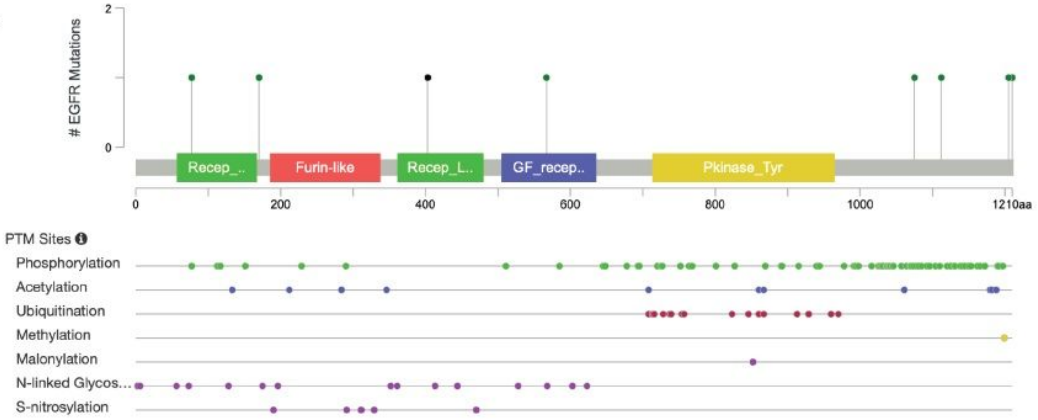

Figure 2

EGFR nonsynonymous rare variants detected in lung tumour samples (title). A. Table listing the eight nonsynonymous rare variants identified in EGFR. Variants are named according to the reference mRNA sequence NM_005228.5. Combined Annotation Dependent Depletion (CADD) score was calculated with the Ensembl Variant Effect Predictor (VEP, http://grch37.ensembl.org/Homo_sapiens/Tools/VEP). Variants with CADD PHRED score >=15 are bolded. The CADD PHRED score ranks a variant relative to all possible substitutions in the human genome: hence, a score of >15 indicates the 5% most deleterious genetic variants. AC, adenocarcinoma; MS, missense; SD, splice donor. B. Lollipop representation of variant localization and the EGFR protein, together with the annotation of Post Translational Modifications (PTMs) obtained from dbPTM (http://dbptm.mbc.nctu.edu.tw). The figure was generated using MutationMapper, available at https://www.cbioportal.org/mutation_mapper.

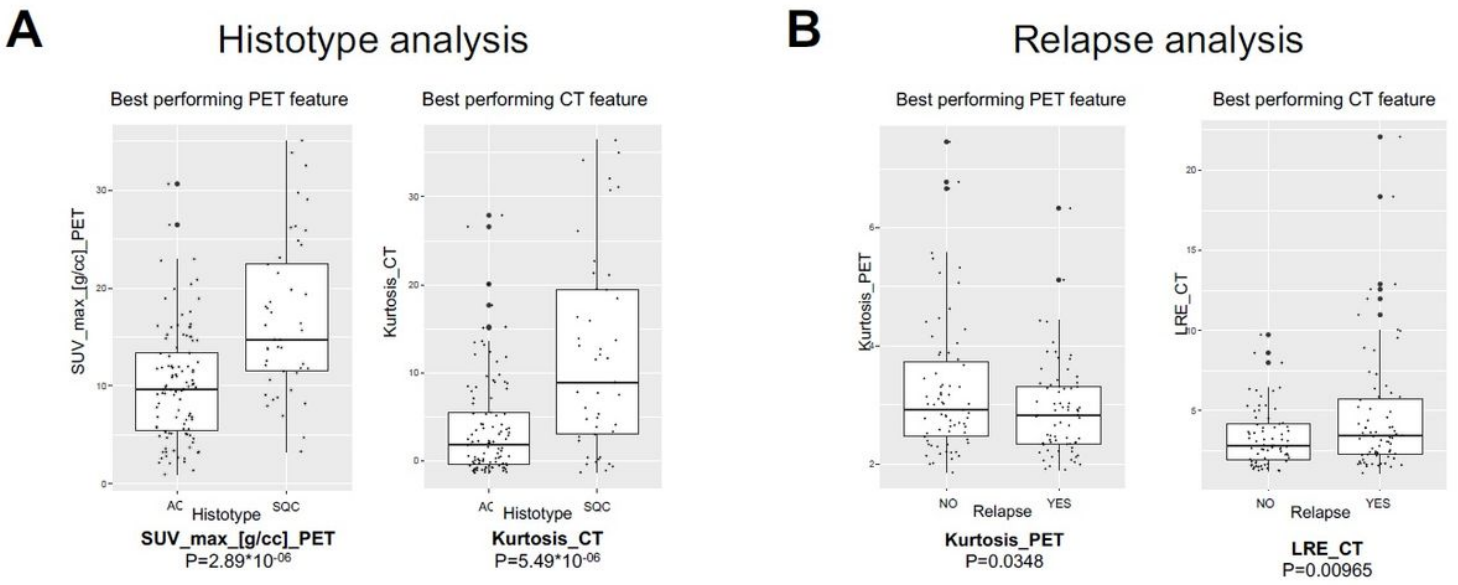

Figure 3

Top PET and CT features discriminating patients based on their lung cancer histotype or their tendency to relapse (title). The best discriminative features were identified by using the generalized linear model approach to predict histology (A) and outcome (B). A) Boxplots show Standardized Uptake Value and Kurtosis, the best performing PET and CT features, respectively in discriminating cancer histotype. B) Boxplots show kurtosis and LRE, the best performing PET and CT features, respectively in discriminating tumour recurrence. Boxes define the interquartile range; thick central lines refer to the median. P values before Bonferroni correction are provided for each feature.

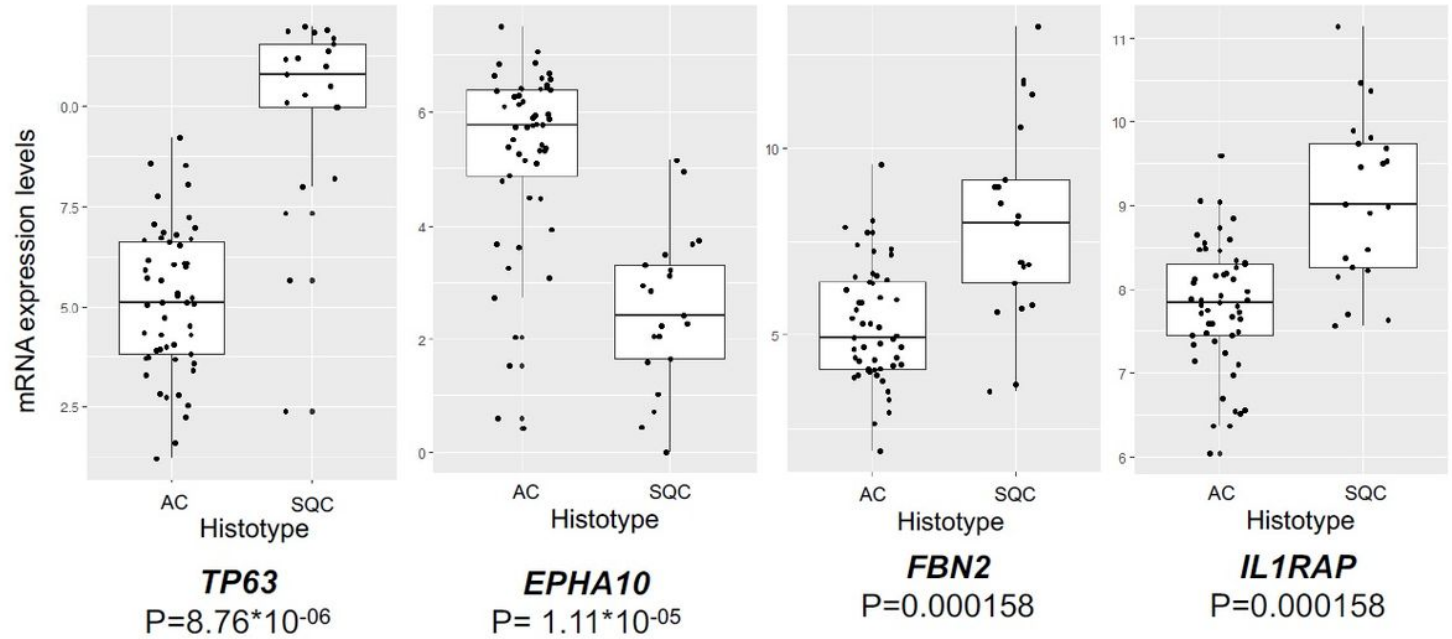

Figure 4

Top differentially expressed genes discriminating patients based on their lung cancer histotype (title). The four box plots show mRNA expression levels of TP63, FBN2, EPHA10, and IL1RAP genes, with lung cancer individuals grouped upon histotype. Boxes define the interquartile range; thick central lines refer to the median. Below each box plot, the P-value for the difference is indicated (t-test; the threshold for Bonferroni correction for multiple testing corresponding to P=0.00021).

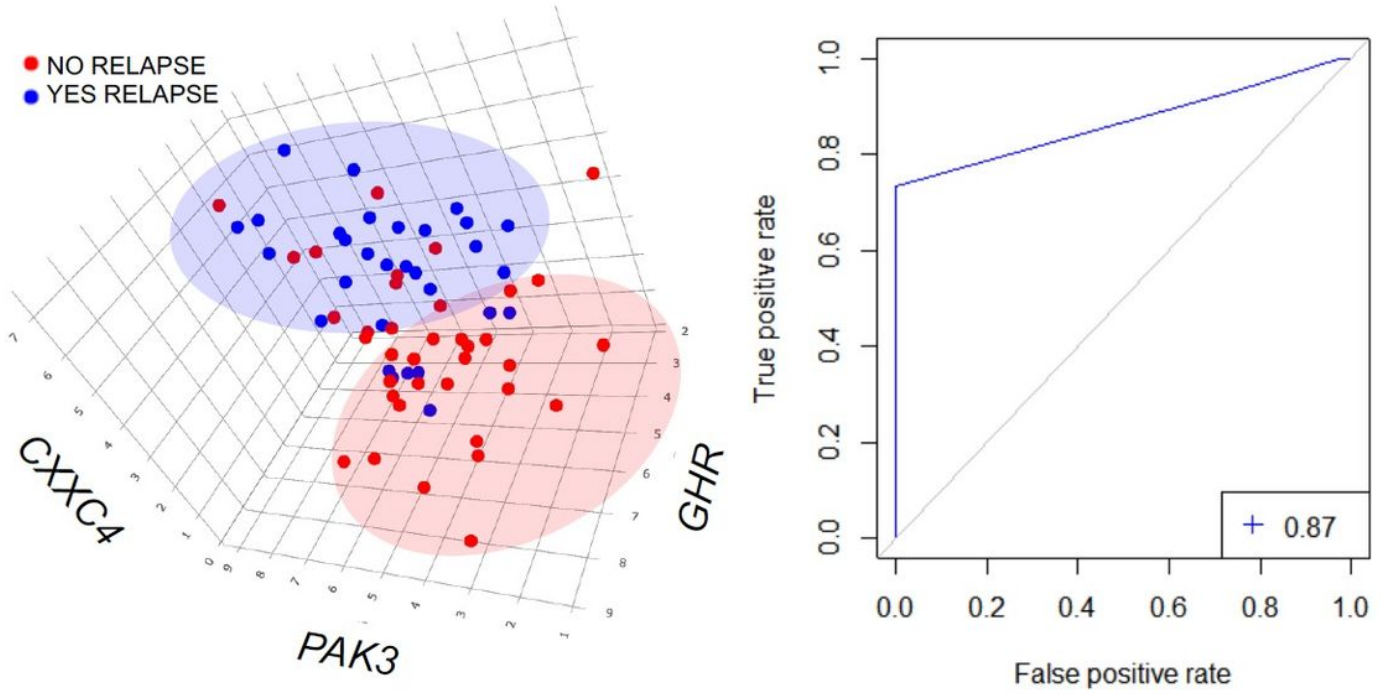

Figure 5

Clusterization of relapsing/non-relapsing patients based on the best-performing prediction rule evidenced by the Rulex LLM analysis (title). On the left: three-dimensional scatter plot of patients experiencing (blu dots) or not (red dots) relapse. Patients were plotted based on the three genes' expression levels evidenced by the Rulex LLM analysis (determining the first three conditions of rule number 4; see Table 6B). On the right: ROC curve for differentiating relapsing and non-relapsing patients based on a “score” including the expression levels of the CXXC4, PAK3, and GHR genes, as well as on the radiomic parameter LRHGE_PET. For each patient, the score was built summing, for each of the four conditions of the rule (Table 6B), 1 or 0 points. At the bottom rich corner of the ROC panel, the AUC value is reported.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RadiogenomicsAIRC2016finalsuppl2101.docx