Predicting Response to Cancer Immunotherapy using Non-invasive Radiomic Biomarkers

Supplementary Material

S. Trebeschi [1, 2, 3, *], S.G. Drago [1, 4], N.J. Birkbak [5], I. Kurilova [1, 2], A.M. Călin [1, 6], A. Delli Pizzi [1, 7], F. Lalezari [1], D.M.J. Lambregts [1], M. Rohaan [8], C. Parmar [3], K.J. Hartemink [9], C. Swanton [5], J.B.A.G. Haanen [8], C.U. Blank [8], E.F. Smit [10], R.G.H. Beets-Tan [1, 2, &, +], H.J.W.L. Aerts [1, 3, &, +]

[1] Department of Radiology, Netherlands Cancer Institute, Amsterdam, The Netherlands. [2] GROW School of Oncology and Developmental Biology, Maastricht, The Netherlands. [3] Departments of Radiation Oncology and Radiology, Dana Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4] Department of Radiology, Milano-Bicocca University, San Gerardo Hospital, Monza, Italy. [5] The Francis Crick Institute & University College London, London, UK. [6] Affidea Romania, Cluj-Napoca, Romania. [7] ITAB Institute of Advanced Biomedical Technologies, University G. d'Annunzio, Chieti, Italy. [8] Department of Medical Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands. [9] Department of Surgery, Netherlands Cancer Institute, Amsterdam. The Netherlands. [10] Department of Thoracic Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands. (*) first author, (&) equally contributed (+) corresponding author

S1. BASELINE CHARACTERISTICS

Study cohort I baseline characteristics stratified per cancer-type.

	All	Melanoma	NSCLC	*p-value* (MEL vs NSCLC)
Patients Characteristics				
Age [median, IQR]	63 (IQR 13)	62 (IQR 22)	63 (IQR 12)	0.60
Gender (female) [N, %]	94 (46.3%)	38 (47.5%)	56 (45.5%)	0.06
NSCLC Squamous [N, %]	-	-	29 (23.6%)	-
Non Squamous	-	-	94 (76.4%)	-
Melanoma Cutaneous [N, %]	-	72 (90.0%)	-	-
Non Cutaneous	-	8 (10.0%)	-	-
1y Survival [N, %]	139 (68.5%)	55 (68.8%)	88 (71.5%)	0.67
1y Best Overall Response (CR, [N, %])	10 (4.9%)	7 (8.8%)	3 (2.4%)	**0.04**
Partial Response [N, %]	54 (26.6%)	16 (20.0%)	38 (30.9%)	0.09
Stable Disease [N, %]	47 (23.2%)	23 (28.7%)	24 (19.5%)	0.13
Progressive Disease [N, %]	84 (41.4%)	30 (37.5%)	54 (43.9%)	0.37
Past Treatments [N, %]	182 (89.7%)	60 (75.0%)	122 (99.2%)	**< 0.01**
Chemotherapy	133 (65.5%)	11 (13.8%)	122 (99.2%)	**< 0.01**
Radiotherapy	86 (42.4%)	37 (46.2%)	49 (39.8%)	0.37
Ipilimumab	55 (27.1%)	55 (68.8%)	0	**< 0.01**
Targeted Therapy	22 (10.8%)	20 (25.0%)	2 (1.6%)	**< 0.01**
Immunotherapy [N, %]	203 (100.0%)	80 (100.0%)	123 (100.0%)	1.00
Nivolumab	145 (71.5%)	22 (27.5%)	123 (100.0%)	**< 0.01**
Pembrolizumab	58 (28.6%)	58 (72.5%)	0	**< 0.01**
Exam Characteristics				
Interval between baseline CT scan and start of treatment [days]	26.4	28.8	24.9	0.19
Interval between start of treatment and follow-up CT scan [days]	69.3	77.9	63.9	**< 0.01**
Lesion Count				
All lesions [N lesions, N patients, median]	1055 (203 pts, 3/pt)	483 (80 pts, 4/pt)	572 (123 pts, 3/pt)	-
Pulmonary	359 (129 pts, 2/pt)	85 (34 pts,1.5/pt)	274 (95 pts, 2/pt)	**< 0.01**
Hepatic	212 (42 pts, 2/pt)	135 (23 pts, 2/pt)	77 (19 pts, 2/pt)	**0.03**
Lymph Nodes	312 (116 pts, 2/pt)	130 (47 pts, 2/pt)	182 (69 pts, 2/pt)	0.82
Adrenal gland	58 (40 pts, 1/pt)	26 (17 pts, 1/pt)	32 (23 pts, 1/pt)	0.79
Subcutaneous	96 (27 pts, 2/pt)	91 (25 pts, 2/pt)	4 (2 pts, 2/pt)	**< 0.01**
Splenic	18 (8 pts, 1/pt)	15 (6 pts, 1/pt)	3 (2pts, 1.5/pt)	0.08
Per Lesion Response Outcomes				
Responding [N, %]	351 (33.3 %)	195 (40.4 %)	156 (27.3 %)	**< 0.01**
Pulmonary	112 (31.2 %)	41 (48.2 %)	71 (25.9 %)	**< 0.01**
Hepatic	66 (31.1 %)	49 (36.3 %)	17 (22.1 %)	**0.04**
Lymph Nodes	105 (33.7 %)	49 (37.7 %)	56 (30.8 %)	0.25
Others	68 (39.5 %)	56 (42.1 %)	12 (30.8 %)	0.26
Stable [N, %]	395 (37.4 %)	176 (36.4 %)	219 (38.3 %)	0.53
Pulmonary	128 (35.7 %)	32 (37.6 %)	96 (35.0 %)	0.75
Hepatic	85 (40.1 %)	59 (43.7 %)	26 (33.8 %)	0.20
Lymph Nodes	130 (41.7 %)	50 (38.5 %)	80 (44.0 %)	0.39
Others	50 (30.2 %)	35 (26.3 %)	17 (43.6 %)	0.07
Progressive [N, %]	309 (29.3 %)	112 (23.2 %)	197 (34.4 %)	**< 0.01**
Pulmonary	119 (33.1 %)	12 (14.1 %)	107 (39.1 %)	**< 0.01**
Hepatic	61 (28.8 %)	27 (20.0 %)	34 (44.2 %)	0.17
Lymph Nodes	77 (24.7 %)	31 (23.8 %)	46 (25.3 %)	0.87
Others	52 (30.2 %)	42 (31.6 %)	10 (25.6 %)	0.59

S2. RADIOGRAPHIC DIFFERENCES

Summary of radiographic differences in different metastatic locations: adrenal (A), hepatic (H), lymph nodes (LN), pulmonary (P), subcutaneous (SUBq) and spleen lesions (S). Reference (*ref*) to Figure 3A is given, along with with feature settings of filter, class, feature name, binning (B) and resampling (R). Association with response are shown by means of mixed model *p-values.* Significance after FDR is marked in bold. Failure of model convergence is reported as N/A.

	Radiographic Feature					**Difference Responding vs Progressive (*p*)**						
ref	**B**	**R**	**Filter**	**Class**	**Feature**	All	A	H	LN	P	Sq	S
f_01	1	1.0	LoG.5.0mm	FirstOrder	Minimum	0.31	**< 0.01**	0.28	0.26	0.09	0.33	**< 0.01**
f_02	5	3.0	LoG.2.5.mm	GLCM	Homogeneity1	0.47	0.80	0.52	**0.02**	0.43	1.00	0.80
f_03	5	3.0	-	GLCM	Homogeneity1	0.44	0.84	0.91	0.25	0.67	0.64	0.84
f_04	25	5.0	Wavelet.HHH	GLCM	Homogeneity1	0.73	0.92	0.42	0.12	0.04	0.44	0.92
f_05	1	1.0	Wavelet.HLH	GLSZM	ZoneEntropy	**< 0.01**	0.99	**< 0.01**	0.73	0.13	0.52	0.99
f_06	5	3.0	Square	FirstOrder	Entropy	0.22	N/A	0.16	**< 0.01**	0.59	0.72	N/A
f_07	5	3.0	LoG.5.0mm	GLCM	DifferenceEntropy	0.43	0.95	0.24	0.17	0.26	0.97	0.95
f_08	5	3.0	SquareRoot	GLRLM	LowGrayLevRunEm.	0.88	0.76	**0.01**	0.37	0.53	0.90	0.76
f_09	25	5.0	-	Shape	SurfaceVolumeRatio	**0.01**	0.90	**0.01**	**0.01**	0.97	0.54	0.90
f_10	25	5.0	Wavelet.LLL	GLCM	MaximumProbability	0.14	0.93	**0.02**	0.56	0.43	0.37	0.93

S3. BIOMARKER PERFORMANCE

Prediction performance of the chosen machine learning classifier on independent validation set. Size of both discovery and validation sets are reported in terms of number of patients (Pts), number of positive samples i.e. non-responding lesions (N+), and number of negative samples (N-).

Cancer	Organ	Discovery Set			Test Set			AUC	*p*
		Pts	N+	N-	Pts	N+	N-		
NSCLC	-	81	135	266	42	62	109	0.75	**< 0.001**
	Lung	61	61	124	34	46	43	0.80	**< 0.001**
	Lung (primary)	29	10	21	16	4	12	0.79	0.05
	Lung (metastases)	43	51	102	25	42	31	0.83	**< 0.001**
	Lymph Nodes	47	37	88	22	9	48	0.78	**< 0.01**
	Liver	16	30	38	3	4	5	0.75	0.14
	Adrenal	15	6	13	8	3	10	0.70	0.18
	Spleen	2	0	3	0	0	0		N/A
	Subcutaneous	1	1	0	1	0	3		N/A
Melanoma	-	52	77	274	28	35	97	0.55	0.20
	Lung	22	6	51	12	6	22	0.55	0.37
	Lymph Nodes	25	14	56	22	17	43	0.64	0.05
	Liver	16	20	88	7	7	20	0.55	0.35
	Adrenal	12	10	8	5	4	4	0.58	0.43
	Spleen	4	1	12	2	1	1		N/A
	Subcutaneous	21	26	59	4	0	7		N/A
All	-	133	212	540	70	97	206	0.66	**< 0.001**

S4. ACQUISITION PROTOCOLS

Immunotherapy Dataset. The CT scans were performed by either covering the chest (n=86) or covering the chest and abdomen (n=117) using multi-slice CT equipment (Toshiba Aquilion CX, Minato, Tokyo, Japan; Siemens Somatom Sensation Open, Erlangen, Germany) with a tube voltage of 120 kVp, slice thickness of 1 mm, and in-plane resolution of 0.75 x 0.75 mm. The bolus injection was performed at 3 ml/s (Omnipaque 300, GE Healthcare, Chicago, Illinois, US) not pre-warmed, with a total amount based on the patient weight + 40 cc (minimum of 90 cc and maximum of 130 cc) followed by a saline flush of 30 cc. The chest CT examinations were performed 40 seconds after contrast injection, whereas the chest and abdomen examinations were performed at 70 seconds.

Genomics Dataset. Contrast-enhanced CT scans were acquired 60 days within diagnosis, as part of the Thoracic Oncology Program protocol, of the L. Lee Moffitt Cancer Center (Tampa, Florida, USA). Gene expression of 60,607 probes was measured on a custom Rosetta/Merk Affymetrix 2.0 microarray chipset (HuRSTA_2a520709.CDF, GEO accession number GPL15048) by the Moffitt. The University of South Florida IRB institutional review board approved and waived the informed consent requirement (IRB#16069); data were collected and handled in accordance with the Health Insurance Portability and Accountability Act. Informed consent for gene expression collection was written and oral. For acquisition of imaging and clinical data USF IRB approved protocol (IRB#108426) provided a waiver of informed consent.

Chemotherapy Dataset. The CT scans were performed covering the chest and abdomen (n=39) using multi-slice CT equipment (Toshiba Aquilion CX, Minato, Tokyo, Japan; Siemens Somatom Sensation Open, Erlangen, Germany) with a tube voltage of 120 kVp, slice thickness of 1 mm, and in-plane resolution of 0.75 x 0.75 mm. Specific of the scanning protocols were identical to the immunotherapy dataset.

S5. LESION DELINEATION

The inclusion criteria were: availability of CE-CT BL and FU and, presence of measurable target lesions at baseline. Measurable lesions were defined as any tumor lesions (primary or metastatic lesions) whose entire border could be identified on both BL and FU scans, as our radiomic feature extraction pipeline requires segmented region of interest to extract features

Lesions that disappeared in the FU were flagged as complete response. Lesions that could not be accurately discriminated from surrounding tissues (e.g. lung nodule within atelectasis), with ill-defined borders (e.g. lung lesions adjacent to atelectasis) and lesions which could not be tracked down from other adjacent tumour lesions at baseline or follow-up CTs (e.g. confluent metastases) were not delineated and excluded. Lesions poorly visualized because of the presence of imaging artefacts (e.g. scattering, motion or breathing artefacts) were excluded as well.

S6. RADIOMICS FEATURE EXTRACTION PIPELINE

To reduce the influence of outlier intensity values in the image, the volume was clipped between -1000 HU and 3000 HU. Radiomic features were extracted from original images as well as from different image transformations including five Laplacian of Gaussian filters (σ = 1.0, 2.5, 5.0, 7.5, 10.0 mm), eight wavelets decompositions, and four non-linearities (exponential, square, square root and logarithm). We also repeated the extraction over three different scales, each defined by a set of radiomic parameters: (1) a fine scale with 1 mm isotropic resolution and 1HU bin width, (2) a medium scale with 3 mm isotropic resolution and bin width of 5HU and (3) coarse scale with 5 mm isotropic resolution and bin width of 25 HU. In this way, the algorithm can choose the best radiomic extraction parameters and/or their combination. Features which resulted in invalid values for more than one lesion were dropped.

S7. MACHINE LEARNING

Dataset Preparation The entire dataset was divided into train, validation and test set based on patient identification numbers (*pid*). Patients whose pid was divisible by three were assigned to the train set, those whose (*pid - 1*) was divisible by three were assigned to the validation set, and those whose (*pid - 2*) was divisible by three were assigned to the test set.

Classifier Pool The first group is composed by three linear classifiers based on logistic regression (LR) models [37], each differentiated by a different feature selection method: (1) *unsupervised* resulting from PCA, (2) *supervised* resulting from wrapper feature selection (WFS), or (3) no feature selection. Similarly, we defined a second group of non-linear classifiers based on random forests (RF) [38]. Finally, we generated two additional classifiers via genetic evolution (GEN-1 and GEN-2) [39]. Each classifier was trained using 2-fold cross validation and optimized via sequential model based optimization [40,41]

Training Strategy Each classifier is trained on the training set using a 2-fold cross validation procedure. To prevent the model from learning to recognize patients rather than the actual lesion-wise classification task, we enforced cross validation at a patient level, avoiding the distribution of lesions of the same patient across different folds. Once trained, the model is evaluated in on the test set to check for under- or overfitting, and model selection.

Classifier Optimization Each classifier comes with a set of tunable parameters, i.e. hyperparameters. We made use of a machine learning procedure, a.k.a. *sequential model based optimization* (SMBO), to tune the hyperparameters of each classifier. SMBO procedure is an iterative procedure, where at each iteration the performance is modelled as a function f of the hyperparameters. The search of the optimal hyperparameters is achieved via optimization of a criterion on f. We chose the commonly used *Expected Improvement* (EI) defined as

$$EI_t(x) = \int max(t - y, 0) p_M(y|x) dy$$

which represents the expectation under some model M of f that $f(x)$ will negatively exceed some threshold t. Parzen estimators were used to approximate the function f.

Hyperparameter Space Logistic regressions had only one tunable hyperparameter representing the weight of the L^2 regularization coefficient. Random forests had four hyperparameters: the max depth of the trees (d), the minimum number of samples in each leaf (mL), the minimum number of samples (mS) and minimum Gini impurity in each split (G). The hyperparameters of the genetic classifiers depend on the specific search result. Finally, wrapper feature selectors had one hyperparameter k, indicating the number of top-performant features selected. For each classifier, we selected the set of hyperparameters resulting in the highest AUC.

Model Selection Once completed, the optimization procedure results in a set of eight trained classifiers. We selected the final classifier by comparing their performance on the validation set. The classifier that achieved the highest AUC score was selected as candidate solution. The following table summarizes in details the results for each classifier on the discovery set.

Classifier	**Feature Selection**	**Train AUC**		**Val. AUC**	***diff***
		Mean	**SD**		
Logistic Regression	*none*	0.62	**0.01**	0.59	0.03
Logistic Regression	Unsupervised (PCA)	0.62	**0.01**	0.59	0.03
Logistic Regression	Supervised (WFS)	0.62	**0.01**	0.58	0.04
Random Forest	*none*	0.64	0.02	0.61	0.03
Random Forest	Unsupervised (PCA)	0.48	0.03	0.54	0.06
Random Forest	Supervised (WFS)	0.62	**0.01**	**0.62**	**0.00**
Genetic Evolution 1		**0.64**	**0.01**	0.60	0.04
Genetic Evolution 2		0.55	0.11	0.55	**0.00**

All algorithms, except for wrapper random forests and the second genetic evolution classifier, reported a certain degree of overfitting quantified by a lower accuracy on the validation set w.r.t the one reported on the training set. During training, all algorithms perform similarly between the two folds of cross validation, except second genetic evolution classifier which showed higher variance. Our

choice of using wrapper random forests as candidate classifier was motivated by the fact that this configuration reached the highest performance with the least amount of overfitting.

S8. CONTROL FOR OVERFITTING

When analysing high dimensional data, overfitting problems might hamper the validity of results. In the context of standard inferential statistics, overfitting is present in the form of type-I error (i.e. false positives). When quantifying radiographic differences between responding and progressive lesions, we made use of standard inferential statistics. To control for overfitting, we applied dimensionality reduction followed by false discovery rate (FDR) adjustment of the p-values. To ensure unbiased dimensionality reduction, we applied an unsupervised method which aims to reduce information redundancy — often found in radiomics data. With this method, we selected the top 10 features that minimize redundancy without using the outcome variable in any way. The selection is made purely based on the feature values, and not on the correlation of the features with outcome, from where the overfitting may originate. In addition, multiple testing correction has been applied on the top 10 features, which further minimizes the probability of overfit and fits our analysis to more common robust analysis made in previous studies.

In the second part of the analysis, we use made use of machine learning. Overfitting in machine learning happens when complex models will start to adapt their parameters so closely to the training data that the trained model will not be able to generalize on unseen data. To control for overfitting in our machine learning pipeline, we employed standard control methods applied in computer science research for artificial intelligence. These methods are fundamentally based on the split of the dataset into three independent sets: training, tuning, and independent testing. Each split has a specific function within the whole analysis. The training set is used by the model to learn the relation between radiomics features and outcome, and fit its parameters accordingly (e.g. leaves splits for random forests). Aside from employing a splitting procedure of the dataset with a training set for learning model parameters, we also apply additional cross-validation within the training set to early detect and discard model parameters that could potentially lead to over-fit. These combined checks in fact result in a machine learning model which selects and uses only 68 features out of a total of 5865. Larger models (using >100 features) were over-fitting already in cross-validation and therefore discarded by the training procedure. Once the training procedure is over, the parameters of the candidate model

are frozen. That is, the model parameters remain completely unchanged when testing the performance on the independent test set.

In summary, to control for over-fitting, we applied data splitting between training, tuning, and independent test set. Model parameters have been fitted only on the training set using cross-validation for early detection of model configuration prone to overfit. Tuning is used during training, aside from the training set, as an additional check for overfitting. Finally, the performance of the candidate model have been evaluated uniquely on the independent test set consisting of unseen data.

S9. Supplementary Figures

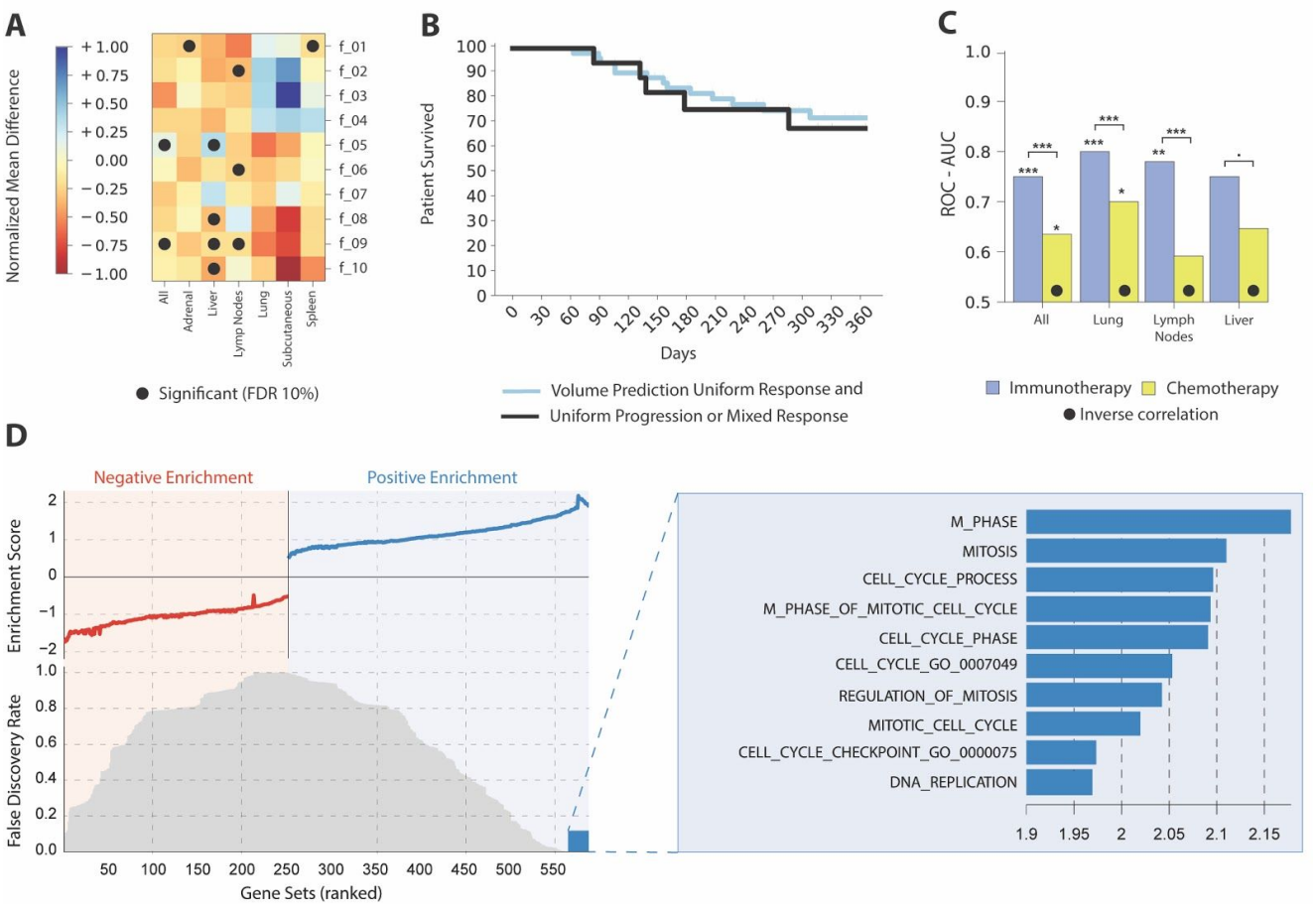

Figure 1. (A) Heatmap of the radiographic differences in different metastatic locations. Values express the *normalized means difference*, where higher values represent higher expression in responding lesions. Significance after FDR is marked by •. Full feature names are reported in S2. **(B)** Overall survival predicted according to volume **(C)** Comparison of lesion radiomic predictive performance between immunotherapy and chemotherapy cohorts **(D)** Association of the biomarker as found by the gene enrichment analysis.