

Test–Retest Reproducibility Analysis of Lung CT Image Features

Yoganand Balagurunathan · Virendra Kumar · Yuhua Gu · Jongphil Kim · Hua Wang · Ying Liu · Dmitry B. Goldgof · Lawrence O. Hall · Rene Korn · Binsheng Zhao · Lawrence H. Schwartz · Satrajit Basu · Steven Eschrich · Robert A. Gatenby · Robert J. Gillies

Published online: 3 July 2014

© Society for Imaging Informatics in Medicine 2014

Abstract Quantitative size, shape, and texture features derived from computed tomographic (CT) images may be useful as predictive, prognostic, or response biomarkers in non-small cell lung cancer (NSCLC). However, to be useful, such features must be reproducible, non-redundant, and have a large dynamic range. We developed a set of quantitative three-dimensional (3D) features to describe segmented tumors and evaluated their reproducibility to select features with high potential to have prognostic utility. Thirty-two patients with NSCLC were subjected to unenhanced thoracic CT scans acquired within 15 min of each other under an approved protocol. Primary lung cancer lesions were segmented using semi-automatic 3D region growing algorithms. Following segmentation, 219 quantitative 3D features were extracted from each lesion, corresponding to size, shape, and texture, including features in transformed spaces (laws, wavelets). The

most informative features were selected using the concordance correlation coefficient across test–retest, the biological range and a feature independence measure. There were 66 (30.14 %) features with concordance correlation coefficient ≥ 0.90 across test–retest and acceptable dynamic range. Of these, 42 features were non-redundant after grouping features with $R^2_{\text{Bet}} \geq 0.95$. These reproducible features were found to be predictive of radiological prognosis. The area under the curve (AUC) was 91 % for a size-based feature and 92 % for the texture features (runlength, laws). We tested the ability of image features to predict a radiological prognostic score on an independent NSCLC (39 adenocarcinoma) samples, the AUC for texture features (runlength emphasis, energy) was 0.84 while the conventional size-based features (volume, longest diameter) was 0.80. Test–retest and correlation analyses have identified non-redundant CT image features with both high

Electronic supplementary material The online version of this article (doi:10.1007/s10278-014-9716-x) contains supplementary material, which is available to authorized users.

Y. Balagurunathan · V. Kumar · Y. Gu · H. Wang · Y. Liu · R. J. Gillies
Department of Cancer Imaging and Metabolism, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

J. Kim · S. Eschrich
Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

D. B. Goldgof · L. O. Hall · S. Basu
Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

R. Korn
Definiens AG, Bernhard-Wicki-Straße 5, 80636 Munchen, Germany

B. Zhao · L. H. Schwartz
Department of Radiology, Columbia University, New York, NY, USA

R. A. Gatenby · R. J. Gillies
Radiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

H. Wang · Y. Liu
Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China

R. J. Gillies (✉)
Experimental Imaging Program, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, SRB-2, Tampa, FL 33612, USA
e-mail: robert.gillies@moffitt.org

intra-patient reproducibility and inter-patient biological range. Thus making the case that quantitative image features are informative and prognostic biomarkers for NSCLC.

Keywords Test–retest reproducibility · Lung cancer · CT · Quantitative image features

Introduction

Classically, CT imaging is routinely used to establish anatomical and macroscopic pathologies in cancer patients. CT images of tumors also depict characteristics that can be related to physiological processes, such as cell density, necrosis, and perfusion, which may not be commonly evaluated. The appearance of the tumor in CT images has been used, qualitatively, to provide information about tumor type, degree of spread, and organ invasion [1, 2]. Such features are typically described subjectively (i.e., “mildly irregular”, “highly spiculated”, “moderate necrosis”). However, to be useful as biomarkers, features must be reproducible, quantifiable, and objective [2]. Thus, there is a need to identify features from CT images that can be reliably extracted and converted into quantifiable, mineable data as potential prognostic, predictive, or response biomarkers. In current clinical practice, only two tumor quantitative CT features; i.e., bi- and uni-dimensional measurements (WHO and RECIST, respectively) are routinely obtained and used to assess response to therapy. While these are satisfactory under some conditions, reduction in tumor size often does not reflect clinico-pathological response [3, 4].

Recent advances in both image acquisition and image analysis techniques allow semi-automated segmentation, extraction, and quantitation of numerous features from images, such as texture. Such features extracted from CT images of lung tumors have been shown to relate to glucose metabolism and stage [5], distinguish benign from malignant tumors [6–9], or differentiate between aggressive and nonaggressive malignant lung tumors [10, 11]. In liver cancer, combinations of 28 image features obtained from CT images could reconstruct 78 % of the global gene expression profiles [12]. As this area of investigation continues to expand, a number of critical questions remain unanswered, including correlated features and reproducibility of individual features. In the present study, we extracted and analyzed a large number of image features describing shape, size, run length encodings, pixel intensity histograms, textures, entropy, and wavelets. In this agnostic approach, we gave equal importance to all features with no prior bias towards radiologist preferences or accepted semantics. Such an analysis of a high dimensional feature space, i.e., “radiomics”, requires standardization and optimization to qualify these potential biomarkers for prognosis, prediction, or therapy response [13, 14]. An important step in the

qualification process is to statistically characterize individual features as being reproducible, non-redundant, and having a large biological range. The most reproducible features are more likely to be able to identify subtle changes with time, pathophysiology, or in response to therapy. Additionally, the reproducibility must be compared to the entire biological range available to that feature across patients. The biological range can be expressed as a dynamic range, DR. It is expected that features will be more useful if they have a large dynamic range. In addition, features must be identified that are not redundant, as it is axiomatic that redundant features can overwhelm learning algorithms and be non-informative for decision support systems.

The inter-scan reproducibility of features may be affected by differences in patient variables, such as positioning, respiration phase, and contrast enhancement, as well as acquisition and processing parameters, including image acquisition power and slice thickness, image reconstruction algorithm, segmentation software, and user input for segmentation. In the present study, the acquisition and processing parameters were fixed, and patient variables were minimized by obtaining two separate CT scans from the same patient on the same machine using the same parameters, within 15 min of each other. Acquisition of these images and reproducibility of tumor uni-dimensional, bi-dimensional, and volumetric measurements has been previously reported [15]. These data have been made publically available under the NCI-sponsored Reference Imaging Database to Evaluate Response (RIDER) project [16]. The objective of the current study is to determine the variability in a large set of agnostic image features extracted from this data set in order to identify the most informative features using empirical filters.

In prior work, we have demonstrated that semi-automatic segmentation had 73 % overlap between operators across a test set of 129 patients [17]. Hence, lesions in the current study were volumetrically segmented using semi-automatic approach (with expert correction) and 219 features were extracted.

Although we began with a large feature set compared to prior conventional radiological analyses [15], it was expected that some features could be redundant due to the sample size. Thus, to reduce the dimensionality of this agnostic data set, we first filtered features based on their reproducibility, searching for those with the highest intra-feature concordance correlation coefficients between the repeats. As a second filter, we used the dynamic range (range of values observed across patients combined with repeatability). Finally, redundancy was assessed by computing an inter-feature coefficient of determination (R^2_{Bet}) between all possible pairs of features. A representative feature set was found by selecting from dependent groups to form an independent set.

We also acquired an existing, independent NSCLC data set (39 adenocarcinoma) of mixed histology, where the tumor

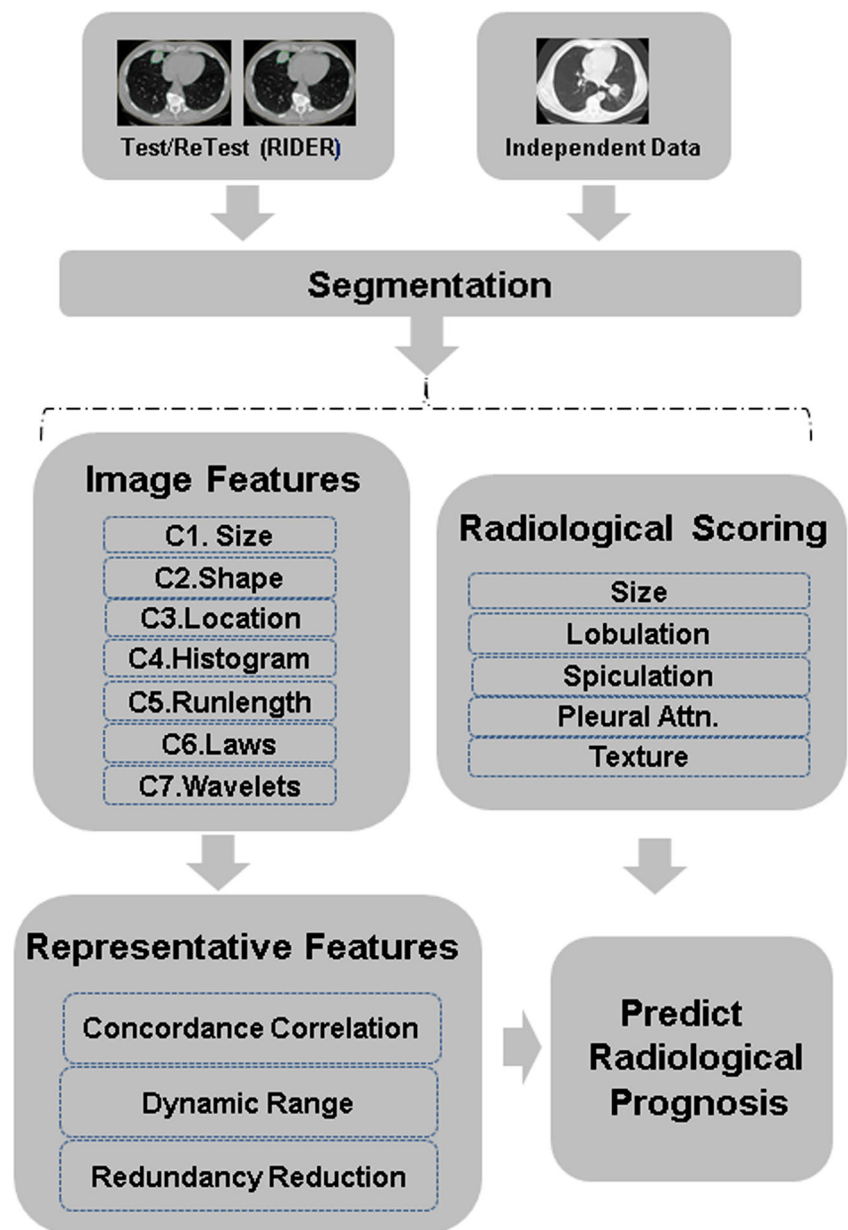
size, node, and metastasis (TNM) stage ranges from I to III. The images features were computed on this subset. Using the known test/retest repeatable features (CCC and DR>0.9), the subset of features obtained and redundancy was reduced ($R^2_{Bet} \geq 0.95$). These image features showed high predictability of radiological prognosis. Figure 1 illustrates the process flow followed in the paper.

Data Collection

The details of patient recruitment have been described in Zhao et al. [15], the samples were collected in an Institutional Review Board (IRB)-approved study. In brief, baseline and

follow up CTs of the thorax for each patient were acquired within 15 min of each other, using the same CT scanner and imaging protocol. Among other possibilities, this enables testing extracted image features for stability. Unenhanced thoracic CT images were acquired using 16-detector (GE Light Speed) or 64-detector (VCT; GE Healthcare) scanners, with 120-kvp tube voltage and image slices thickness of 1.25 mm were reconstructed using the same lung convolution kernel without overlap. The CT scans were acquired from 32 patients (mean age, 62.1 years; range, 29–82 years) with non-small cell lung cancer. There were 16 men (mean age, 61.8 years; range, 29–79 years) and 16 women (mean age, 62.4 years; range, 45–82 years). All patients had a primary pulmonary tumor of 1 cm or larger. The images are available

Fig. 1 Test–retest study workflow to find representative features and its ability to predict radiological prognosis



in the “RIDER Lung CT” collection in NBIA under the “Collections” sections. The RIDER CT image data is on the National Biomedical Imaging Archive (NBIA) site [18], is de-identified without any patient information, complies with HIPPA requirement for sharing data.

Segmentation of Tumors

Definiens Developer XD [19] was used as the image analysis platform. It is based on the Cognition Network Technology[®] [20, 21] which allows the development and execution of image analysis applications. Here, the lung tumor analysis (LuTA) application was used [22]. LuTA contains a semi-automated three-dimensional click-and-grow approach for segmentation of tumors under the guidance of an operator. The semi-automatic segmentation workflow contained the following steps: (a) Preprocessing: The preprocessing performed automated organ segmentation with the main goal of segmenting the aerated lung with correct identification of the pleural wall in order to facilitate the semi-automated segmentation of juxtapleural lesions. (b) Semi-automated correction of the pulmonary boundary: In order to perform seed-based segmentation of a target lesion, the latter has to be completed within the extracted lung image object. In cases where a medical expert (trained radiologist with over 1 year of experience) concluded that the automated preprocessing described above failed to accurately identify the border between a target lesion and the pleural wall, it was necessary to enable correction of the automated lung segmentation. To this end, the image analysts identified the part of the lung that needed modification and placed a seed point manually where the segmentation should be corrected. A seed point outside the lung defined a lung extension, whereas a seed point inside the lung defined a reduction. (c) Click and grow: In order to segment a target lesion the image analysts identified the lesion within the segmented lung and placed a seed point in its interior—typically at the perceived center of the lesion. If the growing process did not sufficiently capture the target lesion, the operator could place additional seed points within the lesion and repeat the growing process outlined above. Upon completion of the segmentation, the individual image objects were merged to form a single image object representing the segmented target lesion. (d) Manual refinement and generation of lesion statistics: Upon completing a seed-based lesion segmentation as described above; the results were viewed by scrolling up and down the stacks of axial images to verify that the segmentation followed the anatomical compartment boundaries properly. To facilitate manual adjustment of the seed-based growing algorithm, tools of two types were constructed. The first type allowed the operator to limit the boundaries beyond which the

region could grow during the “Click-Grow” step by manually placing “blocker” points. Another approach allowed for manual editing of the contour of each segmented lesion on each axial slice by cutting, merging and reclassifying objects and thus enabled the image analysts to perform any desired modifications of the segmented lesion. Image analysts were empowered to override as much or as little of the semi-automatically grown regions as their expertise suggested was indicated. The semi-automatic segmentation process (a)–(d) above required multiple human interactions in order to get the “correct” segmentation boundaries.

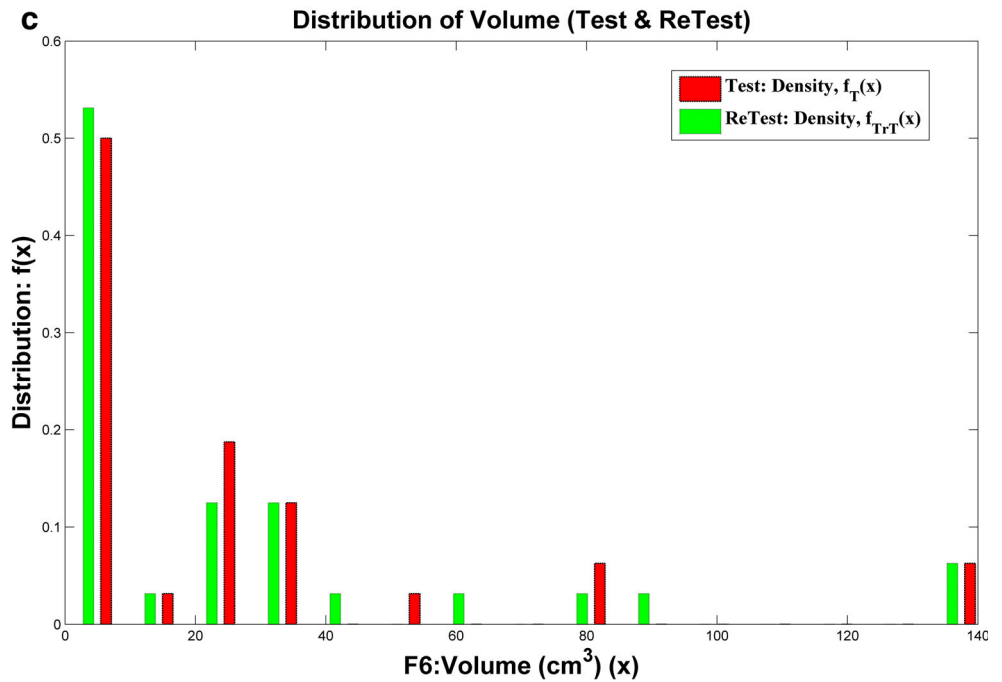
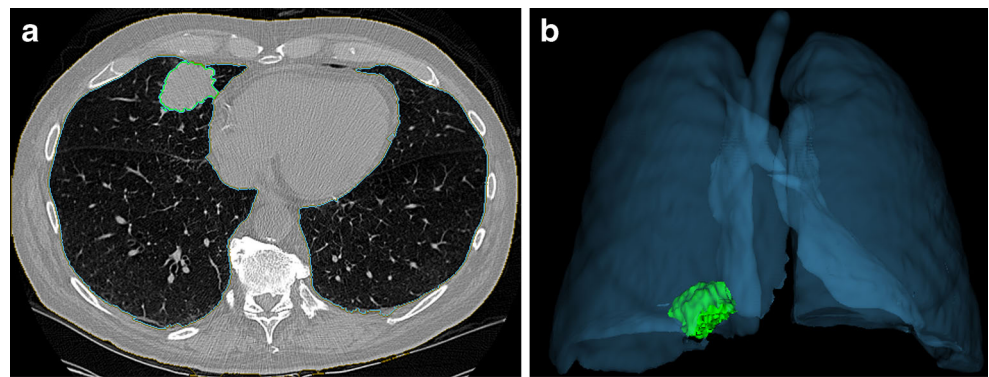
Once the segmentation of all target lesions was complete (e.g., Fig. 2a, b), quantified metrics on each lesion, such as volume, conventional size measurements, and custom feature implementations (texture categories), were extracted. In total, 64 lesions were segmented, i.e., two per patient. Then quantitative values of image features were extracted from each segmented volume. Figure 2c shows the distribution of the volume (in centimeter) measured after segmentations, for both test and retest scans. The volume distribution showed a diverse population, wherein half of the samples had small (volume ≤ 4 cm³) tumors while the rest of the samples are larger in size (largest group close to 140 cm³).

Image Features and Categorization

We extracted several types of image features to describe the tumors heterogeneous shape and structure (details in the subsection below). Note that there are multiple features extracted in some of the categories. As mentioned before, texture features have been shown to be good descriptors of the tumor and have shown relevance for survival prediction [23]. In this study, we have used 219 (including custom) 3D image features. Details of the features are described in Electronic Supplementary Material (ESM) Tables S.1 and description Table S.2. Most size- and shape-based feature computations were implemented within the Definiens XD[®] platform [18], while texture and other derived features were computed from algorithms implemented in C/C++. All the features were obtained from the region of interest (i.e., after the segmentation).

Although 219 features seem like a large set, other effective descriptors may yet need to be added. We categorize our feature set into seven broad categories to describe the lesion, namely: size based, shape based, location based, pixel intensity histogram based, run length and co-occurrence, law’s kernel-based texture and wavelets-based texture descriptors. Table 1 shows the number of features in each of the categories and a detailed description is provided in the ESM Section S.1. Our approach has been driven by the conventional radiologist belief that an ensemble of factors including tumor shape, size, location, and density best describe a heterogeneous tumor

Fig. 2 Representative segmentation of tumor using ensemble semi-automated algorithm: **a** segmented lung tumor in right lung boundaries shown in green outline; **b** 3D view of the lung and segmented tumor. The distribution of tumor volume estimated for the test and retest cases is shown in (c)



lesion. It has also been shown that features are dependent within and across the categories. Our approach is to find representative features in each category so as to “best describe” the tumor in feature space. We assembled comprehensive descriptors to cover most categories.

Feature Concordance and Dynamic Range

The sets of informative features were selected using a three-step process. We first tested the consistency of extracted features between the test and retest experiments. For each

Table 1 Feature categories with the counts in each

Category	Description	Number of descriptors
C1: Tumor size	Size, volume descriptors	13
C2: Tumor shape (roundness)	Roundness/circularity descriptors	12
C3: Tumor location	Relative to pleural wall, boarder flags	14
C4: Pixel intensity histogram	Statistics on the intensity or attenuation values (in HU)	8
C5: Grayscale: runlength and co-occurrence	Run length and co-occurrence patterns	17
C6: Texture: laws features	Laws kernel (energy)	125
C7: Texture: wavelets	Wavelet kernels (entropy and energy)	30
	Total	219

image feature, the concordance correlation coefficient was used to quantify reproducibility between two scans performed on each patient. The Concordance correlation coefficient (CCC) is superior to the Pearson correlation coefficient for repeated experiments [24]. Suppose that $X_{1,k}$ and $X_{2,k}$ are the feature values for k th feature and that $(X_{1,k}(i), X_{2,k}(i))$ are independent and follow a bivariate distribution with means and covariance matrix: μ_{x1}, μ_{x2} , and $([\sigma_{x1,k}^2, \sigma_{x1,k, x2,k}], [\sigma_{x1,k, x2,k}^2, \sigma_{x2,k}^2])$, for the lesions measured in the i th test and retest experiment. Then the CCC [24] is defined as

$$\text{CCC} = \frac{2\sigma_{x1,k, x2,k}}{\left(\sigma_{x1,k}^2 + \sigma_{x2,k}^2 + [\mu_{x1,k} - \mu_{x2,k}]^2\right)} \quad (1)$$

The CCC, a standardization of mean squared deviation, evaluates a deviation from the identity line (i.e., the 45° line through the origin) and is typically used to measure concordance between test and retest. The CCC values ranges from 1 to -1 , implying perfect agreement between the repeated experiments to reverse agreement between them.

On a selected set of highly reproducible features, the next step was to select the features with a large inter-patient variability, using the “dynamic range”. The normalized dynamic range for a feature was defined as the inverse of the average difference between measurements divided by the entire range of observed values in the sample set as in [2]:

$$\text{DR}_k = \left(1 - \frac{1}{n} \sum_{i=1}^n \frac{|\text{Test}_k(i) - \text{Retest}_k(i)|}{\text{Max}_k - \text{Min}_k}\right) \quad (2)$$

where i refers to sample index, for the k th feature, $\text{Test}_k(i)$ or $\text{Retest}_k(i)$ are sample i 's, k th feature values for a test/retest population of n patient cases, the maximum (Max_k) and minimum (Min_k) are computed on the entire sample set. The dynamic range for feature k is, $\text{DR}_k \in [0, 1]$. Values close to 1 are preferred, and imply that the feature has a large biological range relative to reproducibility. Increasing the variation between the test–retest repeats will lead to a reduction in the DR value. Screening for a large DR will eliminate features that show greater variability in the repeat scans compared to the range of the coverage. The dynamic range measure will effectively address the “effect size” by identifying features with a lower value that are either not reproducible (relative to their range), or that are not highly variable across an entire sample set. This metric helps to give features a higher score that have relatively larger coverage (with respect to the repeatable differences), this does not intend to describe the dynamic range of the entire population.

The last step is to eliminate redundancies, based on the calculation of dependencies within the group. We computed the coefficient of determination (R^2_{Bet}) between the features

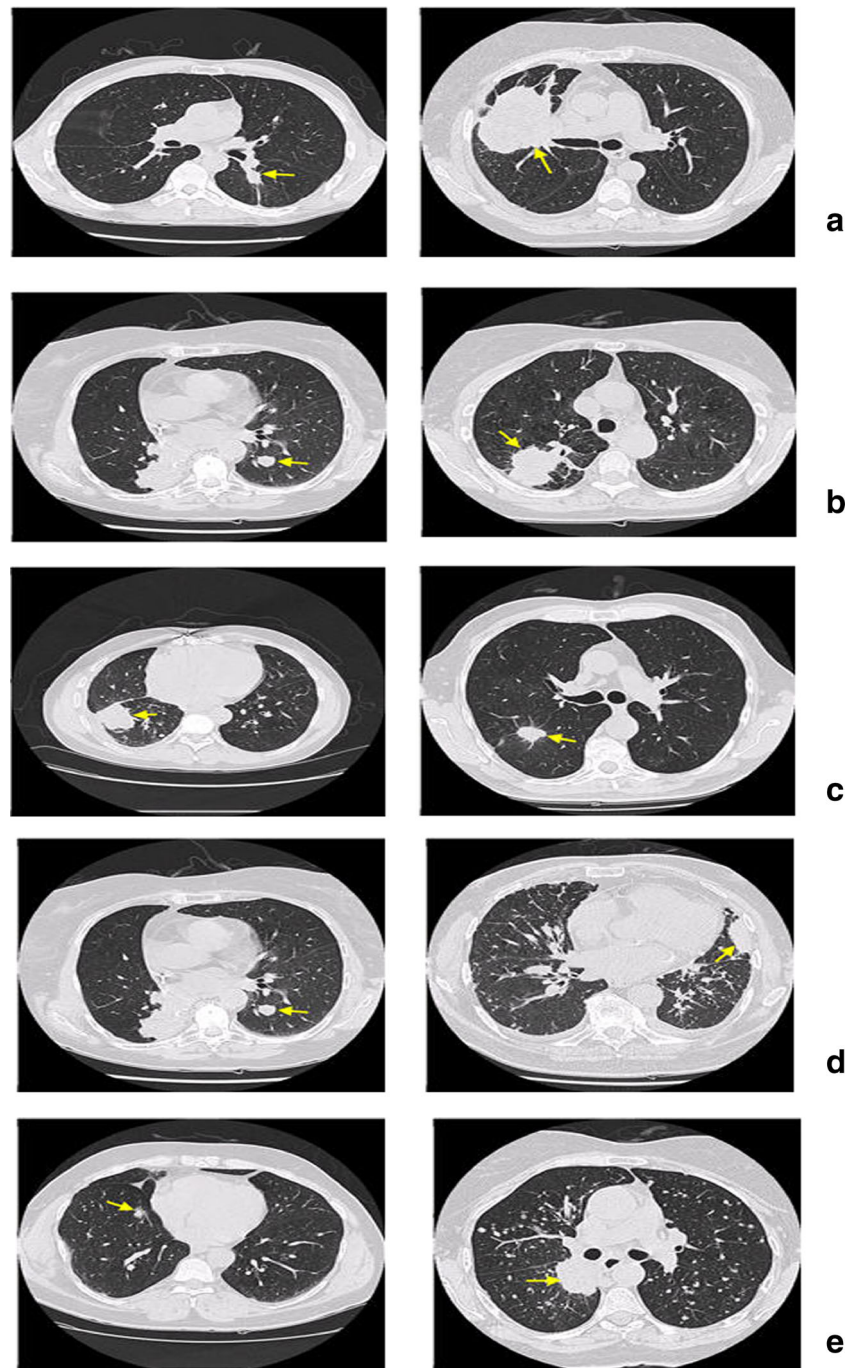
that passed a dynamic range threshold to quantify dependency. It is a linear estimate of the correlation or dependency and has a range of 0 to 1. Values close to 1 would mean that the data points are close to the fitted line (i.e., closer to dependency) [24, 25]. The coefficient of determination of simple regression is equal to the square of the Pearson correlation coefficient [25, 26]. We used different threshold values for R^2_{Bet} to consider each feature as linearly dependent on any other feature(s) in the list. The features that passed the cutoff limit were grouped and replaced by a representative from the group; the one having highest dynamic range. The purpose of this third filter was to eliminate redundancies (and not necessarily identify independence). A range of R^2_{Bet} thresholds were explored. We filtered the features category wise to find the representative in each category to ensure equal representation. Feature reduction taking all the categories of features together was also carried out, see ESM Table S.3.

Prognostic Label

We used our attending radiologist to categorize the lung CT dataset into two broad prognostic groups using quantitative metrics to score the tumor on a point scale. Both the RIDER test/retest data set and the independent NSCLC data set were scored separately using the same semantics. It has been reported that tumor size, differentiation, vascular invasion, margin status: negative vs. positive or close margins, have all been shown to have prognostic value [27–29]. We used five observable features: lobulated margin, size of the tumor lesion, spiculated margin, plueral wall attachment, and texture (e.g., ground glass opacity, GGO) as factors to scale the tumor into high risk to moderate risk individuals. The semantic scheme was first proposed and individual semantics prognostic ability of these metrics studied in NSCLC (Wang et al., in preparation). The observations were given a score of 1 to 5. In order to obtain a single risk score for a sample, the five semantic values were summed, averaged, and standardized to $[0, 1]$ to obtain a prognostic score. A normalized prognostic score over the median prognostic score was considered high risk (or poor prognosis). A sample below the median would indicate relatively lower to moderate risk. In the RIDER data set, two samples could not be scored reliably using the point scale metric due to diffused lesions and one sample was partly scored due to obscured margin. So, three samples were eliminated.

The two created categories were then used to find discriminatory markers between the poor to better prognosis groups. The table in ESM Section S.4 shows the score for individual samples for RIDER set. Figure 3 shows examples of CT images for extreme semantic scores.

Fig. 3 Radiological semantics on the RIDER data set, single central slice is shown. The *left panel cases* represent low scores (1 on 5 scale) and *right panel cases* represents high score (5 on 5 scale) values for **a** size, **b** lobulation, **c** spiculation, **d** pleural attachment, **e** texture



Feature Selection

The reproducibility of radiographic features obtained from CT scans of lung cancer was investigated to establish potential quantitative imaging biomarkers. Most of the features showed high reproducibility using an automated image analysis program with segmentation done by a single reader. Prior work has demonstrated the use of three types of features: univariate, bivariate, and volumetric for automatic and manually

segmented lung lesions, which seems to be limited in describing the complex nature of a tumor [15]. In the current study, the tumor is described by many features using different categories: size (volume, diameter, border length), shape (shape index, compactness, asymmetry), boundary region (border length, spiculation), relation to the lung field, image intensity (relative attenuation) based features (mean, standard deviation, average air space, deviation of airspace, energy, entropy, skewness, etc.), and transformed texture descriptors (wavelet

transform: entropy and energy and laws features). The consistency of this novel set of features in repeat scans (test, retest) was tested and they were filtered to find independent features. The stable, independent features provide an image feature set that may, for example, be used to predict prognosis.

One requirement for an image feature to be qualified as a response biomarker is that the change in its value between pre- and post-therapy scans must be significantly greater than the difference observed in the “Test–Retest” (or “15 minute coffee break”) measurements. In the present study, we can estimate the change of individual features that may be encountered post-therapy to be within the entire pre-therapy biological range. The changes between test and retest for the conventional radiological measurement (Longest diameter) and other related size measurements (Area and Volume) are within the accepted bounds with prior work [15], (see ESM Fig. 1 and ESM Table S. 5).

The ratio of the range to the inter-scan variability is a measure of “dynamic range” (see ESM Fig. 2A). Features showing high dynamic range were considered potentially more informative. The distribution of CCC values between test and retest, which as expected is skewed toward higher end values. It has high concordance between the test and retest cases (see ESM Fig. 2B). There is also a larger peak toward zero values. Investigating the peaks shows some of the laws and higher level wavelet features have low concordance between the repeated test and retest scans. It is hypothesized that the reimaging of the patients resulted in some change in texture (perhaps from small patient movements and segmentation differences). These Laws features compute energy after kernel convolution in a region. Small changes in sub regional textures would make these features vary, as they capture small localized changes. A similar analogy could be made for wavelet features for higher layer decompositions (or higher layers), where discordance can be seen.

Table 2 Features obtained after concordance correlation coefficient (test and retest) and dynamic range procedures

Category	Number of features		
A1. Concordance correlation coefficient (CCC_{TrT}):			
	$CCC_{\text{TrT}} \geq 0.95$	$CCC_{\text{TrT}} \geq 0.90$	$CCC_{\text{TrT}} \geq 0.85$
C1: Tumor size	11 (84.6 %)	12 (92.3 %)	12 (92.3 %)
C2: Tumor shape	3 (25 %)	6 (50 %)	7 (58.3 %)
C3: Tumor location	7 (50 %)	10 (71.4 %)	12 (85.7 %)
C4: Histogram	3 (37.5 %)	5 (62.5 %)	7 (87.5 %)
C5: Grayscale	5 (29.4 %)	6 (35.29)	6 (35.3 %)
C6: Laws	2 (1.6 %)	12 (9.6 %)	34 (27.2 %)
C7: Wavelets	14 (46.6 %)	15 (50 %)	15 (50 %)
All category	45 (20.5 %)	66 (30.1 %)	93 (42.5 %)
A2. Dynamic range (DR)			
	$DR \geq 0.95$	$DR \geq 0.90$	$DR \geq 0.85$
C1: Tumor size	11 (84.6 %)	13 (100 %)	13 (100 %)
C2: Tumor shape	3 (25 %)	8 (66.7 %)	12 (100 %)
C3: Tumor location	9 (64.3 %)	14 (100 %)	14 (100 %)
C4: Histogram	3 (37.5 %)	7 (87.5 %)	8 (100 %)
C5: Grayscale	6 (35.3 %)	12 (70.6 %)	17 (100 %)
C6: Laws	13 (10.4 %)	115 (92 %)	125 (100 %)
C7: Wavelets	14 (46.7 %)	20 (66.7 %)	30 (100 %)
All category	59 (26.9 %)	189 (86.3 %)	219 (100 %)
A3. Concordance correlation coefficient (CCC_{TrT}) and dynamic range (DR)			
	CCC_{TrT} and $DR \geq 0.95$	CCC_{TrT} and $DR \geq 0.90$	CCC_{TrT} and $DR \geq 0.85$
C1: Tumor size	11 (84.6 %)	12 (92.3 %)	12 (92.3 %)
C2: Tumor shape	3 (25 %)	6 (50 %)	7 (58.3 %)
C3: Tumor location	7 (50 %)	10 (71.4 %)	12 (85.7 %)
C4: Histogram	3 (37.5 %)	5 (62.5 %)	7 (87.5 %)
C5: Grayscale	5 (29.4 %)	6 (35.3 %)	6 (35.3 %)
C6: Laws	2 (1.6 %)	12 (9.6 %)	34 (27.2 %)
C7: Wavelets	14 (46.7 %)	15 (50 %)	15 (50 %)
All Category	45 (20.6 %)	66 (30.1 %)	93 (42.5 %)

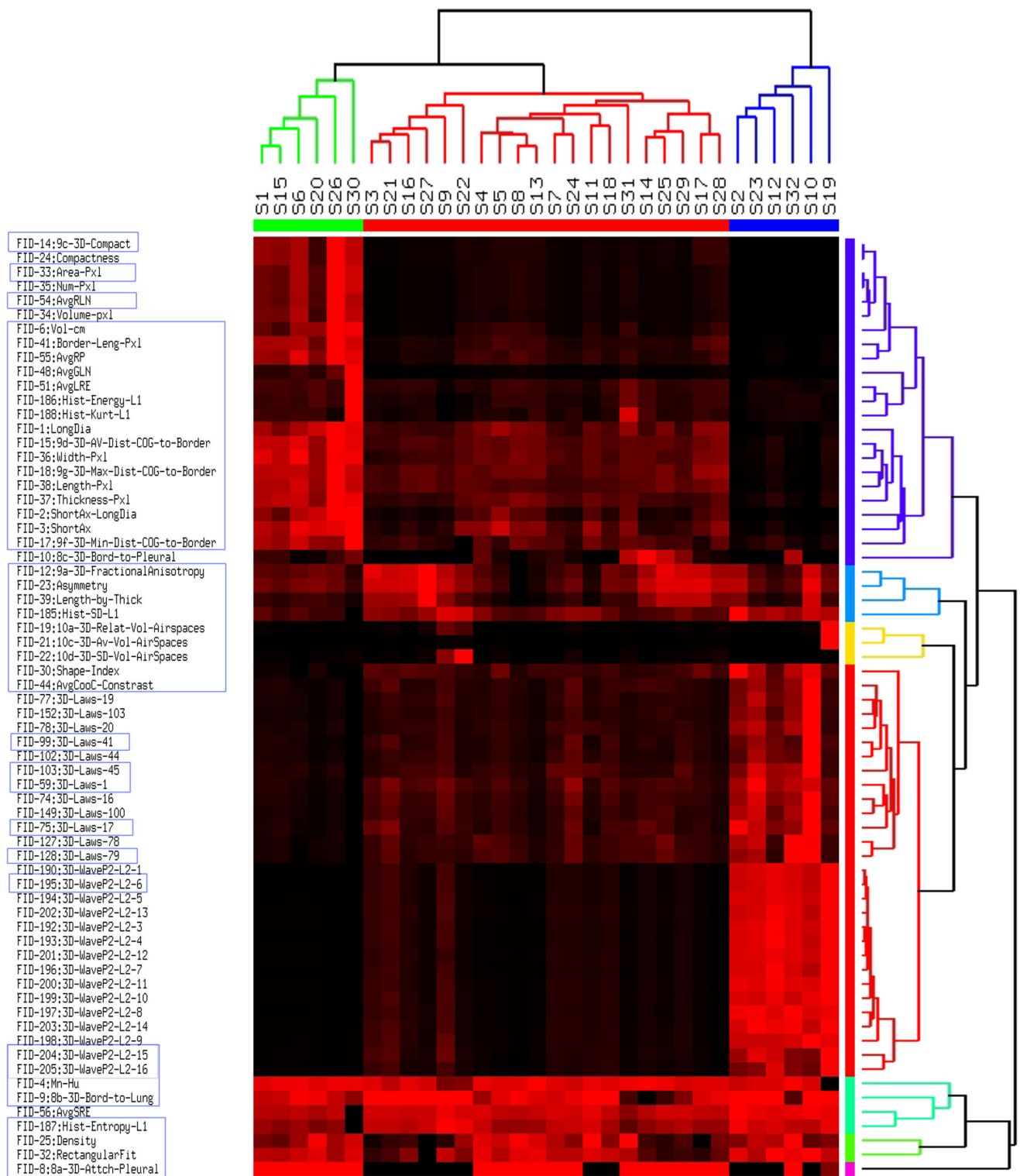


Fig. 4 Heatmap of the normalized feature values [0 1], obtained by the averaging features obtained in the test and retest experiment. The features (*in the rows*) reported are the ones with CCC_{TeT} and dynamic range (DR) ≥ 0.90 . The features that have been picked as representatives are those with R^2 value ≥ 0.95 , outlined with a *solid box*. The representative features were picked category wise, displayed across all categories. The hierarchical

clustering was stopped arbitrarily at six groups of features and three groups on the sample side, represented by the *multicolor bar*. The dendrograms on the top and sides show complete groupings with average linkage. The *color map* of the clustogram ranges from 0 to 1, values close to 1 have a *red shade* and values close to 0 have a *black shade*. The groups with *red shade* mean relatively high feature values

In prior work, Segal et al. have used a correlation coefficient threshold of 0.9 to distinguish highly correlated features [12]. In the current study, we used the coefficient of determination (R^2) between the features to find the dependency.

Feature Reduction

Feature reduction to select an informative feature set is an active research field; metrics that have been used in the past, include: the correlation coefficient, regression methods, and classification accuracy [30–32]. In our study, we propose finding a representative feature set that will eliminate redundancy in terms of information content, as complete independence may not be relevant for our study as texture information is subjective (and affected by sample issues, scanner setting, protocol followed, etc.). We used the coefficient of determination (R^2_{Bet}) between features to quantify dependency. Features were grouped based on R^2_{Bet} between them; in this subset one representative was picked that had the highest dynamic range. The procedure was repeated recursively to cover all the features resulting in a most representative group, done independently for each category.

The test, retest values were averaged before computing R^2_{Bet} . We set different limits to combine the features, R^2_{Bet} from 0.75 to 0.99. For higher thresholds R^2_{Bet} , less features will be grouped together resulting in a larger representative group. Setting the R^2_{Bet} to a lower limit will group more features together resulting in a smaller representative feature set (i.e., set of independent features). The combination of reproducibility, plus informative and independent are needed characteristics for an imaging biomarker. Table 2 shows a number of representative features obtained at different thresholds for concordance and dynamic range. As an example, the midlevel threshold CCC_{TreT} and $\text{DR} \geq 0.90$ yielded 66 features. In this subset, the representative features were found with $R^2_{\text{Bet}} \geq 0.95$ resulting in 42 features. Figure 4 shows a heatmap of coefficient of determination (R^2_{Bet}) between the features, for CCC_{TreT} and $\text{DR} \geq 0.90$. All the representative features with a cutoff of $R^2_{\text{Bet}} \geq 0.95$ are outlined. Tables 3 and 4 lists the feature counts and description category wise.

The image features left out of the representative reduced feature set could also be useful features. The image features are expected to capture different aspects of morphology and texture information. Due to the consistency in samples chosen and a limited sample population, the image features computed may show a higher level of dependency. It is hypothesized that the samples chosen as primary lung tumor may have a limited amount of texture or morphological changes.

Independent Data Set and Reader Variability

The lung tumor samples (NSCLC, 39 lung adenocarcinoma) were collected in an IRB-approved study. This study was

Table 3 List of representative reproducible Image features that was obtained by combining those with high R^2_{Bet} for with CCC and $\text{DR} \geq 0.90$. Number of features

CCC _{TreT} and DR ≥ 0.90	
3A	
Category	Combine features with $R^2_{\text{Bet}} (R^2_{\text{Bet}} \geq 0.95)$
C1: Tumor size	10 (76.92 %)
C2: Tumor shape	5 (41.67 %)
C3: Tumor location	9 (64.28 %)
C4: Histogram	5 (62.5 %)
C5: Grayscale	5 (29.41 %)
C6: Laws	5 (4 %)
C7: Wavelets	3 (10 %)
Total	42

confined to adenocarcinomas with mixed staging (TNM 1A/B, 15 samples; TNM 2A/B, 11 samples; TNM 3 A/B, 11 samples). There were three samples with missing pathology staging, which were replaced with clinical staging (all of them

Table 4 List of representative reproducible image features that was obtained by combining those with high R^2_{Bet} for with CCC and $\text{DR} \geq 0.90$. The feature description

Index	Representative feature ($\text{CCC}_{\text{TreT}} \geq 0.90$ and $\text{DR} \geq 0.90$) obtained at $R^2_{\text{Bet}} \geq 0.95$ (prefix represent feature index in the total list of 219).
1	<p>Category C1: representative features (10): F1:LongDia; F2:ShortAx-LongDia; F3:ShortAx; F6:Vol-cm; F33:Area-Pxl; F36:Width-Pxl; F37:Thickness-Pxl; F38:Length-Pxl; F39:Length-by-Thick; F41:Border-Leng-Pxl</p> <p>Category C2: representative features (5): F14:9c-3D-Compact; F23:Asymmetry; F25:Density; F30:Shape-Index; F32:RectangularFit</p> <p>Category C3: representative features (9): F8:8a-3D-Attch-Pleural; F9:8b-3D-Bord-to-Lung; F12:9a-3D-FractionalAnisotropy; F15:9d-3D-AV-Dist-COG-to-Border; F17:9f-3D-Min-Dist-COG-to-Border; F18:9g-3D-Max-Dist-COG-to-Border; F19:10a-3D-Relat-Vol-Airspaces; F21:10c-3D-Av-Vol-AirSpaces; F22:10d-3D-SD-Vol-AirSpaces</p> <p>Category C4: representative features (5): F4:Mn-Hu; F185:Hist-SD-L1; F186:Hist-Energy-L1; F187:Hist-Entropy-L1; F188:Hist-Kurt-L1</p> <p>Category C5: representative features (5): F44:AvgCooC-Constrast; F48:AvgGLN; F51:AvgLRE; F54:AvgRLN; F55:AvgRP</p> <p>Category C6: representative features (5): F59:3D-Laws-1; F75:3D-Laws-17; F99:3D-Laws-41; F103:3D-Laws-45; F128:3D-Laws-79</p> <p>Category C7: representative features (3): F195:3D-WaveP2-L2-6; F204:3D-WaveP2-L2-15; F205:3D-WaveP2-L2-16</p>

were 1A). The patients had a CT scan before surgical resection or treatment. The tumor samples were analyzed by a board-certified pathologist. The clinical and vital statistics were obtained from the Moffitt Cancer Registry (Tampa, FL). Vital statistics are typically updated on a yearly basis via patient survey. Table 5 shows the sample histology, which indicates a mixed population (tumor stage, gender, vital status, etc.). The mixture is maintained even after the median split based on the overall Radiological prognostic score.

These samples were independently scored on the five-point scale and the agreement between the two radiologists was measured by the Weighted Kappa index for an ordinal variable. The kappa value is typically interpreted as the following: <0: poor agreement; 0–0.2: slight agreement; 0.2–0.4: fair agreement; 0.4–0.6: moderate agreement; 0.6–0.8: substantial agreement; >0.8: almost perfect agreement [33]. The agreement for the five radiological parameters used in the paper had substantial to perfect agreement; the values were in the range of 0.75 to 1.0, details shown in Table 6.

Results and Discussion

As described in an earlier section, the CT data were segmented semi-automatically with user input to obtain the tumor boundaries [3, 14]. The distribution of tumor volumes across the sample set is diverse (see Fig. 2c). In the segmented regions of interest (ROI), 219 3D features were extracted; a comprehensive list is given in ESM Table S.1. The feature names were abbreviated to fit them in the table format, for example: “F78:3D-Laws-20” would mean, feature#78 (from the total of 219 features), it’s a 3D texture feature, computed by the “Laws” kernel of type 20. The kernel reference can be found in ESM Table S.2, which in this case is “E5 S5 W5 Layer 1”. All features can be decoded in this way.

Conventional Radiologist Measures In order to be comparable with previous work [15], we compared the concordance

Table 6 Agreement between two radiological readers in an independent test data (~39 NSCLC samples)

	Scoring characteristics	Kappa index (95 % confidence interval)
1	Size	1 (1–1)
2	Pleural attachment	0.80 (0.59–1)
3	Lobulation	0.75 (0.60–0.90)
4	Spiculation	0.91 (0.82–1)
5	Texture	0.79 (0.62–0.95)

correlation confidence limits for segmentation on three commonly used features: length, area, and volume (ESM Table 5). As before, we found high concordance across test–retest. The difference distribution between test and retest for the three features was consistent with previous findings (See ESM Fig. 1). As the tumor size increased, the difference between test and retest was reduced, as observed in previous analyses [15].

Concordance in the New Features The 219 extracted features were first compared using the CCC, which is a stringent measure of reproducibility. A CCC_{TrT} value ≥0.75 indicates that the data are of acceptable reproducibility. For our data set, we examined various thresholds with a preference for high stringency. These analyses identified 45, 66, and 93 features that had CCC_{TrT} values above thresholds of 0.95, 0.90, and 0.85, respectively (see Table 2).

Dynamic Range in New Features At a second level of analysis, the dynamic range was computed as described in Methods. Features with a dynamic range ≥0.95 have a biological range that is more than 20-fold greater than the test–retest difference. These analyses identified 59, 189, and 219 features above dynamic range thresholds of 0.95, 0.90, and 0.85, respectively. Applying both the filters, we identified features that passed the threshold set by CCC as well as the dynamic range filter. These two filtering procedures result in a set of features that is reproducible and has a large range compared to the variability between the test and retest experiments.

Table 5 Survival and histological details of the independent data samples

Details	Overall	Better prognosis	Poor prognosis
Sample size (histology)	39 (adenocarcinoma)	17	22
Survival (months): median (mean/standard deviation)	47 (39.4/ 23.4)	52 (41.1/23.4)	43.5 (38.1/23.9)
Vital stats (alive/dead)	20/19	11/6	9/13
Gender (male/female)	19/20	6/11	13/9
TNM (1A/1B/2A/2B): pathology/clinical	12/5/3/8	8/1/3/0	5/4/0/8
TNM (3A/3B/unknown): pathology/clinical	9/22	5/0	4/2

Redundancy Reduction It is known that agnostic features may be inter-dependent. To reduce redundancies, we used the coefficient of determination (R^2_{Bet}) between all possible pairwise combinations of features to quantify the level of similarity. The threshold level to flag features as linearly dependent is critical and subject to change with sample size, tumor shape, and texture. Using an R^2_{Bet} threshold of ≥ 0.95 to identify interdependence, there were 42 features that had CCC_{TreT} and DR values ≥ 0.90 . At a lower setting, of CCC_{TreT} and $\text{DR} \geq 0.85$, there were 93 features, with redundancy reduction ($R^2_{\text{Bet}} \geq 0.95$) we obtained 44 features.

Metric Distribution The ordered distribution plot for the dynamic ranges along with the distribution of concordance coefficients shown in ESM Fig. 2. The features' concordance and dynamic range criteria were computed after segmentation and the features obtained for discrete cutoffs are shown in ESM Table 3.

Representative Features Figure 4 shows a heatmap of coefficient of determination values (R^2_{Bet}) for the selected features at CCC_{TreT} and $\text{DR} \geq 0.90$, the highly inter-dependent features tend to be more closely grouped. Features were chosen with CCC_{TreT} and $\text{DR} \geq 0.90$ and $R^2_{\text{Bet}} \geq 0.95$. Most size features are highly reproducible and independent. Among all features in the size category, 69 % of them passed all three filters. At the other end of the spectrum, texture features showed high inter-dependency. Hence, only 4 % of the laws features and 10 % of wavelet features passed all three filters (see Tables 3 and 4). More inferences could be derived by comparing features across various categories. Table 3 lists the number of representative features obtained for different categories and Table 4 lists the feature details. ESM Table S.3 contains exhaustive feature tables and feature listings for other CCC, DR, and R^2_{Bet} threshold. In obtaining discriminatory biomarkers, one may have less interest in categorizing the features, and hence, the analysis was repeated without categorization, results are presented in ESM Table S.6A and B.

Practical Application of Repeatable Markers

Quantitative image features have been shown to predict prognosis in prior studies [34, 35]. Our objective in this work was to find reproducible, non-redundant, and high dynamic range image set of features that could be prognostic or response markers. The following is a practical example to illustrate the potential utility of these markers, for which reproducibility is a required trait to be useful as a prognostic predictor.

Radiological Prognosis Discrimination in the RIDER Dataset As an application, we used the features that passed the concordance and dynamic range filters to test their ability to discriminate groups based on their prognostic score, as described in Methods. We applied statistical tests (two sample t test) to find image features which are associated with the prognostic groups. The unadjusted p value for each feature was computed and the significant features were identified using the false discovery rate (FDR) method in order to guarantee the family wise error rate [36, 37], the feature details are reported in ESM Table 7. Because they were reproducible, the feature values for the test and retest were taken as independent observations. The features that had a FDR of ≤ 0.05 were considered to be prognostic discriminators.

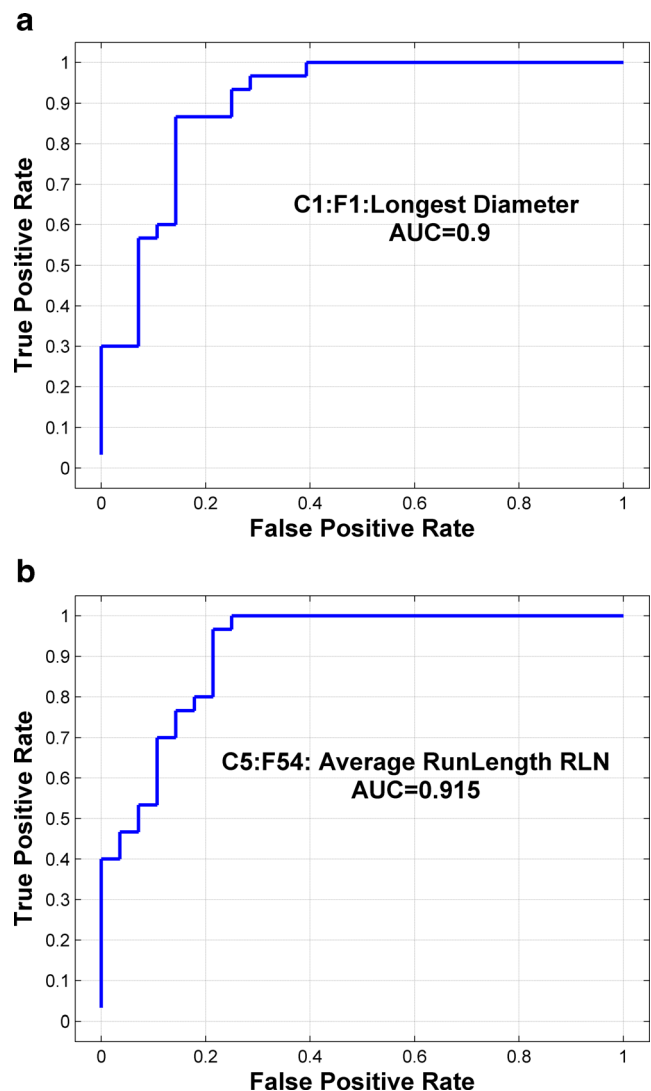


Fig. 5 Receiver operator characteristics graphs for discriminant image feature **a** size based: longest diameter (sensitivity, 0.87; specificity, 0.86), **b** texture: runlength feature (sensitivity, 0.43; specificity, 0.96)

Optimal Threshold Linear discriminant analysis was used to find a cutoff level for the significant features. Using the prognostic labels as ground truth sensitivity, specificity, and area under the curve (AUC) were computed. Some of the texture features (runlength, laws, and wavelets) had to be linearly scaled (by a factor of 1,000) before computation to avoid numerical errors.

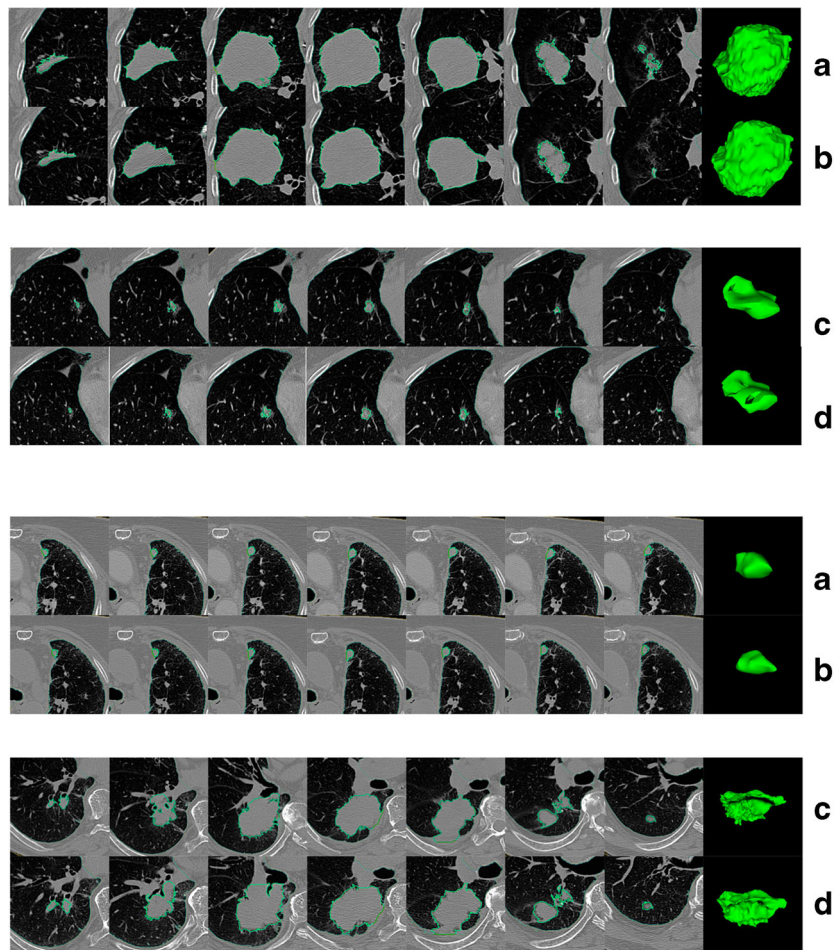
The ROC (receiver operator characteristics) for two features is shown in Fig. 5, the area under the curve is 90 % for the conventional size based feature (longest diameter) and 91 % for the chosen texture (run length) feature. The ESM Fig. 3 shows representative plots of prognostic features (size and texture category) expressed according to the prognostic score for RIDER data. We arbitrarily assigned scores below the median as “good” (in green) and those above the median as “bad” (in red), in a relative sense.

The example in Fig. 6, shows two extreme sample cases with representative slices for run length and a laws kernel feature. The laws 1 (E5 E5 E5 Layer 1) is an edge detecting kernel of length 5, applied across

all directions (x , y , and z) and normalized based on the size of the tumor. It is expected to have an inverse relation to tumor size and expected to measure tumors edges. Figure 7 shows the relationship of the prognostic feature to the conventional measures. It shows in the lower range it tracks the size measurement but deviates as the feature value increases. We hypothesize that these texture features capture more information than traditional size based measurements.

The sample set considered was diverse with more large tumors than small ones, and hence most size-based features were near the top of the prognostic predictor list. Size is a well-known prognostic feature for many tumors [38]. In addition, we also observed that a large number of texture features (histogram, laws, and wavelets) were prognostic. Notably, texture-, size-, and shape-based descriptors showed equal prognostic value (see ESM Table 7). Table 7 reports sensitivity, specificity, and area under the curve for the discriminant features. It is noteworthy that some of the texture features perform better than the conventional size-based features.

Fig. 6 Segmented region for representative slices from top to bottom of the tumor, for samples with extreme values of texture features (also identified to be predictive of prognostic score). The samples in case 1: *A, B* and *C, D* had the highest and lowest average run length (F54: Avg.Run.L) measure for test/retest. The feature value for images in *A, B* was 22,153.85 and 221,262.36. While *C* and *D* was 688.56 and 700.96, respectively. The samples in case 2: *A, B* and *C, D* had the highest and lowest laws kernel measure (F59: laws—1: E5 E5 E5 Layer 1) for test/retest. The feature value for images in *A, B* was 0.07404 and 0.07507. While *C, D* was 0.00864 and 0.00658, respectively.



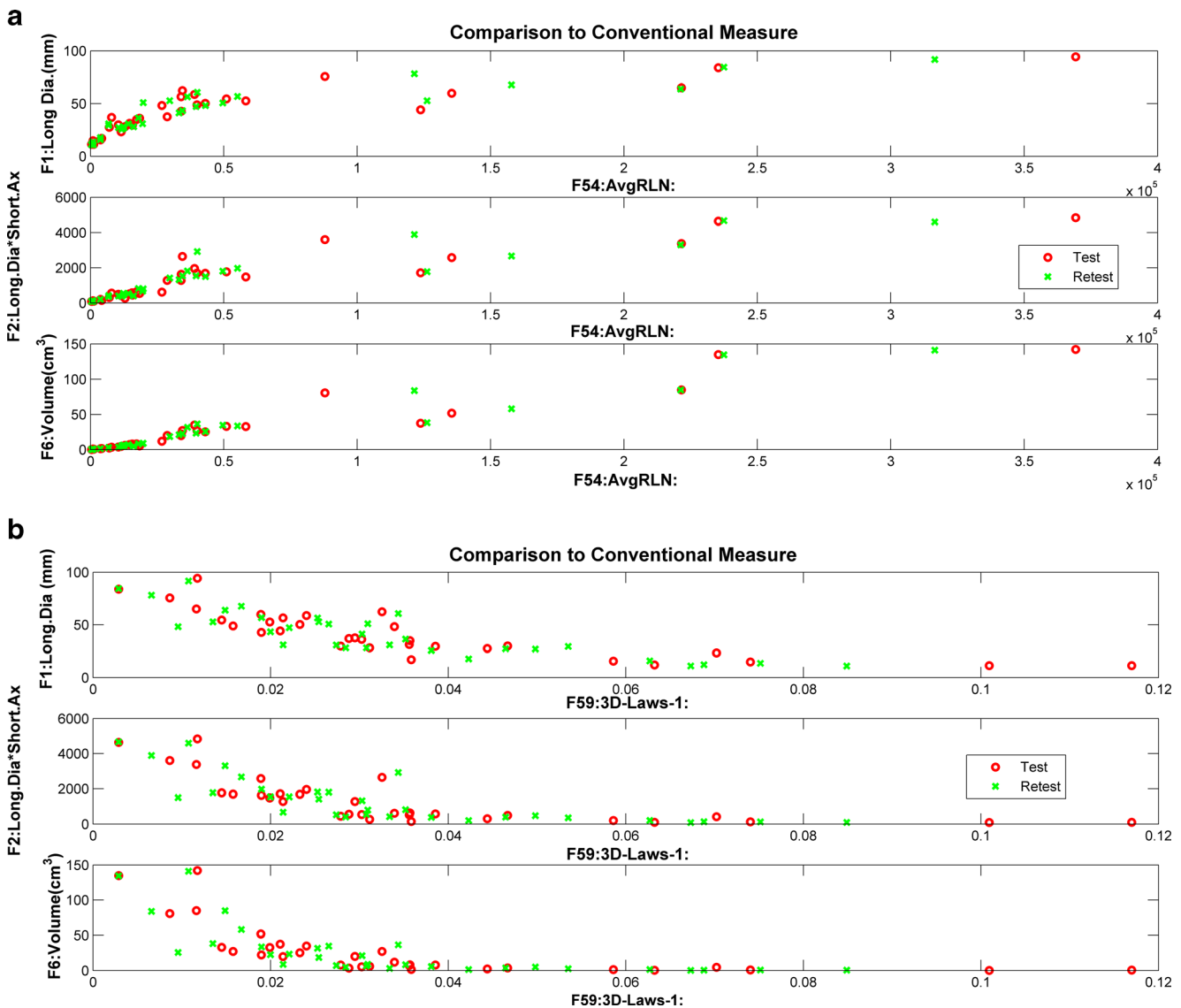


Fig. 7 Prognostic texture features relationship to conventional measures (longest axis, short axis \times longest axis and volume). **a** run length features; **b** laws kernel features

Radiological Prognosis Discrimination Using Independent Data

We obtained 39 patient CT scans of adenocarcinomas with mixed stage; the lesion was marked by a trained radiologist and segmented using a semi-automated tool. From the region of interest, 219 quantitative image features were computed. The image data was log transformed to minimize the effect of the outliers. Based on this test/retest study, we selected 66 image features that are known to be reproducible with CCC and $DR \geq 0.90$ (Table 2 and details in ESM Table S.3.a, b). The redundancy between the features in the 39 independent lung tumor data was removed by computing the R^2 . The features were grouped by a cutoff ($R^2 \geq 0.95$) and the

group of dependent features replaced by a representative that had the largest dynamic range. This reduced the number of features to 41. The resident radiologist scored these images on five known radiological semantics, using the five-point scale described in the previous section. Figure 8 shows two patient samples with better radiological prognosis ($<$ median) and two samples with poor prognosis (\geq median). It is interesting to note that there is a 9-month difference in median survival between samples in the better prognostic group arm compared to those in poor prognostic group arm.

These 41 image features were then used to find discriminators for the radiological prognostic score by computing the false discover rate on the p values (two tail t test), independently repeated in each category. For the discriminant features (FRD \leq

Table 7 Sensitivity, specificity, and area under the curve computed for the significant features that discriminate prognostic score. An optimal threshold was obtained using linear discriminant function

	Feature	Sensitivity ^a	Specificity ^a	AUC ^a
C1: Tumor size				
1	F1:LongDia	0.87	0.86	0.90
2	F36:Width-Pxl	0.77	0.89	0.91
3	F37:Thickness-Pxl	0.70	0.82	0.90
4	F3:ShortAx	0.77	0.79	0.87
5	F2:ShortAx-LongDia	0.77	0.86	0.90
6	F41:Border-Leng-Pxl	0.57	0.89	0.90
7	F38:Length-Pxl	0.63	0.79	0.85
8	F6:Vol-cm	0.57	0.86	0.91
9	F33:Area-Pxl	0.8	0.65	0.79
C2: Tumor size				
1	F30:Shape-Index	0.90	0.71	0.88
2	F14:9c-3D-Compact	0.50	1.00	0.90
C3: Tumor location				
1	F15:9d-3D-AV-Dist-COG-to-Border	0.73	0.93	0.92
2	F17:9f-3D-Min-Dist-COG-to-Border	0.67	0.93	0.89
3	F18:9g-3D-Max-Dist-COG-to-Border	0.67	0.79	0.86
4	F8:8a-3D-Attch-Pleural	0.87	0.61	0.53
5	F9:8b-3D-Bord-to-Lung	0.57	0.86	0.72
C4: Pixel intensity histogram ^b				
1	F185:Hist-SD-L1	0.8	0.79	0.86
2	F187:Hist-Entropy-L1	0.77	0.82	0.83
3	F186:Hist-Energy-L1	0.70	0.86	0.84
4	F188:Hist-Kurt-L1	0.53	0.93	0.82
5	F4:Mn-Hu	1.00	0.43	0.86
C5: Grayscale: run length and co-occurrence ^b				
1	F55:AvgRP	0.57	0.93	0.90
2	F44:AvgCooC-Constrast	0.83	0.50	0.86
3	F54:AvgRLN	0.43	0.97	0.92
4	F48:AvgGLN	0.43	1.00	0.93
5	F51:AvgLRE	0.67	0.827	0.78
C6: Texture: laws features ^b				
1	F59:3D-Laws-1	0.90	0.61	0.92
2	F99:3D-Laws-41	0.90	0.68	0.91
3	F75:3D-Laws-17	0.93	0.57	0.88
4	F103:3D-Laws-45	0.87	0.5	0.80
5	F128:3D-Laws-79	0.83	0.5	0.82
C7: Wavelet features ^b				
1	F195:3D-WaveP2-L2-6	1.00	0.46	0.88
2	F204:3D-WaveP2-L2-15	1.00	0.46	0.82
3	F205:3D-WaveP2-L2-16	1.00	0.46	0.82

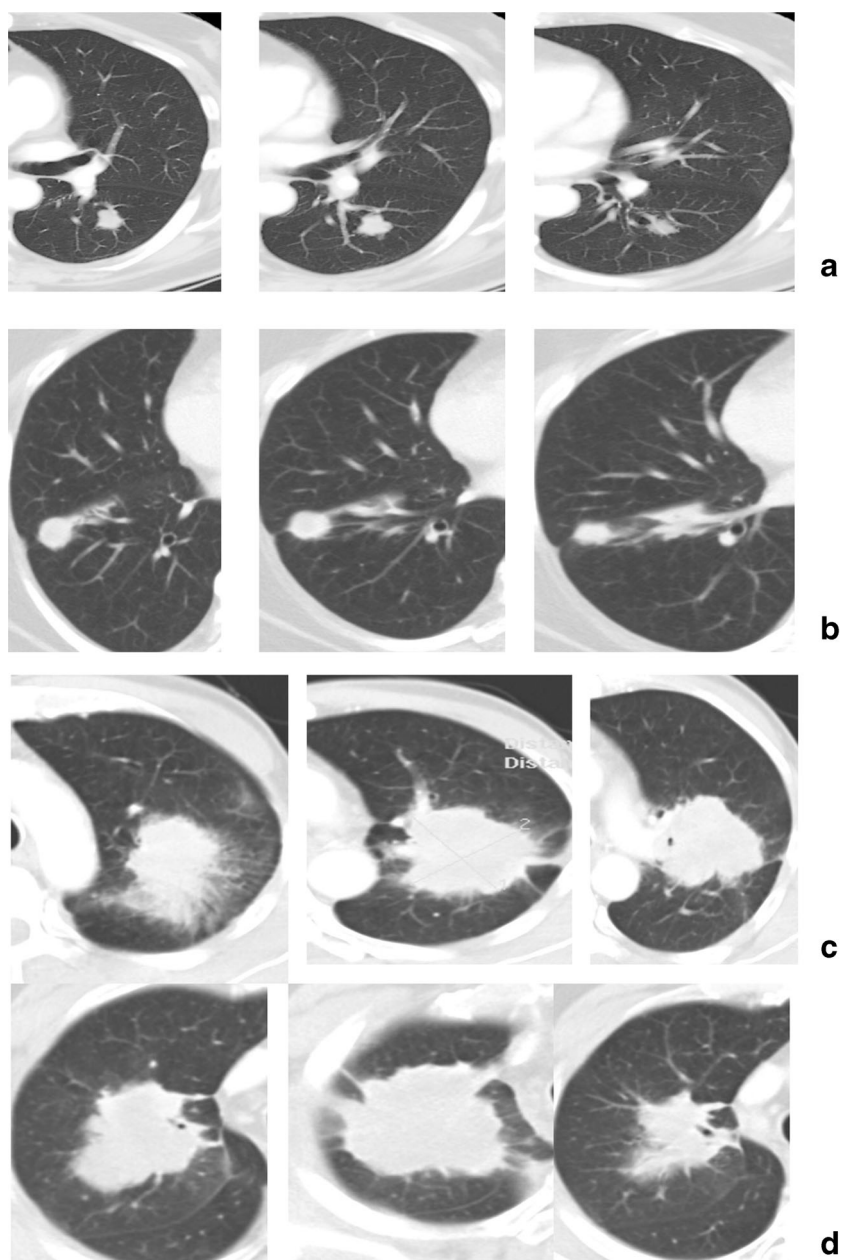
^a Rounded to two decimal precision

^b Features were linearly scaled to avoid numerical errors in AUC computations

0.05), the ability to predict the radiological scores was evaluated by computing the sensitivity, specificity, and area under the curve using linear discriminant analysis. Figure 9 and Table 8 show the discriminant features and their individual predictability. It is interesting to note conventional size-based measures (longest diameter, volume) show an AUC of over 0.8. The pixel

histogram and runlength categories of features have better AUC than size distribution, over 0.84. This indicates more than the size of the tumor; the texture and pixel distributions carry prognostic information. Some of the features are picked up as prognostic in the RIDER data set. It is also interesting to note that laws and wavelets were eliminated due to broader line

Fig. 8 Three representative slices of patient CT scans selected from 39 adenocarcinoma cases with better radiological prognostic score (**a** and **b**, score of 0.44 and 0.48, respectively) and poor radiological prognostic score (**c** and **d**, score of 0.8 and 0.92, respectively)



statistical significance and additional constraint to reduce false discovery [37]. We hypothesize that the runlength and other texture feature describe the heterogeneity of the tumor.

Conclusions

The current study demonstrates that the test–retest reproducibility of most CT features of primary lung cancer is high when using an image analysis approach with semi-automated segmentation. Across all patients, the biological ranges for the majority of individual features were very high. We

propose that the features with lowest inter-scan variance relative to the largest biological range (i.e., dynamic range, DR) should be explored as potentially the most informative for use as imaging biomarkers. Additionally, a co-variance matrix of features identified several redundancies in the feature set that could be combined into a single variable. Combining inter-scan variance, biological range, and co-variance, we have reduced the total number of features from 219 to a most informative set of 42 features identified at a setting of CCC_{TrT} and $DR \geq 0.9$ ($R^2_{\text{Bet}} \geq 0.95$). These reproducible and representative features show high ability to discriminate tumors

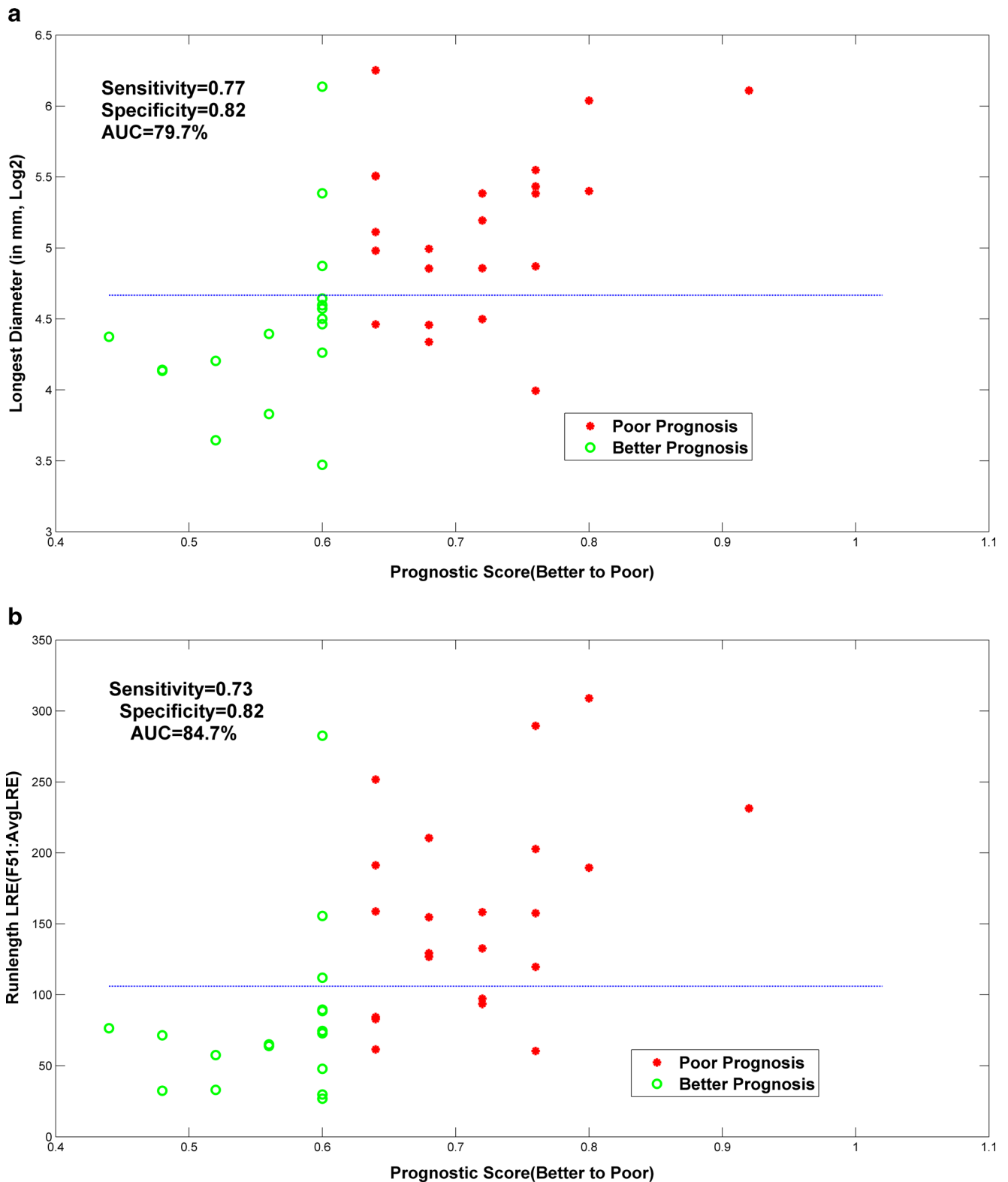


Fig. 9 Image features that are discriminating the radiological prognostic score in the independent NSCLC data set: **a** longest diameter (sensitivity, 0.77; specificity, 0.82); **b** run-length feature (sensitivity, 0.73; specificity, 0.82).

based on prognostic labels. For 69 % of size-based features, 62 % of histogram features and 29 % run length features, it was possible to discriminate between

tumors with low and high prognostic scores with considerable accuracy (area under the curve over 90 %). In the independent data set, a similar trend can be seen.

Table 8 Image features discrimination ability of the prognostic score in the independent data (39 adenocarcinoma). The representative features (CCC and DR \geq 0.90, $R^2\geq$ 0.95) was obtained on the independent data set filtered for false discover rate \leq 0.05. Sensitivity, specificity, and area under curve (AUC) computed for linear discriminant functions

Feature	Sensitivity ^a	Specificity ^a	AUC ^a
C1: Tumor size			
1 F1:LongDia	0.77	0.82	0.80
2 F2:ShortAx-LongDia	0.73	0.71	0.79
3 F3:ShortAx	0.64	0.71	0.76
4 F6:Vol-cm	0.73	0.82	0.80
C2: Tumor size			
1 F30:Shape index	0.73	0.65	0.75
C4: Pixel intensity histogram ^b			
1 F185:Hist-SD-L1	0.86	0.35	0.69
2 F187:Hist-Entropy-L1	0.82	0.72	0.79
3 F186:Hist-Energy-L1	0.77	0.82	0.85
4 F188:Hist-Kurt-L1	0.73	0.76	0.85
5 F4:Mn-Hu	0.77	0.65	0.76
C5: Grayscale: run length and co-occurrence ^b			
1 F44:AvgCooC-Constrast	0.77	0.65	0.76
2 F48:AvgGLN	0.77	0.77	0.85
3 F51:AvgLRE	0.73	0.82	0.85

^a rounded to two decimal precision

^b Features were linearly scaled to avoid numerical errors in AUC computations

The size-based features discriminate the prognostic score with an AUC of 0.8 while the texture-based features (runlength, histogram based) resulted in an AUC of 0.84. The texture-based features provided additional information about the tumor and increased the ability to be prognostic. Multivariate methods with other clinical factors could improve predictability. The current findings shed light on the selection of reproducible, informative, and independent features that are candidate imaging biomarkers to predict prognosis and assess (or predict) better therapy planning and response.

References

- Nguyen T, Rangayyan R: Shape analysis of breast masses in mammograms via the fractal dimension. *Conf Proc IEEE Eng Med Biol Soc* 3:3210–3213, 2005
- Schuster DP: The opportunities and challenges of developing imaging biomarkers to study lung function and disease. *Am J Respir Crit Care Med* 176(3):224–230, 2007
- Suzuki C, Jacobsson H, Hatschek T, et al: Radiologic measurements of tumor response to treatment: practical approaches and limitations. *Radiographics* 28(2):329–344, 2008
- Tuma RS: Sometimes size doesn't matter: reevaluating RECIST and tumor response rate endpoints. *J Natl Cancer Inst* 98(18):1272–1274, 2006
- Ganeshan B, Abaleke S, Young RC, et al: Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* 10:137–143, 2010
- Way TW, Sahiner B, Chan HP, et al: Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Med Phys* 36(7):3086–3098, 2009
- Samala R, Moreno W, You Y, et al: A novel approach to nodule feature optimization on thin section thoracic CT. *Acad Radiol* 16(4):418–427, 2009
- Lee MC, Boroczky L, Sungur-Stasik K, et al: Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. *Artif Intell Med* 50(1):43–53, 2010
- Zhu Y, Tan Y, Hua Y, et al: Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. *J Digit Imaging* 23(1):51–65, 2010
- Al-Kadi O, Watson D: Texture analysis of aggressive and nonaggressive lung tumor CE CT images. *IEEE Trans Biomed Eng* 55(7):1822–1830, 2008
- Kido S, Kuriyama K, Higashiyama M, et al: Fractal analysis of internal and peripheral textures of small peripheral bronchogenic carcinomas in thin-section computed tomography: comparison of bronchioloalveolar cell carcinomas with nonbronchioloalveolar cell carcinomas. *J Comput Assist Tomogr* 27(1):56–61, 2003
- Segal E, Sirlin CB, Ooi C, et al: Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol* 25(6):675–680, 2007
- Buckler AJ, Mozley PD, Schwartz L, et al: Volumetric CT in lung cancer: an example for the qualification of imaging as a biomarker. *Acad Radiol* 17(1):107–115, 2010
- America RSoN: Quantitative imaging biomarker alliance for volumetric CT image analysis: roadmap for a staged validation plan, 2010
- Zhao B, James LP, Moskowitz CS, et al: Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* 252(1):263–272, 2009
- RIDER. The Reference Image Database to Evaluate Therapy Response. Available at: <https://wiki.cancerimagingarchive.net/display/Public/RIDER+Collections;jsessionid=C78203F71E49C7EA3A43E0D213CE555>. Accessed 24 Jun 2014
- Gu Y, Kumar V, Hall LO, et al: Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. *Pattern Recogn* 46(3):692–702, 2013
- NBIA. National Biomedical Imaging Archive. Available at: <https://imaging.nci.nih.gov/ncia>. Accessed 30 June 2014
- Definiens. Definiens AG, Munchen, Germany. Available at: <http://www.definiens.com/product-services/definiens-xd-product-suite.html>. Accessed 30 June 2014
- Athellogou M, Schmidt G, Schaepe A, et al: Cognition network technology—a novel multimodal image analysis technique for automatic identification and quantification of biological image contents. In: Shorte SL, Frischknecht F Eds. *Book cognition network technology—a novel multimodal image analysis technique for automatic identification and quantification of biological image contents*. Springer-Verlag, New York City, 2007, pp 407–422
- Baatz M, Zimmermann J, Blackmore CG: Automated analysis and detailed quantification of biomedical images using Definiens Cognition Network Technology. *Comb Chem High Throughput Screen* 12(9):908–916, 2009
- Bendtsen C, Kietzmann M, Korn R, Mozley P, Schmidt G, Binnig G: X-ray computed tomography: semiautomated volumetric analysis of

- late-stage lung tumors as a basis for response assessments. *Int J Biomed Imaging*, vol 2011, 2011
23. Basu S, Hall LO, Goldgof DB, et al: Developing a classifier model for lung tumors in ct-scan images. *IEEE Intl Conf on Systems, Man and Cybernetics, (SMC 2011)*, Anchorage, Alaska, 2011
 24. Lin LI-K: A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:13, 1989
 25. RGD Steel JT: *Principles and procedures of statistics*. McGraw-Hill, New York, 1960
 26. Colin C, Frank AW, Gramaji H, et al: An R-square measured of goodness of fit for some common nonlinear regression models. *J Econ* 77(2):1790–1792, 1997
 27. Aoki T, Tomoda Y, Watanabe H, et al: Peripheral lung adenocarcinoma: correlation of thin-section CT findings with histologic prognostic factors and survival. *Radiology* 220(3):803–809, 2001
 28. Takashima S MY, Hasegawa M, Saito A, Haniuda M, Kadoya M. High-resolution CT features: prognostic significance in peripheral lung adenocarcinoma with bronchioloalveolar carcinoma components. *Respiration: Int Rev Thorac Dis* 70(1), 2003
 29. Subramanian J, Simon R: Gene expression-based signature in lung cancer: ready for clinical use? *JNCI* 102(7):464–474, 2010
 30. Jain AK, Zongker D: Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal* 19(2):153–158, 1997
 31. Pudil P, Novovičová J, Kittler J: Floating search methods in feature selection. *Pattern Recogn Lett* 15:1119–1125, 1994
 32. Saeys Y, Inza I: A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507–2517, 2007
 33. Landis JR, Koch G: The measurement of observer agreement for categorical data. *Biometrics* 33:159–174, 1977
 34. Ganeshan B, Panayiotou E, Burnand K, et al: Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. *Eur Radiol* 22(4):796–802, 2012
 35. Yanagawa M, Tanaka Y, Kusumoto M, et al: Automated assessment of malignant degree of small peripheral adenocarcinomas using volumetric CT data: correlation with pathologic prognostic factors. *Lung Cancer* 70(3):286–294, 2010
 36. John S: A direct approach to false discovery rate. *J R Stat Soc B* 64(3):479–498, 2002
 37. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 57(1):289–300, 1995
 38. Zhao B, Oxnard G, Moskowitz CS, et al: A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development. *Clin Cancer Res* 16(18): 4647–4653, 2010