Translational Oncology
www.transonc.com

Volume 7 Number 1 February 2014 pp. 72–87 72

Reproducibility and Prognosis of Quantitative Features Extracted from CT Images[1,2]

Yoganand Balagurunathan*, Yuhua Gu*, Hua Wang*,†, Virendra Kumar*, Olya Grove*, Sam Hawkins‡, Jongphil Kim§, Dmitry B. Goldgof‡, Lawrence O. Hall‡, Robert A. Gatenby¶ and Robert J. Gillies*,¶

*Department of Cancer Imaging and Metabolism, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL; †Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China; ‡Department of Computer Science and Engineering, University of South Florida, Tampa, FL; §Department of Biostatistics, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL; ¶Department of Radiology, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL

Abstract

We study the reproducibility of quantitative imaging features that are used to describe tumor shape, size, and texture from computed tomography (CT) scans of non–small cell lung cancer (NSCLC). CT images are dependent on various scanning factors. We focus on characterizing image features that are reproducible in the presence of variations due to patient factors and segmentation methods. Thirty-two NSCLC nonenhanced lung CT scans were obtained from the Reference Image Database to Evaluate Response data set. The tumors were segmented using both manual (radiologist expert) and ensemble (software-automated) methods. A set of features (219 three-dimensional and 110 two-dimensional) was computed, and quantitative image features were statistically filtered to identify a subset of reproducible and nonredundant features. The variability in the repeated experiment was measured by the test-retest concordance correlation coefficient (CCC_{TreT}). The natural range in the features, normalized to variance, was measured by the dynamic range (DR). In this study, there were 29 features across segmentation methods found with $CCC_{TreT} \geq 0.9$ and DR ≥ 0.9 and $R^2_{Bet} \geq 0.95$. These reproducible features were tested for predicting radiologist prognostic score; some texture features (run-length and Laws kernels) had an area under the curve of 0.9. The representative features were tested for their prognostic capabilities using an independent NSCLC data set (59 lung adenocarcinomas), where one of the texture features, run-length gray-level nonuniformity, was statistically significant in separating the samples into survival groups ($P \leq .046$).

Translational Oncology (2014) 7, 72–87

Address all correspondence to: Robert J. Gillies, PhD, Vice-chair (Radiology) and Director, Experimental Imaging Program, H Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, SRB-4 (Imaging), Tampa, FL 33612. E-mail: robert.gillies@moffitt.org

[1]We acknowledge research support to the work from the following grants: National Institutes of Health (NIH) U01CA143062, Radiomics of NSCLC, and Florida Biomedical Research Programs, King Team Science grant 2KT01, Radiomics of Lung Cancer Screening.
[2]This article refers to supplementary materials, which are designated by Tables W1 to W7 and Figure W1 and are available online at www.transonc.com.
Received 13 December 2013; Revised 27 January 2014; Accepted 11 February 2014

DOI 10.1593/tlo.13844

Introduction

Lung cancer has been one of the most common forms of cancer and a leading cause of death in the United States and most of the world. Although a small percentage (about 15%) of the cases are curable when detected early, the 5-year survival rate remains low at about 16.6% [1,2]. The disease has been very visible with the publication of the association of increased risk with tobacco usage [3]. Early detection of lung cancer through screening has resulted in adopting lung computed tomography (CT) as the standard modality for early detection of the disease. In the last decade, enormous advancements in genomics technologies have contributed to the understanding of the biology of lung cancer. Despite these advancements, the survival rate of patients with lung cancer has not changed significantly. There has been great improvement in the imaging technologies in the last decade, especially in CT that has seen increases in the number of detector rows, decreased rotation time, sophisticated radiation dosing methods, helical scanning, and better reconstruction methods. All these improvements have led to better capturing of the anatomic structure for the regions of interest (ROIs).

The tumor regions in CT images have traditionally been described qualitatively to measure size and degree of spread, organ invasion, as well as aggressiveness [4,5]. Such features are typically described and quantified subjectively (i.e., "mildly irregular," "highly spiculated," and "moderate necrosis"). Currently, the standard method to measure tumor response to therapy using CT remains the Response Evaluation Criteria in Solid Tumors (RECIST) that is a unidirectional linear measurement to estimate tumor diameter [6]. The RECIST criteria assume a spherical tumor with linear growth uniformly in all directions. A simple linear measurement allows the practicing clinician to make an easy assessment; however, the linear growth assumption is often violated. This is reflected in high interobserver variability in finding the lesion boundary between radiologic experts due to nonuniform growth (and other anatomic structural factors), resulting in RECIST measurement variability [7]. Typically, clinical response criteria involve using RECIST linear measurements to discretely categorize patients into "complete response," "partial response," "stable disease," and "progressive disease." This categorization is a "coarse" metric, which is seen as a loose bound to categorize growth: partial response is defined by 30% linear sum reduction. Although these metrics are considered satisfactory under ideal conditions, reduction in tumor size often does not reflect clinicopathologic response [8,9].

The CT tumor measurement "bias" and "variance" are critical issues with widespread influence especially in clinical trials that study the effectiveness of drugs in patient treatments [10]. Hence, 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 response biomarkers. However, to be useful as biomarkers, features must be reproducible, quantifiable, and objective [11]. National Cancer Institute (NCI) funded RIDER project [12] has been a tremendous resource and enabled this repeatability study.

In prior work, we have demonstrated that a semiautomated (ensemble) multiseed point segmentation can reliably generate segmented volumes, as defined by the Dice similarity index (SI). The SI between machine-segmented lesions was >0.93, whereas the SI for manual segmentation was 0.73 across a test set of 129 patients [11]. Hence, lesions in the current study were segmented by both manual and ensemble methods, and 219 3D (and 110 2D) features were extracted from these segmented volumes. Although we began with a large feature set compared to prior conventional radiologic analyses [13], it is expected that there may be redundancy in these features due to various limitations on

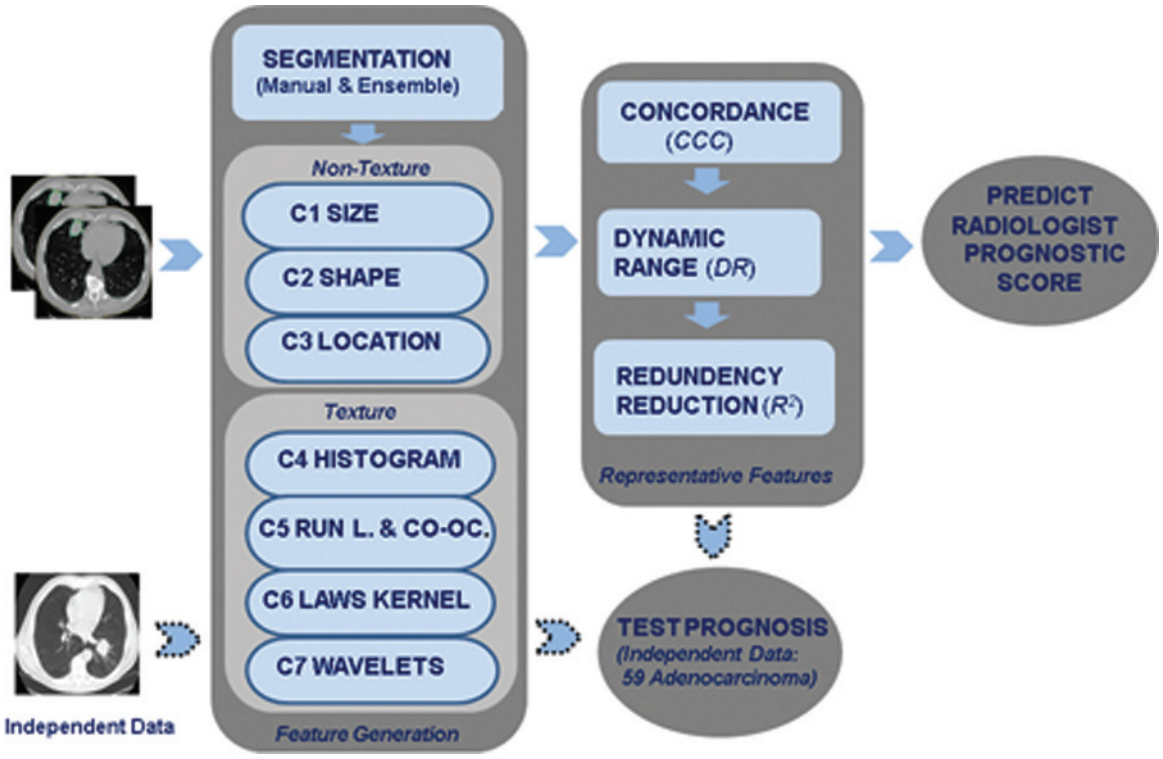

Figure 1. Process flow for finding representative image features using test-retest data set and testing for prognosis using independent data set.

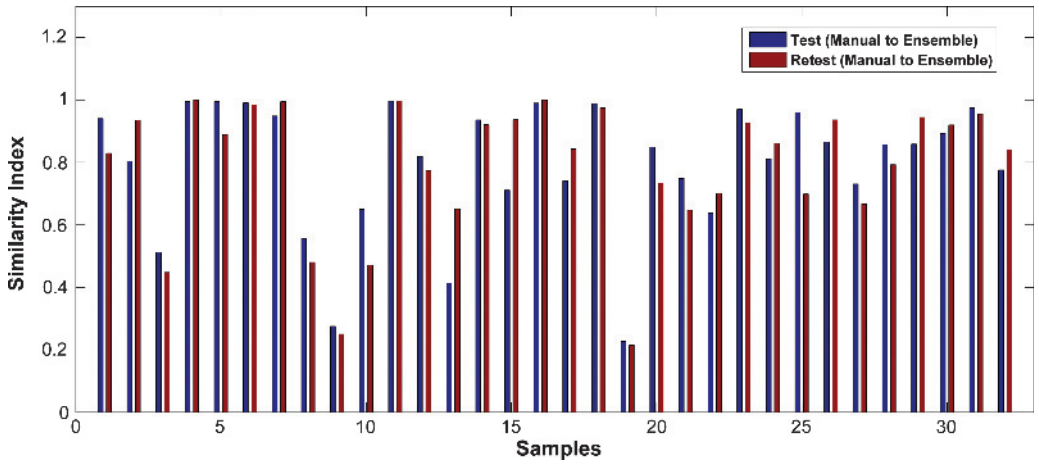

Figure 2. SI of manual to ensemble segmentation. The average SI is 79% and 78% for test and retest data sets.

the sample size, texture, and consistency in the population. Thus, to reduce the dimensionality of this agnostic feature set, we first filtered features on the basis of their reproducibility, i.e., those with the highest intrafeature concordance correlation coefficients (CCC_{TreT}) between the repeats. As a second filter, we used dynamic range (DR) on the basis of the interpatient variability normalized by test-retest difference. Finally, redundancy was assessed by computing an interfeature coefficient of determination (R^2_{Bet}) between all possible pairs of features, and a representative feature set was found by combing dependent groups to form a reduced set. These features were then tested for their ability to predict a radiologist-created prognostic score. In an effort to create a prognostic reproducible biomarker, 59 independent non–small cell lung cancer (NSCLC) samples of adenocarcinoma subtype were curated, and 219 3D-image features were extracted. A subset of repeatable features was obtained on the basis of the current study, and these tested for prognostic ability. Figure 1 pictorially illustrates the process flow.

Materials and Methods

The CT of the thorax for 32 patients in the test/retest (baseline and follow-up) was acquired within 15 minutes of each other, using the same CT scanner and imaging protocol [13]. The patients in the study were asked to get off the scan table between the repeats. Unenhanced thoracic CT images were acquired using Light Speed (GE Medical Systems, Milwaukee, WI) or VCT scanner with 16/64 detectors

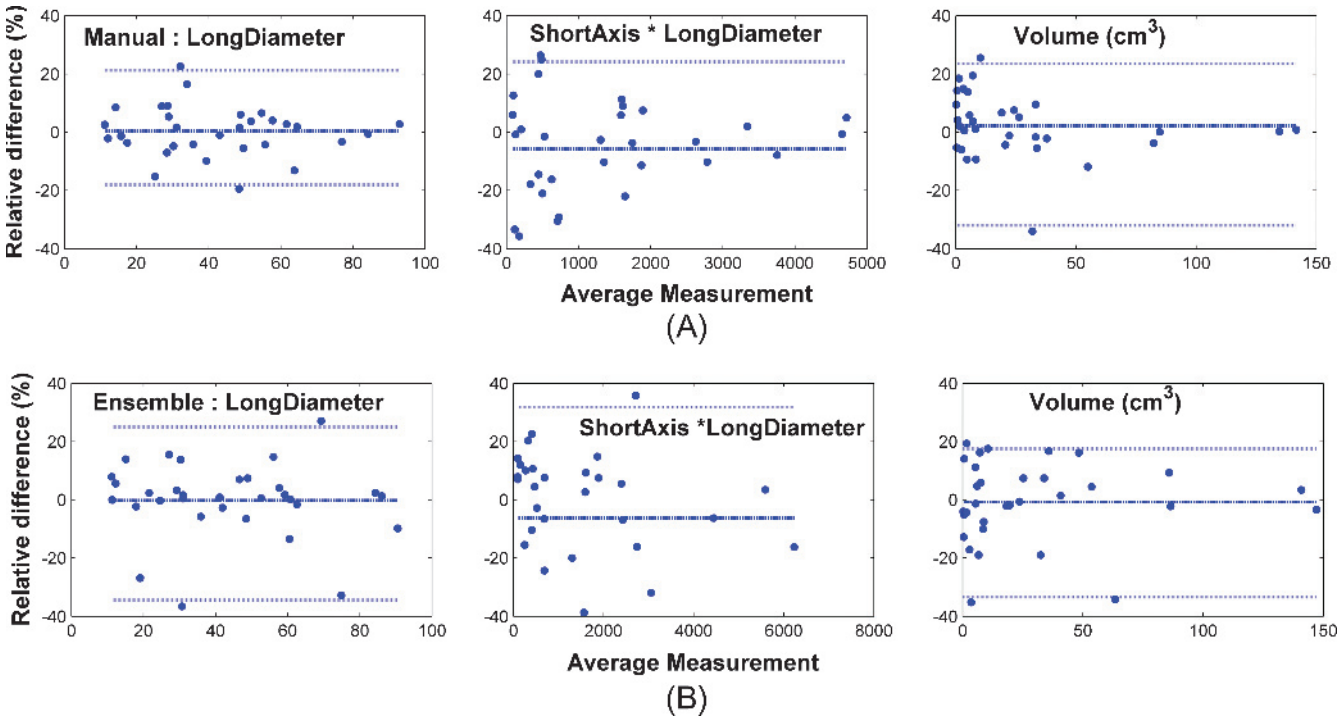

Figure 3. Bland-Altman plot for test and retest to data is shown for conventional univariate, bivariate, and volume features in (A) manual and (B) ensemble segmentations.

Table 1. CCC for Conventional Size-Based Measures (RECIST and WHO) Are Reported with 95% Confidence Limit Estimated by Bootstrap Resampling (n = 5000).

Feature Name	Manual CCC	Ensemble CCC	Manual and Ensemble CCC
Univariate: (RECIST)			
Longest axis (cm)	0.9963 (0.967, 0.9998)	0.9920 (0.9509, 0.9997)	0.9943 (0.9523, 0.9997)
Short axis (cm)	0.9951 (0.9047, 0.9952)	0.9868 (0.9078, 0.9995)	0.9878 (0.9084, 0.9995)
Bivariate: (WHO)			
(Longest axis *Short axis)	0.9855 (0.9384, 0.9997)	0.9757 (0.8342, 0.9988)	0.9807 (0.8848, 0.9992)
Volumetric:			
Volume (cm^3)	0.9934 (0.9571, 0.9997)	0.9913 (0.9552, 0.9997)	0.9924 (0.9574, 0.9998)

in 28/4 patients, respectively. The scan voltage was set at 120 kVp, pitch 1.375:1 (0.984 for VCT) with rotation time of 0.5 second, and image slices at 1.25 mm were reconstructed using a lung convolution kernel without overlap. The CT scans were acquired from patients (mean age = 62.1 years; range = 29-82 years) with NSCLC. 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 CT lung cancer images were downloaded from the Cancer Imaging Archive (http://cancerimagingarchive.net). The images are contained in "RIDER Lung CT," under the "Collections" sections.

Segmentation of Tumors

We used Definiens Developer XD© (Munich, Germany) as the image analysis platform to perform tumor segmentation and feature extraction. Definiens is based on the Cognition Network Technology [14,15] that allows the development and execution of image analysis applications. Here, the Lung Tumor Analysis application was used [16]. Lung Tumor Analysis is a semiautomated three-dimensional "Click&Grow" approach for segmentation of tumors under the guidance of an operator. To perform the seed-based segmentation of a target lesion, the latter has to be completely within a lung-image object. In cases where a medical expert 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.

The manual segmentation process required many human interactions to get the "correct" segmentation boundaries. In our study, we used a trained radiologist to assist in the manual segmentation process. Consequently, we developed an automatic single-click ensemble segmentation (SCES) algorithm [11]. In brief, the SCES algorithm uses the initial seed point to automatically generate multiple seed points with region growing. It makes use of the "Click&Grow" algorithm by using a manually selected initial seed point to define a small circumscribed area within the tumor boundary, within which multiple seed points are automatically generated. An ensemble segmentation is obtained from the multiple regions that were grown from these multiple seed points. In this algorithm, an *ensemble segmentation* refers to a set of different input segmentations (multiple runs using the same segmentation technique but different initializations) that are combined to generate a "consensus" segmentation.

Once the segmentation of all target lesions was deemed sufficiently accurate, statistics for each lesion, such as volume, center of gravity, and average density, all readily available as object features within the commercial cognitive network language, were extracted. Figure 2 shows a comparison of the segmentation masks between segmentations and repeats.

In total, 64 lesions were segmented, i.e., 2 per patient, and quantitative values of image features were extracted from each segmented volume. Figure 3 shows a Bland-Altman plot of conventional size measurements (long diameter, longest diameter*short axis, and volume), estimated after manual and ensemble segmentations (Table 1). The volume distribution showed a diverse population with relatively high variability in midsized to smaller sized tumors. Half the samples had small volume ≤ 4 cm^3 tumors, whereas the rest of them are larger (the largest group close to 120 cm^3). Figure 4 shows an example of a patient tumor delineated with manual segmentation for test/retest scans.

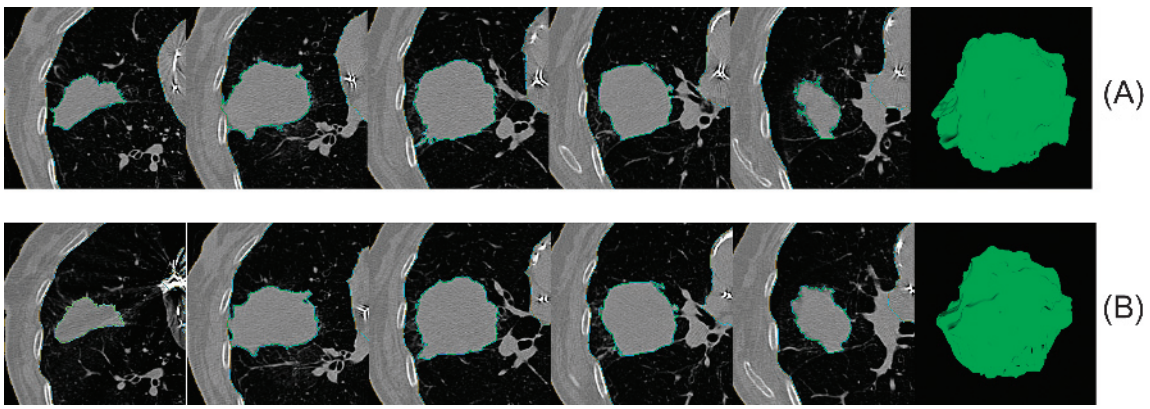

Figure 4. Example of slice and 3D region for a sample segmented using manual method for test and retest of the patient (top and bottom rows). The subset of slices was arbitrarily selected by increasing slice numbers (matched for test and retest) to approximately cover the entire volume.

Table 2. Histologic Sample Details on Both the Arms Split at the Median Value of Feature (Run-Length GLN).

Adenocarcinoma (Dichotomized: Run-Length GLN feature)			
	All	< Median	≥ Median
No. of samples	59	29	30
Median survival (mo):	27	30	21
Mean and SD of survival	26.19 (15.97)	28.86 (15.11)	23.6 (16.59)
Vital statistics (alive/dead)	30/29	19/10	11/19
Gender (male/female)	28/31	15/14	13/17
TNM score (1A/1B)	13/8	10/2	3/6
TNM score (2A/2B)	2/9	2/3	0/6
TNM score (3A/3B/unknown)	9/3/14	7/1/5	2/2/9

Image features. We extracted several types of image features, both in 2D and 3D with most of our analysis directed on using 3D features. In this work, we present 3D features that are broadly divided into the following two classes: nontexture and texture features. Each of these classes can be subdivided into several categories on the basis of their functional description, which also facilitate analysis and presentation. Specifically, nontexture class includes tumor size, tumor shape, and tumor location categories, and texture class includes pixel histogram, run length, co-occurrence, Laws, and wavelet feature categories. We note that texture features have been shown to be good descriptors of the tumor and some have shown to be useful in survival prediction [17]. In this study, we have used 219 three-dimensional and 111 two-dimensional image features. Most of these were implemented within the Definiens platform [18], whereas some were computed by implementing the algorithms in C/C++ (former Bell Labs USA) and MATLAB (Mathworks, Natick, MA). All the features were obtained from the ROI (i.e., after the segmentation). The 2D features are expected to have lower variability in measurement due to limited span of the ROI; in this repeated experiment, matching slices between test/retest has been a challenge, which adds to measurement noise. In this study, our focus has been geared toward 3D features, which provide better description of the tumor volume region.

Texture descriptors provide measure of properties such as smoothness, coarseness, and regularity, as no standard description exists. Typically, the following three principal approaches are used to describe texture: statistical, structural, and spectral [18]. Our features cover all the categories and use most popular approaches for texture computation. A brief description of each feature category is provided, and additional information is provided in the Supplemental section.

Nontexture features. Tumor size, shape, and location descriptors make up our nontexture features.

- *Tumor size* category contains features that can be broadly categorized as univariate (longest diameter, short axis, width, and other size measurements), bivariate (area, length by width, length by thickness, and other size measurements), and volume measurements both in pixel units and in native resolution (centimeters).
- *Tumor shape* category feature measures circularity of the tumor in various forms: compactness, largest elliptical fit in the tumor region, asymmetry, density, and compactness. Asymmetry is a measure of variance from round shape (disproportional length). It is computed as a ratio of smallest and largest Eigen values of the tumor. Density describes spatial distribution of the pixels with respect to cubical object. The density is higher when the volume of the object follows a cubical shape (lower when it is like a filament). Compactness measures the cuboid occupied by the object computed as a ratio of the first three Eigen values to the number of voxels in the *tumor*. The "MacSpic" feature measures the number of countable spiculations in the tumor.
- *Tumor location* category feature measures tumor position with respect to anatomic structure of the lung. The Attached to Pleural is a binary flag that tells if the tumor (in 3D) is attached to lung wall. The Main direction feature is a measure of the angle between the best-fit line on the centers of gravity for each 2D slice to the z-axis.

Texture features. In CT images, texture is typically attributed to gray-level changes seen by a radiologist. In other types of imaging, in addition to gray-level changes, texture is well characterized in transformed domain (kernel based or functional mapping). These features have been shown to be useful in medicine [19]. We categorized

Table 3. Features Obtained after CCC (Test and Retest) and DR Filtering Procedures for Manual and Ensemble Segmentations Using (A) 3D Features and (B) Outlier-Removed 3D Features.

Feature Category	No. of Features: CCC_{TreT} and DR ≥ 0.90		
	Manual Segmentation	Ensemble Segmentation	Common (Manual and Ensemble)
(A) CCC_{TreT} and DR: 3D features			
C1: Tumor size	12 (92.31%)	11 (84.62%)	11 (84.62%)
C2: Tumor shape	6 (50%)	4 (33.33%)	4 (33.33%)
C3: Tumor location	11 (78.57%)	8 (57.14%)	8 (57.14%)
C4: Histogram	6 (75%)	3 (37.5%)	3 (37.5%)
C5: Run length and co-occurrence	6 (35.29%)	6 (35.29%)	6 (35.29%)
C6: Laws	16 (12.8%)	74 (59.2%)	16 (12.8%)
C7: Wavelets	15 (50%)	0 (0%)	0 (0%)
All Categories	72 (32.88%)	106 (48.4%)	48 (21.92%)
(B) CCC_{TreT} and DR: Filtered 3D features			
C1: Tumor size	12 (92.31%)	12 (92.31%)	8 (61.54%)
C2: Tumor shape	6 (50%)	8 (66.67%)	7 (58.33%)
C3: Tumor location	11 (78.57%)	6 (42.86%)	5 (35.71%)
C4: Histogram	3 (37.5%)	1 (12.5%)	1 (12.5%)
C5: Run length and co-occurrence	4 (23.53%)	5 (29.41%)	1 (5.88%)
C6: Laws	13 (10.4%)	68 (54.4%)	8 (6.4%)
C7: Wavelets	19 (63.33%)	23 (76.67%)	19 (63.33%)
All Categories	68 (31.05%)	123 (56.16%)	49 (22.37%)

histogram, run length, co-occurrences, Laws kernel, and wavelet-based features as textures.

- *Pixel intensity histogram features* are computed on the pixel intensity (in Hounsfield units or HU) for the region (voxel) of interest. First and higher-order statistics, entropy, and energy on the tumor volumes are reported as features.
- *Run-length and co-occurrence features* may find some correlation to radiologist-visualized texture. The *run length* is defined as a measure of contiguous gray levels along a specific orientation. Fine textures tend to have short run length, whereas coarser texture will have longer run lengths with similar gray level. These features capture coarseness in 3D-image structure and have been found useful in a number of texture analyses [20,21]. If *R (k,p)* is the run-length matrix n_1 by n_2, at gray-level k, then the number of such lengths equals p, along an orientation, in the volume (x,y,z). One useful measure of run length in this study has been the measure of nonuniformity ($RunL_{\mathrm{GLN}}$) that measures extent of smoothness or similarity in the image.

$$RunL_{GLN} = \left(\frac{1}{n} \sum_{k=1}^{n_1} \left(\sum_{p=1}^{n_2} R(k,p) \right)^2 \right) \quad (1)$$

We compute 11 different run-length metrics, each of which has a property to capture gray-level variations in the tumor. The co-occurrence matrix contains the frequency of one gray-level intensity appearing in a specified spatial relationship with another gray-level intensity in a given

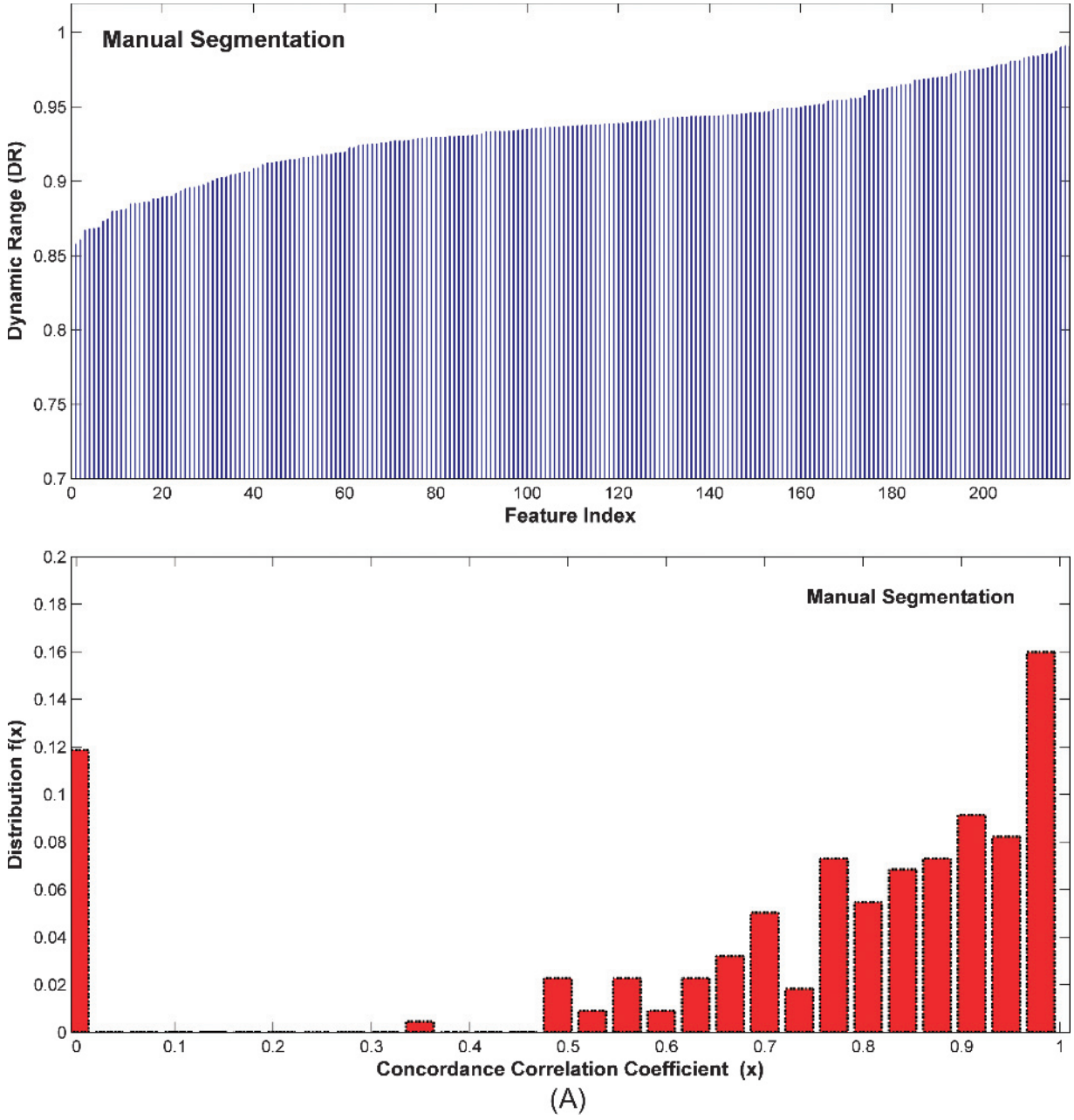

Figure 5. Distribution of DR and CCC computed on test/retest data in (A) manual and (B) ensemble segmentations.

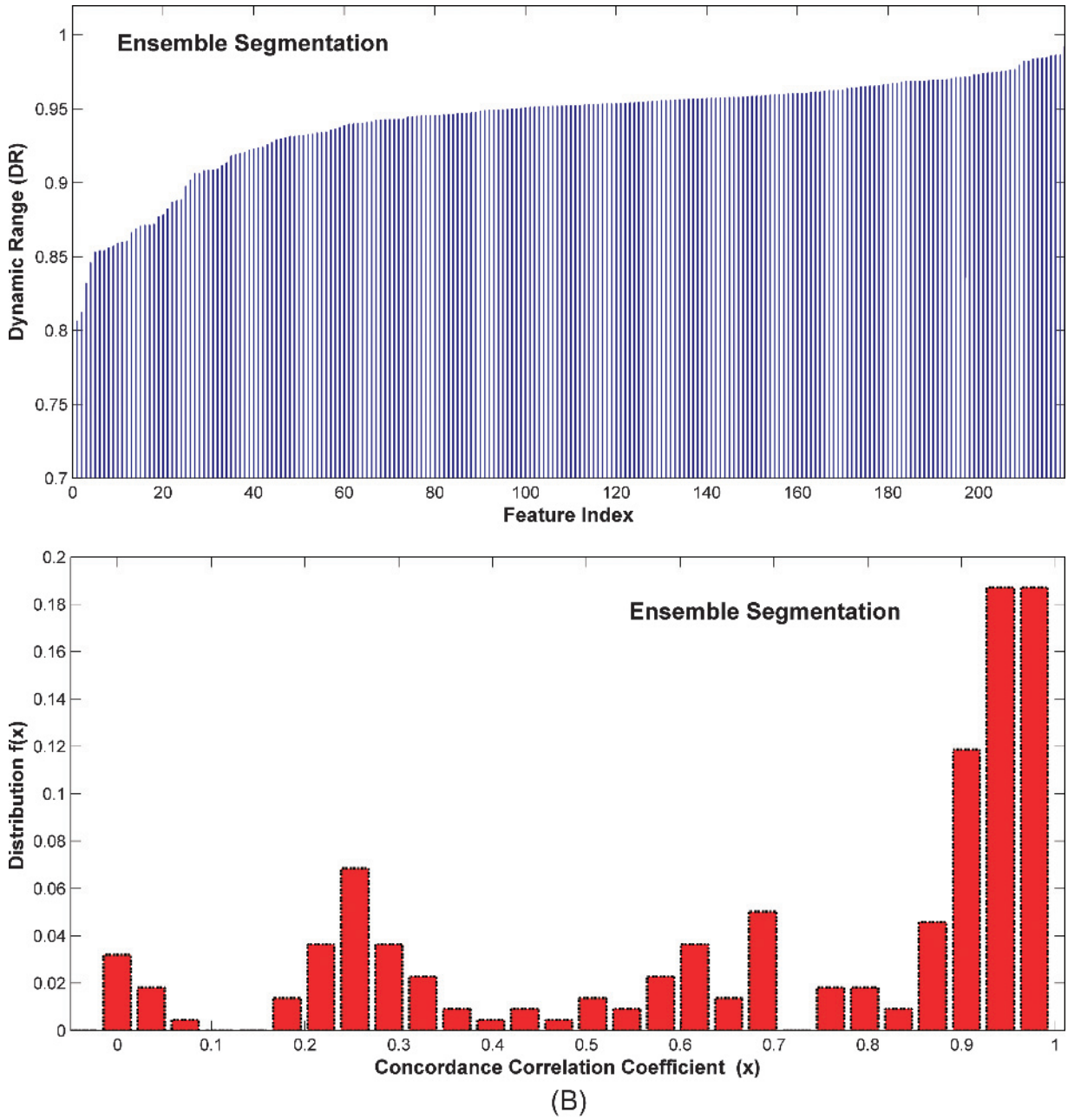

Figure 5. (*continued*).

range. The co-occurrence matrix is first constructed, and different formulations of the values are then calculated where measurements include contrast, energy, homogeneity, entropy, mean, and maximum probability [22].

- *Laws features* are constructed from a set of five one-dimensional kernels each designed to reflect a different type of structure in the image [23]. The kernel has the ability to enhance certain regions of the image. About 125 features are computed on different sets of kernel and orientation.
- *Wavelet features* are kernel-based functions that decompose the image (3D) into orthogonal components. We used *Daubechies* (*Coiflet*) wavelets in this study [24,25]. Statistics on the decomposition have been widely used in image texture identification. In this study, we calculate two metrics (energy and entropy) along each direction of the 3D-image volume with two levels of decomposition yielding 30 features.

Two-dimensional features. Traditionally, image processing has been carried out on 2D gray-scale images; however, most proposed features could be extended to 3D. The advantage of 2D features includes easier interpretability and visualization. In this study, the tumors were delineated in 3D using the segmentation methods described in the previous section. Slices (2D) in the ROI were matched between test/retest scans by a trained radiologist. The identification criteria were based on the anatomic structure of the lung. On a 2D slice, we extracted 110 2D features that describe the shape, size, and texture of the lesion. Additionally, 2D measurements are an estimate of the true tumor size that is

Table 4. Representative Image Features That Were Obtained by Combining Those with High R^2_{Bet}, with CCC and DR ≥ 0.90.

(A) No. of Representative Features (CCC_{TreT} and DR ≥ 0.90; Combine R^2_{Bet} ≥ 0.95)

Category	Redundancy Reduction Categorywise		Redundancy Reduction across All Categories	
	All Samples	After Outlier Removal	All Samples	After Outlier Removal
C1: Tumor size	9	8	7	9
C2: Tumor shape	3	5	2	3
C3: Tumor location	6	5	6	5
C4: Histogram	3	1	2	1
C5: Gray scale	4	1	2	1
C6: Laws	4	6	4	6
C7: Wavelets	0	5	0	5
Total	29	31	23	30

(B) (i) Representative Feature (CCC_{TreT} ≥ 0.90 and DR ≥ 0.90) Obtained at R^2_{Bet} ≥ 0.95 (All Samples)

(F No.: Suffix No. represents feature index in the total list of 219.)
Category C1: Representative features (9):
F1:LongDia; F2:ShortAx-LongDia; F3:ShortAx; F6:Vol-cm; F33:Area-Pxl; F36:Width-Pxl; F37:Thickness-Pxl; F38:Length-Pxl; and F41:Border-Leng-Pxl
Category C2: Representative features (3):
F14:9c-3D-Compact; F25:Density; and F30:Shape-Index
Category C3: Representative features (6):
F8:8a-3D-Attch-Pleural; F9:8b-3D-Bord-to-Lung; F15:9d-3D-AV-Dist-COG-to-Border; F16:9e-3D-SD-Dist-COG-to-Border; F17:9f-3D-Min-Dist-COG-to-Border; and F19:10a-3D-Relat-Vol-Airspaces
Category C4: Representative features (3):
F4:Mn-Hu; F186:Hist-Energy-L1; and F187:Hist-Entropy-L1
Category C5: Representative features (4):
F44:AvgCooC-Constrast; F48:AvgGLN; F51:AvgLRE; and F54:AvgRLN
Category C6: Representative features (4):
F67:3D-Laws-9; F74:3D-Laws-16; F103:3D-Laws-45; and F128:3D-Laws-79
Category C7: None

(B) (ii) Removing Outliers

Representative feature (CCC_{TreT} ≥ 0.90 and DR ≥ 0.90) obtained at R^2_{Bet} ≥ 0.95 (prefix represents feature index in the total list of 219).
Category C1: Representative features (8):
F1:LongDia; F2:ShortAx-LongDia; F3:ShortAx; F6:Vol-cm; F34:Volume-pxl; F36:Width-Pxl; F37:Thickness-Pxl; and F38:Length-Pxl
Category C2: Representative features (5):
F13:9b-3D-Circularity; F14:9c-3D-Compact; F25:Density; F30:Shape-Index; and F32:RectangularFit
Category C3: Representative features (5):
F9:8b-3D-Bord-to-Lung; F12:9a-3D-FractionalAnisotropy; F16:9e-3D-SD-Dist-COG-to-Border; F17:9f-3D-Min-Dist-COG-to-Border; and F18:9g-3D-Max-Dist-COG-to-Border
Category C4: Representative features (1):
F186:Hist-Energy-L1
Category C5: Representative features (1):
F48:AvgGLN; F51
Category C6: Representative features (6):
F62:3D-Laws-4; F68:3D-Laws-10; F69:3D-Laws-11; F72:3D-Laws-14; F143:3D-Laws-94; and F182:3D-Laws-133
Category C7: Representative features (5):
F197:3D-WaveP2-L2-8; F206:3D-WaveP1-L2-17; F208:3D-WaveP1-L2-19; F211:3D-WaveP1-L2-22; and F216:3D-WaveP1-L2-27

(C) Common between Manual and Ensemble: Across Categories (All Samples)

Concordance cutoff: CCC_{TreT} ≥ 0.90 and DR ≥ 0.90
Representative features (23): Combine features with R^2_{Bet} ≥ 0.95
C1: Tumor size:
F1:LongDia; F2:ShortAx-LongDia; F3:ShortAx; F6:Vol-cm; F33:Area-Pxl; F37:Thickness-Pxl; and F38:Length-Pxl
C2: Tumor shape:
F25:Density and F30:Shape-Index
C3: Location:
F8:8a-3D-Attch-Pleural; F9:8b-3D-Bord-to-Lung; F15:9d-3D-AV-Dist-COG-to-Border; F16:9e-3D-SD-Dist-COG-to-Border; F17:9f-3D-Min-Dist-COG-to-Border; and F19:10a-3D-Relat-Vol-Airspaces
C4: Pixel intensity histogram
F4:Mn-Hu and F187:Hist-Entropy-L1
C5: Co-occurrence and run length:
F48:AvgGLN; F51:AvgLRE
C6: Laws features:
F67:3D-Laws-9; F74:3D-Laws-16; F103:3D-Laws-45; and F128:3D-Laws-79

(A) Number of 3D features. (B) Feature description. (C) Feature description across categories.

dependent on segmentation methods (see Supporting Analysis section, Tables O1–O3). The slice matching between test and retest experiments adds additional variations that may not be uniform across the features, making it difficult to discern test/retest variability.

Repeatable and representative features. Finding features that are consistent in repeated experiments is a prerequisite step, which is followed by a redundancy reduction step to obtain an informative set. We tested the consistency between the test and retest experiments. For each image feature, the CCC_{TreT} was computed to quantify reproducibility between two scans performed on each patient. The CCC_{TreT} measures deviation from the 45° line, which is appropriate for repeated experiments and shown to be superior to the Pearson correlation coefficient [26]. On this set of highly reproducible features, the next step was to select the features with a large interpatient variability, using the "dynamic range" metric. The normalized DR for a feature was defined as the inverse of the average difference between measurements to the observed biologic (interpatient) range:

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

where i refers to an individual sample from the n patient cases; the maximum and minimum are computed on the entire sample set. The DR runs from 0 to 1. Values close to 1 are preferred and imply that the feature has a large biologic range relative to reproducibility. Increasing 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 last step is to eliminate redundancies, on the basis of the calculation of dependencies within the group. We computed the R^2_{Bet} between the remaining features to quantify the dependency. The R^2 has a range of 0 to 1 and is a ratio of known variance measured by linear model to total variance between two variables, where one is the outcome and the other is used to form the predictor. Values close to 1 would mean that the data points are close to the fitted line (i.e., closer to dependency) [24,25]. The R^2 of simple regression is equal to the square of the Pearson correlation coefficient [27,28]. The features were grouped on the basis of the R^2_{Bet} between them; in this subset, one representative that had the highest DR was picked. The procedure was repeated recursively to cover all the features, resulting in a most representative group. This was carried out in two ways: done independently for each category and across categories.

We implemented different cutoff values for R^2_{Bet} to consider the feature as linearly dependent with any other features in the list. Because the purpose of this third filter is to eliminate redundancies (and not necessarily identify independence), features with R^2_{Bet} values in the

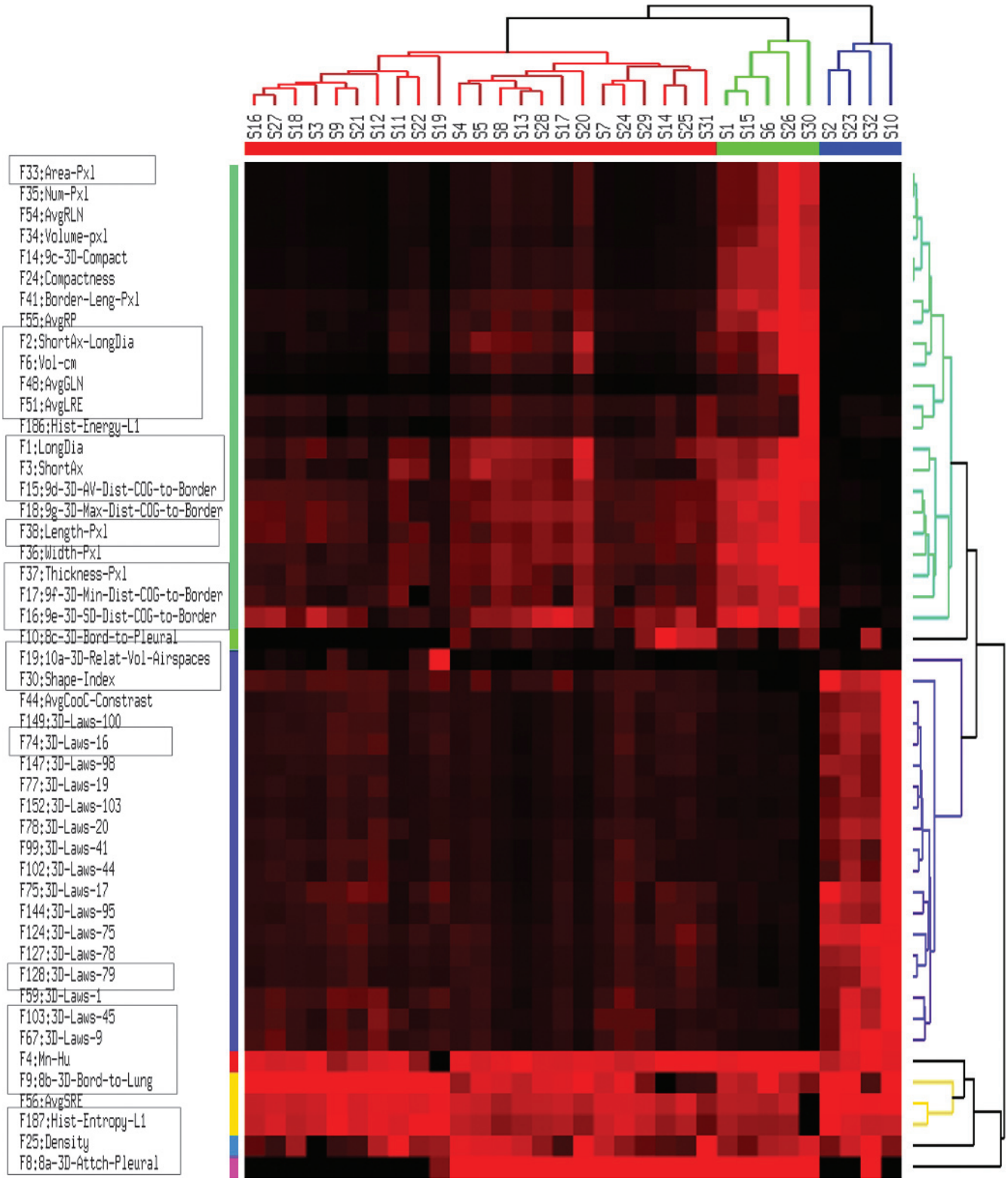

Figure 6. Hierarchial clustering of repeatable image features (CCC and DR > 0.9) in test/retest data and across segmentations. The representative features are obtained by removing features with high dependency ($R^2 \geq 0.95$); those that pass the cutoff are outlined (see Table 4*C*). The feature value was averaged over different segmentations (manual and ensemble) and repeats (test and retest). The features were standardized to 0 to 1. The clustering was arbitrarily stopped at seven and four groups on feature and sample axes, respectively. The F No. indicates feature position in the overall set of 219 3D features.

range ≥0.75 to ≥0.99 were found. This was repeated for different segmentation methods (manual, ensemble, and common between).

Prognostic Scoring and Independent Data Set

The RIDER test/retest data are a unique set that enables us to measure the effect of scanner and patient variability on the extracted features. Due to the unavailability of clinical information, we have created a physician (radiologist)–based prognostic risk score. Putting the samples into prognostic groups enables us to study the ability of image features to predict prognosis under typical scan variations. We used a trained radiologist to categorize the test/retest data set using five conventional observations for prognosis to score the tumor on a point scale. It has been reported that tumor size, differentiation, vascular invasion, and margin status (negative *versus* positive or close margins) have all been shown to have prognostic value [29–31]. We used five observable features—lobulated margin, size of the tumor lesion, spiculated margin, plueral wall attachment, and texture (e.g., ground-glass opacity)—as factors to scale the tumor into high-risk to moderate-risk individuals. The observations were given a score of 1 to 5. The five values were summed, averaged, and standardized to a scale of 0 to 1 to obtain a prognostic score. A normalized prognostic score over the median value was considered high risk (or poor prognosis), whereas the score lower than the median value is considered to be moderate risk (or better prognosis) as first proposed in [32]. Two samples could not be scored reliably using the point-scale metric due to diffused lesions, and one sample was partly scored due to an obscured margin. In total, three samples were eliminated from the scoring. The two created categories were then used to find discriminatory markers between the poor to better prognosis groups. Table W6 shows the score for individual samples.

Independent data set. Non-small cell lung tumor samples were collected in an Institutional Review Board (IRB)-approved study; of the 81 lung adenocarcinoma samples, 22 had mixed histology. In this study, prognostic testing was confined to adenocarcinoma group (59 samples; median split) and late-stage adenocarcinoma *versus* early-stage (TNM ≥ 3A *vs* TNM ≤ 1B; 15 *vs* 35 samples). These patients had a CT scan before surgical resection. The tumor sample was analyzed by a board-certified pathologist. The clinical and vital statistics were obtained from the Moffitt Cancer registry (Tampa, FL). The vital statistics are typically updated on a yearly basis. Table 2 shows broad histology information for the data.

Results

As described in the Materials and Methods section, the CT data were segmented in two ways: ensemble (E) with a single human interaction; and manually (M), with multiple human decisions on the tumor boundaries. In the obtained ROIs, 219 3D (and 110 2D) features (see Supporting Analysis section, Table O1) were extracted and quantified; a comprehensive list is shown in Tables W7 and W8. We computed the variability bound (95% confidence limits) using the concordance correlation confidence for three conventional features (see Table 1), using both manual and ensemble segmentations. A strong confidence bound was obtained for both segmentations for these size measurements, also observed by previous authors [13]. Figure 3 shows the Bland-Altman difference distribution plots between test and retest for the three measures (longest diameter, longest diameter*short axis, and volume). As the tumor size increased, the difference between test and retest was reduced in most cases. Figure 4 shows a sample CT image with a segmentation boundary for test and retest cases. The segmentation methods defined the tumor boundaries and hence influenced the extracted features (most directly size and shape). The SI was computed between manual and ensemble boundaries in the test and retest experiment (see Figure 2). SI is the ratio of common volume to union of the two, which measures extent of similarity between methods. The average SI between manual and ensemble delineation in test (retest) was 79% (78.9%) with an SD of 21% (22%), respectively. The SCES had difficulty delineating tumors that were attached to the pleural wall where the boundary definition was heuristic. In these challenging cases, the boundary found by the methods was based on semioptimal pixel valley, which may disagree with the expert. There are also cases with ground-glass opacity and concave-type tumors for which it was difficult to find accurate boundaries with automated methods. Numerous studies have compared boundaries created by different radiologists. In Meyer et al. [33], an estimated volume difference was more than 31% for pulmonary nodules. In large isolated and solid tumor lesions, segmentation is fairly easy, but issues remain in finding boundaries for attached and diffuse tumors.

Concordance in Repeated Experiment

The 219 extracted features (3D and 110 2D) were first compared in a test/retest experiment using the CCC_{TreT}, which is a stringent measure of reproducibility. A CCC_{TreT} value ≥ 0.75 indicates that the data are of acceptable reproducibility. At a second level of analysis, the DR was computed as described in Materials and Methods section, and this metric will identify those features with the largest biologic

Table 5. Representative Image Features in Test/Retest Data were Used to Predict Prognostic Scores.

	Feature	Accuracy (%)	Sensitivity	Specificity	Area Under the Curve
C1: Tumor size					
1	F1:LongDia	81.03	0.8	0.82	0.88
2	F2:ShortAx-LongDia	80.17	0.75	0.86	0.89
3	F3:ShortAx	77.59	0.77	0.79	0.87
4	F6:Volume(cm^3)	71.55	0.57	0.88	0.91
5	F33:Area-Pixel	68.97	0.45	0.95	0.92
6	F37:Thickness-Pixel	77.59	0.7	0.86	0.90
7	F38:Length-Pixel	75	0.67	0.84	0.83
C2: Tumor shape					
1	F25:Density	63.79	0.73	0.54	0.64
2	F30:Shape-Index	78.45	0.92	0.64	0.90
C3: Tumor location					
1	F8:8a-3D-Attch-Pleural	74.14	0.87	0.61	0.53
2	F9:8b-3D-Bord-to-Lung	69.83	0.57	0.84	0.75
3	F15:9d-3D-AV-Dist-COG-to-Border	81.9	0.72	0.93	0.92
4	F16:9e-3D-SD-Dist-COG-to-Border	61.21	0.65	0.57	0.75
5	F17:9f-3D-Min-Dist-COG-to-Border	80.17	0.72	0.89	0.88
6	F19:10a-3D-Relat-Vol-Airspaces	64.66	0.97	0.30	0.71
C4: Pixel intensity histogram					
1	F4:Mean (HU)	66.38	0.9	0.41	0.79
2	F187:Hist-Entropy-L1	78.45	0.73	0.84	0.85
C5: Gray scale: Run length and co-occurrence					
1	F48:AvgRunL(GLN)	70.69	0.47	0.96	0.93
2	F51:AvgRunL(LRE)	73.28	0.67	0.80	0.79
C6: Texture: Laws features					
1	F67:3D-Laws-9	81.03	0.98	0.63	0.91
2	F74:3D-Laws-16	76.72	0.95	0.57	0.89
3	F103:3D-Laws-45	70.69	0.88	0.52	0.79
4	F128:3D-Laws-79	68.10	0.88	0.46	0.79

An optimal threshold was obtained using linear discriminant function. All values were rounded to two-decimal precision.

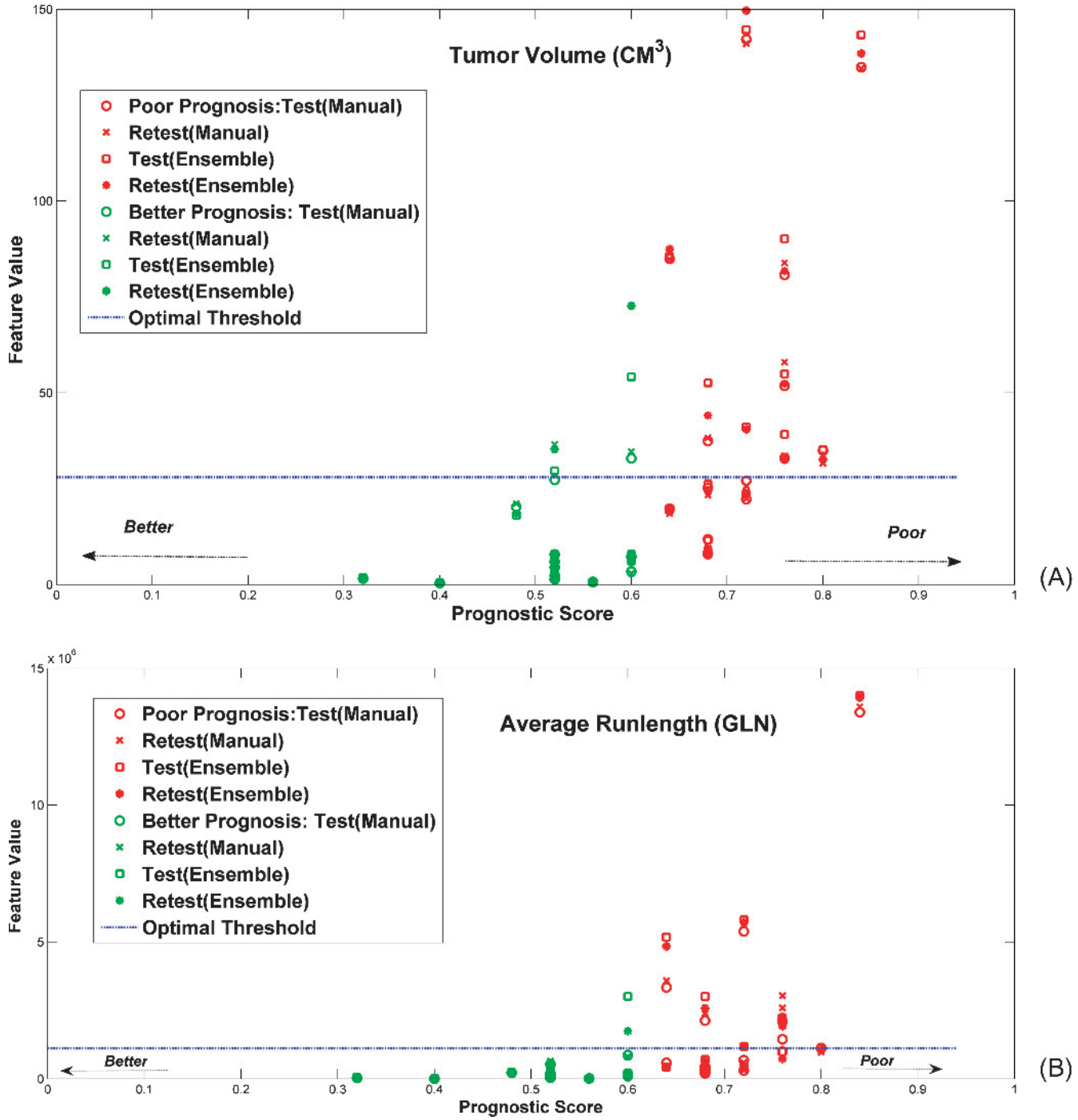

Figure 7. Discrimination of prognostic score with feature value for size- and texture-based features with optimal threshold. (A) Volume feature, (B) run-length GLN, and (C) Laws feature.

range relative to their reproducibility. For our data set, we examined various cutoffs and, with stringent limits, there were 72 (manual, M) and 106 (ensemble, E) 3D features that had a $CCC_{TreT} \geq 0.90$ and DR ≥ 0.90. Of these, 48 (~22%) of the features were common between segmentations (in total of 219 features; see Table 3*A*). In the 2D set, 51 (manual, M) and 28 (ensemble, E) features had a $CCC_{TreT} \geq 0.90$ and DR ≥ 0.90. Of these, 17 (15.3%) of the features were common between the segmentations. Details are reported in Supporting Analysis section, Tables O2 and O3. It was interesting to note that some of the texture features (wavelet and Laws kernel) were not repeated in the ensemble segmentation. This was attributed to feature outliers due to segmentation boundary differences, which resulted in adding different regions in test/retest. When the outlier samples were removed (in the samples with wavelet layer 1 energy > 20%), there was an 8% increase in the number of ensemble features (see Table 3*B*).

These concordance and DR filtering procedures will result in obtaining a set of features that are reproducible with a large range compared to the variability between the test and retest experiments. However, these resulting features may have interdependencies. Therefore, we used the R^2_{Bet} between the features to quantify the

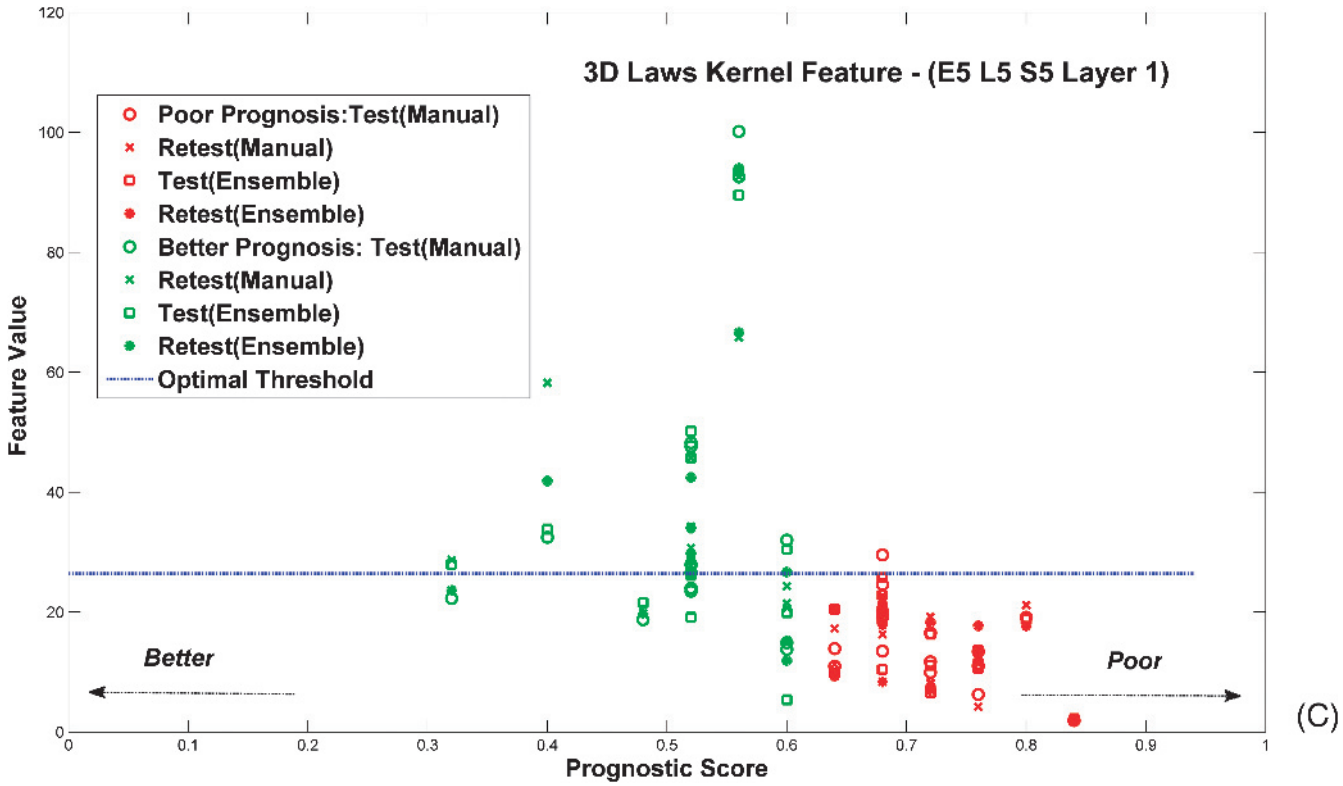

Figure 7. (*continued*).

levels of similarity. In this approach, if a feature of interest is linearly predicted by any other feature in the filtered feature set, the two were grouped together, repeated to cover all pairs. In the group of interdependent features, the one having the largest DR was chosen as the representative feature for the group, and the rest were removed. The procedure was repeated to cover the entire subset to form the reduced set. The cutoff level to reduce features based on linear dependency is critical and is subject to change with the sample size and the tumor shape and texture. This redundancy reduction can be carried out categorywise or by combining the categories; both were attempted. Finding reproducible features categorywise with redundancy reduction will help us form an informative feature set that will translate to a similar range of tumors. We performed the filtering at a few different levels: at $R^2_{\text{Bet}} \geq 0.95$, there were 29 common features between segmentation methods for CCC_{TreT} and DR ≥ 0.90. After removing the category bounds, there were 23 common features.

Figure 5 shows the ordered distribution plot for the DRs along with the distribution of concordance coefficients. The features' concordance and DR criteria were computed for manual and automatic segmentation independently, and the common features were obtained later. We tested interdependency by computing the R^2_{Bet} between the image features in each of the filtered sets and followed

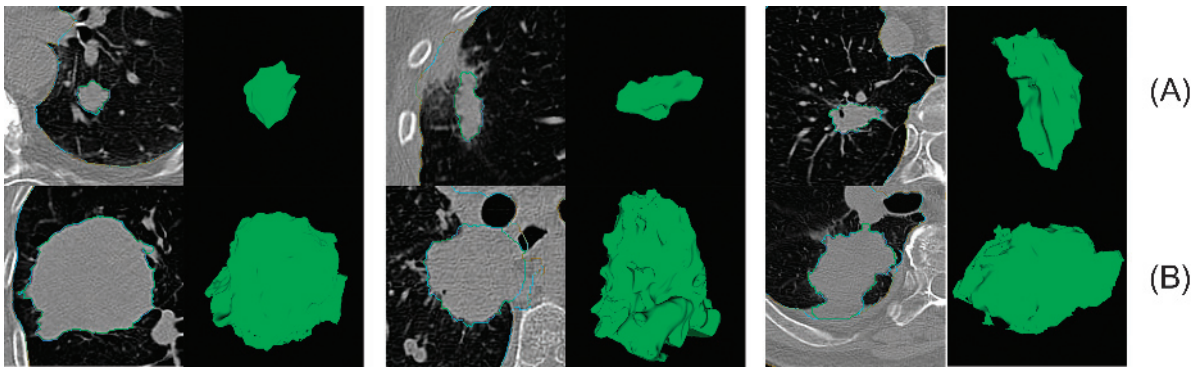

Figure 8. The 2D slice and 3D rendering of three samples selected in the test/retest RIDER data. The samples in A had a better radiologist prognostic score (smaller average run-length GLN), whereas the samples in B had a poor radiologist prognostic score (larger average run-length GLN).

Table 6. Significance Values of the Representative Image Features on Independent NSCLC Images (59 adenocarcinomas, with 15 *vs* 32 Samples Correspond to TNM ≥ 3A *vs* ≤1B).

	Feature	Adenocarcinoma	Adenocarcinoma: Extreme Grades (TNM ≥ 3A *vs* ≤1B)
C1: Tumor size (log-rank test: *P* values)			
1	F1:LongDia	0.233	0.261
2	F2:ShortAx*LongDia	0.079	0.261
3	F6:Volume (cm^3)	0.322	0.516
C2: Tumor shape (log-rank test: *P* values)			
1	F25:Density	0.979545	0.690185
2	F30:Shape-Index	0.205607	0.497
C3: Tumor location (log-rank test: *P* values)			
1	F8:8a-3D-Attch-Pleural	0.079279	0.030118
2	F9:8b-3D-Bord-to-Lung	0.063549	0.085015
3	F15:9d-3D-AV-Dist-COG-to-Border	0.067196	0.898633
4	F16:9e-3D-SD-Dist-COG-to-Border	0.085849	0.516197
5	F17:9f-3D-Min-Dist-COG-to-Border	0.394427	0.982512
6	F19:10a-3D-Relat-Vol-Airspaces	0.960735	0.842027
C4: Pixel intensity histogram (log-rank test: *P* values)			
1	F4:Mean (HU)	0.871203	0.528778
2	F187:Hist-Entropy-L1	0.082696	0.053524
C5: Gray scale: Run length and co-occurrence (log-rank test: *P* values)			
1	F48:AvgRunL(GLN)	0.04596	0.898633
2	F51:AvgRunL(LRE)	0.092659	0.898633
C6: Texture: Laws features (log-rank test: *P* values)			
1	F67:3D-Laws-9	0.702408	0.166091
2	F74:3D-Laws-16	0.748021	0.643091
3	F103:3D-Laws-45	0.993109	0.142433
4	F128:3D-Laws-79	0.587798	0.06058

the described procedure to find a representative feature for a group of highly dependent features (see Discussion section).

Table 4 shows feature details, whereas Table W3 provides features with categorywise reduction. When category boundaries were removed, more feature reduction was observed. Table 4, *A* and *C*, and Table W5 show feature details without category restrictions. Figure 6 shows a clusterogram heat map of representative features that are common between segmentations. The feature value was averaged in the test/retest experiment and across segmentations (manual and ensemble). The concordance and DR was set at: CCC_{TreT} and DR ≥ 0.90. The representative features were outlined for both types of segmentations with feature reduction cutoff of $R^2_{Bet} \geq 0.95$.

Discrimination Ability of Radiologic Prognostic Score

The subset of features with a high DR that were reproducible and nonredundant was subsequently used to test the discrimination ability of a radiologist-determined prognostic score. Each observation in manual/ensemble test/retest was considered an independent measurement, and the optimal threshold for each was computed by linear discriminant analysis. Using the radiologic score as observed truth, the discriminator's performance was tested by computing sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve for each of the image features. Table 5 lists the values for each feature, and Figure 7 shows conventional shape-based (volume) feature compared to texture features (run length and Laws kernel). The ROC curve for the discrimination is presented in Figure W1. Figure 8 shows the sample tumors in the test-retest data set with different values of run-length feature (average run-length nonuniformity or GLN) for the better and poorer prognostic groups. This feature measured the distribution of the similarity of gray-level pattern, which can be considered measure of heterogeneity in the gray levels. It was interesting to note that run-length long run emphasis and run-length nonuniformity (GLN) follow similar trends (but at different scales). Testing the features on the radiologist prognostic scores allowed us to test the ability of features to predict expert observations in the presence of repeatable noise.

Prognostic Ability on Independent Data

The 219 3D-image features were extracted on 59 adenocarcinomas from selected Moffitt patients accrued before 2010, with survival refreshed. The feature set was filtered to select reproducible representative features across segmentations (CCC_{TreT} and DR ≥ 0.90), and these were then tested for prognostic potential in patient data. The data set was dichotomized at the median value of the image feature. The survival function was estimated by the Kaplan-Meier approach, and statistical significance was computed using the log-rank test. Table 6 lists the *P* value for the representative features tested using the adenocarcinoma samples. The texture feature average run-length nonuniformity (GLN) shows statistical significance ($P = .046$). Figure 9 shows the survival plot and sample tumors for the two factions. A large value of run-length GLN indicates a more homogeneous tumor, and this was related to a longer survival, whereas low values of run-length GLN indicate more heterogeneity, and this was related to shorter survival. Notably, this texture feature had better prognostic potential compared to the more conventional measurements of shape and size; including longest diameter, longest diameter*short axis, and volume.

Discussion

The reproducibility of radiographical 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. A key component of our work is application of an ensemble (semiautomatic) segmentation process so that minimal operator input and manual editing were required. Prior work has demonstrated three conventional univariate, bivariate, and volumetric (RECIST and World Health Organization or WHO defined criteria) measures to infer concordance consistency for automatically segmented lung lesions, which seem to be limited in describing the complex nature of the tumor [13]. In the current study, our focus has been to describe the tumor with many features using the following different categories: size (volume, diameter, and border length), shape (shape index, compactness, and asymmetry), boundary region (border length and spiculation), relation to the lung field, image intensity-based features (mean, SD, average air space, deviation of airspace, energy, entropy, skewness, and other features), and transformed texture descriptors (wavelet transform: entropy, energy, and Laws features). For this new set of features, consistency in the repeat scans (test, retest) was tested and filtered for independent features to yield an image feature set to better predict prognosis.

In a parallel study, we demonstrated that a semiautomated (ensemble) multiseed point segmentation can reliably generate segmented volumes, as defined by the SI. The SI between machine-segmented lesions was >0.93, whereas the SI for manual segmentation was 0.73 across a test set of 129 patients [11].

One requirement for a feature to be qualified as a response biomarker is that the change in an image's feature between pretherapy and posttherapy scans must be significantly greater than the difference

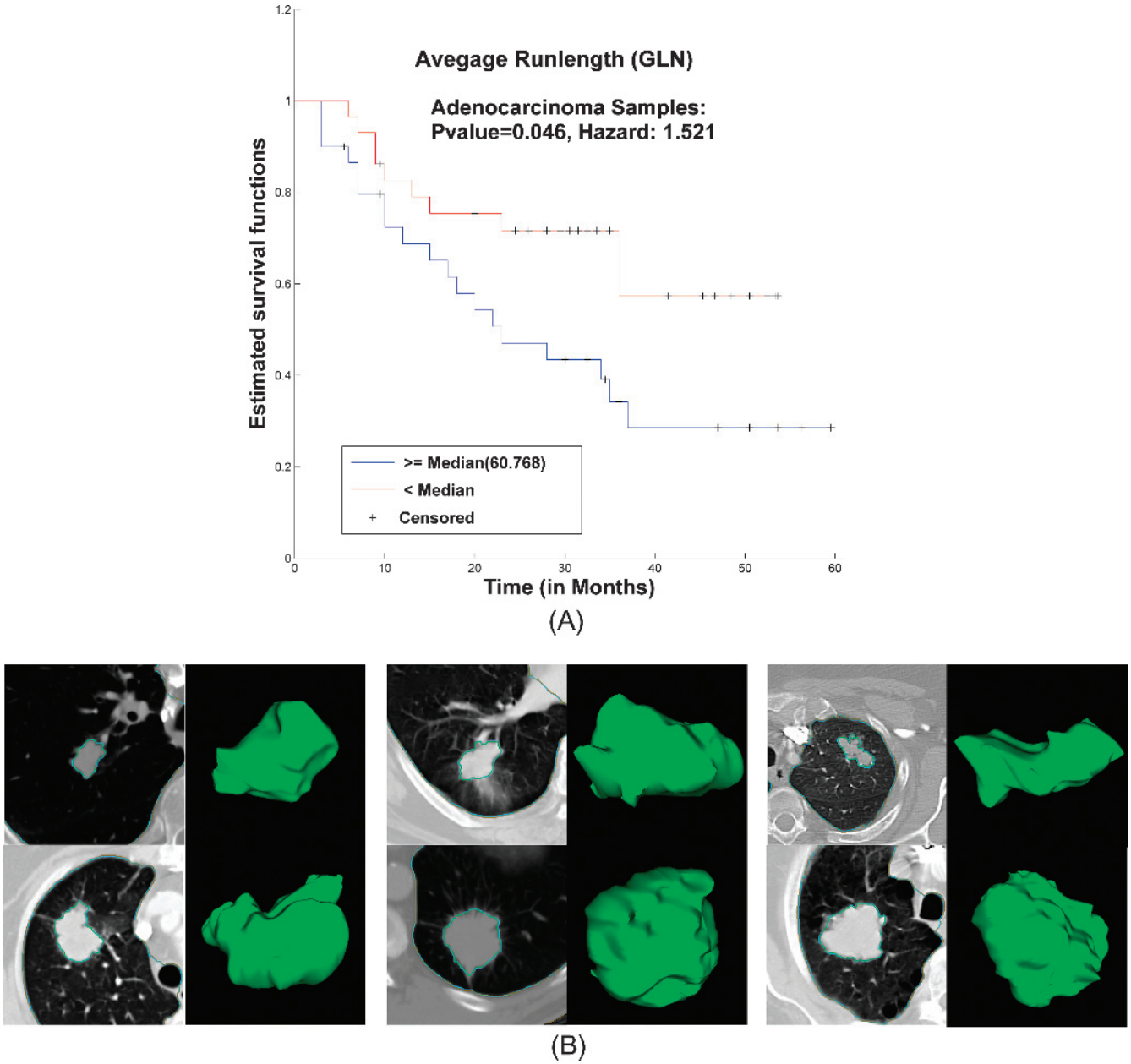

Figure 9. Prognostic test result using run-length nonuniformity (GLN) image feature split at the median value is shown. (A) Kaplan-Meier plot for independent adenocarcinoma cases. (B) Example of three extreme tumor samples with low and high values of average run-length GLN is shown in 2 and 3D plots.

observed in test-retest study (also referred to as "coffee break") measurements. In the present study, we can estimate the change of individual features posttherapy against the entire pretherapy biologic range. The ratio of the range to the interscan variability is a measure of "dynamic range," as shown in Figure 5 (*first panel*). Features showing high DR were considered potentially more informative. Figure 5 (*second panel*) also shows the distribution of concordance coefficients between test and retest, which is skewed toward higher end values as one would expect showing high concordance between the test and retest cases. There is also a larger peak toward zero values in both manual and ensemble segmentation. Investigating the peaks shows that some of the Laws and higher level wavelet features show low concordance between test and retest repeats. It is hypothesized that reimaging the patients resulted in some change in texture. These Laws features compute energy after filtering in a region. Small changes in subregional 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.

In prior work, Segal et al. has used a correlation coefficient cutoff of 0.9 to distinguish highly correlated features [34]. Feature reduction to combine features that convey similar information has been widely investigated; different metrics that have been used in the past are the correlation coefficient and regression methods [35–37]. In our study, we found a representative feature set that will eliminate redundancy in terms of information content, as complete independence may not be as relevant for our study as texture information is subjective (sample issues, scanner settings, protocol followed, etc.). We used the R^2_{Bet} between features to quantify dependency. Features were

grouped on the basis of R^2_{Bet} between them; in this subset, one representative that had the highest DR was picked. The procedure was repeated recursively to cover all the features, resulting in a most representative group both with and without the category labels. The test, retest values were averaged before computing R^2_{Bet}. We set different limits on the R^2_{Bet} to decide when to combine the features, from 0.75 to 0.99 (user choice of acceptable level of dependency). For higher cutoff values of R^2_{Bet}, a relatively smaller number of features will be indicated as dependent set, resulting in a larger representative group. Setting the R^2_{Bet} to a lower limit will group more number of features into a dependent set, resulting in a smaller representative feature set. The combination of reproducible, informative, and independent sets of features is critical to obtain a feature set that may be a good candidate for imaging biomarkers.

Looking closely at the features, the ensemble segmentation has more independent features than manual segmentation with the highest R^2_{Bet} settings. Less than half the features in the manual segmentation appear in the ensemble list. The texture features (wavelet and Laws) show a less than expected concordance in the ensemble segmentation. The difference could be attributed to suboptimal segmentation in tumors with ill-defined boundaries. It is known that the manual methods involve more operator inputs that could include areas around the tumor, creating more visually smooth regions, whereas the ensemble method looks at connected regions with spurious local minima being a stopping condition, which may result in some differences in repeated segmentation. This situation is mitigated by generating multiple seed points (more than 21 seed points spread in eight directions) in an SCES method.

The image features are expected to capture different aspects of morphology and texture information. Due to the consistency in samples chosen, a limited sample population the image features computed may show a higher level of dependency, which was also observed in Figure 6, which could be attributed to a greater reduction in the final feature set.

Prognosis

The radiologic prognostic score provides us a method to capture expert opinion in a quantifiable fashion. This type of scoring scheme has been used in the past [38,39], but in our work, we use it as a metric to group the samples into risk populations. This allows us to develop predictive schemes and quantify the ability of image markers to match the opinion of a conventional radiologist (see Table 4 and Figures 7 and 8). The true prognostic ability has been conventionally tested using survival plots [40] that provided a quantifiable measure of survival differences between groups in the population. We used the reproducible features to test their true prognostic ability in an independent NSCLC set. Each of the representative features across segmentation methods was used to dichotomize the population and the survival groups (or factions) tested for significance (see Table 6 and Figure 9). A texture feature (run-length GLN) that tracks consistency of gray-scale values shows a prognostic significance in adenocarcinoma. Dividing the sample into higher grade (TNM $\geq$ 3A *vs* $\leq$1B) did not help to improve significance. This could be attributed to small sample size for the secondary test.

Conclusions

In the current study, we demonstrated that many CT features of primary lung cancer are repeatable in a controlled test-retest scan and independent of segmentation methods (manual and semiautomatic). Across all patients, the biologic ranges for several individual features were highly variable. Combining interscan variance, segmentation differences, biologic range, and covariance, we have reduced the total number of features from 219 (3D features) to a most informative set of 48 features (in both ensemble and manual) at CCC_{TreT} and DR $\geq$ 0.90. Of these, 29 features are representative features ($R^2_{Bet} \geq 0.95$), with stringent cutoffs. These repeatable features predicted a conventional radiologist prognostic score with area under the curve of 0.9. The sensitivity and specificity could be improved using a multivariate approach. The true prognosis was tested using an independent NSCLC set (59 samples of adenocarcinoma) with mixed stages, where a repeatable and representative texture feature (run-length GLN) showed significance. The current findings will allow selection of reproducible, informative/independent, and prognostic features as the candidate imaging biomarkers to predict or assess therapy response.

References

[1] Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, and Thun MJ (2008). Cancer statistics, 2008. *CA Cancer J Clin* **58**, 71–96.

[2] SEER-NCI (2013). *SEER Cancer Statistics Review (CSR) 1975-2010*. In N Howlader, AM Noone, M Krapcho, J Garshell, N Neyman, SF Altekruse, CL Kosary, M Yu, J Ruhl, Z Tatalovich, et al. (Eds). National Cancer Institute, Bethesda, MD.

[3] USPH-Service (1964). *Smoking and Health: Report of the Advisory Committee to the Surgen General of the Public Health Service*. Government Printing Office, Washington, DC.

[4] Nguyen T and Rangayyan R (2005). Shape analysis of breast masses in mammograms via the fractal dimension. *Conf Proc IEEE Eng Med Biol Soc* **3**, 3210–3213.

[5] Schuster DP (2007). The opportunities and challenges of developing imaging biomarkers to study lung function and disease. *Am J Respir Crit Care Med* **176**, 224–230.

[6] Therasse OP, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, Van Oosterom AT, Christian MC, et al. (2000). New guidelines to evaluate the response to treatment in solid tumor. *J Nat Cancer Inst* **92**, 205–216.

[7] Schwartz LH, Mazumdar M, Brown W, Smith A, and Panicek DM (2003). Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. *Clin Cancer Res* **9**, 4318–4323.

[8] Suzuki C, Jacobsson H, Hatschek T, Torkzad MR, Bodén K, Eriksson-Alm Y, Berg E, Fujii H, Kubo A, and Blomqvist L (2008). Radiologic measurements of tumor response to treatment: practical approaches and limitations. *Radiographics* **28**, 329–344.

[9] Tuma RS (2006). Sometimes size doesn't matter: reevaluating RECIST and tumor response rate endpoints. *J Natl Cancer Inst* **98**, 1272–1274.

[10] McNitt-Gray MF, Bidaut LM, Armato SG, Meyer CR, Gavrielides MA, Fenimore C, McLennan G, Petrick N, Zhao B, Reeves AP, et al. (2009). Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error. *Transl Oncol* **2**(4), 216–222.

[11] Gu Y, Kumar V, Hall LO, Goldgof DB, Li CY, Korn R, Bendtsen C, Velazquez ER, Dekker A, Aerts H, et al. (2013). Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. *Pattern Recognit* **46**, 692–702.

[12] Armato SG III, Meyer CR, McNitt-Gray MF, McLennan G, Reeves AP, Croft BY, and Clarke LP (2008). RIDER Research Group The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software. *Clin Pharmacol Ther* **84**, 448–456.

[13] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, Qin Y, Riely GJ, Kris MG, and Schwartz LH (2009). Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer. *Radiology* **252**, 263–272.

[14] Athelogou M, Schmidt G, Schaepe A, Baatz M, and Binnig G (2007). Cognition Network Technology - a novel multimodal image analysis technique for automatic identification and quantification of biological image contents. In *Imaging Cellular and Molecular Biological Functions, Principles and Practice.* SL Shorte and F Frischknecht (Eds). Springer-Verlag, Berlin, Germany. pp. 407–422.

[15] Baatz M, Zimmermann J, and Blackmore CG (2009). Automated analysis and detailed quantification of biomedical images using Definiens Congnition Network Technology. *Comb Chem High Throughput Screen* **12**, 908–916.

[16] Bendtsen C, Kietzmann M, Korn R, Mozley P, Schmidt G, and Binnig G (2011). X-ray computed tomography: semiautomated volumetric analysis of late-stage lung tumors as a basis for response assessments. *Int J Biomed Imaging* **2011**, 1–11.

[17] Basu S, Hall LO, Goldgof DB, Gu Y, Kumar V, Choi J, Gilles RJ, and Gatenby RA. Developing a classifier model for lung tumors in CT-scan images, Systems, Man and Cybernetics (SMC), 2011 IEEE Conference, Anchorage, AK. pp. 1306–1312.

[18] Jain R, Kasturi R, and Schunck BG (1995). Machine Vision, McGraw-Hill, New York, USA.

[19] Koss JE, Newman FD, Johnson TK, and Kirch DL (1999). Abdominal organ segmentation using texture transforms and Hopfield neural network. *IEEE Trans Med Imaging* **18**, 640–648.

[20] Galloway M (1975). Texture analysis using gray level run lengths. *Computer Graphics Image Process* **4**, 172–179.

[21] Tang X (1998). Texture information in run-length matrices. *IEEE Trans Image Process* **7**, 1602–1609.

[22] Haralic RM and Shanmugam K (1973). Texture features for image classification. *IEEE Trans System Man Cybernet* **6**, 610–621.

[23] Laws K (1980). *Texture Image Segmentation*. University of South California, Los Angeles, CA.

[24] Jafari-Khouzani K, Soltanian-Zadeh H, Elisevich K, and Patel S (2004). Comparison of 2D and 3D wavelet features for TLE lateralization. *Proc of SPIE Medical Imaging 2004: Physiology, Function and Structure from Medical Images* **5369**, 593–601.

[25] Daubechies I (1988). Orthogonal bases of compactly supported wavelets. *Commun Pure Appl Math* **41**, 909–996.

[26] Lin LI-K (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 13.

[27] Steel RGD and Torrie JH (1960). *Principles and Procedures of Statistics*. McGraw-Hill, New York.

[28] Colin Cameron A, Windmeijer F, Gramaji H, Cane DE, and Khosla C (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *J Econom* **77**, 1790–1792.

[29] Aoki T, Tomoda Y, Watanabe H, Nakata H, Kasai T, Hashimoto H, Kodate M, Osaki T, and Yasumoto K (2001). Peripheral lung adenocarcinoma: correlation of thin-section CT findings with histologic prognostic factors and survival. *Radiology* **220**, 803–809.

[30] Takashima S, Maruyama Y, Hasegawa M, Saito A, Haniuda M, and Kadoya M (2003). High-resolution CT features: prognostic significance in peripheral lung adenocarcinoma with bronchioloalveolar carcinoma components. *Respiration* **70**, 36–42.

[31] Subramanian J and Simon R (2010). Gene exression–based signature in lung cancer: ready for clinical use? *J Natl Cancer Inst* **102**, 464–474.

[32] Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Basu S, Korn R, Zhao B, Goldgof DB, Hall LO, et al. Reproducibility of quantitative features extracted from CT images of lung tumors. *J Digital Imaging* (in review).

[33] Meyer CR, Johnson TD, McLennan G, Aberle DR, Kazerooni EA, Macmahon H, Mullan BF, Yankelevitz DF, van Beek EJ, Armato SG III, et al. (2006). Evaluation of lung MDCT nodule annotation across radiologists and methods. *Acad Radiol* **13**, 1254–1265.

[34] Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, Chan BK, Matcuk GR, Barry CT, Chang HY, et al. (2007). Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol* **25**, 675–680.

[35] Jain AK and Zongker D (1997). Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Analysis* **19**, 153–158.

[36] Pudil P, Novovièová J, and Kittler J (1994). Floating search methods in feature selection. *Pattern Recognit Lett* **15**, 1119–1125.

[37] Saeys Y, Inza I, and Larrañaga P (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517.

[38] Kress MA, Collins BT, Collins SP, Dritschilo A, Gagnon G, and Unger K (2012). Scoring system predictive of survival for patients undergoing stereotactic body radiation therapy for liver tumors. *Radiat Oncol* **7**, 148.

[39] Colinet B, Jacot W, Bertrand D, Lacombe S, Bozonnat MC, Daurès JP, and Pujol JL, oncoLR health network (2005). A new simplified comorbidity score as a prognostic factor in non-small-cell lung cancer patients: description and comparison with the Charlson's index. *Br J Cancer* **93**, 1098–1105.

[40] Kaplan EL and Meier P (1958). Nonparametric estimation from incomplete observations. *J Amer Statist Assn* **53**, 457–481.