Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?

Xenia Fave[a)]
Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030 and The University of Texas Graduate School of Biomedical Sciences at Houston, 6767 Bertner Avenue, Houston, Texas 77030

Dennis Mackin, Jinzhong Yang, and Joy Zhang
Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

David Fried
Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030 and The University of Texas Graduate School of Biomedical Sciences at Houston, 6767 Bertner Avenue, Houston, Texas 77030

Peter Balter and David Followill
Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

Daniel Gomez
Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

A. Kyle Jones
Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

Francesco Stingo
Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

Jonas Fontenot
Mary Bird Perkins Cancer Center, 4950 Essen Lane, Baton Rouge, Louisiana 70809

Laurence Court
Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030 and Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

(Received 10 July 2015; revised 10 September 2015; accepted for publication 18 October 2015; published 6 November 2015)

Purpose: Increasing evidence suggests radiomics features extracted from computed tomography (CT) images may be useful in prognostic models for patients with nonsmall cell lung cancer (NSCLC). This study was designed to determine whether such features can be reproducibly obtained from cone-beam CT (CBCT) images taken using medical Linac onboard-imaging systems in order to track them through treatment.

Methods: Test-retest CBCT images of ten patients previously enrolled in a clinical trial were retrospectively obtained and used to determine the concordance correlation coefficient (CCC) for 68 different texture features. The volume dependence of each feature was also measured using the Spearman rank correlation coefficient. Features with a high reproducibility (CCC > 0.9) that were not due to volume dependence in the patient test-retest set were further examined for their sensitivity to differences in imaging protocol, level of scatter, and amount of motion by using two phantoms. The first phantom was a texture phantom composed of rectangular cartridges to represent different textures. Features were measured from two cartridges, shredded rubber and dense cork, in this study. The texture phantom was scanned with 19 different CBCT imagers to establish the features' interscanner variability. The effect of scatter on these features was studied by surrounding the same texture phantom with scattering material (rice and solid water). The effect of respiratory motion on these features was studied using a dynamic-motion thoracic phantom and a specially designed tumor texture insert of the shredded rubber material. The differences between scans acquired with different Linacs and protocols, varying amounts of scatter, and with different levels of motion were compared to the mean intrapatient difference from the test-retest image set.

Results: Of the original 68 features, 37 had a CCC >0.9 that was not due to volume dependence. When the Linac manufacturer and imaging protocol were kept consistent, 4–13 of these 37 features

passed our criteria for reproducibility more than 50% of the time, depending on the manufacturer-protocol combination. Almost all of the features changed substantially when scatter material was added around the phantom. For the dense cork, 23 features passed in the thoracic scans and 11 features passed in the head scans when the differences between one and two layers of scatter were compared. Using the same test for the shredded rubber, five features passed the thoracic scans and eight features passed the head scans. Motion substantially impacted the reproducibility of the features. With 4 mm of motion, 12 features from the entire volume and 14 features from the center slice measurements were reproducible. With 6–8 mm of motion, three features (Laplacian of Gaussian filtered kurtosis, gray-level nonuniformity, and entropy), from the entire volume and seven features (coarseness, high gray-level run emphasis, gray-level nonuniformity, sum-average, information measure correlation, scaled mean, and entropy) from the center-slice measurements were considered reproducible.
Conclusions: Some radiomics features are robust to the noise and poor image quality of CBCT images when the imaging protocol is consistent, relative changes in the features are used, and patients are limited to those with less than 1 cm of motion. [http://dx.doi.org/10.1118/1.4934826]

Key words: texture, quantitative imaging features, reproducibility, cone-beam CT

1. INTRODUCTION

Lung cancer is the leading cause of all cancer deaths in the United States.[1] Nonsmall cell lung cancer (NSCLC) accounts for approximately 85% of all newly diagnosed lung cancer cases.[1,2] The high mortality rate has prompted numerous research studies to identify patient-specific prognostic factors with the goal of individualizing treatment.[3–5] Models built using radiomics features, or imaging features, are one novel approach for identifying patients with the highest risk for disease progression, poor survival, or other clinical outcomes.[6–12] Radiomics features are extracted from the region-of-interest (ROI) in an image in order to assign a quantitative value to that ROI. Data mining and machine learning techniques are then used to build models and capture valuable insights from those imaging features. In the context of tumor analysis, univariate or multivariate models using these features have typically been built to diagnose lesions,[13–15] identify secondary effects,[16] or predict outcome.[6,12]

Recent publications have demonstrated that a wide variety of radiomics features may predict NSCLC patient outcomes when extracted from computed tomography (CT),[6,12,17] contrast-enhanced CT,[18,19] or positron emission tomography (PET)[20–22] images. However, no studies have yet examined whether there is a potential for imaging features extracted from cone-beam CT (CBCT) images to be useful. Unlike CT or PET images, which are used for diagnosis and treatment planning, CBCT images are generally used to check patient positioning before radiation treatment and accordingly are acquired using low imaging doses. The resulting images include substantially more scatter than diagnostic CT images because of the flat-panel detector design. Additionally, the CBCT scan time is relatively long, up to 1 min, which increases the probability of motion artifacts. Despite these limitations on CBCT image quality, these images are often acquired before every fraction of treatment or at least weekly and as a result could provide a source for tracking patient imaging feature changes during the course of treatment. This is important because changes in imaging features could become a timely biomarker for early detection of tumor response.

Reproducibility of radiomics features extracted from CBCT images must be investigated first so that the models built from these features can be consistently and reliably applied to different institutions and cohorts. The purpose of this study was to identify the impact of different imaging protocols, levels of scatter, and amounts of motion on imaging features extracted from CBCT images. Only after the effects of these parameters are characterized and understood can guidelines for the use of texture features in CBCT be developed, and the feasibility of tracking features through treatment established.

2. METHODS

2.A. Patient test-retest CBCT images

CBCT images were acquired from patients who were part of a previous IRB-approved clinical trial.[23] Inclusion criteria for the trial were diagnosis of a stage II–IIIb NSCLC tumor, Karnofsky performance score >70 or ECOG score 0–1, and suitability for treatment with concurrent chemoradiation. Exclusion criteria were small cell tumor histology, prior radiation to the treatment field, pregnancy, and the presence of implanted devices that prohibited radiation treatment. We retrospectively searched the imaging history for each of the patients included in the clinical trial for repeat CBCT images. From this search, only ten patients were found who had two sets of CBCT images obtained within 15 min of each other using the same imager. Patient characteristics are tabulated in Table I.

All CBCT patient images were acquired using the thoracic imaging protocol on a Varian Linac: peak tube voltage of 110 kVp, tube current of 20 mA, and exposure time (total pulsed beam-on time) of 7–14 s. Images were reconstructed as a 512×512 grid with pixel dimensions of 0.8 and 2.5 mm slice thickness. For each of these ten patients, we deformably transferred the GTV contour from the treatment plan to their

TABLE I. Characteristics of the ten patients whose images were used in this study.

Characteristics	Number of patients	Percent of patients (%)
n, number of patients	10	NA
Median age (range)	65.5 (49–80)	NA
Median GTV volume (range) (cm^3)	77.5 (17–315)	NA
Gender		
Male	3	30
Female	7	70
Tumor stage		
II	1	10
III	9	90
Tumor histology		
Squamous cell carcinoma	2	20
Adenocarcinoma/other	8	80

two CBCT image sets using our in-house deformation image registration software, CT-assisted targeting.[24,25] The images and contours were imported into our biomarker software (details below) to extract the values for the imaging features.

The concordance correlation coefficient (CCC) was calculated for each feature using this test-retest image set. Features whose CCC <0.9 were not considered reproducible and were excluded from the rest of our analysis.[26,27] This test removed features that were not reproducible even when measured in images obtained from the same patient within 15 min using the same imager. The cutoff value of 0.9 was chosen based on the recommended criteria of McBride that considered a correlation of 0.9 to reflect moderate strength-of-agreement and all correlations <0.9 to be poor. The Spearman correlation coefficient (r_s) was also calculated between each feature and the ROI volume. The Spearman coefficient was calculated for the test and retest image volumes individually. Any feature with r_s >0.85 in both image sets was excluded from the rest of our analysis in order to remove features whose CCC was high only because of that feature's strong correlation with volume.[28,29] The cutoff value of 0.85 is within the range of values of that has been cited in the literature as representative of strong correlations such as Zou *et al.*[28] who considered any $r_s > 0.8$ to be strongly correlated and Mukaka[29] who interpreted only $r_s > 0.9$ to have very high correlations. Because the purpose of this step was to reduce the likelihood of false positives in the following experiments where each feature was independently examined for its reproducibility under different conditions and not to establish an explicit volume dependence, we selected a relatively high cutoff to remove only the most egregious relationships.

For the remaining features, the mean intrapatient test-retest differences were calculated with Eq. (1). These values were used as benchmarks for reproducibility in our subsequent phantom studies. This criterion was used because we expected a phantom to change substantially less from scan to scan than a patient. In Eq. (1), N_{pats} is the number of patients and $x_{n,t}$ and $x_{n,r}$ are the nth patient's test and retest values, respectively,

$$\text{Mean intrapatient difference} = \frac{\sum_{n=1}^{N_{\text{pats}}} |(x_{n,t} - x_{n,r})|}{N_{\text{pats}}}. \quad (1)$$

2.B. Texture phantom

The Credence Cartridge Radiomics phantom was used to evaluate the impact of different scanners, protocols, scatter levels, and amounts of motion on texture values extracted from CBCT images, Fig. 1. This phantom was designed at our institution specifically for investigating the reproducibility of texture features.[30] The phantom is a hollow, acrylic, rectangular prism. Inside are cartridges of $10.1 \times 10.1 \times 3.2$ cm^3 made of different materials: wood, dense cork, regular cork, shredded rubber, acrylic, and resin. The phantom also contains four 3D printed cartridges with tessellated hexagons of different sizes. A previous analysis of the phantom imaging characteristics conducted at our institution using CT scans demonstrated that the texture values extracted from the shredded rubber and dense cork cartridges were closest to values obtained from patients.[30] For this reason, only these two materials were used in this analysis. A ROI of $6.0 \times 6.0 \times 2.0$ cm^3 was positioned at the center of these two materials for feature extraction in each image, Fig. 1. This size was used because it was close to the median size of the patient GTV volumes (77.5 cm^3) and to avoid including the edges of the phantom in the ROI.

2.C. Texture features and preprocessing

For this study, a comprehensive set of 68 texture features was initially selected (Table II). Features were selected to cover the diverse range of features that have been used in previous texture feature studies using CT images of NSCLC.[6–8,10,11,13–15,17,31,32] The features were all calculated using the open-source Imaging Biomarker EXplorer (IBEX) software, which is available for download at http://bit.ly/IBEX_MDAnderson.[33] Selected features included first-order descriptors from the intensity histogram (Hist); second-order features to describe spatial relationships in gray level intensities from the co-occurrence matrix (COM),[17,34] run-length matrix (RLM),[35] and neighborhood gray-tone difference matrix (NGTDM);[36] and Laplacian of Gaussian (LoG) filtered features, which can highlight tumor characteristics not visible in the original image.[6,31] In subsequent tables and figures, features are named for the feature category and then the feature name (e.g., contrast from the COM is listed as COMcontrast). Features with longer names are abbreviated with the abbreviations listed in Table II (e.g., long run emphasis from the RLM is listed as RLMlre). Specific parameters for the calculation of each feature in IBEX are included in the supplementary material.[37]

The patient ROIs were preprocessed with a thresholding step to exclude air, bones, and normal lung tissue. Values less than −150 HU or greater than 200 HU were excluded for the patient images. For the motion phantom, a lower threshold of −700 HU was used to ensure none of the

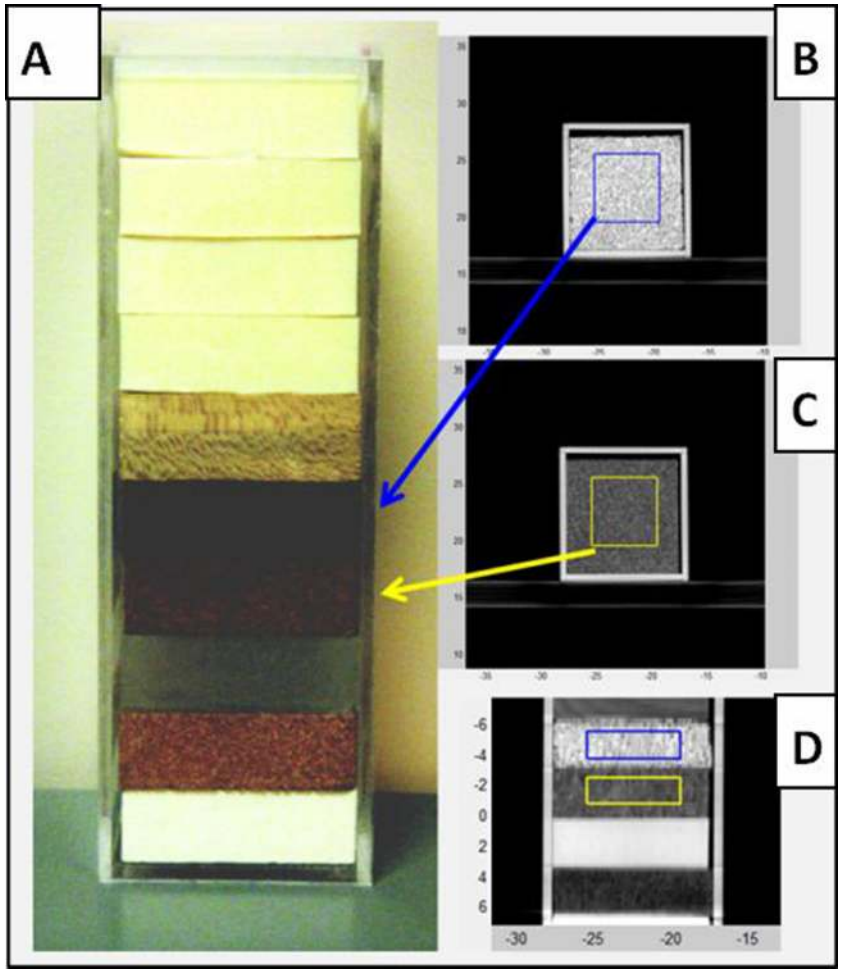

FIG. 1. (A) The texture phantom used in this study and [(B)–(D)] CBCT images of the phantom with the ROIs used. Only the (B) shredded rubber and (C) dense cork cartridges were used for the current analysis.

surrounding lung-equivalent material was included in the ROI. Thresholds were not used for the texture phantom ROIs in order to ensure that all of the voxels within the ROI would be included.

All the images were also rescaled to 8-bit images before calculating the COM, RLM, and NGTDM features; this was done to reduce the effect of noise on the texture features and prevent sparsely populated matrices from being produced. The Hist features were calculated both with and without 8-bit rescaling. The LoG features were calculated without the rescaling step because the Gaussian filter already acts to smooth the images and reduce noise. The LoG features were calculated at two different scales: a fine filter (fineFilt) with a window size of 5 and sigma of 1 and a medium filter (medFilt) with a window size of 7 and sigma of 1.5.

2.D. Effect of scanners

During the course of treatment, a patient may receive some of his or her dose fractions on a different Linac than the one used for the first fraction, and thus the daily or weekly CBCT images would be from separate machines. Additionally, images from different patients that are accumulated for a radiomics study are likely to come from different imagers. To determine whether these differences have an influence on the resulting texture values, we imaged the texture phantom with the CBCT imagers on 19 Linacs, including nine Elekta Linacs and ten Varian Linacs. Two scans were acquired per machine: one with the default head protocol and one with the default thoracic protocol for that machine. The standard image reconstruction was used for all scans. The characteristics of these Linacs and scans are described in Table III. Each scan was classified as a Varian head scan (V-head), Varian thorax scan (V-thorax), Elekta head scan (E-head), or Elekta thorax scan (E-thorax). For each texture feature, the absolute difference between values measured from every possible pair of scanners was calculated. These differences were then categorized by the types of scans being compared (e.g., V-thorax scan vs E-head). The differences were compared individually to the mean intrapatient difference for each feature. If the difference between two scans was less than the mean intrapatient difference, the comparison passed and the feature was considered reproducible between those two scans. The mean intrapatient difference was used as the criteria, because it was assumed that the phantom would demonstrate substantially less variation than the patients when scanned under different conditions. The overall percentage of passing scans for each comparison category was recorded. A high percentage of passing scans

TABLE II. Texture features that were used in this analysis. Abbreviated versions of the feature names are listed in parentheses next to their corresponding feature.

Histogram	Co-occurrence matrix	Run length matrix	Neighborhood difference matrix	LoG filtered features
(*Without rescaling*)	Autocorrelation (autoCorr)	Gray-level nonuniformity (glnuN)	Busyness	(*Fine filter*)
Max	Cluster-prominence (clusProm)	High gray-level run emphasis (hglre)	Coarseness	Entropy
Mean	Cluster-shade (clusShade)	Long run emphasis (lre)	Complexity	Mean
Median	Cluster-tendency (clusTend)	Long run high gray-level emphasis (lrhgle)	Contrast	Standard deviation (std)
Entropy	Contrast	Long run low gray-level emphasis (lrlgle)		Uniformity (unif)
Energy	Correlation (corr)	Low gray-level run emphasis (lglre)		Kurtosis (kurt)
Standard deviation (std)	Difference-entropy (diffEnt)	Run length nonuniformity (rlnuN)		Skewness (skew)
Uniformity (unif)	Dissimilarity (dissim)	Run percentage (runPerc)		
Kurtosis (kurt)	Energy	Short run emphasis (sre)		
Skewness (skew)	Entropy	Short run high gray-level emphasis (srhgle)		
Variance				
(*With rescaling*)	Homogeneity (homog)			(*Medium filter*)
Max	Homogeneity2 (homog2)			Entropy
Mean	Information measure correlation (infoMC)			Mean
Median	Information measure correlation2 (infoMC2)			Standard deviation (std)
Entropy	Inverse difference moment norm (invDiffMN)			Uniformity (unif)
Energy	Inverse difference norm (invDiffN)			Kurtosis (kurt)
Standard deviation (std)	Inverse variance (invVar)			Skewness (skew)
Uniformity (unif)	Max probability (maxProb)			
Kurtosis (kurt)	Sum average (sumAvg)			
Skewness (skew)	Sum entropy (sumEnt)			
Variance	Sum variance (sumVar)			
	Variance (var)			

implies that the feature is reproducible between that subset of Linacs.

2.E. Effect of scatter

CBCT image quality is largely limited by the amount of scatter created by the volume being imaged. The impact of different amounts of scatter from different sized patients on texture values is unknown. The texture phantom used in this study is relatively small and does not well approximate the amount of scatter created by a patient. To determine whether increased scatter would substantially change texture feature values, we imaged the texture phantom on its own, then with one layer (thickness of 2.5–8 cm on each side) of scatter material (solid water equivalent and sandwich size Ziploc bags of rice), and then with two layers (thickness of 5–11 cm on each side) of scatter material, Fig. 2. These three setups were imaged with both the head and thoracic protocols on a Varian Linac.

The absolute differences in each of the features for both protocols with either one layer or two layers of scatter material versus no surrounding scatter material and the difference between one layer of scatter and two layers of scatter were calculated. The log of the ratio of these differences in phantom values to the mean intrapatient differences was then calculated as the metric for this test, Eq. (2). A negative value for the log of the ratio implies the phantom differences were less than the mean intrapatient difference and passed while a positive value implies the phantom differences were larger and thus that the feature failed,

$$\log_{10}\left(\frac{\text{phantom diff}}{\text{mean intrapatient diff}}\right). \quad (2)$$

2.F. Effect of motion

A third source of uncertainty in texture values obtained from CBCT images is the effect of motion. Motion has a larger effect on values measured from CBCT images than conventional CT images because the scans take longer to acquire. To analyze the effect of this motion, we used a CIRS dynamic-motion phantom, Fig. 3 (CIRS, VA). This motion phantom has a width of 30 cm, height of 20 cm, and length of 15 cm. These thicknesses provide some scatter and were on par with the overall thicknesses created in Sec. 2.E studying only scatter. No extra scatter material was added around the phantom.

This anthropomorphic phantom has a rod made of lung-equivalent material that can be programmed to move cyclically

TABLE III. Scan characteristics for the phantom CBCT images used in this study.

ID	Manufacturer	Protocol	Image size (pixels)	Pixel size (mm)	Slice thickness (mm)	Tube voltage (kVp)	Exposure time[a] (ms)	Tube current (mA)
1	Elekta	Head	410	1.0	5.00	120	20	32
2	Elekta	Head	410	1.0	3.00	120	20	32
3	Elekta	Head	410	1.0	4.00	120	20	32
4	Elekta	Head	410	1.0	4.00	120	20	32
5	Elekta	Head	410	1.0	5.00	120	20	32
6	Elekta	Head	410	1.0	4.00	120	20	32
7	Elekta	Head	410	1.0	4.00	120	20	32
8	Elekta	Head	410	1.0	2.00	120	40	40
9	Elekta	Head	410	1.0	2.00	120	40	40
1	Elekta	Thorax	270	1.0	5.00	100	10	10
2	Elekta	Thorax	270	1.0	3.00	100	10	10
3	Elekta	Thorax	270	1.0	4.00	100	10	10
4	Elekta	Thorax	270	1.0	4.00	100	10	10
5	Elekta	Thorax	270	1.0	5.00	100	10	10
6	Elekta	Thorax	270	1.0	4.00	100	10	10
7	Elekta	Thorax	270	1.0	4.00	100	10	10
8	Elekta	Thorax	270	1.0	2.00	100	10	10
9	Elekta	Thorax	270	1.0	2.00	100	10	10
10	Varian	Head	512	0.5	2.50	100	7 300	20
11	Varian	Head	512	0.5	2.50	100	7 480	20
12	Varian	Head	384	0.7	2.50	100	7 160	20
13	Varian	Head	512	0.5	2.50	100	7 460	20
14	Varian	Head	512	0.5	2.50	100	7 560	20
15	Varian	Head	512	0.5	2.50	100	7 300	20
16	Varian	Head	512	0.5	2.50	110	13 180	20
17	Varian	Head	512	0.5	1.98	100	7 470	20
18	Varian	Head	512	0.5	1.98	100	7 350	20
19	Varian	Head	512	0.5	1.98	100	7 350	20
10	Varian	Thorax	512	0.9	2.50	110	13 160	20
11	Varian	Thorax	512	0.9	2.50	110	13 480	20
12	Varian	Thorax	384	1.2	2.50	110	12 900	20
13	Varian	Thorax	512	0.9	2.50	110	13 600	20
14	Varian	Thorax	512	0.9	2.50	110	13 660	20
15	Varian	Thorax	512	0.9	2.50	110	12 980	20
16	Varian	Thorax	512	0.9	2.50	110	13 180	20
17	Varian	Thorax	512	0.9	1.98	125	13 395	20
18	Varian	Thorax	512	0.9	1.98	125	13 275	20
19	Varian	Thorax	512	0.9	1.98	125	13 275	20

[a]For Elekta machines, the exposure time represents the pulse length, for Varian machines, the exposure time represents the full beam-on time.

with different respiratory rates through the phantom lungs. This rod has a cavity designed for the placement of different tumor-equivalent or dose measurement inserts. For our analysis, we created a block of the shredded rubber material that had the size and shape of this cavity (height of 4.5 cm and diameter of 4.1 cm). Then the phantom was programmed to move the rod using a $1-2\cos^4(t)$ waveform with peak-to-peak amplitudes of either 0, 2, 4, 6, 8, 10, 15, 20, or 25 mm. This waveform has been shown to be representative of respiratory motion.[38,39] A new CBCT scan was acquired using the same thoracic protocol on a Varian Linac for each programmed motion.

A 3D ROI encapsulating the shredded rubber material was delineated manually on the images of the phantom acquired with no motion. This ROI was copied to the images with motion. Texture values were then calculated for each image set. Equation (2) was used to determine at which amplitude of motion features ceased being reproducible. Here, the absolute difference between the texture value measured from images with motion and the values measured without motion was calculated and used as the phantom difference values in the numerator of Eq. (2). The values were compared to the mean intrapatient difference values, and as before, negative values implied passing, while positive values implied failing. This test was repeated using only the center slice of the motion phantom's original 3D ROI. The values from only the center slice were expected to be more reproducible because the average density change in the center of the image is less than at the edges, especially as the tumor motion increases. This test was done to determine if texture features could be reliably measured from tumors with large motion if the edges were excluded.

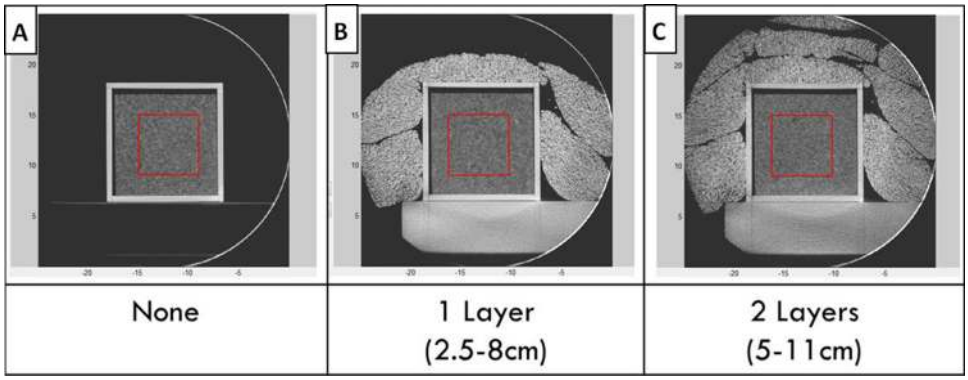

FIG. 2. To measure the effect of scatter, the texture phantom was imaged with and without surrounding scatter material.

3. RESULTS

3.A. Patient test-retest CBCT images

In the first part of this study, we examined whether any features should be excluded because they were not reproducible even when measured from two images of the same patient acquired on the same scanner within 15 min. Of the original 68 features, 23 had a CCC <0.9 and were excluded from further analysis (Table IV). Of the remaining 45 features, eight were excluded because the absolute value of their r_s with volume was greater than 0.85 for both the test and retest image sets, and thus might only be reproducible because they are volume-dependent (Table IV). Thus, a total of 37 features remained for the subsequent analyses. These included five features from the histogram without eight-bit scaling, five features from the histogram with eight-bit scaling, 16 features from the COM, eight features from the RLM, one feature from the NGTDM, and two features filtered with the LoGMedFilt.

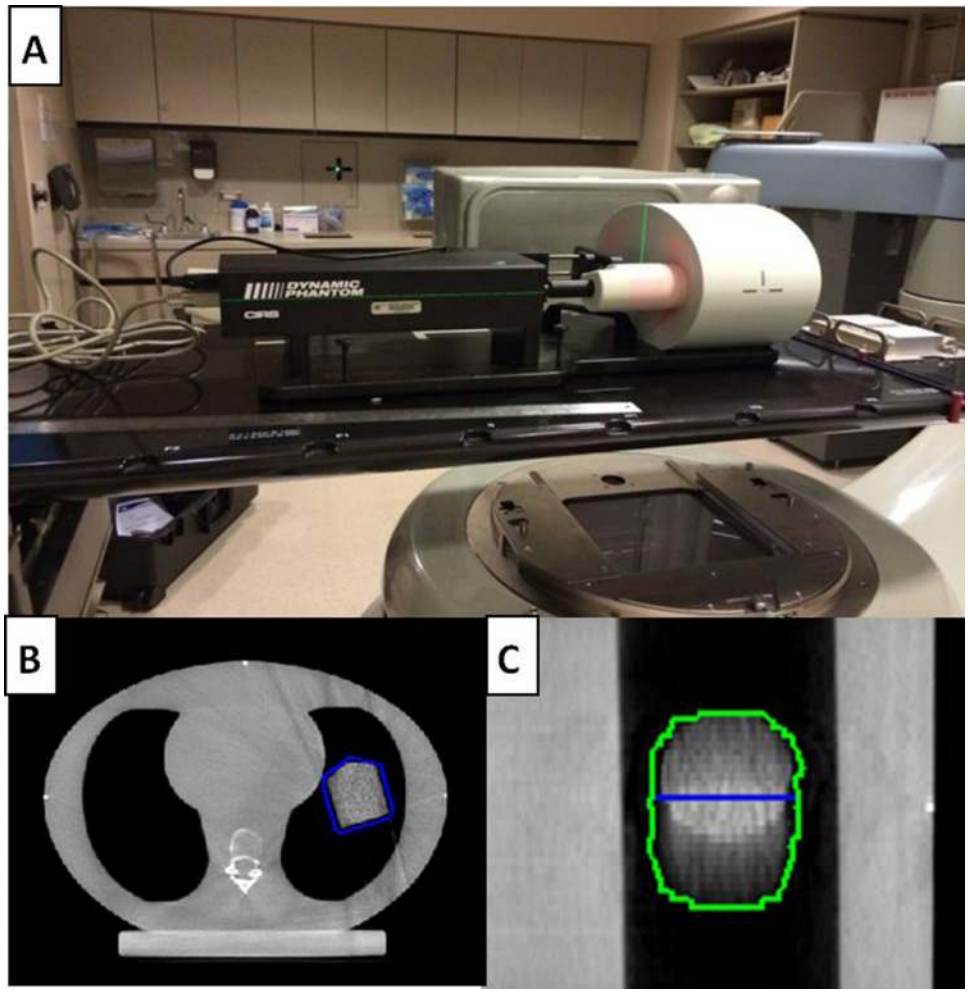

FIG. 3. (A) Setup for taking CBCT images of the CIRS dynamic-motion phantom with the shredded rubber insert in place. (B) An axial slice of the CBCT scan of the phantom with the insert visible and (C) a zoomed-in coronal slice of the insert with the largest motion of 25 mm.

TABLE IV. The results of the CCC and r_s tests for the patient test-retest data.

Feature	CCC	VolDepA	VolDepB	Feature	CCC	VolDepA	VolDepB
HistEnergy	0.992	1.000	0.988	[a]COMinvvar	0.984	0.709	0.673
[a]HistEntropy	0.926	0.624	0.539	COMmaxProb	0.531	−0.091	−0.103
[a]HistMax	0.908	0.689	0.693	[a]COMsumAvg	0.965	0.782	0.903
HistMean	0.980	0.891	0.915	[a]COMsumEnt	0.903	0.491	0.503
[a]HistMedian	0.982	0.782	0.903	[a]COMsumVar	0.980	0.745	0.745
[a]HistStd	0.963	0.370	0.467	[a]COMvar	0.981	0.491	0.406
HistUnif	0.784	−0.624	−0.588	[a]RLMglnuN	0.900	−0.758	−0.697
HistKurt	0.697	−0.176	0.200	[a]RLMhglre	0.982	0.842	0.915
[a]HistSkew	0.967	−0.661	−0.721	[a]RLMlre	0.988	0.891	0.842
ScaledHistEnergy	0.992	1.000	0.988	RLMlrhgle	0.990	0.939	0.939
ScaledHistEntropy	0.885	0.430	0.467	[a]RLMlrlgle	0.977	0.782	0.600
[a]ScaledHistMax	0.909	0.830	0.834	[a]RLMlglre	0.981	−0.830	−0.915
[a]ScaledHistMean	0.980	0.842	0.915	[a]RLMrlnuN	0.986	−0.830	−0.891
[a]ScaledHistMedian	0.965	0.841	0.902	RLMrunPerc	0.991	−0.867	−0.891
[a]ScaledHistStd	0.963	0.358	0.467	[a]RLMsre	0.984	−0.818	−0.891
ScaledHistUnif	0.866	−0.394	−0.503	[a]RLMsrhgle	0.968	0.358	0.358
ScaledHistKurt	0.728	−0.115	0.164	RLMsrlgle	0.986	−0.964	−0.964
[a]ScaledHistSkew	0.967	−0.661	−0.721	NGTDMbusyness	0.609	−0.933	−0.933
[a]COMautoCorr	0.977	0.867	0.818	[a]NGTDMcoarseness	0.943	0.750	0.667
[a]COMclusProm	0.992	0.479	0.430	NGTDMcomplexity	0.764	0.433	0.317
[a]COMclusShade	0.940	−0.370	−0.358	NGTDMcontrast	0.838	−0.867	−0.867
[a]COMclusTend	0.981	0.491	0.406	NGTDMtexstrength	0.765	−0.950	−0.950
[a]COMcontrast	0.978	−0.673	−0.636	FineFiltEntropy	0.859	0.150	0.300
COMcorr	0.966	0.939	0.903	FineFiltMean	0.774	−0.583	−0.567
[a]COMdiffEnt	0.976	−0.673	−0.564	FineFiltStd	0.703	−0.667	−0.583
[a]COMdissim	0.977	−0.733	−0.624	FineFiltUnif	0.736	−0.033	−0.217
COMenergy	0.671	0.042	−0.127	FIneFiltKurt	0.788	0.867	0.883
COMentropy	0.874	0.115	0.297	FineFiltSkew	0.602	0.850	0.917
[a]COMhomog	0.980	0.733	0.733	MedFiltEntropy	0.815	0.190	0.167
[a]COMhomog2	0.983	0.733	0.733	[a]MedFiltMean	0.969	−0.857	−0.690
[a]COMinfoMC	0.944	−0.770	−0.624	MedFiltStd	0.118	−0.548	−0.405
[a]COMinfoMC2	0.933	0.842	0.818	MedFiltUnif	0.780	−0.095	−0.238
COMinvDiffMN	0.885	0.939	0.952	[a]MedFiltKurt	0.932	0.833	0.786
COMinvDiffN	0.936	0.952	0.964	MedFiltSkew	0.895	0.810	0.762

[a]Indicates features that have passed all three of these initial tests and were used in the rest of the analysis.

3.B. Effect of different scanners

In the second part of this study, we examined whether changing the scanner or protocol resulted in changes in the texture features that were larger than the mean intrapatient difference. Features that changed less than the mean intrapatient difference passed a comparison and the overall passing percentages were recorded for each scanner/protocol combination. The results of the interscanner analysis for each feature are shown in Table V for the shredded rubber cartridge and in the supplementary material for the dense cork cartridge.[37] Features were most likely to be reproducible when scans using the same protocol and the same manufacturer were compared. This result is highlighted in Fig. 4, where the results for entropy measured from the histogram are shown as an illustrative example.

For shredded rubber, when the same protocol and manufacturer were used, three features had a passing percentage of 100%, the average across all features was 31%, and three features had a 0% passing percentage. When scans from the same manufacturer but different protocols were compared, the highest passing percentage was 90%, the average was only 19%, and 23 features had a 0% passing percentage. When scans from different manufacturers were compared, the highest passing percentage was only 36%, the average was less than 1%, and 36 features had a 0% passing percentage.

For dense cork, the results were slightly better for each category and one feature, clustershade from the co-occurrence matrix, had a 100% passing percentage for every category. When the same protocol and manufacturer were used, six features had a passing percentage of 100%, the average across all features was 43%, and two features had a 0% passing percentage. When scans from the same manufacturer but different protocols were compared, four features had a passing percentage of 100%, the average was 26%, and 13 features had a 0% passing percentage. When scans from different manufacturers were compared, two features had a passing percentage of 100%, the average was 10%, and 33 features had a 0% passing percentage.

TABLE V. Results of the interscanner variability test for the shredded rubber ROI. When categories from different manufacturers (E = Elekta and V = Varian) were compared, essentially no comparisons passed the mean intrapatient difference threshold. When both the manufacturer and protocol were the same, many of the features had a passing rate above 50%.

Feature	E-head vs E-head	E-head vs E-thorax	E-head vs V-head	E-head vs V-thorax	E-thorax vs E-thorax	E-thorax vs V-head	E-thorax vs V-thorax	V-head vs V-head	V-head vs V-thorax	V-thorax vs V-thorax
HistEntropy (%)	55.6	44.4	0.0	0.0	38.9	0.0	0.0	84.4	0.0	80.0
HistMax (%)	2.8	0.0	0.0	0.0	5.6	3.3	3.3	13.3	0.0	11.1
HistMedian (%)	16.7	0.0	0.0	0.0	11.1	0.0	0.0	8.9	10.0	8.9
HistStd (%)	44.4	33.3	0.0	0.0	22.2	0.0	0.0	20.0	0.0	28.9
HistSkew (%)	83.3	49.4	0.0	21.1	50.0	0.0	6.7	80.0	39.0	51.1
ScaledHistMax (%)	0.0	0.0	0.0	0.0	5.6	1.1	3.3	11.1	0.0	4.4
ScaledHistMean (%)	16.7	0.0	0.0	0.0	13.9	0.0	0.0	8.9	8.0	11.1
ScaledHistMedian (%)	27.8	0.0	0.0	0.0	11.1	0.0	0.0	17.8	10.0	20.0
ScaledHistStd (%)	44.4	33.3	0.0	0.0	22.2	0.0	0.0	17.8	0.0	28.9
ScaledHistSkew (%)	83.3	53.1	0.0	21.1	50.0	0.0	6.7	80.0	43.0	53.3
COMautoCorr (%)	8.3	4.9	0.0	0.0	2.8	0.0	0.0	2.2	0.0	0.0
COMclusProm (%)	44.4	38.3	0.0	0.0	25.0	0.0	0.0	2.2	0.0	4.4
COMclusShade (%)	100.0	90.1	0.0	10.0	72.2	0.0	4.4	4.4	0.0	31.1
COMclusTend (%)	44.4	37.0	0.0	0.0	19.4	0.0	0.0	15.6	0.0	13.3
COMcontrast (%)	47.2	32.1	0.0	0.0	16.7	0.0	0.0	4.4	5.0	11.1
COMdiffEnt (%)	16.7	18.5	0.0	0.0	11.1	0.0	0.0	20.0	17.0	42.2
COMdissim (%)	30.6	23.5	0.0	0.0	13.9	0.0	0.0	4.4	9.0	24.4
COMhomog (%)	16.7	16.1	0.0	0.0	11.1	0.0	0.0	28.9	25.0	42.2
COMhomog2 (%)	13.9	13.6	0.0	0.0	11.1	0.0	0.0	31.1	26.0	51.1
COMinfoMC (%)	58.3	55.6	0.0	0.0	38.9	0.0	0.0	35.6	9.0	84.4
COMinfoMC2 (%)	77.8	69.1	35.6	0.0	61.1	31.1	0.0	35.6	6.0	57.8
COMinvVar (%)	16.7	12.4	0.0	0.0	11.1	0.0	0.0	26.7	24.0	44.4
COMsumAvg (%)	19.4	12.4	0.0	0.0	2.8	0.0	0.0	13.3	0.0	6.7
COMsumEnt (%)	47.2	38.3	0.0	0.0	27.8	0.0	0.0	71.1	0.0	64.4
COMsumVar (%)	8.3	4.9	0.0	0.0	2.8	0.0	0.0	4.4	0.0	0.0
COMvar (%)	44.4	37.0	0.0	0.0	19.4	0.0	0.0	15.6	0.0	13.3
RLMglnuN (%)	61.1	49.4	0.0	0.0	44.4	0.0	0.0	100.0	0.0	91.1
RLMhglre (%)	22.2	0.0	0.0	0.0	11.1	0.0	0.0	11.1	10.0	13.3
RLMlre (%)	13.9	12.4	0.0	0.0	11.1	0.0	0.0	48.9	60.0	84.4
RLMlrlgle (%)	16.7	0.0	0.0	0.0	11.1	23.3	16.7	37.8	25.0	20.0
RLMlglre (%)	5.6	0.0	0.0	0.0	16.7	0.0	0.0	8.9	13.0	6.7
RLMrlnuN (%)	22.2	16.1	0.0	0.0	13.9	0.0	0.0	35.6	40.0	51.1
RLMsre (%)	16.7	14.8	0.0	0.0	11.1	0.0	0.0	44.4	54.0	73.3
RLMsrhgle (%)	13.9	0.0	0.0	0.0	2.8	0.0	0.0	11.1	6.0	8.9
NGTDMcoarseness (%)	36.1	27.2	0.0	0.0	22.2	0.0	0.0	86.7	56.0	100.0
MedFiltMean (%)	47.2	39.5	0.0	0.0	25.0	0.0	0.0	26.7	0.0	37.8
MedFiltKurt (%)	36.1	11.1	15.6	12.2	16.7	1.1	16.7	91.1	0.0	75.6
Features with a passing rate >50%	7	4	0	0	4	0	0	7	3	13

3.C. Effect of scatter

In the third part of this study, we examined whether adding scatter material created changes in the texture features that were larger than the mean intrapatient difference. Features that changed less than the mean intrapatient difference passed the comparison. Results from the comparisons of features calculated with and without scatter material are in Fig. 5. For the dense cork cartridge imaged with the thoracic protocol, 25 of the 37 features were reproducible with one layer of scatter material. When a second layer of scatter material was added, 16 features were still reproducible. However, for the shredded rubber cartridge imaged with the thoracic protocol, only four features were reproducible (regardless of the amount of scatter material added).

For the head protocol, only ten features from dense cork and 11 features from shredded rubber were reproducible with one layer of scatter material. These features were not consistent, and only four appeared in both groups. With two layers of scatter material, the number of reproducible features from dense cork dropped to 4, while the number of reproducible features from shredded rubber remained at 11, 9 of which were the same as before. The most reproducible feature was skewness from the histogram.

The differences between one and two layers of scatter were smaller than the patient test-retest differences for 23 of the

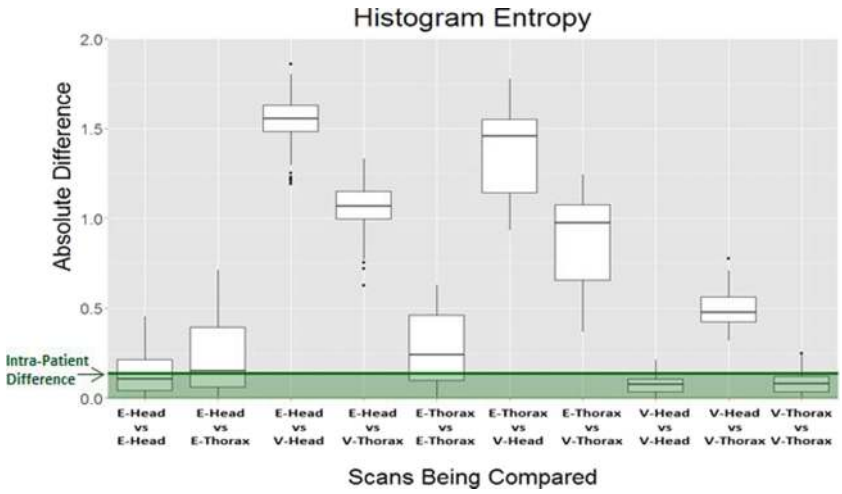

FIG. 4. The absolute differences between pairs of scans were plotted by the types of groups being compared. Scans from machines of different manufacturers had the largest absolute differences which were above our criteria of the mean intrapatient difference (the horizontal green line).

features measured from the thoracic scan of dense cork and 11 for the head scan of dense cork. For the thoracic scan of shredded rubber, only five features passed and for the head scan of shredded rubber only eight features passed.

3.D. Effect of motion

In the fourth part of this study, we examined whether adding motion produced changes in the texture features that were larger than the mean intrapatient difference. Features that changed less than the mean intrapatient difference passed this comparison. The number of features that were reproducible decreased with increasing motion amplitude, Fig. 5. Three features: LoG filtered kurtosis, gray-level nonuniformity from the RLM, and entropy from the histogram, were reproducible for motions of 6–8 mm when measured from the entire volume. At 4 mm of motion, 12 of the 37 features were reproducible for the entire volume. When only the center image slice was used for feature calculation, seven features were reproducible for up to 6–10 mm of motion. The most consistent features measured from only the center slice were coarseness from the NGTDM, high gray-level run emphasis and gray-level nonuniformity from the RLM, sum-average and information measure correlation from the COM, and scaled mean and entropy from the histogram. At 4 mm of motion, 14 of the 37 features were reproducible for the center slice measurements.

4. DISCUSSION

4.A. Patient test-retest

The goal of this study was to determine whether any texture feature can be reproducibly measured from CBCT images so that features could be tracked periodically through treatment. In order to investigate this question, we initially considered a large number of features. Two tests were used to eliminate features that were not reproducible in a patient test-retest dataset or that were only reproducible due to their volume dependence. Features that were not reproducible even for the same patient on the same machine are extremely unlikely to be useful in future models and could lead to erroneous results. Features that are volume dependent will appear to be reproducible especially in patient datasets where the volume range is large. However, because volume is already known to be prognostic and is easy to extract from patient images without the rigor of a texture analysis, volume-dependent features would not add meaningful information to future models and could have led to misleading results in this investigative study. Approximately half the features initially considered passed both of these qualifying tests. Interestingly, at least one feature from every feature category was successful in passing these tests. The large number and wide variety of features that passed offer preliminary support for the possibility of texture analysis in CBCT images. This relatively high pass rate occurred despite the strict criteria for reproducibility (CCC $\geq$ 0.9). We deliberately adopted very strict and conservative cutoffs here in order to minimize the possibility of false-positives in this analysis. Many of the excluded features had CCC values in the 0.75–0.89 range representing medium reproducibility and thus we may have excluded features that could potentially be useful in the future.

In order to determine the effects of scanner, scatter, and motion on this reproducibility, phantom measurements were compared to the mean intrapatient difference for each feature. This choice of threshold is one limitation of our study since it is partly arbitrary, and may not accurately reflect the amount of variability in a texture that would significantly influence an eventual prognostic model. However, because the purpose of this paper was only to introduce the magnitudes of variability that are created by changes in scanner, motion, and scatter, we feel that our choice of threshold is justified. Furthermore, by using the mean of the differences measured from patient test-retest data rather than the maximum, it is more likely that our choice of threshold is overly conservative than lenient.

Another limitation of this part of the analysis was the small number of patients with available repeat images. A larger patient dataset may have increased the variability we saw in

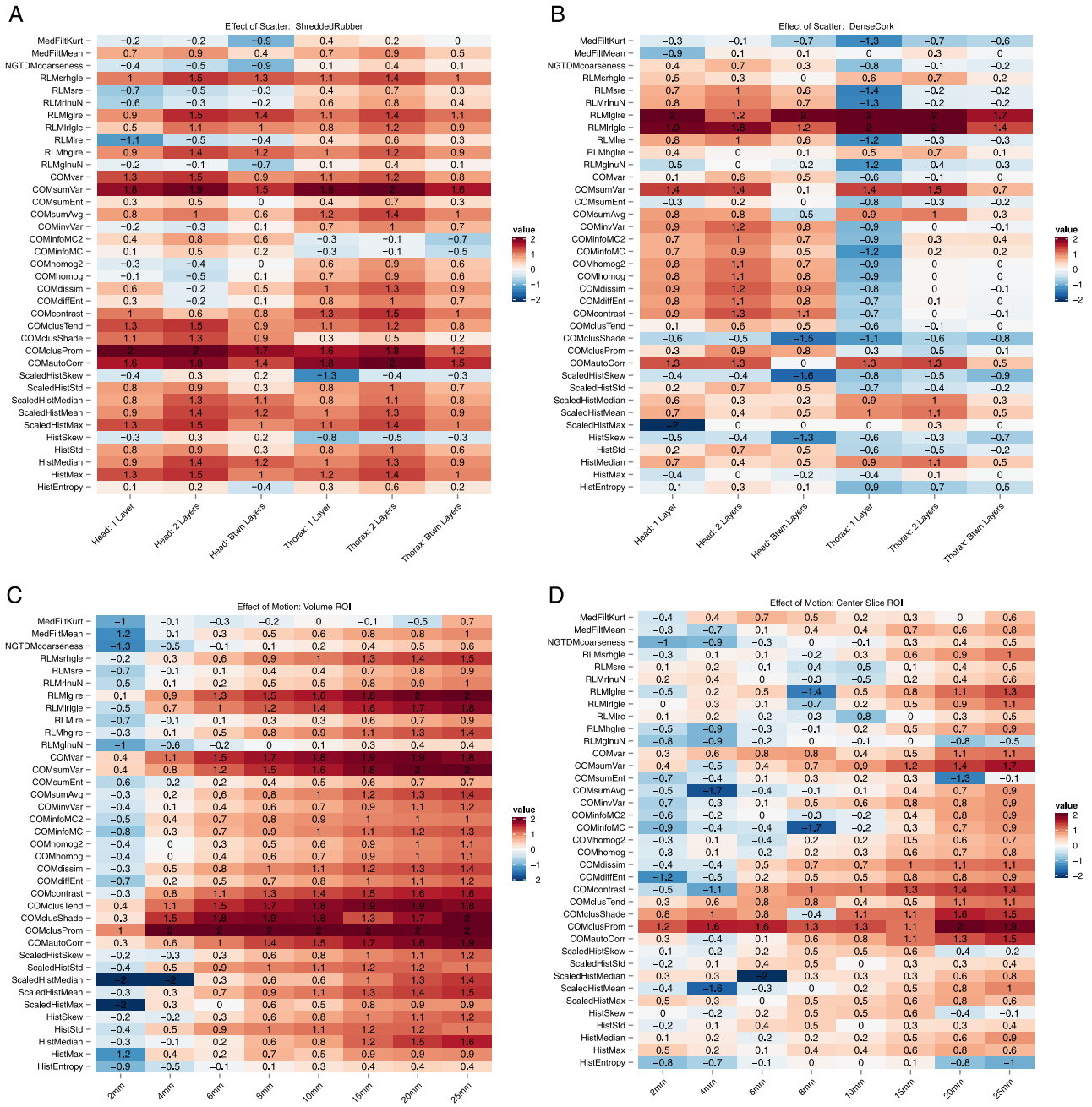

FIG. 5. Results for the scatter [(A) and (B)] and motion tests [(C) and (D)]. Negative values imply smaller differences in the phantom measurements than in the patient test-retest values and thus a "pass." Positive values imply that the phantom difference was larger than the mean intrapatient difference and thus a "fail." For the scatter graphs, comparisons are described by protocol (head or thorax) and the amount of surrounding scatter material. For the motion graphs, comparisons are between different peak-to-peak amplitudes of motion (2–25 mm) and no motion.

patient test-retest values which in turn may have increased the number of features that passed each test.

4.B. Interscanner analysis

The results of our interscanner analysis strongly indicated that texture values obtained from different imaging protocols or different Linac manufacturers should not be compared. This is a useful result for anyone considering extracting texture features from CBCT images in order to produce a model. The Elekta values may have differed from the Varian values because of manufacturer differences in Hounsfield unit scaling. CBCT pixel values tend to be less accurate than the pixels in regular CT images because CBCT images are not used for dose calculation. Thus, differences between manufacturers in HU mapping from CBCT images could play a role in the observed differences seen between scans of the same phantom.

The interscanner analysis also revealed that more features were likely to pass when measured from the dense cork cartridge than the shredded rubber cartridge. This is likely because the dense cork cartridge is physically more uniform than the shredded rubber cartridge. For the same scan, the dense cork standard deviation was typically one-half to

one-third the value of the shredded rubber standard deviation. Thus, even when magnitude shifts or varying levels of noise are introduced by using a different scanner or protocol, the dense cork cartridge individual voxels are less varied than those of the shredded rubber cartridge. The patients' standard deviations fell within the range of both cartridges so the values from the shredded rubber cartridge can be assumed to approximate the variability in a patient with heterogeneous texture while the dense cork cartridge may approximate a patient with homogenous texture.

Some features were not reproducible even when both the manufacturer and protocol were kept consistent. These features may be overpowered by the noise in the image making them essentially random. This is probably the reason why features such as the maximum value from the histogram failed each comparison. In other cases, such as low gray-level run emphasis from the RLM, the texture values from the patient and phantom images are essentially always zero because the feature searches for specific patterns that do not exist in images of tumors (such as straight lines).

This analysis also demonstrated that reproducible features could come from any of the feature categories, e.g., skewness from the histogram, cluster shade from the co-occurrence matrix, normalized gray level nonuniformity from the run length matrix, and the mean after LoG filtration. This broad spectrum of reproducible features is helpful, because features from different categories may provide independent information about an image and when combined, may be able to provide a more complete picture than one feature alone.

4.C. Effect of scatter

When the texture phantom was surrounded by scatter material, most of the texture values changed more than the mean of the patient test-retest differences. This result was not surprising, as we know that more surrounding material will result in more scatter and thus a larger amount of noise in the image as well as artifacts from beam hardening and cupping. The differences between 1 layer and 2 layers of scatter material were also in general larger than the mean intrapatient difference. This is a problem because it suggests that two patients with physiologically alike tumors (i.e., similar levels of heterogeneity) could have very different values for their computed textures if the patients are dissimilar in size. The impact of this problem may be limited because texture features measured from CBCT images would only be used to observe how the features change for a single patient over time. For that analysis, the relative difference in a texture value could be measured for each patient. The change in the amount of scatter would likely be substantially less than shown here if each patient acted as his or her own control. A recent study investigating c-arm CBCT demonstrated that relative changes in mean Hounsfield units were consistent when measured within patients.[40] Thus, it is possible that relative changes in texture may still be used for future prognostic models despite the effect of changes in scatter levels on the absolute value measured from any one patient.

4.D. Effect of motion

Most of the features changed substantially with increasing motion of the tumor texture insert. The main reason for this result was hypothesized to be the slices of the ROI at the edge of the texture insert, where the density changes were greatest. This hypothesis was supported by our data for many of the features which did not significantly change with small motion if only the center slice was used for their calculation.

While a majority of features was no longer reproducible beyond 2 mm of motion when the entire tumor volume was used, 12 of the 37 features did still pass at 4 mm of motion and 4 of these even passed at 6 mm of motion. The number of reproducible features dropped to zero at 10 mm when the entire tumor volume was used and to 1 when only the center slice was used. Therefore, we recommend a threshold of at most 10 mm and potentially as low as 5 mm for future studies. This threshold is not unduly restrictive since a recent study showed a majority of patients with NSCLC had tumor motion less than 5 mm and only 10% had motion greater than 10 mm.[41] Thus, we think a future study limited to patients with little motion and selecting only these most reproducible features for further investigation would be feasible. Additionally, either 4D CBCT or breath-hold CBCT could be used to mitigate tumor motion in future studies and may be more successful than shrinking the tumor contour.

Several features, when measured from the center slice, were reproducible at large motions while not being reproducible at smaller motions. For example, sumEntropy from the co-occurrence matrix was reproducible with 2–4 mm and 20–25 mm but not with 6–15 mm of motion. This inconsistency suggests that at large motions the feature may be returning reproducible values by coincidence or because of artifacts. Thus, we would not consider this feature reproducible beyond 4 mm.

It should also be noted that while the motion phantom was larger than the texture phantom (32 vs 10 cm diameter), it is still smaller than many patients. Thus, in a clinically realistic scenario the effects of motion and additional scatter would be combined and may further reduce the number of features or the range of motion that could be considered reproducible.

4.E. Overall best performing features

From our results, it appears that select features are reproducible under certain circumstances. Several of these reproducible features have been found useful in studies using CT or contrast-enhanced CT images. One study found skewness, which we showed to be robust to scatter, may aid in identifying tumors with genetic mutations[10] while another study demonstrated it was prognostic for overall survival.[42] LoG filtered kurtosis was useful for identifying tumors with genetic mutations[10] and we showed it was robust to scatter and motion. Gray-level nonuniformity from the run-length matrix was the feature we tested that was most robust to the effects of motion and it has been shown to be useful for predicting survival in NSCLC (Ref. 17) and differentiating between

benign and malignant lymph nodes.[14] Cluster shade from the co-occurrence matrix was able to pass all of the scatter tests for the dense cork material, and recently was shown to be useful for prognosis when used in a radiomics signature of three features for NSCLC patients.[8] These links are encouraging but an independent study will still be needed to determine if models built on CBCT features alone can be prognostic. However, it should be clear that the features which did change more dramatically when measured from a phantom than the calculated mean intrapatient differences in our study are unlikely to be useful in future analyses using CBCT images of NSCLC unless the patient cohort was highly restricted to patients exhibiting low intrascan variability (e.g., negligible tumor motion and minimal weight change).

5. CONCLUSION

The goal of this study was to determine if texture features could be reliably extracted from CBCT images under a variety of conditions. A total of 68 features was originally considered. However, 31 of these features were excluded from our analysis because they did not have a high CCC value when measured from a test-retest dataset or had a strong volume dependence that might be responsible for their high CCC. The remaining 37 features included at least one feature from each feature category that had been studied. These features were then investigated for susceptibilities to differences in scanners, imaging protocols, scatter, and motion. Features changed significantly if they were calculated from images acquired with different protocols or with scanners from different manufacturers. Future studies should attempt to keep their imaging protocols as uniform as possible to avoid this source of error. Almost every feature changed more than the mean intrapatient difference with the addition of scatter. Thus, values of features may not be comparable between patients of different sizes while remaining insensitive to small changes in size of each individual patient. Finally, no feature can be reliably measured if the tumor motion is greater than 1 cm. For motion less than 1 cm, reproducibility is improved if the edges of the tumor are excluded from the ROI for texture calculation. In summary, certain texture features can be reliably measured from CBCT images as long as the imaging protocol is consistent, relative differences are used, and patients are limited to those with less than 1 cm of tumor motion.

ACKNOWLEDGMENTS

The authors would like to acknowledge the medical physics residents at Mary Bird Perkins Cancer Center for their help acquiring many of the texture phantom scans; Kelly Thorp for his help shaping the texture insert for the motion phantom; Ramesh Tailor, Ph.D. for the loan of rice as scatter material for the scans, and Kathryn Carnes from the Department of Scientific Publications for her help in revising this manuscript. This project was funded in part by a grant from the NCI (No. R03CA178495-01). Xenia Fave is a recipient of the AAPM & RSNA graduate fellowship.

[a]Author to whom correspondence should be addressed. Electronic mail: xjfave@mdanderson.org

[1]National Cancer Institute, *SEER Stat Fact Sheets: Lung and Bronchus Cancer* (National Cancer Institute, Bethesda, 2014), http://seer.cancer.gov/statfacts/html/lungb.html.

[2]J. R. Molina, P. Yang, S. D. Cassivi, S. E. Schild, and A. A. Adjei, "Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship," Mayo Clin. Proc. **83**(5), 584–594 (2008).

[3]T. Kishi, Y. Matsuo, N. Ueki, Y. Iizuka, A. Nakamura, K. Sakanaka, T. Mizowaki, and M. Hiraoka, "Pretreatment modified Glasgow prognostic score predicts clinical outcomes after stereotactic body radiation therapy for early-stage non-small cell lung cancer," Int. J. Radiat. Oncol., Biol., Phys. **92**(3), 619–626 (2015).

[4]S. K. Jabbour, S. Kim, S. A. Haider, X. Xu, A. Wu, S. Surakanti, J. Aisner, J. Langenfeld, N. J. Yue, B. G. Haffty, and W. Zou, "Reduction in tumor volume by cone-beam computed tomography predicts overall survival in non-small cell lung cancer treated with chemoradiation therapy," Int. J. Radiat. Oncol., Biol., Phys. **92**(3), 627–633 (2015).

[5]M. G. Kris, B. E. Johnson, L. D. Berry, D. J. Kwiatkowski, A. J. Iafrate, I. I. Wistuba, M. Varella-Garcia, W. A. Franklin, S. L. Aronson, P. Su, Y. Shyr, D. R. Camidge, L. V. Sequist, B. S. Glisson, F. R. Khuri, E. B. Garon, W. Pao, C. Rudin, J. Schiller, E. B. Haura, M. Socinski, K. Shirai, H. Chen, G. Giaccone, M. Ladanyi, K. Kugler, J. D. Minna, and P. A. Bunn, "Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs," JAMA, J. Am. Med. Assoc. **311**(19), 1998–2006 (2014).

[6]B. Ganeshan, E. Panayiotou, K. Burnand, S. Dizdarevic, and K. Miles, "Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: A potential marker of survival," Eur. Radiol. **22**, 796–802 (2012).

[7]T. Win, K. A. Miles, S. M. Janes, B. Ganeshan, M. Shastry, R. Endozo, M. Meagher, R. I. Shortman, S. Wan, I. Kayani, P. J. Ell, and A. M. Groves, "Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non-small cell lung cancer," Clin. Cancer Res. **19**(13), 3591–3599 (2013).

[8]T. P. Coroller, P. Grossmann, Y. Hou, E. Rios velazquez, R. T. H. Leijenaar, G. Hermann, P. Lambin, B. Haibe-Kains, R. H. Mak, and H. J. W. L. Aerts, "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma," Radiother. Oncol. **114**(3), 345–350 (2015).

[9]O. Grove, A. E. Berglund, M. B. Schabath, H. J. W. L. Aerts, A. Dekker, H. Wang, E. R. Velazquez, P. Lambin, Y. Gu, Y. Balagurunathan, E. Eikman, R. A. Gatenby, S. Eschrich, and R. J. Gillies, "Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma," PLoS One **10**(3), e0118261 (2015).

[10]G. J. Weiss, B. Ganeshan, K. A. Miles, D. H. Campbell, P. Y. Cheung, S. Frank, and R. L. Korn, "Noninvasive image texture analysis differentiates K-ras mutation from pan-wildtype NSCLC and is prognostic," PLoS One **9**(7), e100244 (2014).

[11]Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, L. O. Hall, R. A. Gatenby, and R. J. Gillies, "Reproducibility and prognosis of quantitative features extracted from CT images," Transl. Oncol. **7**(1), 72–87 (2014).

[12]D. V. Fried, S. L. Tucker, S. Zhou, Z. Liao, O. Mawlawi, G. Ibbott, and L. E. Court, "Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer," Int. J. Radiat. Oncol., Biol., Phys. **90**(4), 834–842 (2014).

[13]H. Wang, X. H. Guo, Z. W. Jia, H. K. Li, Z. G. Liang, K. C. Li, and Q. He, "Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image," Eur. J. Radiol. **74**(1), 124–129 (2010).

[14]H. Bayanati, R. E. Thornhill, C. A. Souza, V. Sethi-Virmani, A. Gupta, D. Maziak, K. Amjadi, and C. Dennie, "Quantitative CT texture and shape analysis: Can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer?," Eur. Radiol. **25**(2), 480–487 (2015).

[15]S. Basu, L. O. Hall, D. B. Goldgof, Y. Gu, V. Kumar, J. Choi, R. J. Gillies, and R. A. Gatenby, "Developing a classifier model for lung tumors in CT-scan images," in *2011 IEEE International Conference on Systems, Man, and Cybernetics* (IEEE, Anchorage, AK, 2011), pp. 1306–1312.

[16]A. Cunliffe, S. G. Armato, R. Castillo, N. Pham, T. Guerrero, and H. A. Al-Hallaq, "Lung texture in serial thoracic computed tomography scans: Correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development," Int. J. Radiat. Oncol., Biol., Phys. **91**(5), 1048–1056 (2015).

[17]H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," Nat. Commun. **5**(4006), 1–8 (2014).

[18]B. Ganeshan, V. Goh, H. C. Mandeville, P. J. Hoskin, and K. A. Miles, "Non–small cell lung cancer: Histopathologic correlates for texture," Radiology **266**(1), 326–336 (2013).

[19]O. S. Al-Kadi and D. Watson, "Texture analysis of aggressive and nonaggressive lung tumor CE CT images," IEEE Trans. Biomed. Eng. **55**, 1822–1830 (2008).

[20]G. J. R. Cook, C. Yip, M. Siddique, V. Goh, S. Chicklore, A. Roy, P. Marsden, S. Ahmad, and D. Landau, "Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy?," J. Nucl. Med. **54**(1), 19–26 (2013).

[21]M. Machtay, F. Duan, B. A. Siegel, B. S. Snyder, J. J. Gorelick, J. S. Reddin, R. Munden, D. W. Johnson, L. H. Wilf, A. Denittis, N. Sherwin, S. Kim, G. Videtic, D. R. Neumann, R. Komaki, H. Macapinlac, J. D. Bradley, and A. Abass, "Prediction of survival by [18F]fluorodeoxyglucose positron emission tomography in patients with locally advanced non-small-cell lung cancer undergoing definitive chemoradiation therapy: Results of the ACRIN 6668/RTOG 0235 trial," J. Clin. Oncol. **31**(30), 3823–3830 (2013).

[22]D. V. Fried, O. Mawlawi, L. Zhang, X. Fave, S. Zhou, S. Tucker, G. Ibbott, Z. Liao, and L. E. Court, "Stage III non-small cell lung cancer: Prognostic value of FDG-PET quantitative imaging features combined with clinical prognostic factors," Radiology (E-pub ahead of print).

[23]MD Anderson Cancer Center, *Image-Guided Adaptive Conformal Photon Versus Proton Therapy* (National Library of Medicine, Bethesda, MD, 2000), https://clinicaltrials.gov/ct2/show/record/NCT00915005.

[24]K. S. C. Chao, S. Bhide, H. Chen, J. Asper, S. Bush, G. Franklin, V. Kavadi, V. Liengswangwong, W. Gordon, A. Raben, J. Strasser, C. Koprowski, S. Frank, G. Chronowski, A. Ahamad, R. Malyapa, L. Zhang, and L. Dong, "Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach," Int. J. Radiat. Oncol., Biol., Phys. **68**(5), 1512–1521 (2007).

[25]H. Wang, L. Dong, M. F. Lii, A. L. Lee, R. de Crevoisier, R. Mohan, J. D. Cox, D. A. Kuban, and R. Cheung, "Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy," Int. J. Radiat. Oncol., Biol., Phys. **61**(3), 725–735 (2005).

[26]G. B. McBride, "A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient," NIWA Client Report No. HAM2005–062 (2005), pp. 1–14.

[27]L. A. Hunter, S. Krafft, F. Stingo, H. Choi, M. K. Martel, S. F. Kry, and L. E. Court, "High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images," Med. Phys. **40**(12), 121916 (12pp.) (2013).

[28]K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," Radiology **227**(3), 617–622 (2003).

[29]M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," Malawi Med. J. **24**(3), 69–71 (2012).

[30]D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones, and L. Court, "Measuring computed tomography scanner variability of radiomics features," Invest. Radiol. **50**(11), 757–765 (2015).

[31]B. Ganeshan, S. Abaleke, R. C. D. Young, C. R. Chatwin, and K. A. Miles, "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).

[32]O. Gevaert, J. Xu, C. D. Hoang, A. N. Leung, Y. Xu, A. Quon, D. L. Rubin, S. Napel, and S. K. Plevritis, "Non–small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—Methods and preliminary results," Radiology **264**(2), 387–396 (2012).

[33]L. Zhang, D. Fried, X. Fave, L. Hunter, J. Yang, and L. E. Court, "IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics," Med. Phys. **42**(3), 1341–1353 (2015).

[34]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Trans. Syst., Man, Cybern. **3**(6), 610–621 (1973).

[35]M. M. Galloway, "Texture analysis using gray level run lengths," Comput. Graphics Image Process. **4**(2), 172–179 (1975).

[36]M. Amadasun and R. King, "Textural features corresponding to textural properties," IEEE Trans. Syst., Man, Cybern. **19**(5), 1264–1274 (1989).

[37]See supplementary material at http://dx.doi.org/10.1118/1.4934826 for feature calculation parameters and the interscanner results of the dense cork cartridge.

[38]A. E. Lujan, E. W. Larsen, J. M. Balter, and R. K. Ten Haken, "A method for incorporating organ motion due to breathing into 3D dose calculations," Med. Phys. **26**(5), 715–720 (1999).

[39]Y. Seppenwoolde, H. Shirato, K. Kitamura, S. Shimizu, M. van Herk, J. V. Lebesque, and K. Miyasaka, "Precise and real-time measurement of 3D tumor motion in lung due to breathing and heartbeat, measured during radiotherapy," Int. J. Radiat. Oncol., Biol., Phys. **53**(4), 822–834 (2002).

[40]A. K. Jones and A. Mahvash, "Evaluation of the potential utility of flat panel CT for quantifying relative contrast enhancement," Med. Phys. **39**(7), 4149–4154 (2012).

[41]H. H. Liu, P. Balter, T. Tutt, B. Choi, J. Zhang, C. Wang, M. Chi, D. Luo, T. Pan, S. Hunjan, G. Starkschall, I. Rosen, K. Prado, Z. Liao, J. Chang, R. Komaki, J. D. Cox, R. Mohan, and L. Dong, "Assessing respiration-induced tumor motion and internal target volume using four-dimensional computed tomography for radiotherapy of lung cancer," Int. J. Radiat. Oncol., Biol., Phys. **68**(2), 531–540 (2007).

[42]S. Y. Ahn, C. M. Park, S. J. Park, H. J. Kim, C. Song, S. M. Lee, H. P. Mcadams, and J. M. Goo, "Prognostic value of computed tomography texture features in non–small cell lung cancers treated with definitive concomitant chemoradiotherapy," Invest. Radiol. **50**(10), 719–725 (2015).