High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images

Luke A. Hunter and Shane Krafft
Department of Radiation Physics, The University of Texas MD Anderson Cancer Centre, 1515 Holcombe, Houston, Texas 77030

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

Haesun Choi
Department of Diagnostic Radiology, The University of Texas MD Anderson Cancer Centre, 1515 Holcombe, Houston, Texas 77030

Mary K. Martel, Stephen F. Kry, and Laurence E. Court
Department of Radiation Physics, The University of Texas MD Anderson Cancer Centre, 1515 Holcombe, Houston, Texas 77030

(Received 10 August 2013; revised 23 October 2013; accepted for publication 24 October 2013; published 2 December 2013)

Purpose: For nonsmall cell lung cancer (NSCLC) patients, quantitative image features extracted from computed tomography (CT) images can be used to improve tumor diagnosis, staging, and response assessment. For these findings to be clinically applied, image features need to have high intra and intermachine reproducibility. The objective of this study is to identify CT image features that are reproducible, nonredundant, and informative across multiple machines.

Methods: Noncontrast-enhanced, test-retest CT image pairs were obtained from 56 NSCLC patients imaged on three CT machines from two institutions. Two machines (“M1” and “M2”) used cine 4D-CT and one machine (“M3”) used breath-hold helical 3D-CT. Gross tumor volumes (GTVs) were semiautomatically segmented then pruned by removing voxels with CT numbers less than a prescribed Hounsfield unit (HU) cutoff. Three hundred and twenty eight quantitative image features were extracted from each pruned GTV based on its geometry, intensity histogram, absolute gradient image, co-occurrence matrix, and run-length matrix. For each machine, features with concordance correlation coefficient values greater than 0.90 were considered reproducible. The Dice similarity coefficient (DSC) and the Jaccard index (JI) were used to quantify reproducible feature set agreement between machines. Multimachine reproducible feature sets were created by taking the intersection of individual machine reproducible feature sets. Redundant features were removed through hierarchical clustering based on the average correlation between features across multiple machines.

Results: For all image types, GTV pruning was found to negatively affect reproducibility (reported results use no HU cutoff). The reproducible feature percentage was highest for average images (M1 = 90.5\%, M2 = 94.5\%, M1\cap M2 = 86.3\%), intermediate for end-exhale images (M1 = 75.0\%, M2 = 71.0\%, M1\cap M2 = 52.1\%), and lowest for breath-hold images (M3 = 61.0\%). Between M1 and M2, the reproducible feature sets generated from end-exhale images were relatively machine-sensitive (DSC = 0.71, JI = 0.55), and the reproducible feature sets generated from average images were relatively machine-insensitive (DSC = 0.90, JI = 0.87). Histograms of feature pair correlation distances indicated that feature redundancy was machine-sensitive and image type sensitive. After hierarchical clustering, 38 features, 28 features, and 33 features were found to be reproducible and nonredundant for M1\cap M2 (average images), M1\cap M2 (end-exhale images), and M3, respectively. When blinded to the presence of test-retest images, hierarchical clustering showed that the selected features were informative by correctly pairing 55 out of 56 test-retest images using only their reproducible, nonredundant feature set values.

Conclusions: Image feature reproducibility and redundancy depended on both the CT machine and the CT image type. For each image type, the authors found a set of cross-machine reproducible, nonredundant, and informative image features that would be useful for future image-based models. Compared to end-exhale 4D-CT and breath-hold 3D-CT, average 4D-CT derived image features showed superior multimachine reproducibility and are the best candidates for clinical correlation.

© 2013 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution 3.0 Unported License. [http://dx.doi.org/10.1118/1.4829514]

Key words: quantitative image features, lung cancer, reproducibility
1. INTRODUCTION

In computed tomography (CT) imaging, nonsmall cell lung cancer (NSCLC) tumors are often described using qualitative descriptors (spiculated, cavitated, heterogeneous, etc.). However, recently there has been a move toward generating additional quantitative, objective features to describe these tumors. Traditional quantitative measures have included the one-dimensional tumor size, mean voxel intensity, intensity standard deviation, etc. Newer quantitative measures are usually statistical, model, or transform based and reflect variations in tumor morphology, heterogeneity, and/or texture. Motivated by the hypothesis that these quantitative image features are related to gene expression and phenotype, the field of radiomics aims to achieve their automated, high-throughput extraction from standard of care images. Ganeshan et al. have shown that quantitative measures of NSCLC CT heterogeneity are associated with tumor stage, metabolism, hypoxia, and angiogenesis. Other quantitative image features have been applied to classify NSCLC as adenocarcinoma or squamous cell carcinoma, and lung tumor image fractal dimension has been shown to correlate with tumor stage and can classify pulmonary nodules. Thus, several lines of evidence support the radiomics hypothesis for NSCLC.

For NSCLC radiomics to be applied clinically, the intra and intermachine reproducibility of image features must be addressed. Toward this goal, the Quantitative Imaging Biomarker Alliance has created a technical committee dedicated toward multivendor CT scan standardization for the measurement of lung nodules. In a CT phantom, Ganeshan et al. showed that quantitative measures of entropy and uniformity have lower coefficient of variation values than the mean intensity. For human study, public data are available from 32 NSCLC patients each scanned twice (15 min apart) using the same CT machine and protocols (Fig. 1). Kumar et al. recently used this so-called “test-retest” dataset to identify 39 out of 327 image features that were reproducible, informative, and nonredundant.

The purpose of our study is to examine reproducibility with multiple machines and image types. By studying different image types (cine 4D-CT end-exhale phase vs average cine 4D-CT vs breath-hold helical 3D-CT), we can determine which image type will give the most reproducible image features. By studying different machines, we can also investigate how much machine variation affects reproducibility. Additionally, we propose a way to integrate reproducibility and redundancy findings across multiple machines to identify a cross-machine reproducible and nonredundant image feature set. This robust set of NSCLC image features could be useful for the development and validation of multimachine and multiinstitutional NSCLC radiomics models.

2. MATERIALS AND METHODS

2.A. Test-retest datasets

We obtained noncontrast-enhanced cine 4D-CT test-retest scans from 31 patients with NSCLC who were imaged under a prospective, IRB-approved study at the University of Texas M.D. Anderson (MDA) Cancer Center in Houston, TX. Sixteen of these test-retest pairs were acquired on a GE (General Electric Healthcare, Milwaukee, WI) Discovery ST model machine (“M1”) and 15 were acquired on a GE LightSpeed RT 16 model machine (“M2”). Imaging parameters are shown in Table I. We also obtained breath-hold, noncontrast-enhanced helical 3D-CT test-retest scans from 25 patients with NSCLC who were imaged at the Memorial Sloan-Kettering Cancer Center in New York, NY. This is the dataset used by Kumar et al. and is publically available as part of the Reference Image Database to Evaluate Therapy Response (RIDER) (Ref. 16) collection hosted by the Cancer Imaging Archive (TCIA) at Washington University in Saint Louis, MO.

The average tumor volume was 63.1 cm³ (standard deviation: 79.5, range: 1.5–296 cm³), and 28.5 cm³ (standard deviation: 38.7 cm³, range: 0.5–163 cm³) and for the MD Anderson and RIDER datasets, respectively. The tumors were solid, with cavitation visible in 12/31, and 8/25 tumors. The distribution of tumor location apex/mid/base was 9/12/10 for the MD Anderson dataset and RIDER dataset, respectively. The tumors were solid, with cavitation visible in 12/31, and 8/25 tumors. The distribution of tumor location apex/mid/base was 9/12/10 for the MD Anderson dataset and RIDER dataset, respectively.

The complete RIDER dataset has 32 test-retest patients and was collected with two machines: a GE LightSpeed 16 and a
GE LightSpeed VCT 64. However, to isolate machine variation we omitted test-retest pairs from the GE LightSpeed VCT 64 machine; we also removed scans with GTVs that had image artifacts or appeared too difficult to contour accurately. This resulted in 25 test-retest pairs all acquired on the same machine, a GE LightSpeed 16 (“M3”). These scans had variable axial pixel dimensions because the field of view was independently set for each scan based on its scout image. Therefore, a “virtual” fourth machine (“M4”) was created by taking each M3 scan and performing 3D cubic spline interpolation to generate a new scan with the same voxel size as the MDA images. This was done to test if features could be extracted more reproducibly from scans interpolated to a uniform voxel size.

2.B. Feature extraction

To minimize interoperator contour variation, a single operator\(^1\) was responsible for contouring the 174 GTVs used in our study. All contouring was performed with the Pinnacle\(^3\) treatment planning system (TPS; Philips Medical Systems, Andover, MA) using its slice by slice auto-contour function with a lower bound of 0 and an upper bound of 500 (values selected to increase automation and reproducibility). For 4D-CT scans, GTVs were contoured on the T50 phase (end-exhale) images and the average (AVG) images. Clinical treatment plans with physician-generated GTVs were used to identify lesion locations. For the 3D-CT scans, lesions were located and contoured using dataset annotations from the TCIA website. In the interpolated dataset, the binary mask from original voxel size scans underwent a 3D cubic spline interpolation and thresholding to determine the new GTV mask.

For quantitative image feature extraction, we developed the in-house imaging biomarker explorer (IBEX) software package. IBEX was developed using MATLAB (Mathworks Inc., Natick, MA) and has a graphical user interface which can directly access images and regions of interest (ROIs) from the Pinnacle\(^3\) TPS file system. Using this information, it can extract image features from the following 3D feature sources: geometry, intensity histogram, absolute gradient image, co-occurrence matrix (COM),\(^19\) and run-length matrix (RLM).\(^20\) For a complete list of features extracted, refer to Table II. Additionally, for NSCLC GTVs, it is common to remove voxels from the GTV region of interest that are below a certain Hounsfield unit (HU) cutoff before performing feature extraction.\(^9,10\) The goal of thresholding the GTV, which includes some lung tissue surrounding the tumor, is to make it more reproducible by focusing feature extraction on tissues other than air. IBEX allows for the specification of an HU cutoff to support this capability.

In our analysis, we used IBEX to extract image features from each GTV using the four 3D sources described. RLM features were extracted for 13 different 3D directions, and COM features were extracted for 3 different 3D displacements. For many features, the resulting values can vary substantially depending on the bit depth used to represent the image. Therefore, in order to get a representative sampling of feature values, all absolute gradient image, RLM, and COM features were extracted with three different image bit depths: 2, 6, and 10. The intensity histogram features were extracted using a bit depth of 12, and the geometric features are unaffected by bit depth. Using these parameters, for every GTV, 328 image features were extracted for 11 different HU cutoffs that were equally spaced between \(-1000\) and 0 HU.

The workflow in IBEX is as follows:

1. Select the location where the contoured images are saved. This can be an Institution in Pinnacle, or a folder location containing DICOM CT images and plan data.
2. Select the first patient.
3. Select the correct treatment plan. At this point the CT is displayed (axial slice), together with a list of available ROIs.
4. Select the appropriate ROI (e.g., the GTV). At this point, the central CT slice for that ROI is displayed, together with its contours. The user can scroll through the image, checking the ROI. This is an important QA step, ensuring that the correct ROI is chosen, and that the contours are reasonable. If the contours are not suitable, or an incorrect ROI is chosen, the user can...
restart this step (e.g., select a different ROI), or go back to a different step (e.g., select a different patient). If the patient contours need to be redone, this is done outside IBEX, typically in the treatment planning system.

(5) If the user approves the ROI, this is then saved, and the user goes back to any step (typically 2) until all the ROIs have been uploaded and saved. The process can be saved and recontinued at any time.

(6) Select the feature parameters, such as threshold and bit depth, and calculate the image features for each setting. This is finally saved as a text file.

2.C. Quantifying reproducibility

For a particular feature, let \( x \) be a vector of that feature’s values across the patients’ first scan. Let \( y \) be a vector of that feature’s values across the patients’ second scan. The concordance correlation coefficient (CCC),\(^{21}\) a commonly used measure of reproducibility and interrater reliability, is then defined as

\[
\text{CCC} \equiv 1 - \frac{\langle (x - y)^2 \rangle}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}.
\] (1)

As suggested by McBride,\(^{22}\) we defined reproducible features as those features with \( \text{CCC} \geq 0.90 \). Kumar et al.\(^{8}\) additionally quantified features using their dynamic range (DR). However, we noticed a strong linear relationship between CCC and DR\(^{-2}\) and found an analytical relationship between them. Therefore, to avoid redundant cutoffs, we chose to only use the CCC constraint.

For each of the six machine/image type combinations (M1 T50, M2 T50, M1 AVG, M2 AVG, M3, and M4), we used the CCC constraint to identify which of the 328 features extracted were reproducible. This was done for each of the 11 HU cutoffs between –1000 and 0 HU. In order to find features that were reproducible across multiple machines, the intersection of individual machine reproducible feature sets was taken.

To quantify how machine-sensitive feature reproducibility is, for two machines with identical image types, let \( A \) be the set of reproducible features on one machine and \( B \) be the set of reproducible features on the other. In terms of these variables, the Dice similarity coefficient (DSC) (Ref. 23) and the Jaccard index (JI) (Ref. 24) are defined as follows:

\[
\text{DSC}(A, B) \equiv \frac{2|A \cap B|}{|A| + |B|}.
\] (2)

\[
\text{JI}(A, B) \equiv \frac{|A \cap B|}{|A \cup B|}.
\] (3)

To quantify intermachine feature selection agreement, each of these values was calculated using machine M1 and machine M2 for both T50 images and AVG images.

2.D. Finding reproducible, nonredundant features

Applying the CCC reproducibility constraint and taking the feature set intersection across multiple machines substantially reduced the total number of candidate features. However, many of the remaining features were strongly correlated with one another, and it would be useful to find a nonredundant subset of these features. Therefore, for a set of features \( F \) that is reproducible across \( N \) machines, let \( \rho_k(i) \) = concordance correlation coefficient for feature \( i \) on machine \( k \), and let \( r_k(i, j) \) be the sample Spearman’s rank correlation coefficient between feature \( i \) and feature \( j \) on machine \( k \). Next, define:

\[
\bar{\rho}(i) = \frac{1}{N} \sum_{k=1}^{N} \rho_k(i),
\] (4)

\[
d_k(i, j) = 1 - |r_k(i, j)| ,
\] (5)

\[
\bar{d}(i, j) = \frac{1}{N} \sum_{k=1}^{N} d_k(i, j),
\] (6)

where \( \bar{\rho}(i) \) is the mean CCC for feature \( i \) across all \( N \) machines, \( d_k(i, j) \) is the similarity distance between features \( i \) and \( j \) on machine \( k \), and \( \bar{d}(i, j) \) is the mean similarity distance between features \( i \) and \( j \) across all \( N \) machines.

Using these definitions, the similarity distance between all of the feature pairs of \( F \) can be calculated for each machine, and the mean similarity distances between the feature pairs of \( F \) can be established. The features present in \( F \) can then be hierarchically clustered based on the \( \bar{d}(i, j) \) distance function. By clustering until a threshold value is reached, several nonredundant, reproducible clusters can be identified. By picking the feature from each cluster with the highest average reproducibility across machines [Eq. (4)], a set of features that are nonredundant and reproducible across multiple machines can be identified. This process was performed for several different HU cutoffs and machine/image type combinations. In each case, the clustering algorithm used an average linkage function and an average mean similarity distance clustering threshold of 0.1.

3. RESULTS

3.A. Single machine reproducibility

Figure 2 shows the percent of the 328 image features that had a certain CCC value or higher when no voxels are removed from the GTV prior to feature extraction. Intersections with \( \text{CCC} = 0.90 \) indicate the number of features that were considered reproducible. In the figure, the three image types track each other initially, but then diverge to have considerably different levels of reproducibility. The cine 4D-CT average images were the most reproducible, with more reproducible features compared with the other scans; 90.5% of the features were reproducible for M1 AVG and 94.5% of the features were reproducible for M2 AVG. The cine 4D-CT T50 scans were the next most reproducible (M1 T50 = 75.0% pass; M2 T50 = 71.0% pass), and the helical 3D-CT breath-hold scans had the fewest number of reproducible features (M3 = 61.0% pass). Images interpolated with uniform voxel sizes did not improve the number of reproducible features (M4 = 57.3% pass). Figure 2 was generated
multiple times using various HU cutoffs to generate Fig. 3. In Fig. 3, it is apparent that for all machines and image types, as the HU cutoff goes down the number of reproducible features tends to go up. This trend implies that for maximum feature reproducibility, no HU cutoff should be used.

3.B. Multimachine reproducibility

For each of the data points in Fig. 3, each machine has a set of reproducible features. By taking the intersection of these sets we found features that were reproducible across multiple machines. Figure 4 shows the reproducible feature percentage for several different multimachine combinations. The same trend is noted: the number of multimachine reproducible features increases as the HU cutoff is decreased. At an HU cutoff of $-1000$ (i.e., no cutoff), for average scans, 86.3% of features were reproducible across both machines (M1 and M2), for T50 scans, 52.1% of features were reproducible across both machines, and for “single phase” scans (M1 T50, M2 T50, and M3), 42.1% of features were reproducible across all three machines. Exchanging M3 with M4 did not appreciably affect the number of multimachine reproducible features (41.5%).

To investigate if feature reproducibility is machinesensitive, Fig. 5 displays M1 CCC vs M2 CCC for both T50 and AVG scans. No HU cutoff was used. Points falling to the right of the vertical lines correspond to reproducible features on M1, and points falling above the horizontal lines correspond to reproducible features on M2. Venn diagram areas indicate the relative number of reproducible features that were common or unique for the two machines. The Dice similarity coefficients and Jaccard indices indicate that T50 feature

![Fig. 2. Percent of GTV image features with a CCC greater than or equal to the indicated value.](image1)

![Fig. 3. Percent of features that are reproducible (CCC ≥ 0.90) when voxels with CT numbers less than the specified HU cutoff are removed from the GTV before feature extraction.](image2)

![Fig. 4. Percent of features that are reproducible across all of the machines specified given a particular HU cutoff.](image3)

![Fig. 5. M1 CCC vs M2 CCC. Panel (a) depicts the relationship for T50 scans and panel (b) depicts the relationship for AVG scans. Venn diagrams, Dice similarity coefficients (DSC), and the Jaccard indices (JI) are displayed. The CCC ≥ 0.90 cutoff is shown for both axes. No HU cutoff was used.](image4)
reproducibility is relatively machine-sensitive (DSC = 0.71, JI = 0.55) while AVG feature reproducibility is relatively machine-insensitive (DSC = 0.93, JI = 0.87). However, if HU cutoffs are applied, both image types begin to show machine-sensitivity. For example, an HU cutoff of 0 gives DSC = 0.61 and JI = 0.44 for T50 images and DSC = 0.54 and 0.37 for AVG images.

3.C. Reproducible, nonredundant feature sets

Across single phase scans (i.e., M1 T50, M2 T50, and M3), 42.1% or 138 out of 328 image features were reproducible on all three machines (no HU cutoff). The cluster heat map25 in Fig. 6 shows the values of these 138 reproducible features for the 25 M3 test-retest pairs (50 scans). Repeated horizontal patterns present in the cluster heat map indicate that there is considerable feature redundancy, especially for the RLM gray level nonuniformity (GLNU) features and RLM run length nonuniformity (RLNU) features.

To address this, Eq. (5) was used to quantify how redundant two features are on a particular machine. The distribution of feature similarity distances \( d(i, j) \) for different machines is shown in Fig. 7 (no HU cutoff). Feature pairs with distances near zero are very redundant; feature pairs with distances near one are nonredundant. The different shape and increased number of nonredundant feature pairs for Fig. 7(b) (M2 T50) compared to Fig. 7(a) (M1 T50) indicate that feature redundancy is moderately machine-sensitive. Similar histograms (data not shown) show different feature similarity distance distributions both for M1 T50 vs M1 AVG and M2 T50 vs M2 AVG, indicating that feature redundancy is also image type sensitive. In Fig. 7(c) (M3 data), both the machine and image type are changed compared to Figs. 7(a) and 7(b). Thus, its feature pair distance distribution appears quite different; relatively more feature pairs are either very redundant or very nonredundant, with fewer feature pairs in between. Figure 7(d) uses Eq. (6) and shows the distribution of the mean similarity distances \( d(i, j) \) across the three machines. Its final (0.9 to 1.0) distance bin is relatively empty, implying that it is rare for a feature to be strongly nonredundant across all three machines.

Figure 8 shows a dendrogram generated from the hierarchical clustering of the 138 features that were reproducible across all single phase scans (no HU cutoff). It used the mean similarity distance of a feature pair as a distance function [distribution shown in Fig. 7(d)]. It also used an average linkage function, so the distance between two feature clusters is the average mean similarity distance, and a cutoff value of 0.1 was selected to find a finite number of relatively nonredundant feature clusters (in this case 23).

This whole process (identifying multimachine reproducible features and clustering to a threshold) was done for 4 different multimachine combinations and 11 different HU cutoffs. All used the same 0.1 clustering threshold. The results are displayed in Fig. 9, and show a similar trend to those in Fig. 4: as the HU cutoff goes down, the number of nonredundant feature clusters goes up. Therefore, we used HU cutoff = −1000 and generated a list of reproducible, nonredundant features for each of the four multimachine combinations in Fig. 9. See Table III. Each entry in the table is a representative feature from one of the reproducible, nonredundant clusters of the multimachine combination indicated. For cine 4D-CT average images and for cine 4D-CT T50 images, we recommend columns 1 and 2, respectively. For “single phase” mixed cine 4D-CT T50 images and helical 3D-CT breath-hold images, we recommend column 3. For helical 3D-CT breath-hold images we recommend column 4.

To confirm that the recommended feature sets are nonredundant, we generated a cluster heat map (Fig. 10) for all single phase scans (32 M1 T50 scans, 30 M1 T50 scans, and 50 M3 scans) using the 23 features from Table III, column 3. The absence of repeated vertical patterns in the heat map indicates that the selected features are nonredundant. Furthermore, blinded to the fact that test-retest pairs were present, 55 out of 56 test-retest pairs were correctly placed adjacent to one another in the figure. This indicates that the selected features were sufficiently informative to accurately match and distinguish patients.

4. DISCUSSION

Our study was strongly inspired by Kumar et al., but there are several key differences. First, we applied their techniques to several different machines and image types and studied how different machines and image types affected feature reproducibility and redundancy. We also studied feature reproducibility as a function of HU cutoff (i.e., GTV pruning) and omitted the feature dynamic range requirement. Additionally, we developed a strategy to integrate the findings to produce a multimachine reproducible and nonredundant feature set. Finally, we also omitted some of the patients used by Kumar et al. used a different set of image features, and had a different semiautomated contouring process. These methodological differences could explain why our CCC values were relatively higher, leading us to choose to use a reproducibility cutoff of 0.90 instead of 0.85.

CT image type was found to strongly affect feature reproducibility (Figs. 2–4). Features from cine 4D-CT average images were substantially more reproducible than those associated with T50 images, and T50 associated features were somewhat more reproducible than breath-hold helical 3D-CT features. This may seem counter-intuitive since T50 images are intended to “freeze” motion. However, because average images are an average of 10 image phase reconstructions, they are less susceptible to photon noise and various 4D-CT artifacts. As for the breath-hold CTs, it is difficult to draw conclusions since the machine, image type, and convolution kernel varied simultaneously compared to the other datasets. However, M4 did not appear to have better reproducibility than M3. That is, no significant reproducibility improvements were observed when scans were interpolated to have a uniform voxel size. If raw CT data were available it may have been possible to reconstruct images with a standard voxel size and achieve better results.

As the HU cutoff goes down (i.e., as fewer low-intensity GTV voxels are pruned), we found (1) reproducibility
FIG. 6. Cluster heat map representation of the M3 scans using the 138 features that passed the reproducibility cutoff on single phase scans. No HU cut-off was used. Green cells correspond to a feature value above the mean, red cells correspond to feature values below the mean, and the color intensity indicates the deviation magnitude. Image features (rows) were clustered using the Pearson’s distance metric, and patient scans (columns) were clustered using a Euclidean distance metric. Both used an average linkage function. Scans are labeled such that the number indicates the test-retest pair; “A” and “B” represent scan #1 and scan #2, respectively. Dendrograms indicate the hierarchical relationships both between the features and between the scans.
Fig. 7. Panels (a), (b), and (c) show histograms of the feature pair similarity distances for the 9453 feature pairs possible from the 138 features that passed the reproducibility cutoff for all single phase scans with no HU cutoff. Panel (d) shows the distribution of mean feature similarity distances. It becomes less machine-sensitive (Fig. 5), (2) individual machine reproducibility goes up (Fig. 3), (3) multimachine reproducibility goes up (Fig. 4), and (4) the number of multimachine, nonredundant clusters goes up (Fig. 9). This last finding is particularly interesting since it implies that the low intensity voxels contain nonredundant information. Since the vast majority of the low intensity voxels appear at the tumor margins, these findings indicate that the tumor margin contains valuable nonredundant information which can be used to discriminate tumors. This is understandable, since noncontrast-enhanced CT has limited soft tissue contrast; if the margins are removed there are fewer intensity variations to extract independent features from. Thus, we recommend against removing voxels based on their intensity values as a preprocessing step to feature extraction. This is in opposition to the NSCLC CT protocol of Ganeshan et al.,9,10 but we believe that it is likely due to methodological differences. In their studies, directionally independent first-order statistics (e.g., entropy and uniformity) were extracted from the filtered 2D image slice that had the largest transverse tumor length. In our work, features are extracted from the intensity histogram, absolute gradient image, RLM, and COM values evaluated over the entire unfiltered GTV. Therefore, as a caveat, we should emphasize that our tumor preprocessing (i.e., pruning) protocols and feature set recommendations are only applicable to similar datasets (noncontrast-enhanced NSCLC CT images) using similar image features (Table II).

One limitation of our study is that all semiautomated contouring was done by a single individual.1 Ideally, multiple individuals could contour each GTV and operator sensitivity could be studied. Alternatively, one of several automated segmentation techniques could be used to possibly increase reproducibility.29 In the future, it would also be useful to study the reproducibility of additional image features and the effect of image convolution kernels. Also, we did not examine the effect of lesion size, morphology, or anatomic location (e.g., apex vs base). Another limitation is that all of the machines used in our study are GE machines. Ideally, multiple machines from various manufacturers should be tested to see if there are any systematic or random feature value variations between them. Finally, a key point to note is that our study only directly assessed intramachine reproducibility. Multimachine reproducible features were defined as features...
that have good intramachine reproducibility on multiple machines. Because individual patients were not test-retest imaged on multiple scanners, we cannot directly assess intermachine reproducibility (i.e., agreement of feature values between machines). For ethical reasons, this necessitates a separate application of rapid prototyping, another group has developed a way to independently control the CT number of every voxel of a phantom.31 This could be used to study image rapid prototyping, placed it into an anthropomorphic phantom, and moved it to match recorded tumor motion trajectories. This phantom would be very useful to study how motion affects feature reproducibility across multiple machines.

### TABLE III. Multimachine reproducible and nonredundant feature lists.

| M1 AVG and M2 AVG (38 features) | M1 T50 and M2 T50 (28 features) | M1 T50 and M2 T50 and M3 (23 features) | M3 (33 features) |
|---------------------------------|---------------------------------|--------------------------------------|------------------|
| IHIST entropy GEO VA ratio     | GEO VA ratio GEO volume IHIST kurtosis IHIST p10tile IHIST skewness | IHIST skewness IGR BD10 PNonZero | IHIST p99tile RLM BD10 FxR (1,0,1) |
| IHIST kurtosis IHIST entropy   | IHIST kurtosis IHIST p10tile IHIST skewness | IGR BD10 skewness RLM BD6 Mean | IGR BD2 variance IGR BD6 Mean RLM BD10 FxR (1,0,1) |
| IHIST max IHIST kurtosis       | IHIST kurtosis IHIST p10tile IHIST skewness | IGR BD10 skewness RLM BD6 Mean | IGR BD2 variance IGR BD6 Mean RLM BD10 FxR (1,0,1) |
| IHIST mean IHIST p10tile       | IHIST skewness IGR BD10 skewness | IGR BD10 skewness RLM BD6 Mean | IGR BD2 variance IGR BD6 Mean RLM BD10 FxR (1,0,1) |
| IHIST skewness IHIST p10tile   | IHIST skewness IGR BD10 skewness | IGR BD10 skewness RLM BD6 Mean | IGR BD2 variance IGR BD6 Mean RLM BD10 FxR (1,0,1) |
| IHIST p99tile IHIST variance   | IHIST mean IGR BD2 variance | IGR BD2 variance IGR BD6 Mean RLM BD10 FxR (1,0,1) | IHIST skewness IHIST p10tile IHIST skewness |
| IHIST variance RLM BD10 RLNU (0,1,0) | RLM BD10 Correlation (0,−1,0) | RLM BD10 Correlation (0,−1,0) | RLM BD10 RLNE (1,0,1) |
| IGR BD10 variance RLM BD10 SRE (1,0,0) | COM BD10 correlation (0,−1,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| IGR BD2 kurtosis COM BD10 AngScMom (−1,0,0) | COM BD10 correlation (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| IGR BD2 variance COM BD10 correlation (−1,0,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| IGR BD2 skewness COM BD10 correlation (−1,0,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD10 LRE (0,−1) COM BD10 DiFEntrp (−1,0,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD10 RLNU (0,0,1) COM BD10 Entropy (−1,0,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD2 FxR (1,1,0) COM BD10 InvDiMom (0,−1,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD2 FxR (−1,1,−1) COM BD10 InvDiMom (−1,0,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD2 RLNU (1,0,−1) COM BD10 SumEntrp (−1,0,0) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD2 SRE (1,1,0) COM BD10 SumOfSqs (0,0,−1) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD2 SRE (−1,1,0) COM BD2 SumOfSqs (0,0,−1) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| RLM BD6 GLNU (0,0,1) COM BD2 SumVarnc (0,0,−1) | COM BD10 AngScMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD10 AngScMom (0,0,−1) COM BD6 contrast (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD10 correlation (0,−1,0) COM BD6 correlation (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD10 DiFVarnc (−1,0,0) COM BD6 DiFEntrp (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD10 entropy (0,−1,0) COM BD6 DiFEntrp (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD10 InvDiMom (0,−1,0) COM BD6 DiVarnc (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD10 SumEntrp (0,0,−1) COM BD6 DiVarnc (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD2 contrast (0,0,−1) COM BD6 InvDiMom (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD2 contrast (−1,0,0) COM BD6 InvDiMom (0,0,−1) | COM BD6 InvDiMom (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 AngScMom (0,−1,0) | COM BD6 AngScMom (0,−1,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 DiFEntrp (0,0,−1) | COM BD6 DiFEntrp (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 DiFEntrp (0,−1,0) | COM BD6 DiFEntrp (0,−1,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 DiFEntrp (−1,0,0) | COM BD6 DiFEntrp (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 DiVarnc (0,0,−1) | COM BD6 DiVarnc (0,0,−1) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 DiVarnc (0,−1,0) | COM BD6 DiVarnc (0,−1,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 entropy (0,−1,0) | COM BD6 entropy (0,−1,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 InvDiMom (0,−1,0) | COM BD6 InvDiMom (0,−1,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |
| COM BD6 InvDiMom (−1,0,0) | COM BD6 InvDiMom (−1,0,0) | RLM BD10 SRE (0,1,0) | RLM BD10 RLNE (1,0,1) |

Coordinates indicate either a direction (RLM) or a displacement (COM); BD indicates bit depth; VA ratio, volume to area ratio; GEO, geometric; IHIST, intensity histogram; IGR, absolute gradient image; RLM, run length matrix; COM, co-occurrence matrix; p10tile, 10th percentile (similarly for p1tile, etc.); FxR, fraction in runs; SRE, short run emphasis; LRE, long run emphasis; RLNU, run length nonuniformity; GLNU, gray level nonuniformity; PNonZero, probability nonzero; InvDfMom, inverse difference moment; AngScMom, Angular 2nd moment; and DiFVarnc, difference variance.
FIG. 10. Cluster heat map representation of all scans from M1 T50, M2 T50, and M3 (no HU cutoff). Green cells correspond to a feature value above the mean, red cells correspond to feature values below the mean, and the color intensity indicates the deviation magnitude. For readability, orientations are reversed from Fig. 6. Each of the 23 features shown had the highest average CCC from one of the 23 independent clusters shown in Fig. 9. Features (columns) were clustered using the Pearson’s distance metric, and patient scans (rows) were clustered using a Euclidean distance metric. Both used an average linkage function. Scans are labeled such that the number indicates the test-retest pair; “A” and “B” represent scan #1 and scan #2, respectively. Dendrograms indicate the hierarchical relationships between the features and between the scans.
feature fidelity in the presence of subtle changes in voxel intensity and indistinct tissue boundaries. Although these phantoms cannot move, previous test-retest images can be “printed” into a phantom to somewhat account for deviations due to patient setup and organ motion.

An important aspect of this work is that we have demonstrated an approach to evaluate whether image features are reproducible and informative. Such an analysis is important prior to trying to use image features to predict patient outcomes (such as overall survival), genetic expression, or other clinically useful outcomes. Our experience is that these experiments are already extremely difficult, with many sources of uncertainty, so it can be very advantageous to focus on image features which have already been shown to be reproducible and informative. As such, we suggest that this approach (or something similar) should be used when investigators develop new image features.

In conclusion, this study integrated multiple NSCLC test-retest CT datasets to identify informative, nonredundant image features with high intramachine reproducibility on multiple machines. These feature sets are strong candidates for future image-based modeling. Image feature quality was best for average 4D-CT images, and the tumor margins were found to play an important role in tumor discrimination. Further advanced phantom studies are needed to investigate intermachine image feature reproducibility.

ACKNOWLEDGMENT

Luke Hunter was funded by a Hertz Applied Science Fellowship.
construction from four-dimensional respiration-correlated CT scans for radiotherapy planning of lung cancer patients,” Int. J. Radiat. Oncol., Biol., Phys. 65(5), 1560–1571 (2006).

29N. Sharma and L. M. Aggarwal, “Automated medical image segmentation techniques,” J. Med. Phys. 35(1), 3–14 (2010).

30L. E. Court, J. Seco, X. Q. Lu, K. Ebe, C. Mayo, D. Ionascu, B. Winey, N. Giakoumakis, M. Aristophanous, R. Berbeco, J. Rottman, M. Bogdanov, D. Schofield, and T. Lingos, “Use of a realistic breathing lung phantom to evaluate dose delivery errors,” Med. Phys. 37(11), 5850–5857 (2010).

31T. S. Yoo, T. Hamilton, D. E. Hurt, J. Caban, D. Liao, and D. T. Chen, Toward quantitative X-ray CT phantoms of metastatic tumors using rapid prototyping technology, Biomedical Imaging: From Nano to Macro, 2011, pp. 1770–1773.