Original Article

Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer

Xenia Fave[1,2], Lifei Zhang[1], Jinzhong Yang[1], Dennis Mackin[1], Peter Balter[1], Daniel Gomez[3], David Followill[1], A. Kyle Jones[4], Francesco Stingo[5], Laurence E. Court[1,2]

[1]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [2]The University of Texas Graduate School of Biomedical Sciences at Houston, Houston, TX, USA; [3]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [4]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [5]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: LE Court; (III) Provision of study materials or patients: D Gomez, L Zhang; (IV) Collection and assembly of data: X Fave; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Xenia Fave. Department of Radiation Physics, Unit 1420, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030, USA. Email: xjfave@mdanderson.org.

Background: Radiomics features have been used in a variety of studies to predict patient outcomes or aid in the diagnosis of non-small cell lung cancer. However, no guidelines exist for the best way to calculate these features to maximize their prognostic potential. The purpose of the current study was to evaluate how different image pre-processing techniques may impact both the volume dependence and prognostic potential of the features in univariate analyses.

Methods: Radiomics features from the histogram, co-occurrence matrix, neighborhood gray-tone difference matrix and run-length matrix were calculated from a set of computed tomography (CT) images of 107 non-small cell lung cancer tumors with volumes ranging from 5 to 567 cm^3. Features were calculated from the images with no pre-processing, 8 bit depth resampling, Butterworth smoothing, or both 8 bit depth resampling and Butterworth smoothing. To determine which features were correlated with volume, we calculated the Spearman rank correlation coefficient (r_s) for each feature and preprocessing combination. For features that had very high volume correlations (r_s >0.95) regardless of which preprocessing algorithm was used, we normalized the algorithm for volume and recalculated the volume correlation. To determine whether the preprocessing technique affected the usefulness of the feature, we fitted univariate Cox proportional hazards models for all four preprocessing techniques for each feature and calculated the P value. Additionally, univariate cox models were recalculated using leave-one-out cross validation to generate risk predictions for each patient. As a result, each patient had a predicted outcome for each model in which the patient was not involved in the model building. The prediction accuracy was assessed using Harrell's concordance index (c-index). Finally, the ability of each feature to improve model fit was examined using the P value of the log-likelihood ratio between a model built using volume only and a model built using volume and one radiomics feature. The Benjamini-Hochberg procedure was used for multiplicity correction.

Results: Five features were entirely volume dependent (busyness, coarseness, grey-level non-uniformity, run-length non-uniformity, and energy) and new algorithms were proposed for these features. Both the correlation with volume and the prognostic value of individual features changed substantially with different preprocessing techniques. In general, preprocessed features that were at least slightly correlated with volume (r_s >0.5) were more likely to be significant in the univariate analysis. Additionally, Butterworth smoothing, used either alone or in conjunction with 8 bit depth resampling, most often yielded features that were significant in univariate analysis.

Conclusions: Preprocessing can have a strong impact on the volume dependence of a feature, and its

significance in univariate models. To create standardized features useful for multivariate modeling, it will be important to balance the usefulness of features with their volume dependence.

Keywords: Computer-assisted image-interpretation; cox models; diagnostic imaging; non-small cell lung cancer; radiomics

Submitted Mar 26, 2016. Accepted for publication Jun 28, 2016.
doi: 10.21037/tcr.2016.07.11
View this article at: http://dx.doi.org/10.21037/tcr.2016.07.11

Introduction

Recently, several research groups have explored the potential of quantitative imaging (radiomics) features to predict patient outcomes prior to treatment (1-10). These quantitative features are calculated from image data already being acquired for clinical purposes. The calculated features may represent the relative heterogeneity of a tumor, capture the spatial relationships of pixels within the tumor, or be calculated from a histogram of image pixel values. Models built upon these features and clinical factors, if successful, could aid physicians in identifying high- and low-risk patients and thus help inform treatment decisions.

The body of literature suggesting that radiomics features may be prognostic in patients with non-small cell lung cancer (NSCLC) has been steadily growing over the past few years. Features extracted from computed tomography (CT) images have been claimed to be correlated with overall survival (1,2,4,5), gene expression patterns (1,4,6), pathologic findings (7), and stage (3). A recent study showed that models built on texture features and clinical factors can improve patient risk stratification for overall survival, local regional control, and freedom from distant metastases compared with models built on clinical factors alone (10).

Although these results are intriguing, the rush to determine whether radiomics features have a useful role in tumor analysis has left many of the fundamental questions surrounding them overlooked or only partially answered. Chief among these is how to determine whether a feature is being calculated correctly. In radiomics, no ground truth exists for the features themselves, and as a consequence most studies have settled for selecting features with high reproducibility in patient test-retest sets and then using a machine-learning algorithm to determine which features are useful for a particular research question. However, this approach can lead to high false-positive rates (11), and has resulted in variability in both the features that are used and how they are calculated (e.g., feature parameters and image preprocessing).

Further, while a feature should be reproducible, reproducibility itself does not guarantee that a feature is informative. For example, a highly smoothed image is much more likely to return the same value for a feature on a retest, but it is also likely to have lost the original spatial differences that the feature was selected to identify. Also, because most of the quantitative imaging features used in radiomics today were initially developed to analyze aerial photographs (12-14), in which only two-dimensional rectangular photographic images of the same pixel dimensions were compared, normalization for area or volume differences was not originally necessary. However, in tumor analysis, the regions of interest (ROIs) are the irregular contours of three-dimensional tumors. As a result, the volumes of tumor ROIs have substantial inter-patient variability. A feature that is correlated with volume would be likely to have high reproducibility when tested and retested in a set of patients with a wide range of ROI volumes. This correlation can dominate the useful spatial distribution or intensity information in the feature that we hope to measure. The impact of these volume differences on features measured from CT images has never been systematically investigated. However, a few recent studies have demonstrated the effects of volume on features in fluorodeoxyglucose-positron emission tomography (FDG-PET) images in which the total number of voxels per tumor was much smaller, and have concluded that radiomics features offer complementary information only above volume thresholds as large as 10 or 45 cm^3 (15,16).

In order to determine how best to calculate features, it is first necessary to determine which features are susceptible to changes in image preprocessing or wide variations in tumor volume. This work has investigated this issue by assessing changes in the correlations between features and volume, as well as the univariate prognostic potential of features, as a function of three different preprocessing techniques.

Table 1 Summary of the clinical characteristics of the study population, n=107.

Characteristic	No. (%)
Median age (range)	66 years (47–80 years)
Median gross tumor volume (range)	39.6 cm^3 (5.4–567 cm^3)
Sex	
Male	62 [58]
Female	45 [42]
Tumor stage	
II	12 [11]
III	93 [87]
IV	2 [2]
Tumor histologic findings	
Squamous cell carcinoma	46 [43]
Adenocarcinoma/other	61 [57]

Methods

Images

We retrospectively reviewed clinical data and outcomes for 134 NSCLC patients treated at our institution with radiation therapy and concurrent chemotherapy as part of a clinical trial with IRB approval (17). The pretreatment simulation 4DCT images and the three-dimensional treatment plan gross tumor volume (GTV) contours were acquired and used as the ROIs for radiomics feature extraction. Tumors were removed from this dataset if they had an ROI volume of less than 5 cm^3 (n=18) or if they were imaged using breath hold instead of 4DCT (n=9) which left 107 images for analysis. The characteristics for the resulting study population are summarized in *Table 1*. The scan parameters for the 4DCT images included a peak tube voltage of 120 kVp, tube current of 100 or 200 mA, and rotation time of 0.5 or 0.8 second. The end-of-exhale phase images were used in our analysis because they are considered the most reproducible (18) and have been used in other texture analyses (10). Reconstructed axial images were 512×512 pixels with an in-plane resolution of 0.98 mm and image thickness of 2.5 mm. These acquisition and reconstruction parameters are our institutional standards for CT imaging.

Image pre-processing

A lower intensity threshold of –100 and upper intensity threshold of 200 Hounsfield Units were applied to all images before feature calculation to ensure that no voxels of the surrounding normal lung tissue or bone were included in the GTV. Images were then further pre-processed with either 8 bit depth resampling, a Butterworth smoothing filter with an order of 2 and a cutoff frequency of 125, both Butterworth smoothing and 8 bit depth resampling, or no additional pre-processing. When both Butterworth smoothing and 8 bit depth resample were used, the Butterworth smoothing was performed first. The general effect of each of these techniques should be to reduce noise in the image and thus improve the overall signal to noise ratio of any radiomics feature. These particular preprocessing options were selected to represent some of the variety available and are not meant to be exhaustive. The bit depth resample changes the images from 12 bit to 8 bit, effectively creating pixel intensity bins of 16 as 12 bit images have 4,096 possible intensity values and 8 bit images have 256 possible intensity values. Because several of the radiomics features are calculated from matrices that track how often pixels of one intensity are next to each of the other intensities, this resample also removes the need to select an appropriate bin size for these matrices, and instead bins of 1 can be used. For example, a 12 bit image would have a COM of 4,096 by 4,096 while an 8 bit image would have a COM of 256 by 256. The range of values in NSCLC tumors is typically much less than the range of values in the entire image. Thus for a hypothetical tumor with values from only 1 to 100 HU, only a 100 by 100 subsection of the 12 bit COM would be used to calculate the feature since the rest of the COM would be filled with zeros. For the 8 bit image, the 1–100 HU range would be resampled to 1–7, and the informative subsection of the COM would thus be a 7 by 7 matrix. By using resampling, the 7 by 7 COM would be less likely to be sparsely populated than the 100 by 100 COM especially if the tumor is small and thus may better represent the spatial patterns in the image and be more informative. The choice of 8 bit was selected because the effect of noise in CT for soft tissue and tumor should be less than 16 HU and because this value had been used in other radiomics analyses (10,16,19,20). The Butterworth smoothing filter acts as a low pass filter to remove high frequency noise. This filter has the advantage of acting

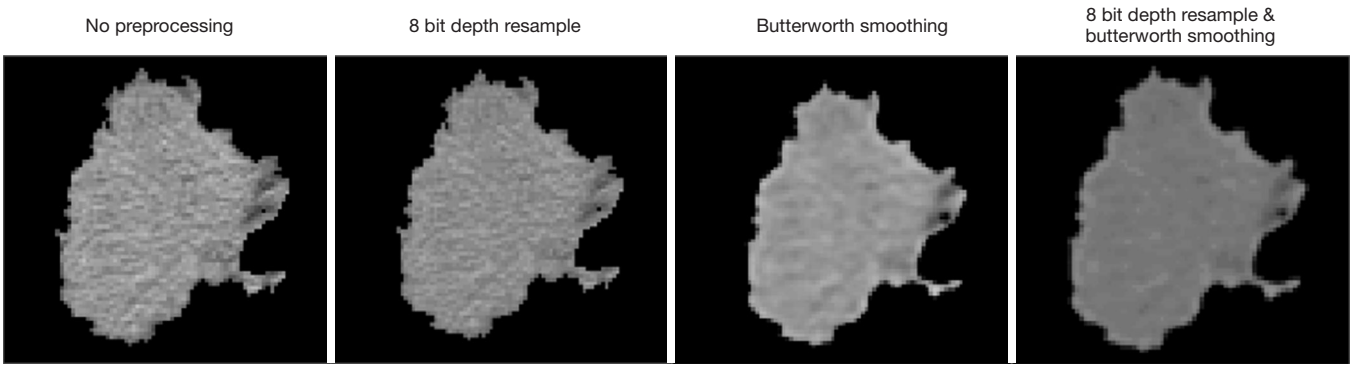

Figure 1 Sample image of a patient tumor ROI using each preprocessing technique.

in the frequency domain, so it is not limited by the size of the filter matrix. Additionally, Butterworth filters have the benefits of reduced ringing and gradual attenuation of higher frequencies. In comparison, Gaussian filters would likely have a similar impact on the image but with a less steep cutoff. *Figure 1* shows the visible result of each of these pre-processing techniques on a sample tumor ROI. The radiomics features were calculated from the tumor ROIs using each of these pre-processing techniques.

Features

Radiomics features from the histogram, co-occurrence matrix (COM), neighborhood gray-tone difference matrix (NDM), and run-length matrix (RLM) were calculated and are summarized in *Table 2*. The abbreviations used for each feature for the figures are also listed in *Table 2*. The histogram features summarize characteristics of the intensity distribution for each tumor. The COM, NDM, and RLM features all contain information about the spatial distribution of the pixel intensities within a tumor. All textures were calculated using the open-source Imaging Biomarker Explorer (IBEX) software (21).

Volume-dependence

To determine whether the features became more or less correlated with volume as a result of image pre-processing, we used the spearman rank correlation coefficient (r_s) to calculate the correlation with volume of each feature after each preprocessing technique. The Spearman rank correlation coefficient ranges from –1 to 1 and evaluates whether a value decreases or increases monotonically; 1 and –1 represent a perfect correlation and 0 represents no correlation. Because the feature algorithms used for tumor analysis in current radiomics studies were originally designed for comparing equally sized photographs (12), it was possible that some algorithms might be inherently dependent on volume and may require correction for the number of voxels in the image. Features with extremely high values (r_s >0.95) for all four preprocessing techniques were identified and new normalized versions of the algorithms for these features were created and added to the feature set for analysis. The Spearman correlation coefficient for the normalized features was calculated for each preprocessing technique. For completeness, we did not remove the features that exhibited these extremely strong correlations with volume from the remainder of the analysis.

Prognostic potential

To determine the impact of preprocessing on the usefulness of radiomics features, we fitted univariate Cox proportional hazards models for overall survival. P values for the fit using the likelihood-ratio test were calculated for each model. P values were corrected using the Benjamini-Hochberg process to control for the false discovery rate (type 1 error). Corrected P values <0.05 were considered significant.

Each univariate model was then recalculated using leave-one-out cross validation to generate predictions for each patient in each model. In this framework, each patient's prediction is calculated using a model in which that patient was left out of the coefficient fitting process. Harrell's concordance index (c-index) was then calculated using the predicted risks. The c-index is similar to the area under the curve and evaluates, for each combination of two patient predictions, how often the patient with the higher predicted risk actually experiences the event (death) before

Table 2 Radiomics features analyzed in this study and their abbreviations

Feature category	Feature	Feature abbreviation
Histogram	Variance	HISTvar
	Uniformity	HISTunif
	Standard deviation	HISTstd
	Skewness	HISTskew
	Minimum	HISTmin
	Median	HISTmed
	Mean	HISTmean
	Maximum	HISTmax
	Kurtosis	HISTkurt
	Entropy	HISTentropy
	Energy	HISTenergy
Run-length matrix	Short run low gray-level emphasis	RLMsrlgle
	Short run emphasis	RLMsre
	Run percentage	RLMrunperc
	Run-length non-uniformity	RLMrlnu
	Long run low gray level emphasis	RLMlrlgle
	Long run high gray level emphasis	RLMlrhgle
	Long run emphasis	RLMlre
	Low gray-level run emphasis	RLMlglre
	High gray-level run emphasis	RLMhglre
	Gray-level non-uniformity	RLMglnu
Neighborhood gray-tone difference matrix	Texture strength	NDMtexstr
	Contrast	NDMcontrast
	Complexity	NDMcomp
	Coarseness	NDMcoarse
	Busyness	NDMbusy
Co-occurrence matrix	Variance	COMvar
	Sum variance	COMsumvar
	Sum entropy	COMsument
	Sum average	COMsumavg
	Max probability	COMmaxprob
	Inverse variance	COMinvvar
	Inverse difference norm	COMinvdifn
	Inverse difference moment norm	COMinvdifmn
	Information measure correlation	COMinfomc
	Information measure correlation 2	COMinfomc2
	Homogeneity	COMhomog
	Homogeneity 2	COMhomog2
	Entropy	COMentropy
	Energy	COMenergy
	Dissimilarity	COMdissim
	Difference entropy	COMdiffent
	Correlation	COMcorrel
	Contrast	COMcontrast
	Cluster tendency	COMclustend
	Cluster shade	COMclusshade
	Autocorrelation	COMautocorrel

NDM, neighborhood gray-tone difference matrix; RLM, run-length matrix; COM, co-occurrence matrix.

the patient with the lower predicted risk. A c-index value of ≤0.5 indicates that the model does not perform better than random chance and a value of 1.0 indicates a perfect model. The c-index for a model with volume as the only covariate was also calculated for comparison.

Lastly, to determine whether features actually outperformed volume alone, we calculated the log-likelihood ratios between Cox proportional hazards models for overall survival fitted with volume only and models fitted with volume and one radiomics feature at a time. The P values for the log-likelihood ratios were then determined and corrected using the Benjamini-Hochberg process. A P value <0.05 would mean that including that particular feature significantly improved the model's fit to the data

when compared with the fit of a model using volume only.

Results

Volume dependence

The absolute values of the Spearman correlation coefficients for each feature after each tested preprocessing technique are plotted in *Figure 2*. Five features demonstrated a very strong volume correlation, with Spearman correlation coefficients absolute values >0.95 regardless of which preprocessing technique was used. These features included energy from the histogram, coarseness and busyness from the NDM, and grey-level non-uniformity and run-length non-uniformity from the RLM. After close investigation, we found that these high levels of volume correlation were due to the feature algorithms, which did not normalize for the number of voxels or matrix elements summed. This means that if these features were measured from two ROIs of different sizes but with pixels of only one intensity, two different values would be obtained. Factors that mitigate this source of volume dependence were introduced for each of these feature algorithms (*Table 3*), and new Spearman correlation coefficients were calculated. The algorithms for energy, grey-level non-uniformity, and run-length non-uniformity were edited by dividing their values by the total number of voxels in the ROI. The algorithms for busyness and coarseness were changed by normalizing the sums of the average difference around each intensity [the NDM values, $s(i)$] by the number of voxels of that intensity. The maximum of the Spearman correlation coefficients for the normalized features was 0.79 (*Figure 2*). All of the normalized features preprocessed with resampling had correlations <0.5.

In general, features were more correlated with volume after either Butterworth smoothing or both Butterworth smoothing and bit depth resampling, and were less correlated with volume after bit depth resampling (*Figure 3*). A few features demonstrated strong (r_s >0.85) correlations with volume for only one or two pre-processing techniques. Both information measure correlation and information measure correlation 2 from the COM had high correlations with volume after no preprocessing or Butterworth smoothing (>0.90), but were not correlated with volume when bit depth resampling was used, either alone or with Butterworth smoothing (r_s <0.5). Inverse difference moment norm from the COM had a correlation of 0.88 with volume after Butterworth smoothing or after Butterworth smoothing and bit depth resampling. Texture strength from the NDM had correlation coefficients of –0.94, –0.87, –0.90, for Butterworth smoothing, bit depth resampling, and both Butterworth smoothing and bit depth resampling respectively, but when no pre-processing was used the coefficient was less than 0.5.

Prognostic potential

The P values for the Cox proportional hazard models are plotted in *Figure 4* for each feature and pre-processing combination, as well as for volume. Almost every feature (39/55) had at least one preprocessing technique that resulted in statistically significant stratification (P value <0.05 after Benjamini-Hochberg correction). A few features from each category were never significant in this univariate analysis: normalized busyness, the original (volume-dependent) busyness, complexity, and contrast from the NDM; maximum, minimum, and the original energy from the histogram; long-run emphasis, run percentage, and the original forms of grey-level non-uniformity and run-length non-uniformity from the RLM; correlation, energy, information measure correlation 2, and max probability from the COM; and volume. Features that were always significant regardless of the preprocessing technique were high and low gray-level run emphasis from the RLM, mean from the histogram, and the original algorithm for coarseness. In general features were more likely to have a significant P value after Butterworth smoothing or both Butterworth smoothing and 8 bit depth resampling.

The c-indices calculated from the predicted values for each univariate model are plotted in *Figure 5*. The c-index for volume was 0.56. The largest calculated c-index was 0.65 for the median from the histogram after both Butterworth smoothing and 8 bit depth resampling. The next highest c-index was 0.60 for both high gray-level run emphasis and short run high gray-level run emphasis from the RLM. In general, using Butterworth smoothing either alone or with 8 bit depth resampling resulted in c-indices close to or slightly larger than the c-index for volume, whereas using 8 bit depth resampling on its own or not using any image pre-processing resulted in c-indices <0.5. With the exception of minimum intensity, for every feature at least one preprocessing technique resulted in a c-index greater than 0.5.

The Benjamini-Hochberg corrected P values for the log-likelihood ratios comparing Cox proportional hazards models for overall survival fitted with volume only and fitted with volume and one of the radiomics features are plotted in *Figure 6*. Of the 54 features, 25 had at least one

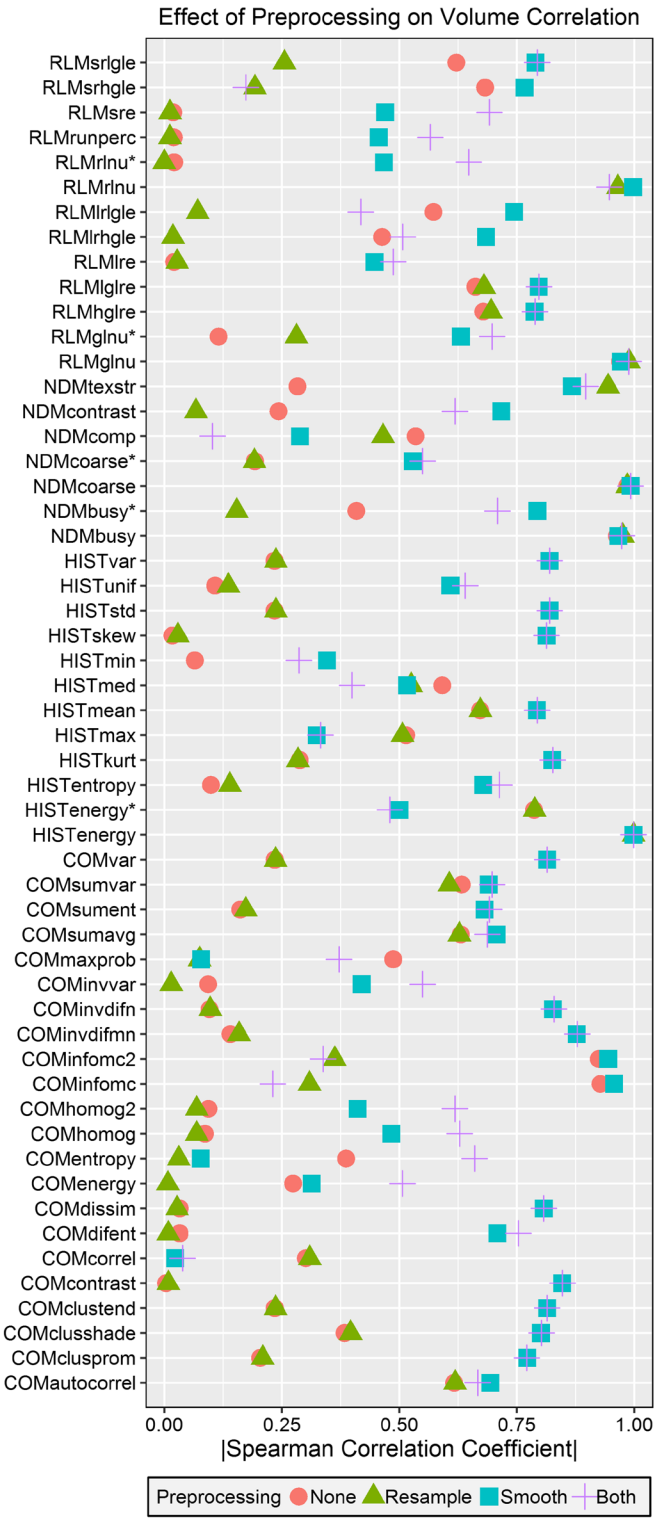

Figure 2 The volume correlation for the features was measured using the spearman rank correlation coefficient. The absolute value of the coefficients for each feature and preprocessing technique are plotted here. The volume correlation for most of the features was substantially changed with different image pre-processing. Five features were extremely correlated with volume regardless of the preprocessing technique used and were recalculated with a normalizing factor. The normalized version of these algorithms are noted with an asterisk.

Table 3 Algorithms for features highly correlated with volume before and after normalization

Feature	Original algorithm		Corrected algorithm	
Busyness (13)	$\frac{\left[\sum_{i=0}^{G_h} p_i s(i)\right]}{\left(\sum_{i=0}^{G_h}\sum_{j=0}^{G_h} ip_i - jp_i\right)}$	[1]	$\frac{\left[\sum_{i=0}^{G_h} p_i \frac{s(i)}{N(i)}\right]}{\left(\sum_{i=0}^{G_h}\sum_{j=0}^{G_h} ip_i - jp_i\right)}$	[2]
Coarseness (13)	$\left[\in + \sum_{i=0}^{G_h} p_i s(i)\right]^{-1}$	[3]	$\left[\in + \sum_{i=0}^{G_h} p_i \frac{s(i)}{N(i)}\right]^{-1}$	[4]
Gray-level nonuniformity (14)	$\frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} p(i,j)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$	[5]	$\frac{1}{N_v}\frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} p(i,j)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$	[6]
Run-length nonuniformity (14)	$\frac{\sum_{j=1}^{N_r}\left[\sum_{i=1}^{N_g} p(i,j)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$	[7]	$\frac{1}{N_v}\frac{\sum_{j=1}^{N_r}\left[\sum_{i=1}^{N_g} p(i,j)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$	[8]
Energy	$\sum_{i}^{N_v} p(i)^2$	[9]	$\frac{1}{N_v}\sum_{i}^{N_v} p(i)^2$	[10]

The algorithms for the volume-dependent features from the literature were changed by introducing a normalization term for the number of voxels of each intensity, i, in the image, N(i), or the total number of voxels Nv. Other terms are: pi-probability of intensity i in the image; s(i)-sum of the average difference value around voxels of intensity i; Gh-Highest gray-level intensity; Ng-Number of gray levels; Nr-Number of run levels; p(i,j)- probability of gray-level i having a run of length j; X(i)-the intensity of the ith voxel in the image, X.

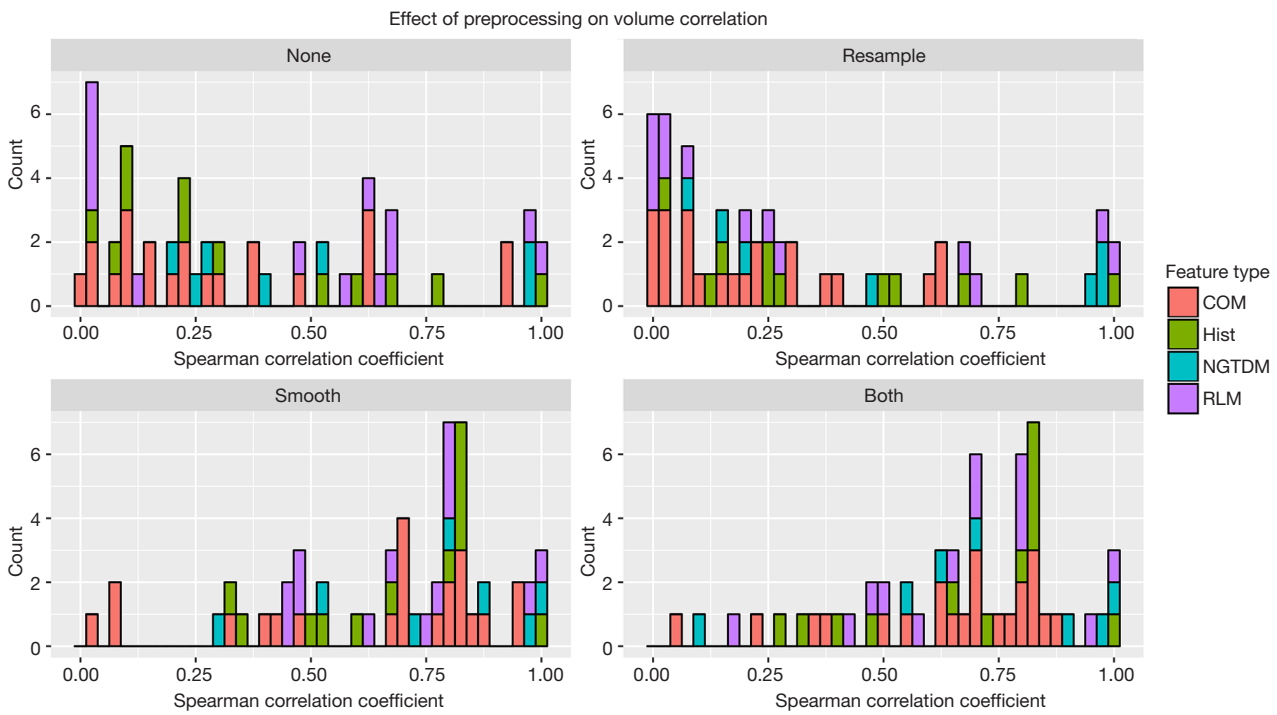

Figure 3 The absolute value of Spearman correlation coefficients are plotted as a histogram here for each preprocessing technique. Butterworth smoothing either alone or combined with 8 bit depth resample increased the strength of the correlation between most of the features and volume. When only 8 bit depth resample was used, the overall volume correlation decreased for the features.

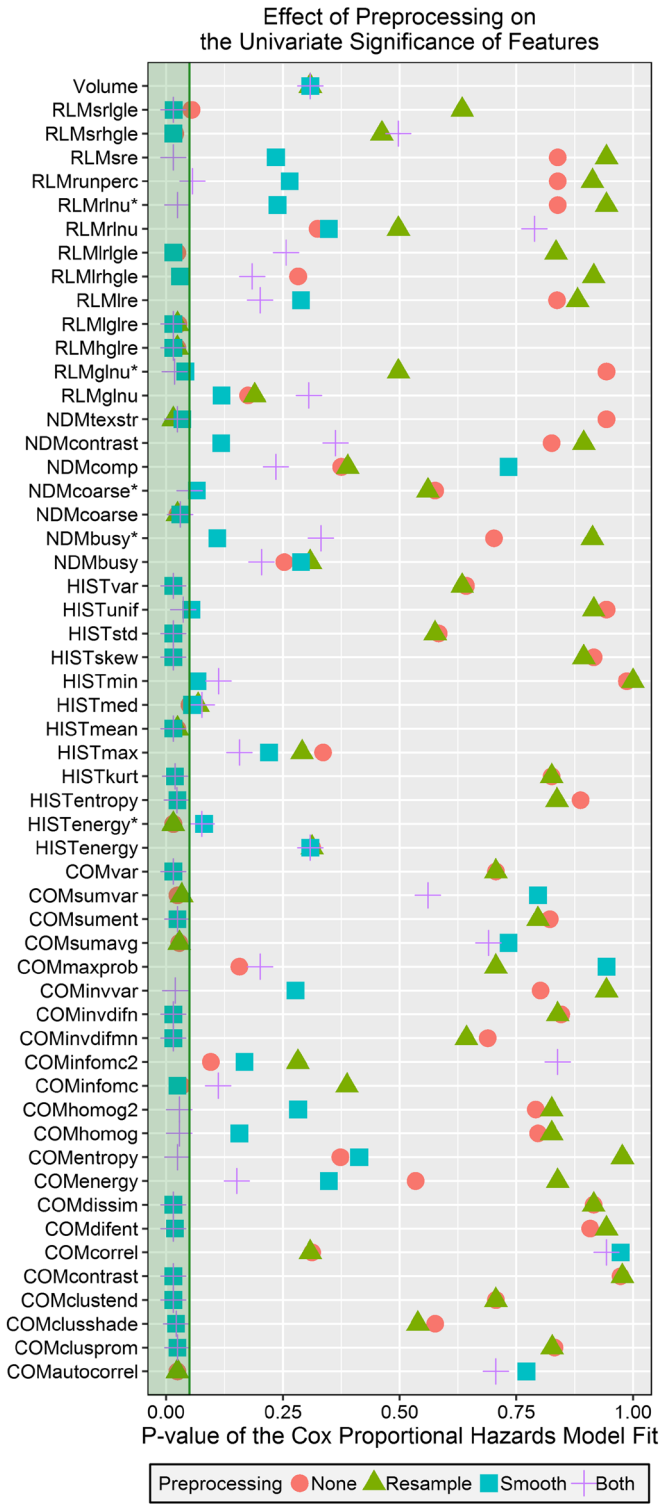

Figure 4 The P values after Benjamini-Hochberg correction for the univariate cox proportional hazards model for each feature and preprocessing combination. The green region of the plot indicates significant values: P value <0.05. Volume was included in the feature set for comparison.

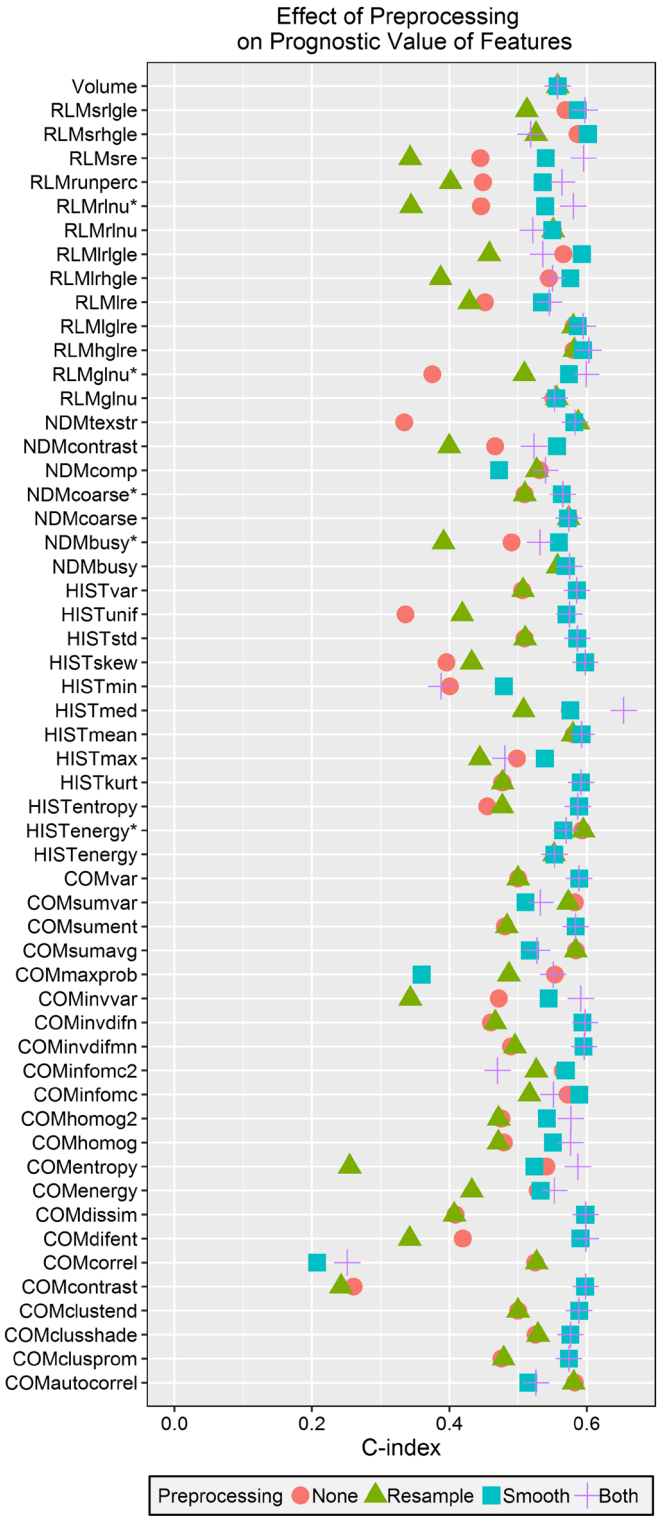

Figure 5 Harrell's concordance index (c-index) was calculated for each of the features after each type of pre-processing using the predictions generated for each patient during leave-one out cross validation. With the exception of minimum intensity, for every feature at least one preprocessing technique resulted in a c-index value greater than 0.5, though only one was larger than 0.6.

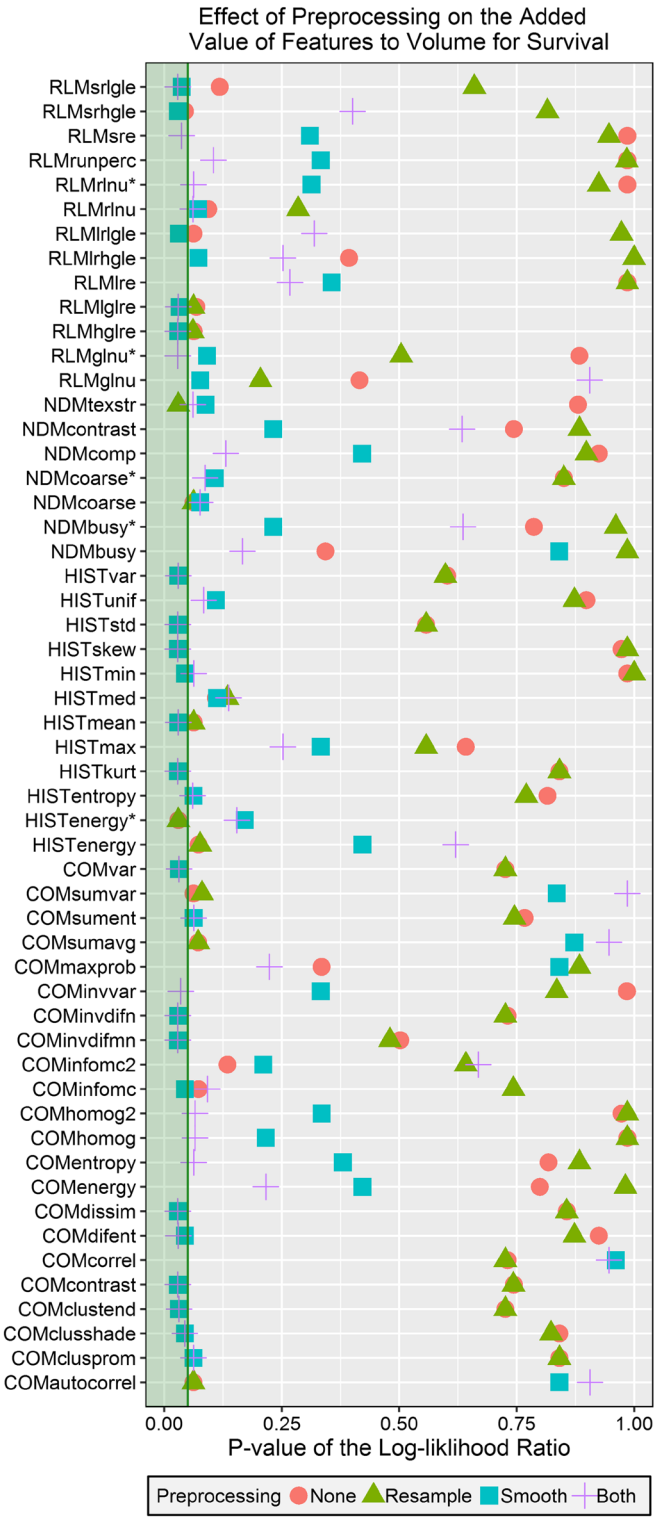

Figure 6 The Benjamini-Hochberg corrected P values for the log-likelihood ratio between cox proportional hazards models with only volume as a covariate and models with volume and one radiomics feature at a time are plotted. The region in green highlights significant P values <0.05. Features were most likely to be significant if they had been preprocessed with Butterworth smoothing either alone or in conjunction with 8 bit depth resample.

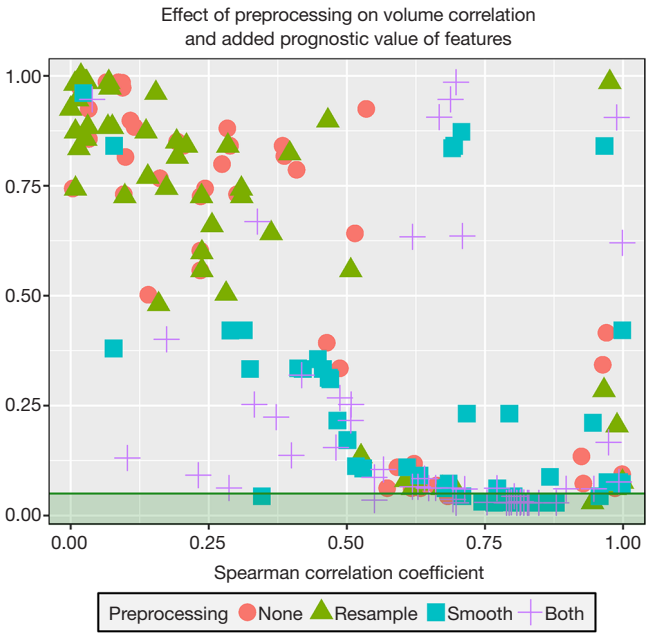

Figure 7 This graph shows the relationship between the volume dependence measured with the spearman correlation coefficient and the added prognostic value of the features (measured with the log-likelihood ratio between cox proportional hazards models for overall survival using only volume and models using volume and one radiomics feature). The green area highlights features whose P value was <0.05 and thus significant.

significant P value from this test. Most of the significant features were calculated with either Butterworth smoothing or Butterworth smoothing and 8 bit depth resampling. Short run high gray-level emphasis energy added significant value to the model when no preprocessing was used, texture strength from the NDM added significant value when 8 bit depth resampling was used, and the volume-corrected version of energy from the histogram was significant when either no preprocessing or 8 bit depth resampling was used. Approximately half of the features (29/54) were not significant regardless of which preprocessing technique was used. This subset included at least one feature from each of the feature categories and all 5 of the uncorrected, volume correlated features identified in the previous section.

Prognostic potential versus volume dependence

The corrected P values for the log-likelihood ratio between Cox proportional hazards models fit with volume as their only covariate and models fit with both volume and one radiomics feature are plotted against the volume correlation for each feature after each preprocessing technique (*Figure* 7). All but one of the features with a significant P value for the log-likelihood ratio had at least a slight correlation with volume (r_s >0.5). However, many features with equally high or higher correlations with volume did not have significant P values. Thus, features with significant P values and some correlation with volume are likely providing complementary information. The only feature with a significant P value and a correlation coefficient less than 0.5 was the minimum of the histogram after Butterworth smoothing. Features with very high volume (r_s >0.95) correlations were not able to add significant value to models built using volume. These features included all of the preprocessing versions of the five original, volume correlated algorithms from the first section of the results, as well as the unpreprocessed versions of the information measure correlation and information measure correlation 2 from the COM and the smoothed version of the information measure correlation 2.

Discussion

Our analysis demonstrated that preprocessing can have a strong impact on the volume dependence and univariate significance of many radiomics features. Specifically, Butterworth smoothing increased the likelihood that a feature was significant in univariate Cox proportional hazards models and significantly improved the model fit in Cox proportional hazards models that included volume as a second covariate. This may be because smoothing removes some of the noise in the image and thus allows the measured features to better represent the tumor's relative heterogeneity and thus its likelihood of responding to treatment. However, this preprocessing technique also increased the correlation with volume of a feature, suggesting that preprocessing techniques must be chosen carefully.

Recent studies of patients with NSCLC have demonstrated the potential of image texture analysis to aid physicians in identifying high-risk patients (4), diagnosing lesions (7,22,23), and predicting overall survival (1,3,5). By providing prognostic or diagnostic information that is complementary to current clinical data, quantitative imaging features could impact clinical decisions prior to treatment. Quantitative imaging techniques have the added benefits of being non-invasive and not time-intensive because they use images that are routinely acquired as the standard of care.

However, although various studies have identified features that on their own or as part of a model may yield prognostic information, very little research has been done on the physical basis for high or low feature values. As observed in the present study, one tumor characteristaaic that can influence feature values is tumor volume. We identified 5 features that were highly correlated with volume owing to terms in the feature algorithms that are directly affected by the number of voxels in the entire image. This dependence would not have been an issue in the original design of these features, which were used only to compare aerial photographs that were the same size. However, in tumor analysis, patients with the same relative heterogeneity can have substantially different tumor sizes and thus widely different values for a radiomics feature if it is dependent on the number of voxels. In our analysis, simple normalizations of the original algorithms were able to lower these correlations. Additionally, we showed that the original versions of the algorithms for these 5 features were not able to add significant value to outcome models that already had volume as a covariate. While for two of these features (energy and grey-level non-uniformity), normalizing them did result in significant P values. Because the direct dependencies we discovered were inherent to the texture equations and not the images, the same relationships are likely to exist in images of different types of cancer, especially those that span a large range of volumes. Similarly, although we used three-dimensional ROIs to capture the full heterogeneity of the tumor, several previous studies have used only the largest axial image slice when determining their ROI (23,24). The strong dependencies we found for these five features will also apply to two-dimensional slice studies because the algorithms are inherently volume-dependent. Thus, we recommend that future studies consider including these modified algorithms in their future feature sets in place of the original volume-dependent features.

A large fraction of the features studied in this work both with and without image preprocessing were at least slightly (r_s >0.5) correlated with volume. These relationships are not necessarily problematic, as the features may still provide information that is complementary to volume. For example, surface area is known to be correlated with volume, yet provides important new information. This idea was supported by the fact that almost all of the features with a significant P value for the log-likelihood test comparing models with volume as a covariate to models with volume and a radiomics feature were at least slightly correlated to volume (r_s >0.5). These correlations may be due to actual differences in the heterogeneities of large versus small tumors on average which the features are designed to measure. To reiterate, a feature correlated with volume should not necessarily be excluded from a dataset, but a feature calculated with an algorithm that is inherently dependent on the number of voxels should be changed or removed. Otherwise, that feature would return two different values when measured from two ROIs of different sizes even if both have the same intensity in each pixel, e.g., two circles filled with pixels of intensity 20 but one has a radius of 5 pixels and one has a radius of 10 pixels.

In this study we also examined the impact of different preprocessing techniques on both the correlation with volume and prognostic significance of each radiomics feature. For some features an increase in the correlation with volume due to preprocessing may represent the amount of information lost in the image. For example, an image that has been overly smoothed eventually has only one intensity value in all of its voxels. Then, because all of the texture information captured by radiomics features has been erased, the feature could represent only the volume information, which is not affected by image preprocessing. However, using no preprocessing at all can also result in meaningless feature values because the values can be dominated by noise in the image. The ideal preprocessing technique for a particular feature would reduce this image noise while maintaining the tumor's actual relative heterogeneity to generate useful information for modeling. Because a ground truth is not known for radiomics features, we used the significance of the features in univariate analysis to evaluate the usefulness of each feature. If a feature was significant in the univariate analysis, then the preprocessing was concluded to have helped it. We found that, in general, using a Butterworth smoothing filter, either on its own or in conjunction with 8 bit depth resampling, resulted in the ability to extract statistically significant features from tumor ROIs. However, the specific trends were feature- dependent. Thus, feature-specific image preprocessing may be required to maximize the usefulness of each radiomics feature. This is perhaps not surprising considering the differences in specific features. For example, the mean intensity from the histogram would change less with smoothing than a feature from the COM, which could benefit from appropriate bin sizes in the calculation of the matrix and thus the right choice for bit depth rescale.

One limitation of the current study is that only 3 different preprocessing techniques were tested. It is possible that superior preprocessing techniques could exist, such as using voxel size resampling or edge-detection filtering, or that fine-tuning the parameters could improve these techniques, such as a 6 bit depth resample in place of 8. The preprocessing techniques used in this study did not comprise an exhaustive set but instead were selected to demonstrate the large changes in a feature's univariate significance that can occur by using different methods for noise-reduction before feature calculation. Because studies have been published with the same features but different preprocessing techniques and parameters for their feature matrices (COM, NDM, and RLM) this is an important result that must be investigated in order for these features to eventually be standardized and used clinically. Similarly, although it is likely that many of the specific trends described in the current study will be different for other image modalities or tumor sites, the overall conclusion that image preprocessing can substantially affect the overall usefulness of a feature should apply in any case. Thus, we highly recommend that future studies examine the most appropriate features to be used for a particular patient population and the calculation parameters accompanying those features before including the features in prognostic models.

Conclusions

Radiomics features calculated from a variety of imaging modalities are being widely studied for potential to help predict patient outcomes or aid physicians in diagnosis. However, so far studies have calculated features using a wide range of software, parameters, and pre-processing techniques. The aim of the current study was to demonstrate the effect that different pre-processing techniques can have on the usefulness of radiomics features by measuring the volume-dependence and prognostic value of each feature in univariate models. We proposed normalization factors for five features that were highly volume-dependent regardless of the preprocessing technique used. Additionally, we found that most features benefited from image smoothing using a Butterworth filter, either alone or in conjunction with 8 bit depth resampling. While smoothing was more likely to make a feature statistically significant in a univariate model, smoothing also tends to increase the volume dependence of the feature. It is important to balance these two effects in order to determine the optimal preprocessing technique for each feature.

Acknowledgments

The authors would like to acknowledge Erica Goodoff for help with manuscript preparation.

Funding: This work was supported by the National Institutes of Health (grant number 5U19CA021239); and the Cancer Prevention Research Institute of Texas (grant number RP110562-P2).

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Translational Cancer Research* for the series "Radiomics in Radiation Oncology". The article has undergone external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tcr.2016.07.11). The series "Radiomics in Radiation Oncology" was commissioned by the editorial office without any funding or sponsorship. Court LE served as the unpaid Guest Editor of the series. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study received institutional review board approval and individual consent for this retrospective analysis was waived.

References

1. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun

2014;5:4006.
2. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Transl Oncol 2014;7:72-87.
3. Ganeshan B, Panayiotou E, Burnand K, et al. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol 2012;22:796-802.
4. Weiss GJ, Ganeshan B, Miles KA, et al. Noninvasive image texture analysis differentiates K-ras mutation from pan-wildtype NSCLC and is prognostic. PLoS One 2014;9:e100244.
5. Win T, Miles KA, Janes SM, et al. 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 2013;19:3591-9.
6. Gevaert O, Xu J, Hoang CD, et al. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data--methods and preliminary results. Radiology 2012;264:387-96.
7. Basu S, Hall LO, Goldgof DB, et al. editors. Developing a classifier model for lung tumors in CT-scan images. Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference On. Anchorage:IEEE;2011:1306-12.
8. Ganeshan B, Goh V, Mandeville HC, et al. Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. Radiology 2013;266:326-36.
9. Al-Kadi OS, Watson D. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. IEEE Trans Biomed Eng 2008;55:1822-30.
10. Fried DV, Tucker SL, Zhou S, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2014;90:834-42.
11. Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. PLoS One 2015;10:e0124165.
12. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. IEEE Trans Syst Man Cybern B-Syst Man Cybern 1973;6:610-21.
13. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 1989;19:1264-74.
14. Galloway MM. Texture analysis using gray level run lengths. Comput Graph Image Process 1975;4:172-9.
15. Hatt M, Majdoub M, Vallières M, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med 2015;56:38-44.
16. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med 2014;55:37-42.
17. Image-Guided Adaptive Conformal Photon Versus Proton Therapy. Available online: https://clinicaltrials.gov/ct2/show/study/NCT00915005?view=record
18. Seppenwoolde Y, Shirato H, Kitamura K, et al. 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 2002;53:822-34.
19. Fave X, Mackin D, Yang J, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Med Phys 2015;42:6784-97.
20. Yang J, Zhang L, Fave XJ, et al. Uncertainty analysis of quantitative imaging features extracted from contrast-enhanced CT in lung tumors. Comput Med Imaging Graph 2016;48:1-8.
21. Zhang L, Fried DV, Fave XJ, et al. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42:1341-53.
22. Wang H, Guo XH, Jia ZW, et al. Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image. Eur J Radiol 2010;74:124-9.
23. Bayanati H, E Thornhill R, Souza CA, et al. Quantitative CT texture and shape analysis: can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? Eur Radiol 2015;25:480-7.
24. Cunliffe A, Armato SG 3rd, Castillo R, et al. 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 2015;91:1048-56.

Cite this article as: Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Court LE. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl Cancer Res 2016;5(4):349-363. doi: 10.21037/tcr.2016.07.11

