Integration of risk survival measures estimated from pre- and post-treatment CT scans improves stratification of early stage non-small cell lung cancer patients treated with stereotactic body radiation therapy

Supplementary data

1. Radiomics features

Imaging data used for extracting radiomics features consists of both original CT scans and data generated from wavelet decomposition of them. In this study, we applied a one-level, discrete, and undecimated three-dimensional wavelet transform to each CT image, which decomposes the original image into 8 decompositions. According to the low pass (*L*) or high pass (*H*) wavelet function in each dimension, the decomposed data sequences are labeled as *LLL*, *LLH*, *LHL*, *HLL*, *HHL*, *LHH*, *HLH*, and *HHH*. We totally defined 680 radiomics features to characterize the tumor properties of one patient, and these features can be categorized as *shape features*, *first-order features*, *gray level co-occurrence matrix* (GLCM) *features*, *gray level size zone matrix* (GLSZM) *features*, and *gray level run length matrix* (GLRLM) *features*. Specifically, the *shape features* are shared across all the 9 sequences (1 original sequence and 8 decomposed sequences), while the *first-order features*, GLCM, GLSZM, and GLRLM separately calculated from each of the 9 sequences. Detailed names of all the calculated radiomics features are summarized in Supplementary Table 1.

Supplementary Table 1. Detailed names of radiomics features.

shape features	shape_Elongation	**GLCM features**	glcm_Autocorrelation
	shape_Flatness		glcm_ClusterProminence
	shape_LeastAxisLength		glcm_ClusterShade
	shape_MajorAxisLength		glcm_ClusterTendency
	shape_Maximum2DDiameterColumn		glcm_Contrast
	shape_Maximum2DDiameterRow		glcm_Correlation
	shape_Maximum2DDiameterSlice		glcm_DifferenceAverage
	shape_Maximum3DDiameter		glcm_DifferenceEntropy
	shape_MeshVolume		glcm_DifferenceVariance
	shape_MinorAxisLength		glcm_Id
	shape_Sphericity		glcm_Idm
	shape_SurfaceArea		glcm_Idmn
	shape_SurfaceVolumeRatio		glcm_Idn
	shape_VoxelVolume		glcm_Imc1
first-order features	firstorder_10Percentile		glcm_Imc2
	firstorder_90Percentile		glcm_InverseVariance
	firstorder_Energy		glcm_JointAverage
	firstorder_Entropy		glcm_JointEnergy
	firstorder_InterquartileRange		glcm_JointEntropy
	firstorder_Kurtosis		glcm_MCC
	firstorder_Maximum		glcm_MaximumProbability
	firstorder_MeanAbsoluteDeviation		glcm_SumAverage
	firstorder_Mean		glcm_SumEntropy

|	Feature		GLRLM features
	firstorder_Median		glcm_SumSquares
	firstorder_Minimum		glrlm_GrayLevelNonUniformity
	firstorder_Range		glrlm_GrayLevelNonUniformityNormalized
	firstorder_RobustMeanAbsoluteDeviation		glrlm_GrayLevelVariance
	firstorder_RootMeanSquared		glrlm_HighGrayLevelRunEmphasis
	firstorder_Skewness		glrlm_LongRunEmphasis
	firstorder_TotalEnergy		glrlm_LongRunHighGrayLevelEmphasis
	firstorder_Uniformity		glrlm_LongRunLowGrayLevelEmphasis
	firstorder_Variance		glrlm_LowGrayLevelRunEmphasis
	glszm_GrayLevelNonUniformity	**GLRLM features**	glrlm_RunEntropy
GLSZM features	glszm_GrayLevelNonUniformityNormalized		glrlm_RunLengthNonUniformity
	glszm_GrayLevelVariance		glrlm_RunLengthNonUniformityNormalized
	glszm_HighGrayLevelZoneEmphasis		glrlm_RunPercentage
	glszm_LargeAreaEmphasis		glrlm_RunVariance
	glszm_LargeAreaHighGrayLevelEmphasis		glrlm_ShortRunEmphasis
	glszm_LargeAreaLowGrayLevelEmphasis		glrlm_ShortRunHighGrayLevelEmphasis
	glszm_LowGrayLevelZoneEmphasis		glrlm_ShortRunLowGrayLevelEmphasis
	glszm_SizeZoneNonUniformity		
	glszm_SizeZoneNonUniformityNormalized		
	glszm_SmallAreaEmphasis		
	glszm_SmallAreaHighGrayLevelEmphasis		
	glszm_SmallAreaLowGrayLevelEmphasis		
	glszm_ZoneEntropy		
	glszm_ZonePercentage		
	glszm_ZoneVariance		

2. Parameter settings of random forest models

To optimize and select parameters of our random survival forest structures, we applied a 3-fold cross validation on the 100-subject cohort. In this process, the RSF models are characterized by four parameters of (1) n_estimators: number of trees in the forest; (2) max_depth: maximum depth of the tree; (3) min_samples_split: minimum number of samples required to split an internal node; and (4) min_samples_leaf: minimum number of samples required to be at a leaf node. Values of all these parameters are optimized by a 3-fold cross-validation procedure of RSF models on the 100-patient cohort. The candidate ranges of parameters are: (1) n_estimators: [10, 500]; (2) max_depth: [1, 5]; (3) min_samples_split: [1, 10]; (4) min_samples_leaf: [1, 10]. The optimized values for the 3-fold cross validation are summarized in Supplementary Table 2.

Supplementary Table 2. Optimal parameter values for the survival prediction modeling.

Name of Parameters	Value of parameter
n_estimators	40
max_depth	3
min_samples_split	2
min_samples_leaf	8

## 3. GTV size of all patient cohorts

The GTV size of all patient cohorts is shown in Supplementary Figure 1. Specifically, the GTV size of the 100-patient cohort and the 60-patient cohort before and after treatment is 1.99±0.87 cm^3, 1.77±0.80 cm^3, and 1.19±0.56 cm^3, respectively.
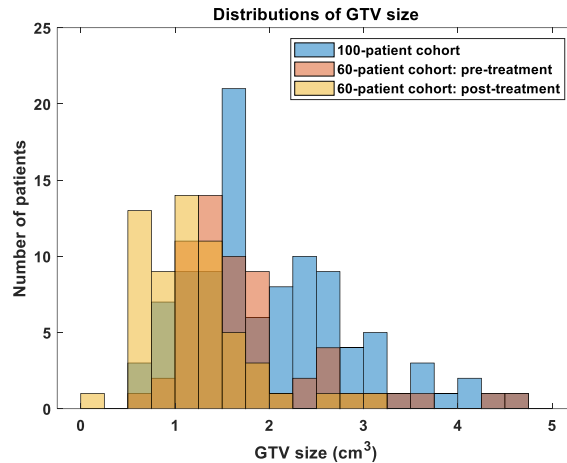
Supplementary Figure 1. Distribution of GTV size of all patient cohorts.

4. Selection of radiomic features

To improve the prediction performance, we adopted four feature selection methods (Parmar, et al., 2015) to select the most informative radiomic features based on the training data, including Wilcoxon test based feature selection (WLCX), Mutual information based feature selection (MIFS), Minimum redundancy maximum relevance (MRMR), and T-test score based feature selection (TSCR). The prediction performance of survival analysis models built upon selected features by these methods is shown in Supplementary Figure 2. Based on these results, we chose the MRMR method to select top 70% radiomics features and built a survival analysis model which was finally combined with a survival analysis model of clinical data based on the training data. The integrated survival analysis model obtained a c-index of 0.744 on the training cohort and a c-index of 0.734 on the testing cohort.
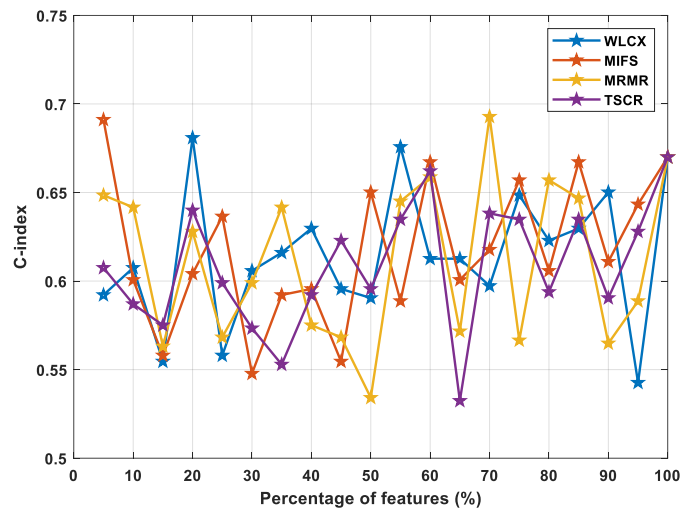
Supplementary Figure 2. Prediction performance of survival analysis models built upon radiomic features selected by different feature selection methods.

References:
Parmar, Chintan, et al. "Machine learning methods for quantitative radiomic biomarkers." Scientific reports 5 (2015): 13087.