Supplementary

Table S1 Baseline characteristics of patients with and without micropapillary/solid pattern in primary cohort

Variables	Total (n=268)	Without MP/S (n= 170)	With MP/S (n=98)	P
Gender, n				<0.001
Female	150 (56.0%)	110 (64.7%)	40 (40.8%)	
Male	118 (44%)	60 (35.3%)	58 (59.2%)	
Age, years	61 (30–87)	60 (30–85)	62 (34–87)	0.19
Smoking history				0.21
Smoker	56 (20.9%)	31 (18.2%)	25 (25.5%)	
Non-smoker	212 (79.1%)	139 (81.8%)	73 (74.5%)	
Lobular location, n				0.73
RUL	84 (31.3%)	51 (30.0%)	33 (33.7%)	
RML	18 (6.7%)	12 (7.1%)	6 (6.1%)	
RLL	55 (20.5%)	35 (20.6%)	20 (20.4%)	
LUL	71 (26.5%)	43 (25.3%)	28 (28.6%)	
LLL	40 (14.9%)	29 (17.1%)	11 (11.2%)	
Predominant subtype, n				<0.001
Lepidic	90 (33.6%)	83 (48.8%)	7 (7.1%)	
Acinar	94 (35.1%)	58 (34.1%)	36 (36.7%)	
Papillary	50 (18.7%)	29 (17.1%)	21 (21.4%)	
Solid	25 (9.3%)	0	25 (25.5%)	
Micropapillary	9 (3.4%)	0	9 (9.2%)	
Lymph node metastasis, n	27 (10.1%)	9 (5.3%)	18 (18.4%)	<0.001
N1 metastasis, n	20 (7.5%)	7 (4.1%)	13 (13.3%)	–
N2 metastasis, n	17 (6.3%)	3 (1.8%)	14* (14.3%)	–
CEA, n				0.02
≥5 ng/mL	34 (14.6%)	15 (10.2%)	19 (22.1%)	
<5 ng/mL	234 (85.4%)	155 (91.2%)	79 (80.6%)	
Size (cm)	2.47±0.88	2.34±0.84	2.66±0.91	0.03

*, 5 patients had multiple mediastinal lymph node metastasis. RUL, right upper lobe; RML, right middle lobe; RLL, right lower lobe; LUL, left upper lobe; LLL, left lower lobe; VPI, visceral pleural invasion; CEA, carcinoembryonic antigen.

http://dx.doi.org/10.21037/tlcr-21-44

Appendix 1

Radiomics feature extraction

90 CT-based radiomics features of five categories were extracted: (I) Shape; (II) First order statistics; (III) gray level co-occurrence matrix (GLCM); (IV) gray level run length matrix (GLRLM); (V) gray level size zone matrix (GLSZM).

Group 1. Shape (n=14)

1: Volume (**V**): determined by counting the number of voxels in the nodule region and multiplying this value by the voxel size.

2: Surface area (**S**)=$\sum_{i=1}^{N}\frac{1}{2}|a_i b_i \times a_i c_i|$, where ***N*** is the number of triangles covering the surface and **a**, **b** and **c** are edge vectors.

3: Surface to volume ratio=$\frac{S}{V}$

4: Sphericity=$\frac{\sqrt[3]{36\pi V^2}}{S}$

5: Spherical disproportion=$\frac{S}{4\pi R^2}$, where R is the radius of a sphere with the same volume as the tumor, and equal to $\sqrt[3]{\frac{3V}{4\pi}}$.

6: Maximum 2D Diameter Slice: is defined as the largest pairwise Euclidean distance between tumor surface mesh vertices in the row-column (generally the axial) plane.

7: Maximum 2D Diameter Row: is defined as the largest pairwise Euclidean distance between tumor surface mesh vertices in the column-slice (usually the sagittal) plane.

8: Maximum 2D Diameter Column: is defined as the largest pairwise Euclidean distance between tumor surface mesh vertices in the row-slice (usually the coronal) plane.

9: Maximum 3D diameter: measured as the largest pairwise Euclidean distance between voxels on the surface of the tumor volume.

10: Major axis=$4\sqrt{\lambda_{major}}$

11: Minor Axis=$4\sqrt{\lambda_{minor}}$

 http://dx.doi.org/10.21037/tlcr-21-44

12: Least axis=$4\sqrt{\lambda_{least}}$

13: Elongation=$\sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$.

14: Flatness= $\sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$. Here, λ_{major},λ_{minor} and λ_{least} are the lengths of the largest, second largest and smallest principal component axes.

Group 2. First-order statistics/ Intensity (n=19)

Histogram features describe the distribution of voxel intensities within the CT image commonly used and basic metrics. Let **X** denote the three-dimensional image matrix with **N** voxels and **P** first order histogram with N_l discrete intensity levels.

1: 10 Percentile: the 10th percentile of X.

2: Maximum: the maximum intensity of the **X**

3: Minimum: the minimum intensity of the **X**

4: Median: the median intensity of the **X**

5: Range: the range of intensity values of **X**

6: Mean ($\overline{X}$)=$\frac{1}{N}\sum_i^N X(i)$

7 90Percentile: the 90th percentile of X.

8 Interquartile range=P_{75}-P_{25}, P_{25} and P_{75} are the 25th and 75th percentile of the X.

9: Mean absolute deviation=$\frac{1}{N}\sum_{i=1}^N |X(i) - \overline{X}|$

10 Robust Mean Absolute Deviation=$\frac{1}{N_{10-90}}\sum_{i=1}^{N_{10-90}} |X_{10-90}(i) - \overline{X}_{10-90}|$

11. Standard deviation=$\sqrt{\frac{1}{N}\sum_{i=1}^N (X(i) - \overline{X})^2}$

12: Root mean square=$\sqrt{\frac{\sum_i^N (X(i)+c)^2}{N}}$

13: Energy=$\sum_{i=1}^N (X(i) + c)^2$

14 Total Energy=V$\sum_{i=1}^N (X(i) + c)^2$

Here, c is optional value, defined by voxel Array Shift, which shifts the intensities to prevent negative values in X. This ensures that voxels with the lowest gray values contribute the least to Energy, instead of voxels with gray level intensity closest to 0.

 http://dx.doi.org/10.21037/tlcr-21-44

15: Entropy= $\sum_{i=1}^{N_l} P(i)\log_2(P(i)+\epsilon)$, ϵ is an arbitrarily small positive number ($\approx 2.2\times10^{-16}$).

16: Kurtosis=$\frac{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\overline{X})^4}{(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\overline{X})^2})^2}$

17: Skewness==$\frac{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\overline{X})^3}{(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\overline{X})^2})^3}$

18: Uniformity=$\sum_{i}^{N_l} P(i)^2$

19: Variance=$\frac{1}{N}\sum_{i=1}^{N}(X(i)-\overline{X})^2$

Group 3: Gray Level Co-occurrence Matrix (GLCM) based features (n=25)

A GLCM is defined as $\mathrm{p}(i,j,\delta,\alpha)$, a matrix with size $N_g \times N_g$ describing the second order joint probability function of an image, where the (i,j) th element represents the number of times the combination of intensity levels i and j occur in two pixels in the image, that are separated by a distance of δ pixels in direction α, and N_g is the number of discrete gray level intensities. In our study, distance δ was set to 1 and direction α to each of the 13 directions in three-dimensions.

Each 3D gray level co-occurrence-based feature was calculated as the mean of the feature calculations for each of the 13 directions.

Let:

$\mathrm{P}(i,j)$ be the co-occurrence matrix for an arbitrary δ and α,

N_g be the number of discrete intensity levels in the image,

μ be the mean of $\mathrm{P}(i,j)$,

$p_x(i)=\sum_{j=1}^{N_g}\mathrm{P}(i,j)$ be the marginal row probabilities,

$p_y(i)=\sum_{i=1}^{N_g}\mathrm{P}(i,j)$ be the marginal column probabilities,

μ_x be the mean of p_x,

μ_y be the mean of p_y,

σ_x be the standard deviation of p_x,

σ_y be the standard deviation of p_y,

$p_{x+y}(k)=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\mathrm{P}(i,j), i+j=k, k=2,3\ldots,2N_g$,

http://dx.doi.org/10.21037/tlcr-21-44

$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \mathrm{P}\,(i,j)\,, |i-j| = k, k = 0,1\ldots, N_g - 1,$

$\mathrm{HX} = -\sum_{i=1}^{N_g} p_x(i) \log_2 [p_x(i)]$ be the entropy of p_x,

$\mathrm{HY} = -\sum_{i=1}^{N_g} p_y(i) \log_2 [p_y(i)]$ be the entropy of p_y,

$\mathrm{H} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \mathrm{P}\,(i,j) \log_2 [P(i,j)]$ be the entropy of $\mathrm{P}\,(i,j)$,

$\mathrm{HXY1} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \mathrm{P}\,(i,j) \log (p_x(i)\, p_y(j))$

$\mathrm{HXY2} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i) p_y(j) \log (p_x(i)\, p_y(j))$

1: Inverse difference moment normalized (IDMN) $= \sum_{=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\frac{k^2}{N^2}}$

2: Joint energy $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P\,((i,j))^2$

3: Difference average (DA) $= \sum_{k=0}^{N_g-1} k P_{x-y}(k)$

4: Difference variance $= \sum_{k=0}^{N_g-1} (k - DA) P_{x-y}(k)$

5: Sum squares $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [i - \mu_x]^2\, P\,(i,j)$

6: Joint entropy $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)(k) \log_2 [p(i,j) + \epsilon]$

7: Inverse difference (ID) $= \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+(\frac{k}{N_g})}$

8: Joint average $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} iP\,(i,j)$

9: IDM $= \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+(\frac{k^2}{N_g{}^2})}$

10: Autocorrelation $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} ijP\,(i,j)$

11: Cluster prominence $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [i + j - \mu_x - \mu_y]^4\, P\,(i,j)$

12: Cluster shade $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [i + j - \mu_x - \mu_y]^3\, P\,(i,j)$

13: Cluster tendency $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [i + j - \mu_x - \mu_y]^2\, P\,(i,j)$

14: Correlation $= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} ij\, P\,(i,j) - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$

15: Contrast $= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} |i-j|^2 P\,(i,j)$

16: Difference entropy $= \sum_{k=0}^{N_g-1} P_{x-y}(k) \log_2 [P_{x-Y}(k) + \epsilon]$

http://dx.doi.org/10.21037/tlcr-21-44

17: Homogeneity1=$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{P(i,j)}{1+|i-j|}$

18: Homogeneity2=$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{P(i,j)}{1+|i-j|^2}$

19: Informational measure of correlation 1(IMC1) =$\frac{HXY-HXY1}{\max\{HX,HY\}}$

20: Informational measure of correlation 2(IMC2) =$\sqrt{1-e^{-2(HXY2-HXY)}}$

21: Maximal Correlation Coefficient = $\sqrt{\sum_{k=0}^{N_g}\frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}}$

22: Inverse difference normalized (IDN)= $\sum_{k=0}^{N_{g-1}}\frac{p_{x-y}(k)}{1+(\frac{k}{N_g})}$

23: Inverse variance=$\sum_{k=1}^{N_g}\frac{p_{x-y}(k)}{k^2}$,

24: Maximum probability=max{ P (i,j)}

25: Sum entropy=$-\sum_{k=2}^{2N_g}P_{x+y}(k)\log_2[P_{x+y}(k)+\epsilon]$

Group 4: Gray-Level Run-Length matrix based features (n=16)

Run-Length metrics quantify gray level runs in an image. A gray level run is defined as the length in numbers of pixels, of consecutive that have the same gray level value. In a gray level run-length matrix p (i,j,θ), the (i,j)th element describes the number of times j a gray level i appears consecutively in the direction specified by θ, and N_g is the number of discrete gray level intensities.

Let:

N_g : the number of discrete intensity values in the image

N_r: the number of different run lengths

N_p : the number of voxels in the mage

$N_z(\theta)=\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}$[p (i,j,θ)] : the number of runs in the image along angle θ

P (i,j,θ): the run length matrix for an arbitrary direction θ

p(i,j,θ): the normalized run length matrix, p$(i,j,\theta)=\frac{\text{P }(i,j,\theta)}{N_z(\theta)}$

1: Run entropy (RE)= $\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j|\theta)\,p((i,j|\theta)+\epsilon)$

http://dx.doi.org/10.21037/tlcr-21-44

2: Run variance (RV)= $\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j|\theta)\,(j-u)^2$

3: Grey level variance (GLV)= $\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j|\theta)\,(i-u)^2$

4: Gray level non-uniformity Normalized (GLNN)= $\frac{\sum_{j=1}^{N_r}[\sum_{i=1}^{N_g} \mathrm{p}\,(i,j,\theta)]^2}{N_z(\theta)^2}$

5: Run length non-uniformity Normalized (RLNN)= $\frac{\sum_{j=1}^{N_r}[\sum_{i=1}^{N_g} \mathrm{p}\,(i,j,\theta)]^2}{N_z(\theta)^2}$

6: Short run emphasis (SRE)=$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}[\frac{\mathrm{p}\,(i,j,\theta)}{j^2}]}{N_z(\theta)}$

7: Long run emphasis (LRE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} j^2\,\mathrm{p}\,(i,j,\theta)}{N_z(\theta)}$

8: Gray level non-uniformity (GLN)= $\frac{\sum_{i=1}^{N_g}[\sum_{j=1}^{N_r} \mathrm{p}\,(i,j,\theta)]^2}{N_z(\theta)}$

9: Run length non-uniformity (RLN)= $\frac{\sum_{j=1}^{N_r}[\sum_{i=1}^{N_g} \mathrm{p}\,(i,j,\theta)]^2}{N_z(\theta)}$

10: Run percentage (RP)= $\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathrm{p}\,(i,j,\theta)}{N_p}$

11: Low gray level run emphasis(LGLRE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}[\frac{\mathrm{p}\,(i,j,\theta)}{i^2}]}{N_z(\theta)}$

12: High gray level run emphasis (HGLRE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2\,\mathrm{p}\,(i,j,\theta)}{N_z(\theta)}$

13: Short run low gray level emphasis(SRLGLE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}[\frac{\mathrm{p}\,(i,j,\theta)}{i^2j^2}]}{N_z(\theta)}$

14: Short run high gray level emphasis(SRHGLE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}[\frac{\mathrm{p}\,(i,j,\theta)i^2}{j^2}]}{N_z(\theta)}$

15: Long run low gray level emphasis(LRLGLE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}[\frac{\mathrm{p}\,(i,j,\theta)j^2}{i^2}]}{N_z(\theta)}$

16: Long run high gray level emphasis(LRHGLE)= $\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2j^2\,\mathrm{p}\,(i,j,\theta)}{N_z(\theta)}$

Group 5: Gray-level size zone matrix (n=16)

A Gray Level Size Zone Matrix (GLSZM) quantifies gray level zones in an image. A gray level zone is defined as the number of connected voxels that share the same gray level intensity. In a gray level size zone matrix $\mathrm{p}\,(i,j)$, the $(i,j)^{\text{th}}$ element equals the

http://dx.doi.org/10.21037/tlcr-21-44

number of zones with gray level i and size j appear in image.

Let:

N_g: the number of discreet intensity values in the image

N_s : the number of discreet zone sizes in the image

N_p : the number of voxels in the image

$N_z = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \mathrm{p}\ (i,j)$: the number of zones in the tumor

$\mathrm{P}\ (i,j)$ be the size zone matrix

$\mathrm{p}\ (i,j)$ be the normalized size zone matrix, defined as $\mathrm{p}\ (i,j) = \frac{\mathrm{P}\ (i,j)}{\sum \mathrm{P}\ (i,j)}$

1: Large Area Low Gray Level Emphasis (LALGLE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\mathrm{P}\ (i,j) j^2 i^2]}{N_z}$

2: Gray Level Variance (GLV)= $\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} p(i,j)\ (i-u)^2$

3: High Gray Level Zone Emphasis (HGLZE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\mathrm{P}\ (i,j) i^2]}{N_z}$

4: Large Area High Gray Level Emphasis (LAHGLE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\frac{\mathrm{P}\ (i,j) j^2}{i^2}]}{N_z}$

5: Gray Level Non-Uniformity Normalized (GLNN)= $\frac{\sum_{i=1}^{N_g} [\sum_{j=1}^{N_s} \mathrm{p}\ (i,j)]^2}{{N_z}^2}$

6: Small Area High Gray Level Emphasis (SAHGLE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\frac{\mathrm{P}\ (i,j) i^2}{j^2}]}{N_z}$

7: Gray Level Non-Uniformity (GLN)= $\frac{\sum_{i=1}^{N_g} [\sum_{j=1}^{N_s} \mathrm{p}\ (i,j)]^2}{N_z}$

8: Low Gray Level Zone Emphasis (LGLZE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\frac{\mathrm{P}\ (i,j)}{i^2}]}{N_z}$

9: Small Area Low Gray Level Emphasis (SALGLE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\frac{\mathrm{P}\ (i,j)}{i^2 j^2}]}{N_z}$

10: Small area emphasis (SAE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\frac{\mathrm{P}\ (i,j)}{j^2}]}{N_z}$

11: Large area emphasis (LAE)= $\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} [\mathrm{P}\ (i,j) j^2]}{N_z}$

12: Zone percentage (ZP)= $\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \frac{\mathrm{P}\ (i,j)}{N_p}$

http://dx.doi.org/10.21037/tlcr-21-44

13: Zone Variance(ZV)= $\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} \mathrm{p(i,j)(j-\mu)^2}$, $\mu=\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} \mathrm{p(i,j)j}$

14: Zone entropy (ZE)= $\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} \mathrm{p(i,j)}\log_2[p(i,j)+\epsilon]$

15: Size-zone non-uniformity (SZN)= $\frac{\sum_{j=1}^{N_s}[\sum_{i=1}^{N_g} \mathrm{P}(i,j)]^2}{N_z}$

16: Size-zone non-uniformity normalized (SZNN)=$\frac{\sum_{j=1}^{N_s}[\sum_{i=1}^{N_g} \mathrm{P}(i,j)]^2}{{N_z}^2}$

The correlation matrix of the radiomics feature was displayed in *Fig S5,* which revealed high multicollinearity between the 90 extracted features.

http://dx.doi.org/10.21037/tlcr-21-44

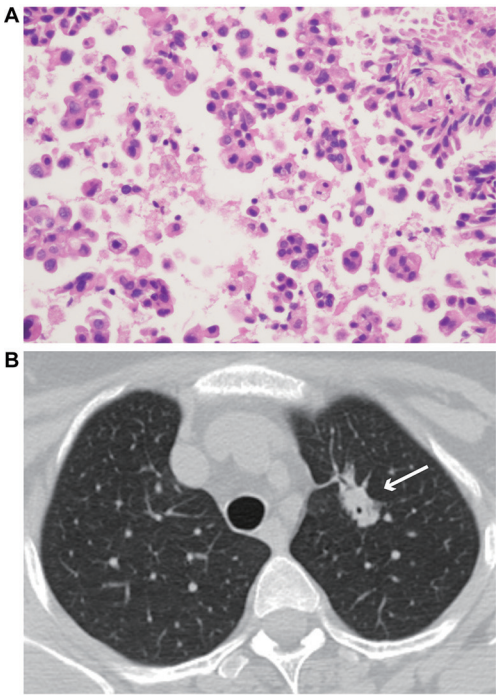

Figure S1 Exemplary pictures of micropapillary growth pattern. (A) Representative hematoxylin and eosin-stained tumor slide of micropapillary pattern. Original magnification, ×20. (B) Radiologic image (axial) of tumor with micropapillary pattern.

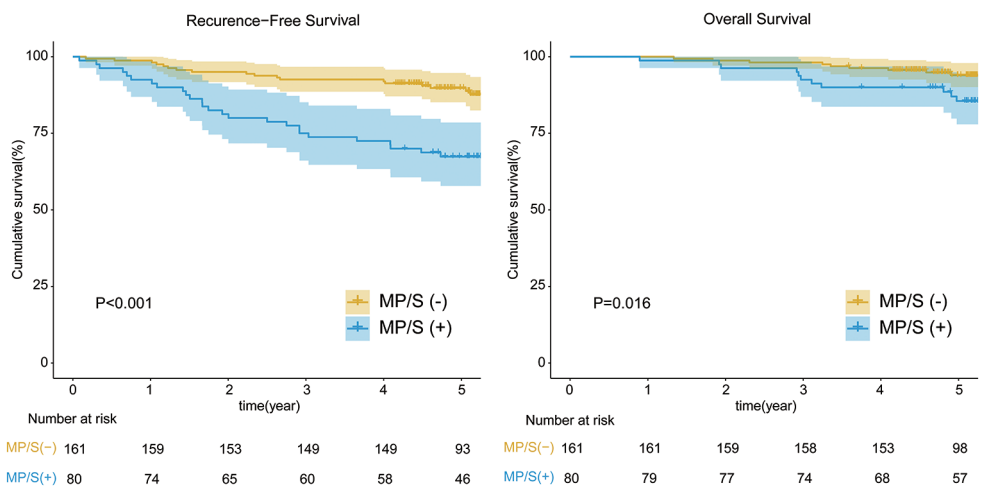

Figure S2 Recurrence-free survival and overall survival curves for patients without lymph node metastasis in the training cohort. MP/S, lung adenocarcinoma containing micropapillary or solid growth pattern.

http://dx.doi.org/10.21037/tlcr-21-44

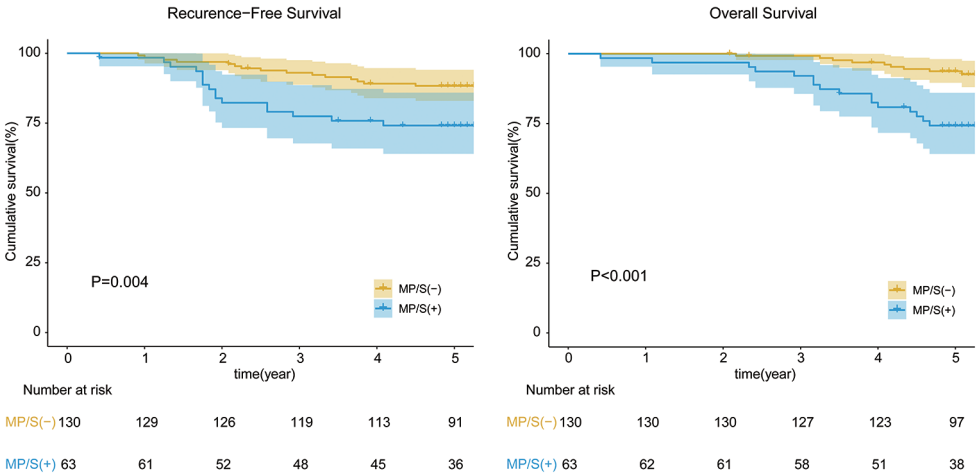

Figure S3 Recurrence-free survival and overall survival curves between patients having MP/S growth pattern and those without in the validation cohort. MP/S, lung adenocarcinoma containing micropapillary or solid growth pattern.

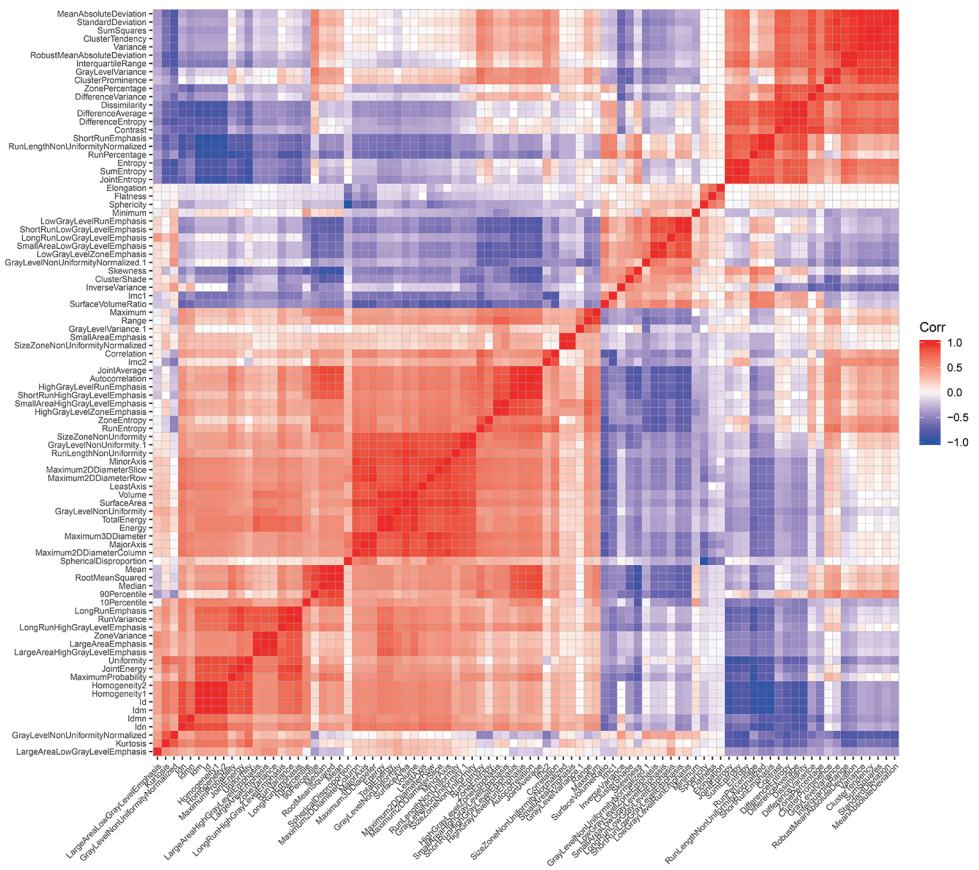

Figure S4 Heatmap of correlation matrix of the 90 extracted radiomics features in primary cohort.

http://dx.doi.org/10.21037/tlcr-21-44

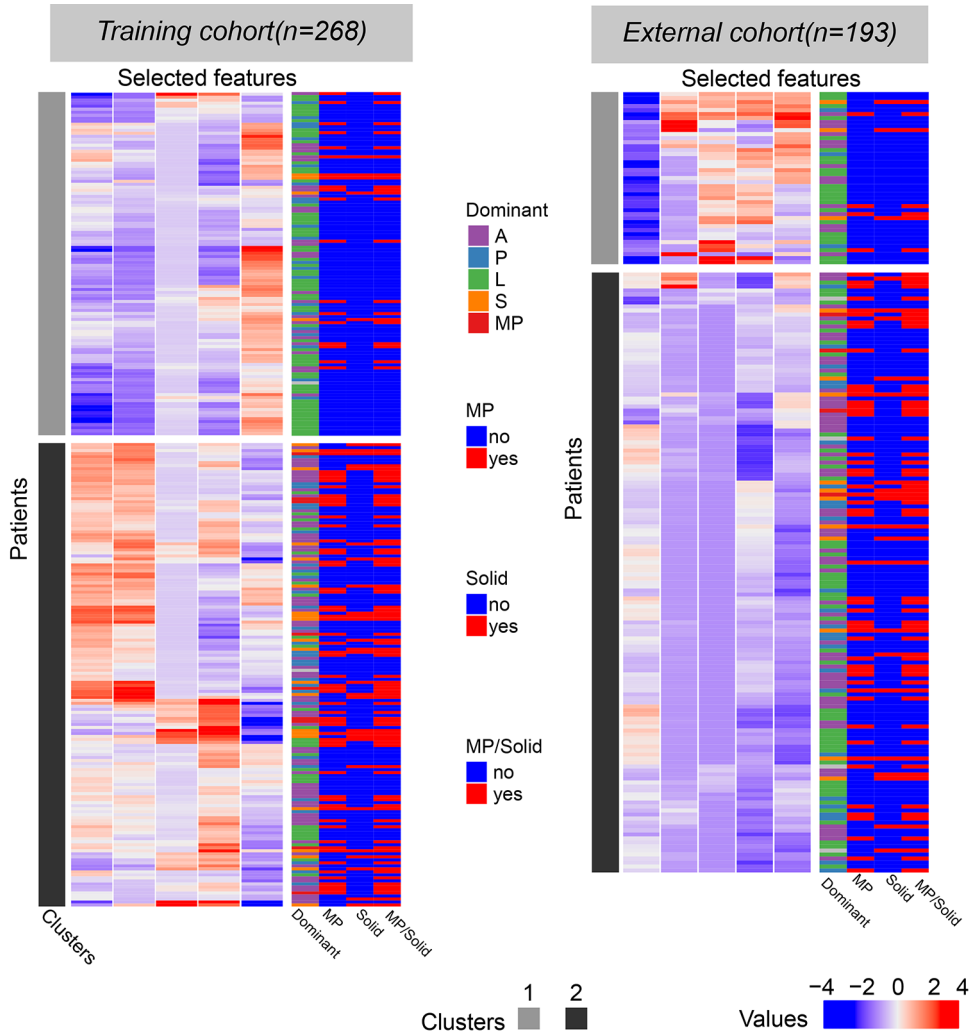

Figure S5 Unsupervised clustering analysis of the five selected radiomics features in both the training and validation cohort.

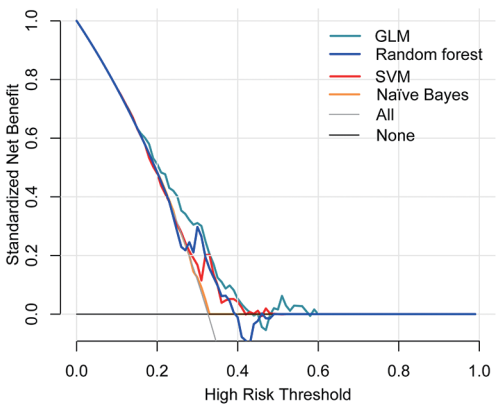

Figure S6 The decision curve analysis of four proposed models in the validation set.

http://dx.doi.org/10.21037/tlcr-21-44