Supplementary

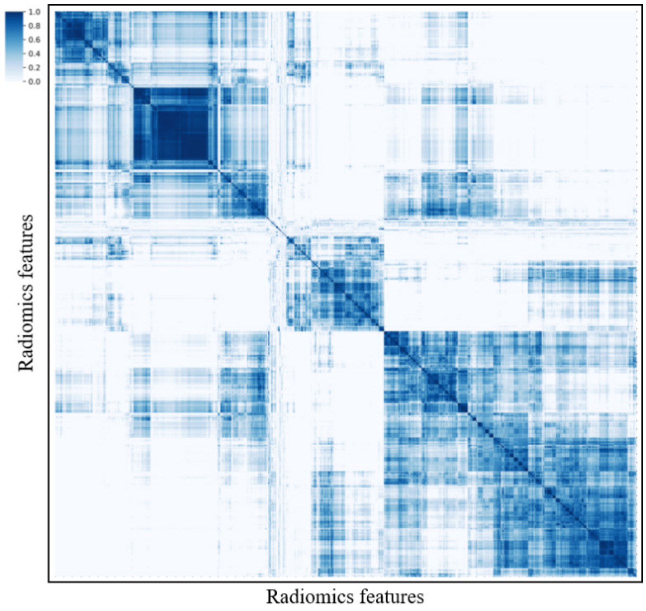

Figure S1 The interrelationship between radiomic features.

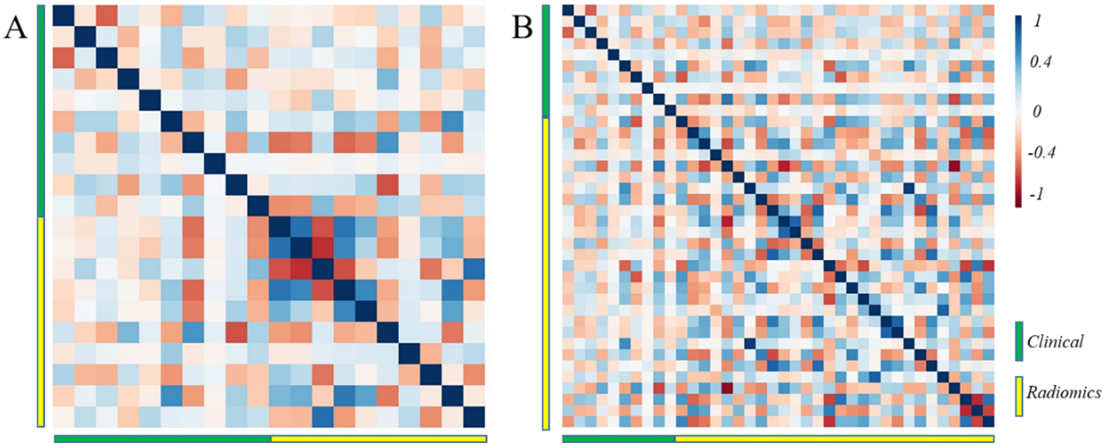

Figure S2 Correlation analysis between radiomic features and clinical characteristics. (A) Correlation map in task [1]. (B) Correlation map in task [2].

https://dx.doi.org/10.21037/qims-21-980

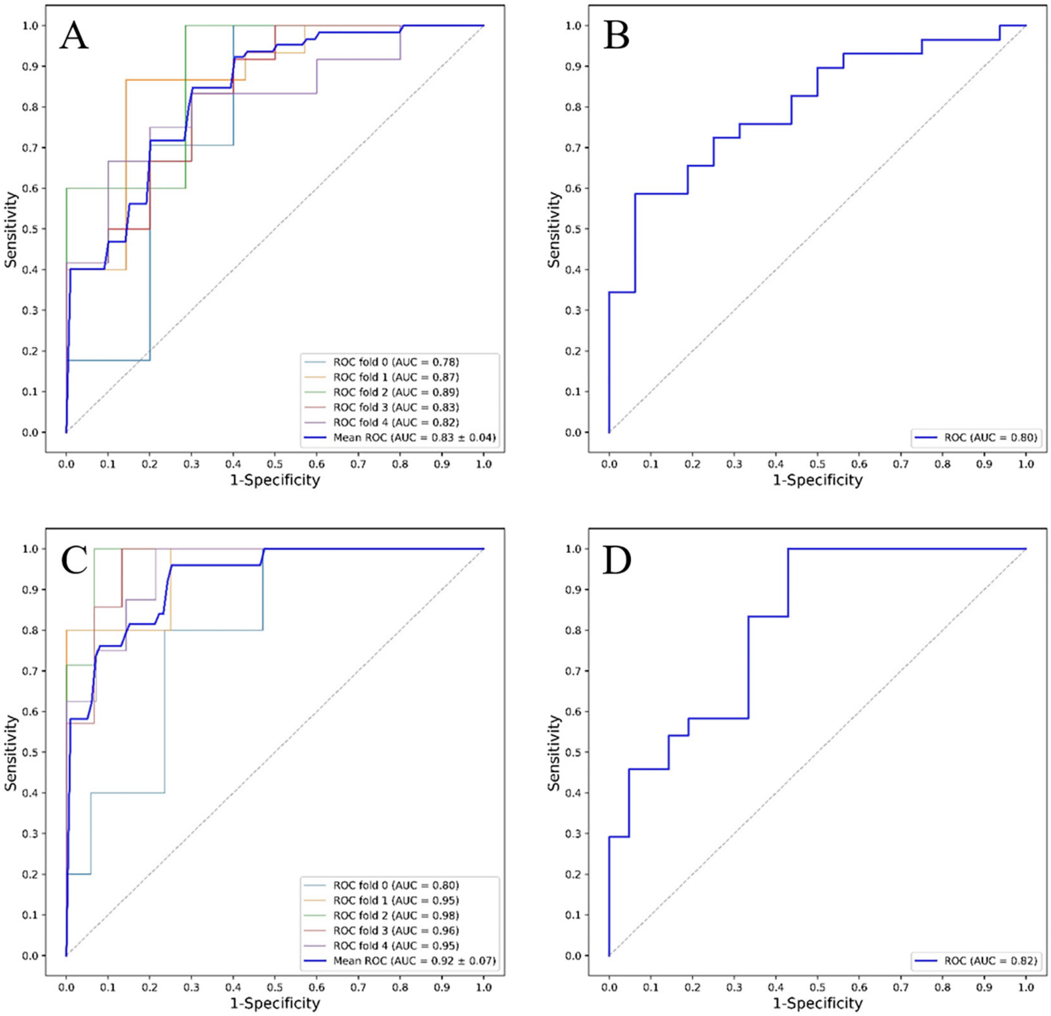

Figure S3 Diagnostic performance of Rad-Scores. (A,B) ROC curves of Rad-Score 1 for *EGFR* prediction in the primary and validation cohort. (C,D) ROC curves of Rad-Score 2 for Ki-67 PI prediction in the primary and validation cohort. Rad-Score, radiomic score; ROC, receiver operating characteristic; AUC, the area under the ROC curve; *EGFR*, epidermal growth factor receptor; PI, proliferation index.

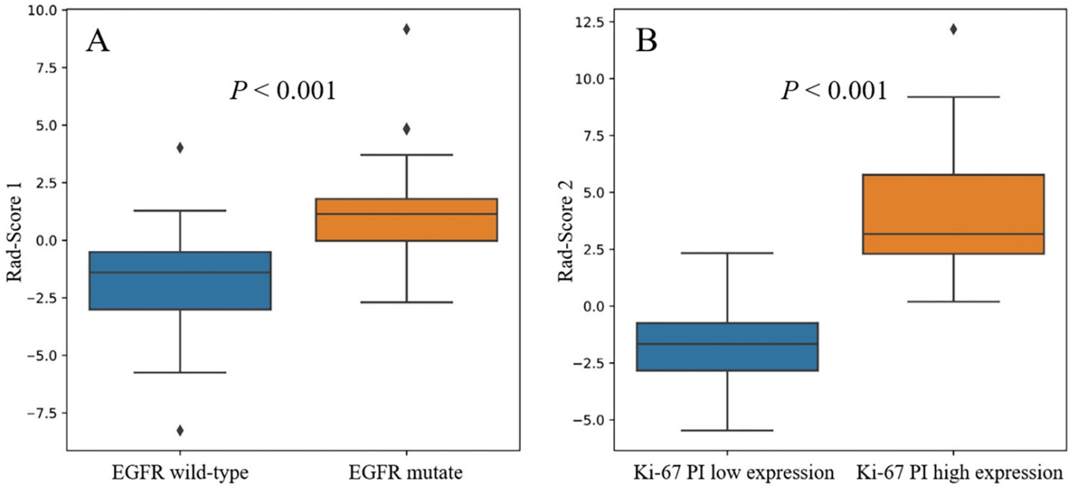

Figure S4 Distribution of Rad-Scores of all patients. (A) The tumors with *EGFR* mutant had significantly higher score than those with *EGFR* wild-type (P<0.001). (B) The tumors with high Ki-67 PI expression had significantly higher score than those with low expression (P<0.001). Rad-Score, radiomic score; *EGFR*, epidermal growth factor receptor; PI, proliferation index.

https://dx.doi.org/10.21037/qims-21-980

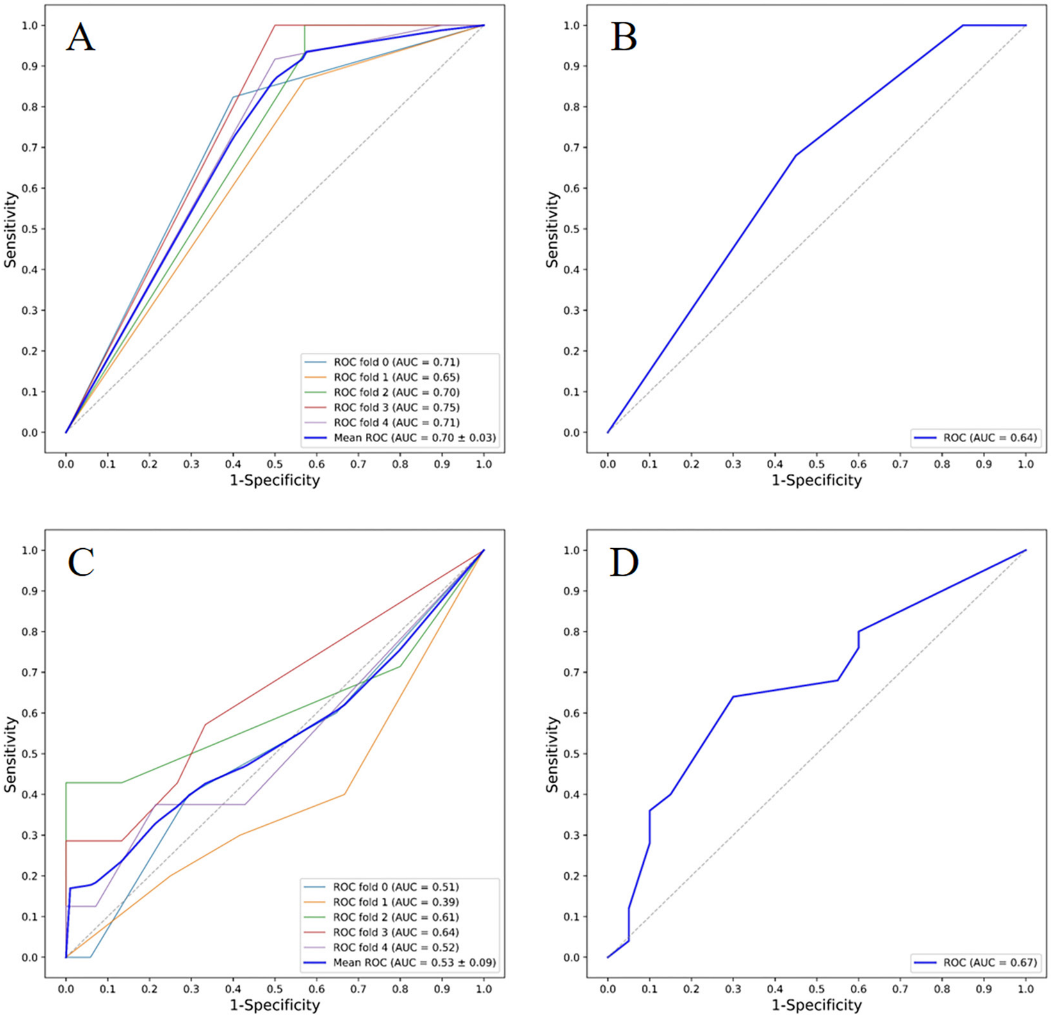

Figure S5 Performance of clinical models. (A,B) ROC curves of clinical models for *EGFR* prediction in the primary and validation cohort. (C,D) ROC curves of clinical models for Ki-67 PI prediction in the primary and validation cohort. ROC, receiver operating characteristic; AUC, the area under the ROC curve; *EGFR*, epidermal growth factor receptor; PI, proliferation index.

Table S1 The details of image preprocessing before feature extraction

Details of the image preprocessing settings
Without normalization
The bin size was 25
The voxel array shift was 1,000
Resampled pixel spacing as [1, 1, 1]

Table S3 The details of Boruta parameters

The details of Boruta parameters	Details
Version	4.0.4
doTrace	2
maxRuns	200
ntree	500

Table S2 The details of radiomic features

Feature type	Feature description*
Shape-based features	Elongation, Flatness, Least Axis Length, Major Axis Length, Maximum 2D Diameter Column, Maximum 2D Diameter Row, Maximum 2D Diameter Slice, Maximum 3D Diameter, Mesh Volume, Minor Axis Length, Sphericity, Surface Area, Surface Volume Ratio, Voxel Volume
Firstorder features	10 Percentile, 90 Percentile, Energy, Entropy, Interquartile Range, Kurtosis, Maximum, Mean Absolute Deviation, Mean, Median, Minimum, Range, Robust Mean Absolute Deviation, Root Mean Squared, Skewness, Total Energy, Uniformity, Variance
glcm features	Autocorrelation, Joint Average, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Joint Energy, Joint Entropy, Imc1, Imc2, Idm, Idmn, Id, Idn, Inverse Variance, Maximum Probability, Sum Entropy, Sum Squares
glrlm features	Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Gray Level Variance, High Gray Level Run Emphasis, Long Run Emphasis, Long Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Low Gray Level Run Emphasis, Run Entropy, Run Length Non-Uniformity, Run Length Non-Uniformity Normalized, Run Percentage, Run Variance, Short Run Emphasis, Short Run High Gray Level Emphasis, Short Run Low Gray Level Emphasis
glszm features	Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Gray Level Variance, High Gray Level Zone Emphasis, Large Area Emphasis, Large Area High Gray Level Emphasis, Large Area Low Gray Level Emphasis, Low Gray Level Zone Emphasis, Size Zone Non-Uniformity, Size Zone Non-Uniformity Normalized, Small Area Emphasis, Small Area High Gray Level Emphasis, Small Area Low Gray Level Emphasis, Zone Entropy, Zone Percentage, Zone Variance
gldm features	Dependence Entropy, Dependence Non-Uniformity, Dependence Non-Uniformity Normalized, Dependence Variance, Gray Level Non-Uniformity, Gray Level Variance, High Gray Level Emphasis, Large Dependence Emphasis, Large Dependence High Gray Level Emphasis, Large Dependence Low Gray Level Emphasis, Low Gray Level Emphasis, Small Dependence Emphasis, Small Dependence High Gray Level Emphasis, Small Dependence Low Gray Level Emphasis
ngtdm features	Busyness, Coarseness, Complexity, Contrast, Strength
Wavelet features	The above 91 features without shape-based features were calculated again after wavelet transformation (LLH, LHL, LHH, HLL, HLH, HHL, HHH, LLL). A total of 728 features
LoG features	The above 91 features without shape-based features were calculated again after LoG transformation with sigma of 1.0, 2.0, 3.0, 4.0, 5.0, respectively. A total of 455 features

*, some of them follow the Imaging Biomarker Standardization Initiative (https://pyradiomics.readthedocs.io/en/latest/index.html) (25). glcm, gray level co-occurrence matrix; glrlm, gray level run length matrix; glszm, gray level size zone matrix; gldm, gray level dependence matrix; ngtdm, neighbouring gray tone difference matrix; LoG, Laplacian of Gaussian; L, low-pass filtering; H, high-pass filtering.

Table S4 Rad-Scores for predicting *EGFR* mutations and Ki-67 PI expression

Rad-Score	Formula
Rad-Score 1	−0.26637586 × (wavelet-HLL_glcm_MaximumProbability)
	−0.36415474 × (wavelet-LLL_glcm_MaximumProbability)
	+0.88922437 × (original_glcm_SumEntropy)
	+0.35225551 × (log-sigma-1-0-mm-3D_glcm_MaximumProbability)
	−0.63638105 × (wavelet-LHL_firstorder_Kurtosis)
	+0.3393241 × (wavelet-LLL_firstorder_Skewness)
	+0.87795737 × (log-sigma-2-0-mm-3D_firstorder_Kurtosis)
	−0.30623909 × (original_shape_Sphericity)
	−0.59557871 × (wavelet-LHL_glszm_LargeAreaHighGrayLevelEmphasis)
	−0.26788923× (original_glcm_ClusterTendency)
Rad-Score 2	−0.33933662 × (wavelet-HLL_gldm_LargeDependenceHighGrayLevelEmphasis)
	−0.24501119 × (log-sigma-1-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis)
	+0.06149887 × (log-sigma-5-0-mm-3D_glcm_Idm)
	+0.66555733 × (log-sigma-5-0-mm-3D_glcm_InverseVariance)
	+0.5809558 × (original_firstorder_Median)
	+0.41924958 × (log-sigma-2-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis)
	+0.23932351 × (wavelet-LHL_glszm_LargeAreaLowGrayLevelEmphasis)
	+0.77765524 × (wavelet-LHL_firstorder_Maximum)
	−0.08881509 × (wavelet-HLL_glszm_GrayLevelNonUniformityNormalized)
	−0.63052453 × (log-sigma-3-0-mm-3D_firstorder_Median)
	−0.57778378 × (log-sigma-3-0-mm-3D_firstorder_90Percentile)
	+0.14979279 × (log-sigma-4-0-mm-3D_glszm_SmallAreaEmphasis)
	−0.19062698 × (wavelet-LHL_glszm_GrayLevelNonUniformityNormalized)
	+0.08261748 × (original_shape_SurfaceVolumeRatio)
	+0.48340228 × (wavelet-LHH_firstorder_Kurtosis)
	+0.68332893 × (wavelet-HHL_glrlm_LongRunHighGrayLevelEmphasis)
	+0.6523309 × (log-sigma-1-0-mm-3D_glcm_Correlation)
	+0.16729185 × (wavelet-LHH_glcm_Correlation)
	−0.24171584 × (original_gldm_LargeDependenceLowGrayLevelEmphasis)
	−0.02439305 × (wavelet-LLL_gldm_LargeDependenceLowGrayLevelEmphasis)
	+0.29861927 × (wavelet-HLL_glszm_LargeAreaLowGrayLevelEmphasis)
	−0.29627875 × (wavelet-HLL_firstorder_Maximum)
	+0.41290256 × (log-sigma-1-0-mm-3D_glrlm_GrayLevelNonUniformityNormalized)
	+0.78760983 × (wavelet-HHH_glcm_Correlation)
	−0.56817862 × (original_firstorder_Skewness)
	+0.38850808 × (wavelet-LHH_glcm_Idn)
	−0.18475777 × (wavelet-LHH_gldm_SmallDependenceLowGrayLevelEmphasis)
	−0.08823505 × (wavelet-HLH_glcm_Idn)

The intraclass correlation coefficient of selected features all >0.6. Rad-Score, radiomic score; glcm, gray level co-occurrence matrix; glszm, gray level size zone matrix; gldm, gray level dependence matrix; glrlm, gray level run length matrix.

Table S5 Comparison of different radiomics studies

Task/author	Years	Sample size	AUC (training)	AUC (validation)	Best modeling algorithm
Task [1]					
Liu *et al.* (34)	2016	298	0.709	_	Logistic regression
Zhang *et al.* (32)	2018	180	0.8618	0.8725	Logistic regression
Jia *et al.* (35)	2019	503	_	0.828	Random forest
Zhao *et al.* (36)	2019	579	_	0.758	3D DenseNet
Zhao *et al.* (37)	2020	637	_	0.757	Logistic regression
Hong *et al.* (39)	2020	201	_	0.851	Logistic regression
Tu *et al*. (38)	2019	404	0.798	0.818	Logistic regression
Lu *et al.* (40)	2020	104	0.90	0.894	Logistic regression
Koyasu *et al.* (41)	2020	138	_	0.843	XGB
Nair *et al.* (42)	2021	50	_	0.87	Logistic regression
Wang *et al.* (43)	2021	52	0.987	0.871	Logistic regression
Rossi *et al.* (44)	2021	109	0.85	0.833	SVM
Zhang *et al.* (45)	2020	914	_	0.910	SE-CNN
Le *et al.* (46)	2021	179	0.89	_	Genetic algorithm
Our study	2021	132	0.891	0.798	Logistic regression
Task [2]					
Zhou *et al.* (47)	2018	110	0.77	_	Logistic regression
Gu *et al.* (48)	2019	245	0.782	_	Random forest
Our study	2021	132	0.981	0.828	Logistic regression

AUC, area under the curve; DenseNet, dense convolutional network; XGB, eXtreme Gradient Boosting; SVM, supported vector machine; SE, squeeze-and-excitation; CNN, convolutional neural network.

Table S6 The RQS (18) analysis of our study

Criteria	Points	RQS
Image protocol quality—well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability	+1 (if protocols are well-documented) +1 (if public protocol is used)	+2
Multiple segmentations—possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities	1	1
Phantom study on all scanners—detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability	1	0
Imaging at multiple time points—collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)	1	0
Feature reduction or adjustment for multiple testing—decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features	–3 (if neither measure is implemented) +3 (if either measure is implemented)	3

Table S6 (*continued*)

Table S6 (*continued*)

Criteria	Points	RQS
Multivariable analysis with non-radiomics features (for example, *EGFR* mutation)—is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non-radiomics features	1	1
Detect and discuss biological correlates-demonstration of phenotypic differences (possibly associated with underlying gene-protein expression patterns) deepens understanding of radiomics and biology	1	1
Cut-off analyses—determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results	1	1
Discrimination statistics—report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+1 (if a discrimination statistic and its statistical significance are reported) +1 (if a resampling method technique is also applied)	+2
Calibration statistics—report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+1 (if a calibration statistic and its statistical significance are reported) +1 (if a resampling method technique is also applied)	+2
Prospective study registered in a trial database—provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	+7 (for prospective validation of a radiomics signature in an appropriate trial)	+0
Validation—the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance	−5 (if validation is missing) +2 (if validation is based on a dataset from the same institute) +3 (if validation is based on a dataset from another institute) +4 (if validation is based on two datasets from two distinct institutes) +4 (if the study validates a previously published signature) +5 (if validation is based on three or more datasets from distinct institutes) *Datasets should be of comparable size and should have at least 10 events per model feature	5
Comparison to 'gold standard' —assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics	2	0
Potential clinical utility—report on the current and potential application of the model in a clinical setting (for example, decision curve analysis)	2	2
Cost-effectiveness analysis—report on the cost-effectiveness of the clinical application (for example, QALYs generated)	1	0
Open science and data—make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study	+1 (if scans are open source) +1 (if region of interest segmentations are open source) +1 (if code is open source) +1 (if radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source)	+4
Total points (36 =100%)	24 ≈66.67%	

RQS, radiomics quality score; *EGFR*, epidermal growth factor receptor; ROC, receiver operating characteristic; AUC, area under the ROC curve; QALY, quality-adjusted life year.