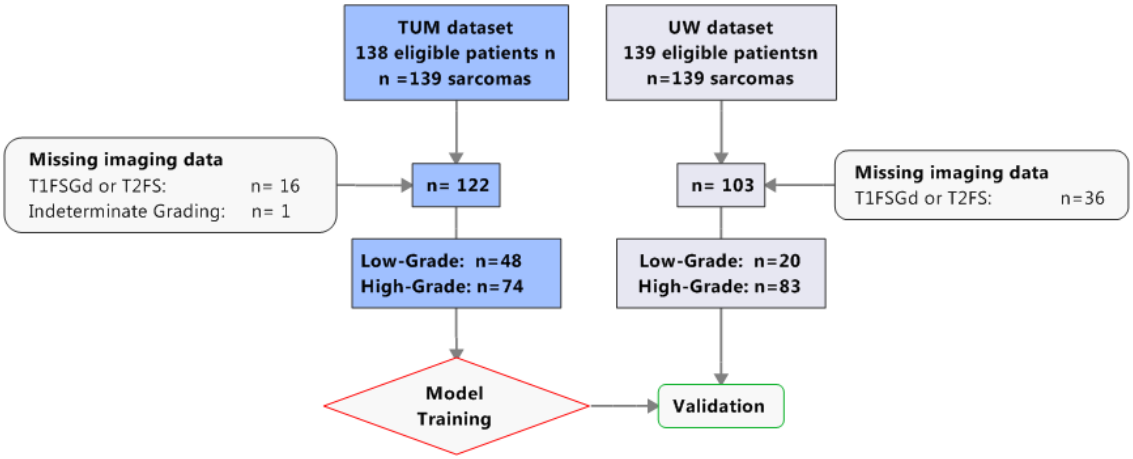
Supplemental Data

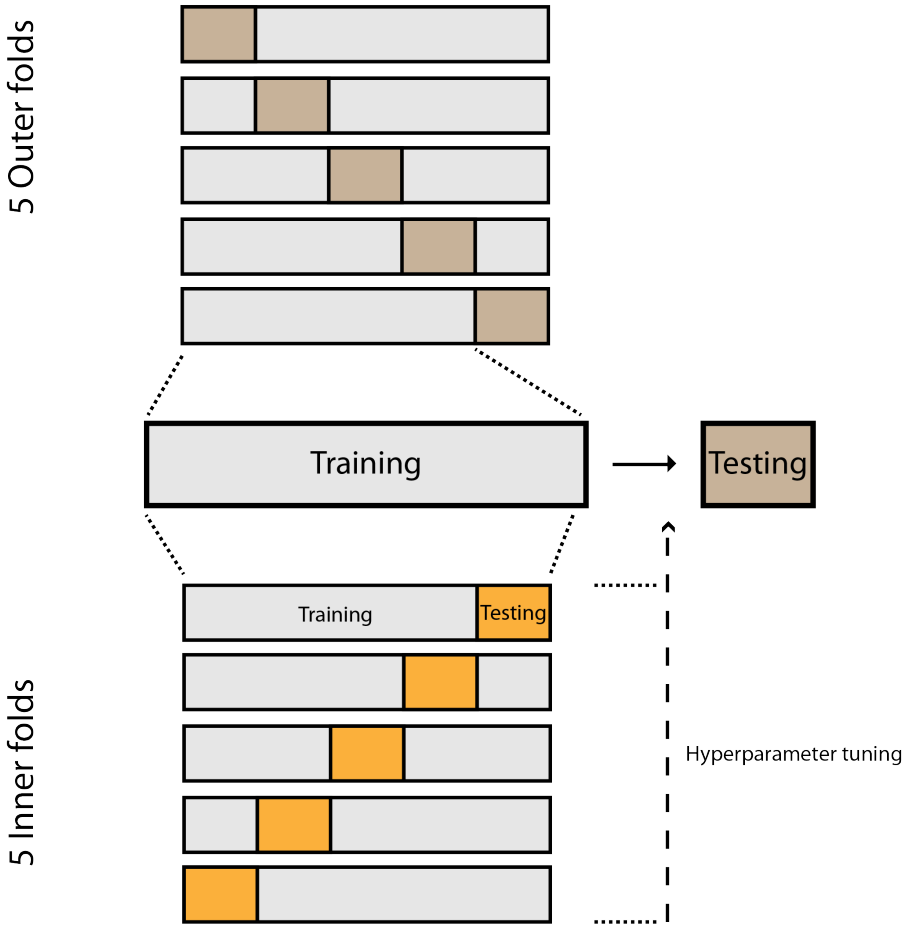
Supplemental Figures

Supplemental Figure 1: Patient workflow

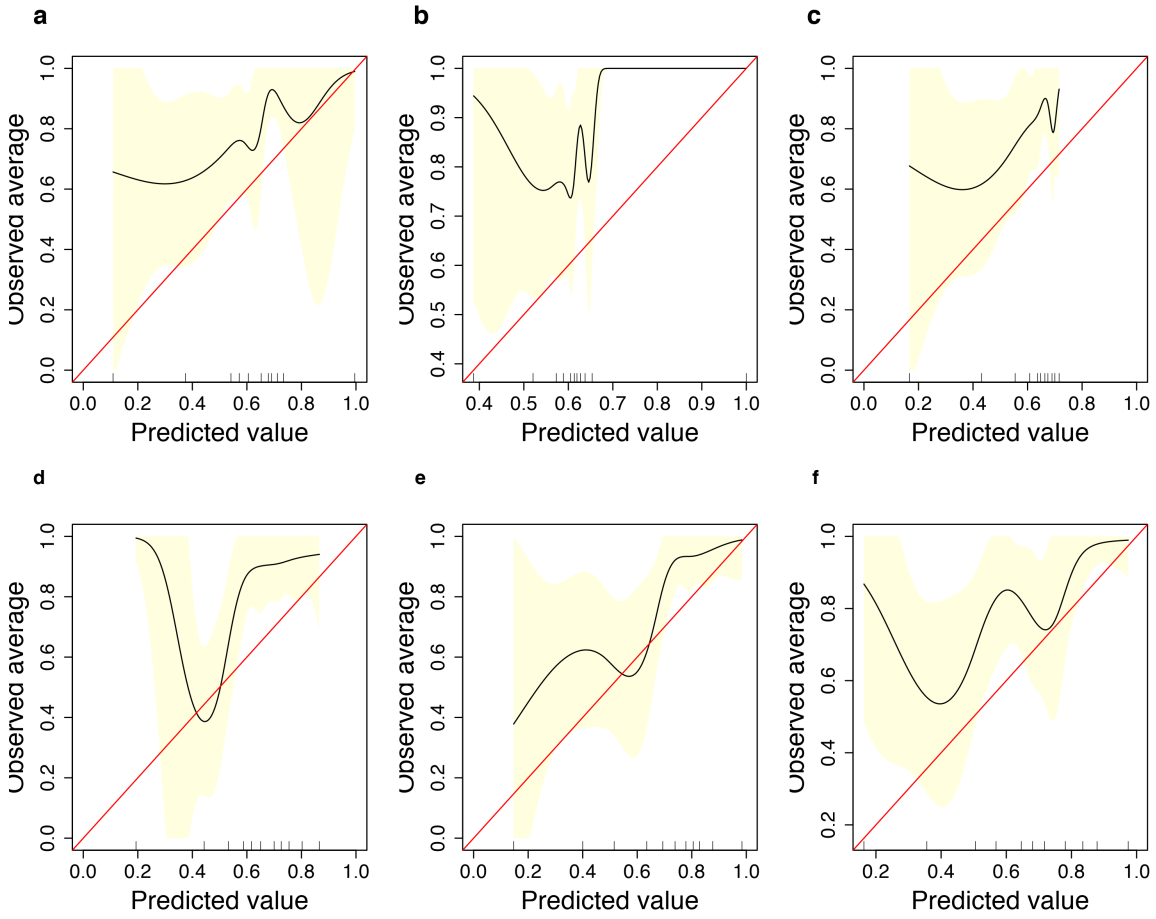
Representation of the patient workflow of the two independent patient cohorts from the Technical University of Munich (TUM) (training set) and the University of Washington (UW) (validation set).



Supplemental Figure 2: Nested cross validation scheme



Supplemental Figure 3: Calibration Curves
Calibration curves for the *Clinical* model (a), the *Tumor-Volume* model (b), the *Radiomics-T1FSGd* model (c), the *Radiomics-T2FS* model (d), the *Radiomics-combined* model (e), and *Clinical-Volume-combined* (f) on the validation set are shown.



Supplemental Table

Supplemental Table 1: Histologies of Soft Tissue Sarcomas
Sorted by patient number. The distribution of histologies was significantly different ($p<0.001$, determined by the Chi-squared test.).

Histological subtype (%)	TUM Training	UW Testing
Liposarcoma	43 (35%)	6 (6%)
Pleomorphic Sarcoma	24 (20%)	38 (37%)
Myxofibrosarcoma	19 (16%)	5 (5%)
Myxoid liposarcoma	9 (7%)	10 (10%)
Synovialsarcoma	8 (7%)	0
Dedifferenciated liposarcoma	7 (6%)	2 (2%)
Leiomyosarcoma	5 (4%)	10 (10%)
Alveolar sarcoma	2 (2%)	0
Rhabdomyosarcoma	2 (2%)	1 (1%)
Angiosarcoma	1 (1%)	1 (1%)
Myofibroblastic sarcoma	1 (1%)	3 (3%)
Synovial sarcoma	1 (1%)	8 (8%)
Fibromyxoid	0	1 (1%)
MPNST	0	5 (5%)
Spindle cell sarcoma	0	4 (4%)
Solitary Fibrous Tumor	0	2 (2%)
Myxoinflammatory fibroblastic sarcoma	0	2 (2%)
Other	0	5 (5%)

abbreviations: TUM: Technical University of Munich, UW: University of Washington.

Supplemental Table 2: MRI acquisition parameters

Median values are depicted. Numbers in brackets represent the range.

	Sequence	TUM[a] Training	UW[b] Validation
Magnetic field strength		1·5 T (1·0 - 3·0)	1·5 (0·7 - 3·5)
In-Plane resolution (mm)	T1FSGd	0·81 x 0·81 (0·31x0·31 -1·4x1·4)	0·56 x 0·56 (0·19 x 0·19 - 1·56 x 1·56)
Slice thickness (mm)		7 (3·3 - 10)	6 (1·0 - 11·0)
Matrix		320 x 320 (176 x 176 - 1152 x 1152)	512 x 512 (190 x 224 - 1200 x 1200)
TR (ms)		683·5 (102 - 1430)	630 (400- 1038)
TE (ms)		12 (4 - 39)	11·8 (3 - 40)
In-Plane resolution (mm)	T2FS	0·81 x 0·81 (0·31x0·31 - 1·4x1·4)	0·625 x 0·625 (0·17 x 0·17 - 1·71 x 1·71)
Slice thickness (mm)		7 (3·3 - 10)	6·5 (3·0 - 11·6)
Matrix		448 x448 (224 x 224 - 1536 x 1536)	480 x 480 (169 x 192 - 880 x880)
TR (ms)		6520 (1086 - 18516)	3933 (9 -11098)
TE (ms)		60 (30 - 133)	68·8 (9-120)

Abbreviation: p: patients, T1wFS: T1-weighted fat saturated with contrast enhancement, T2FS: T2-weighted fat saturated, TE: echo time, TR: repetition time, TUM: Technical University of Munich, UW: University of Washington.

[a]MRI Scanner: GE (Chicago, USA): Signa HdX, Signa Excite; Philips (Amsterdam, Netherlands): Achieva, Ingenia, Intera; Siemens (Munich, Germany): Aera, Avanto, Biograph_mM, Espree, Magnetom Vision, Verio, Symphony,Verio.

[b]MRI Scanner: GE (Chicago, USA): Discovery MR750, Genesis Signa, Signa HdX, Optima 450, Signa Excite; Philips (Amsterdam, Netherlands): Achieva, Gyroscan NT Ingenia, NT Intera; Siemens (Munich, Germany): Avanto, Espree, Harmony, Symphony,TrioTim; Hitachi (Tokio, Japan): Altaire, Oasis.

Supplemental Table 3: MRI plane orientation distribution

Plane orientation	Sequence	TUM[a]	UW[b]
Axial	**T1FSGd**	121 (99%)	101 (98%)
Coronar		1 (1%)	2 (2%)
Sagittal		0	0
Axial	**T2FS**	19 (16%)	101 (98%)
Coronar		98 (80%)	2 (2%)
Sagittal		5 (4%)	0

Abbreviation: p: patients, T1wFS: T1-weight fat saturated with contrast enhancement, T2FS: T2-weight fat saturated, TUM: Technical University of Munich, UW: University of Washington.

Supplemental Table 4: Extracted radiomics features and image discretization

Image discretization was performed using a fixed bin width of ten as previously described. [1] Following the implementation in the pyradiomics package (version 2.1) bins were defined in that way, that the bin edges are equally spaced from zero. [2] The leftmost edge ($\min X_{gl}$) was defined as follows:

$$X_{b,i} = \left(\frac{X_{gl,i}}{W}\right) - \left(\frac{\min X_{gl}}{W}\right) + 1$$

where W is the bin width value. $X_{gl,i}$ and $X_{b,i}$ correspond to gray level intensities before and after discretization, respectively. If the range intensity values is equally dividable by the bin width, the maximum number of number of gray levels corresponds to the number of bins plus one. [3] Also see the pyradiomics documentation [https://pyradiomics.readthedocs.io/en/latest/radiomics.html#module-radiomics.imageoperations].

All over 1394 features per sequence were extract from the labelmap (shape features), the original images or the respective filtered reconstructions (intensity and texture features), summing up to a total of 2788 features per patients (see Supplemental Table 1). The pyradiomics package (version 2.1) implemented in python was used to calculate all features. Most of features were calculated according to the Imaging Biomarker Standardization Initiative (IBSI). [4] Differences to the IBSI definitions are listed below. The online documentation of the pyradiomics package (http://pyradiomics.readthedocs.io/en/latest/features.html) displays further details such as formulae.

	Shape Features
1.)	Voxel Volume (IBSI: Approximate Volume)
2.)	Mesh Volume (IBSI: Volume)
3.)	Surface Area
4.)	Surface Area to Volume ratio
5.)	Sphericity
6.)	Spherical Disproportion
7.)	Maximum 3D Diameter
8.)	Maximum 2D Diameter Slice
9.)	Maximum 2D Diameter Column
10.)	Maximum 2D Diameter Row
11.)	Major Axis Length
12.)	Minor Axis Length
13.)	Least Axis Length
14.)	Elongation
15.)	Flatness
	First Order Features
1.)	Energy
2.)	Total Energy[a]
3.)	Entropy (IBSI: Intensity Histogram Entropy)
4.)	Minimum
5.)	10th Percentile
6.)	90th Percentile
7.)	Maximum
8.)	Mean
9.)	Median
10.)	Interquartile Range
11.)	Range
12.)	Mean Absolute Deviation (MAD)
13.)	Robust Mean Absolute Deviation (rMAD)
14.)	Root Mean Squared (RMS)

15.)	Skewness
16.)	Kurtosis[b]
17.)	Variance
18.)	Uniformity (IBSI: Intensity Histogram Uniformity)
	Gray Level Co-occurrence Matrix (GLCM) Features
1.)	Autocorrelation
2.)	Joint Average
3.)	Cluster Prominence
4.)	Cluster Shade
5.)	Cluster Tendency
6.)	Contrast
7.)	Correlation
8.)	Difference Average
9.)	Difference Entropy
10.)	Difference Variance
11.)	Joint Energy (IBSI: Angular Second Moment)
12.)	Joint Entropy
13.)	Informal Measure of Correlation (IMC) 1
14.)	Informal Measure of Correlation (IMC) 2
15.)	Inverse Difference Moment (IDM)
16.)	Inverse Difference Moment Normalized (IDMN)
17.)	Inverse Difference (ID)
18.)	Inverse Difference Normalized (IDN)
19.)	Inverse Variance
20.)	Maximum Probability (IBSI: Joint maximum)
21.)	Sum Entropy
22.)	Sum of Squares (IBSI: Sum of Squares)
23.)	Maximal Correlation Coefficient (MCC)
	Gray Level Size Zone Matrix (GLSZM) Features
1.)	Small Area Emphasis (SAE)
2.)	Large Area Emphasis (LAE)
3.)	Gray Level Non-Uniformity (GLN)
4.)	Gray Level Non-Uniformity Normalized (GLNN)
5.)	Size-Zone Non-Uniformity (SZN)
6.)	Size-Zone Non-Uniformity Normalized (SZNN)
7.)	Zone Percentage (ZP)
8.)	Gray Level Variance (GLV)
9.)	Zone Variance (ZV)
10.)	Zone Entropy (ZE)
11.)	Low Gray Level Zone Emphasis (LGLZE)
12.)	High Gray Level Zone Emphasis (HGLZE)
13.)	Small Area Low Gray Level Emphasis (SALGLE)
14.)	Small Area High Gray Level Emphasis (SAHGLE)
15.)	Large Area Low Gray Level Emphasis (LALGLE)
16.)	Large Area High Gray Level Emphasis (LAHGLE)
	Gray Level Run Length Matrix (GLRLM) Features
1.)	Short Run Emphasis (SRE)

2.)	Long Run Emphasis (LRE)
3.)	Gray Level Non-Uniformity (GLN)
4.)	Gray Level Non-Uniformity Normalized (GLNN)
5.)	Run Length Non-Uniformity (RLN)
6.)	Run Length Non-Uniformity Normalized (RLNN)
7.)	Run Percentage (RP)
8.)	Gray Level Variance (GLV)
9.)	Run Variance (RV)
10.)	Run Entropy (RE)
11.)	Low Gray Level Run Emphasis (LGLRE)
12.)	High Gray Level Run Emphasis (HGLRE)
13.)	Short Run Low Gray Level Emphasis (SRLGLE)
14.)	Short Run High Gray Level Emphasis (SRHGLE)
15.)	Long Run Low Gray Level Emphasis (LRLGLE)
16.)	Long Run High Gray Level Emphasis (LRHGLE)
	Neighbouring Gray Tone Difference Matrix (NGTDM) Features
1.)	Coarseness
2.)	Contrast
3.)	Busyness
4.)	Complexity
5.)	Strength
	Gray Level Dependence Matrix (GLDM) Features
1.)	Small Dependence Emphasis (SDE)
2.)	Large Dependence Emphasis (LDE)
3.)	Gray Level Non-Uniformity (GLN)
4.)	Dependence Non-Uniformity (DN)
5.)	Dependence Non-Uniformity Normalized (DNN)
6.)	Gray Level Variance (GLV)
7.)	Dependence Variance (DV)
8.)	Dependence Entropy (DE)
9.)	Low Gray Level Emphasis (LGLE)
10.)	High Gray Level Emphasis (HGLE)
11.)	Small Dependence Low Gray Level Emphasis (SDLGLE)
12.)	Small Dependence High Gray Level Emphasis (SDHGLE)
13.)	Large Dependence Low Gray Level Emphasis (LDLGLE)
14.)	Large Dependence High Gray Level Emphasis (LDHGLE)

[a]Not present in the IBSI feature definition.
[b]The feature "kurtosis" calculated by the pyradiomics package is not corrected by a factor of three as it is for "kurtosis" following the IBSI feature definition.

Supplemental Table 5: Comparison of machine learning techniques.

	Details	Hyper-parameter	Training Performance (AUC)			Testing Performance (AUC)		
			T1FSGd	T2FS	Combined	T1FSGd	T2FS	Combined
Elastic net logistic regression	**0<α<1**	**α, λ**	0·78	0·80	0·83	0·67	0·79	0·78
LASSO	**α = 1**	**λ**	0·78	0·77	0·84	0·69	0·78	0·73
Random Forest	**1001 trees**	**mtry**	0·78	0·88	0·86	0·63	0·70	0·73
LogitBoost		**Iteration number**	0·75	0·88	0·86	0·69	0·69	0·72
Decision tree		**Complexity parameter (cp)**	0·70	0·73	0·74	0·68	0·55	0·64
Neural Network	**Single hidden layer**	**size, decay**	0·73	0·80	0·81	0·65	0·67	0·74
Support vector machine	**Radial basis function kernel transform**	**C (cost) Sigma**	0·78	0·80	0·83	0·56	0·6	0·68

The models were trained and compared using ten iteration of five-fold nested cross validation as described in the main text by applying the code published by Deist et al.. [5] Hyperparemeter optimization was performed using grid search. The training performance was evaluated by five-fold cross validation using the optimal hyperparameter combination. The final model was retrained on the whole training dataset using the optimal hyperparameter and tested on the test set. All performance measures were determined following ComBat correction as described in the manuscript.

abbreviations: AUC: area under the receiver operator characteristic curve, LASSO: least absolute shrinkage and selection operator

Supplemental Table 6: Influence of feature reduction and resampling techniques on predictive model performance

Pre-processing	ML Technique	Training Performance (AUC)			Testing Performance (AUC)		
		T1FSGd	T2FS	Combined	T1FSGd	T2FS	Combined
Principal component analysis	**Net logistic regression**	0·68	0·79	0·84	0·63	0·79	0·73
Principal component analysis	**LASSO**	0·67	0·78	0·84	0·62	0·78	0·68
Principal component analysis	**Random Forest**	0·76	0·83	0·84	0·58	0·68	0·74
SMOTE resampling[#]	**LASSO**	0·78	0·77	0·84	0·64	0·74	0·66
Weighting	**LASSO**	0·78	0·78	0·84	0·69	0·76	0·74

[#] A Synthetic resampling techniques implemented in the "DMwR" package that down samples the majority class and synthesizes new minority instances by interpolating between existing ones was tested. [6,7] Resampling was performed within the cross validation.

abbreviation: AUC: area under the receiver operator characteristic curve, LASSO: least absolute shrinkage and selection operator, ML: machine learning, SMOTE: Synthetic minority sampling technique

Supplemental Table 7: Hyperparameter optimization of penalized regression models

For modeling of the prediction models, we applied the penalized regression method. It introduces a penalty into regression models for the inclusion of a large number of input features. As consequence, the coefficient values of unimportant features are reduced towards zero. The amount of shrinkage is modulated by two hyperparameters. The elasticnet mixing parameter α defines the type of penalization. We chose an α value of one which is referred to as "least absolute shrinkage and selection operator" (LASSO). As consequence, feature coefficients are penalized with the sum of the absolute coefficients (L1-norm) forcing coefficients of unimportant features to zero leading to the most profound feature reduction compared to the alternative penalized regression models "ridge regression" ($\alpha = 0$, shrinks coefficients maximally close to zero) or "elastic net regression" ($0<\alpha<1$). The hyperparameter λ fine tunes the amount of coefficient shrinkage. [8] Previous studies have demonstrated a competitive performance compared to other machine learning methods. [9,10] At the same time, it allows for better interpretability due to the direct linear relationship between input features and outcome. The optimal values showed the best performance over all outer cross validation folds and were used for the training of the final models.

Model		**Radiomics-T1FSGd**	**Radiomics-T2FS**	**Radiomics-combined**	**Clinical**	**Clinical-Volume-combined**
Lambda	**Tested values**	0·001-1, by 0·002	0·001-1, by 0·002	0·001-1, by 0·002	0·001-1, by 0·002	0·001-1, by 0·002
	Optimal Value	0·33	0·043	0·091	0·060	0·02

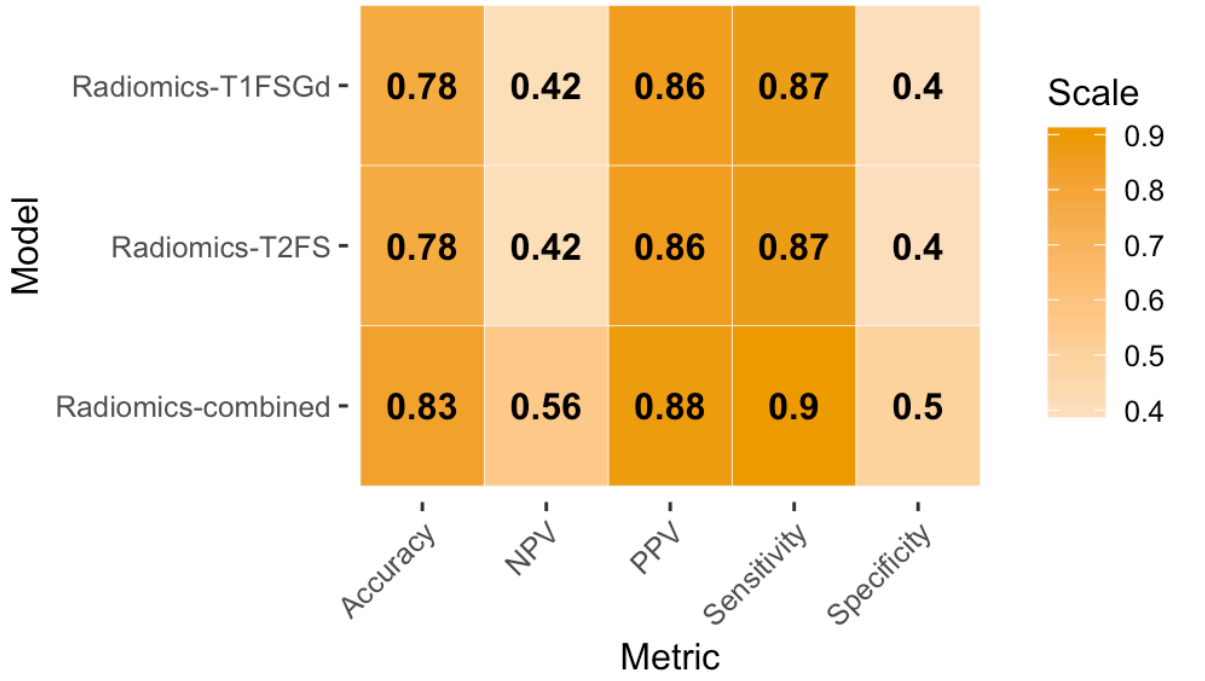
Supplemental Table 8: Feature ranking of prediction models

Features and the respective importance ranking of the developed prediction models.

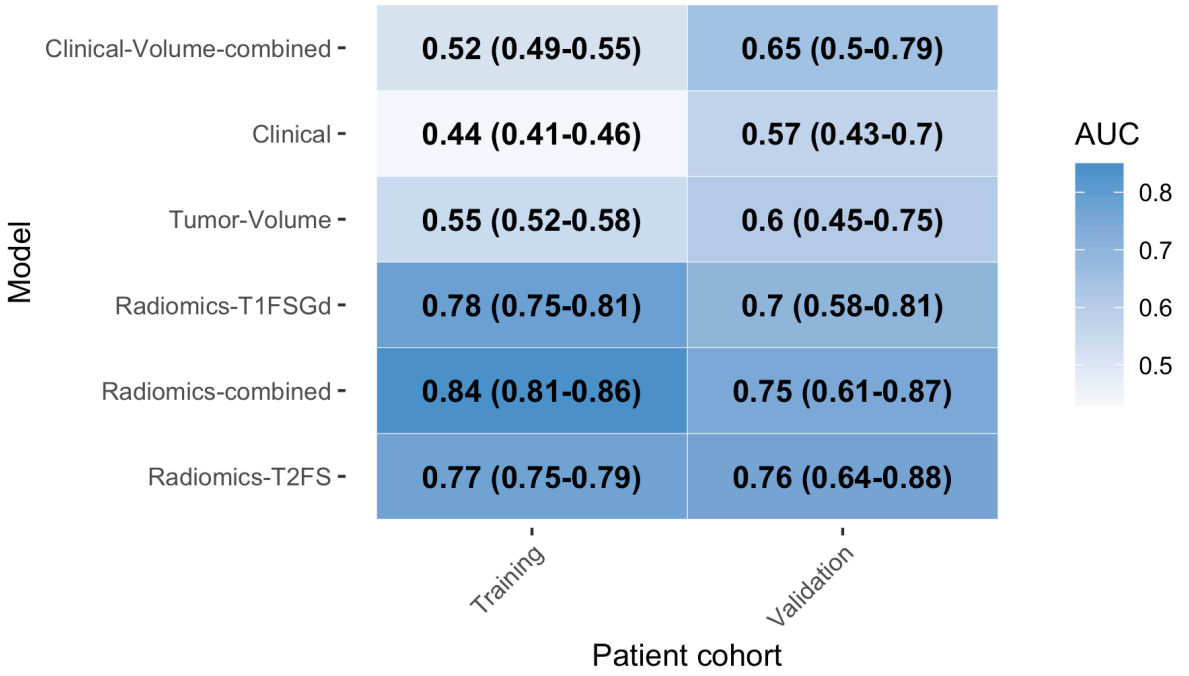
Feature	Importance ranking
Radiomics-T2FS	
wavelet.LHL_firstorder_Kurtosis	1
wavelet.HLH_firstorder_Mean	2
lbp.3D.k_ngtdm_Busyness	3
log.sigma.5.0.mm.3D_glcm_Correlation	4
log.sigma.3.0.mm.3D_firstorder_90Percentile	5
original_glrlm_LongRunLowGrayLevelEmphasis	6
lbp.3D.m2_glcm_DifferenceVariance	7
wavelet.HHL_firstorder_Kurtosis	8
original_shape_Sphericity	9
wavelet.HLH_glcm_ClusterProminence	10
original_shape_Elongation	11
original_shape_Maximum2DDiameterRow	12
lbp.3D.k_glszm_ZoneEntropy	13
original_shape_Maximum2DDiameterSlice	14
log.sigma.5.0.mm.3D_glcm_MaximumProbability	15
lbp.3D.m2_gldm_DependenceEntropy	16
lbp.3D.m2_glcm_MCC	17
original_firstorder_Minimum	18
lbp.3D.m2_glcm_SumEntropy	19
lbp.3D.m2_glszm_GrayLevelVariance	20
wavelet.LHH_firstorder_Skewness	21
original_shape_Flatness	22
log.sigma.4.0.mm.3D_firstorder_90Percentile	23
log.sigma.5.0.mm.3D_glcm_Imc2	24
Radiomics-T1FSGd	
wavelet.HLH_firstorder_Mean	1
wavelet.HLH_glcm_MCC	2
wavelet.HLH_gldm_DependenceEntropy	3
wavelet.LLL_glrlm_LongRunLowGrayLevelEmphasis	4
lbp.3D.m1_firstorder_T1FSGd0Percentile	5
lbp.3D.m1_ngtdm_Complexity	6
lbp.3D.m2_firstorder_RootMeanSquared	7
lbp.3D.m2_glrlm_HighGrayLevelRunEmphasis	8
lbp.3D.k_firstorder_Kurtosis	9
lbp.3D.k_glcm_DifferenceEntropy	10
lbp.3D.k_glrlm_ShortRunHighGrayLevelEmphasis	11
original_shape_Flatness	12
original_shape_Maximum2DDiameterRow	13
original_shape_Sphericity	14
original_firstorder_Skewness	15
original_glcm_JointEnergy	16
original_glcm_MCC	17
original_gldm_DependenceEntropy	18
log.sigma.3.0.mm.3D_firstorder_Skewness	19
log.sigma.5.0.mm.3D_firstorder_Skewness	20
log.sigma.5.0.mm.3D_glszm_LargeAreaEmphasis	21
wavelet.LHH_firstorder_Kurtosis	22
Radiomics-combined	
log.sigma.3.0.mm.3D_firstorder_90Percentile_T2FS	1
wavelet.LHL_firstorder_Kurtosis_T2FS	2
log.sigma.3.0.mm.3D_firstorder_Skewness_T1FSGd	3
wavelet.HHL_firstorder_Kurtosis_T2FS	4
original_firstorder_Skewness_T1FSGd	5
wavelet.LLL_glrlm_LongRunLowGrayLevelEmphasis_T1FSGd	6
wavelet.HLH_firstorder_Mean_T2FS	7
lbp.3D.m2_glcm_SumEntropy_T2FS	8

wavelet.LHH_firstorder_Skewness_T2FS	9
log.sigma.5.0.mm.3D_glszm_LargeAreaEmphasis_T1FSGd	10
original_shape_Flatness_T2FS	11
log.sigma.4.0.mm.3D_firstorder_90Percentile_T2FS	12
original_shape_Flatness_T1FSGd	13

Supplemental Table 9: Performance metrics of the radiomic classifiers
Performance metrics on the validation set for the prediction of high-grade sarcomas including accuracy, negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity of the three radiomics models *Radiomics-T1FSGd*, *Radiomics-T2FS*, and *Radiomics-combined* are shown.



Supplemental Table 10: Predictive performance metrics of radiomic classifiers after ComBat correction with histology as biological covariate.
Area under the receiver operator characteristic curve (AUC) values for the differentiation of high-grade and low-grade soft tissue sarcomas. 95 % confidence intervals are shown in parentheses.



Supplemental Table 11: Predictive performance of prognostic models for overall survival
95% confidence intervals are depicted in parentheses.

	Training	Validation
Grading prediction models		
AJCC clinical stage[a]	0·74 (0·67-0·81)	0·69 (0·60-0·78)
Radiomics-T1FSGd + AJCC	0·76 (0·68-0·85)	0·71 (0·61-0·81)
Radiomics-T2FS + AJCC	0·76 (0·68-0·84)	0·74 (0·64-0·84)
Radiomics-combined + AJCC	0·77 (0·68-0·86)	0·72 (0·62-0·83)
Tumor volume + AJCC	0·78 (0·71-0·85)	0.71 (0·61-·0.81)
Prognostic models		
T1FSGd features	0·70 (0·58-0·80)	0·55 (0·45-0·65)
T2FS features	0·74 (0·66-0·83)	0·60 (0·49-0·70)
Combined features	0·88 (0·82-0·94)	0·60 (0·5-0·69)
Tumor volume	0·59 (0·48-0·71)	0·54 (0·43-0·65)
T1FSGd features + AJCC	0·80 (0·73-0·86)	0·67 (0·57-0·77)
T2FS features + AJCC	0·81 (0·75-0·87)	0·70 (0·59-0·80)
Combined features + AJCC	0·89 (0·84-0·94)	0·64s (0·55-0·74)

a AJCC cancer staging manual 7th edition. [11]

References

1 Peeken JC, Bernhofer M, Spraker MB, *et al.* CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. *Radiother Oncol* 2019; **135**: 187–96.

2 van Griethuysen JJM, Fedorov A, Parmar C, *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* 2017; **77**: e104–7.

3 Leijenaar RTH, Nalbantov G, Carvalho S, *et al.* The effect of SUV discretization in quantitative FDG-PET Radiomics: The need for standardized methodology in tumor texture analysis. *Sci Rep* 2015; **5**: 1–10.

4 Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *ArXiv* 2016; **1612.07003**. http://arxiv.org/abs/1612.07003.

5 Deist TM, Jochems A, van Soest J, *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 2017; **4**: 24–31.

6 Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2011; **16**: 732–5.

7 Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. *Data Min Knowl Discov* 2014; **28**: 92–122.

8 Bruce P, Bruce A. Practical Statistics for Data Scientists - O'Reilly Media. O'Reilly Media, 2017 http://shop.oreilly.com/product/0636920048992.do (accessed March 27, 2019).

9 Deist TM, Dankers FJWM, Valdes G, *et al.* Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Med Phys* 2018; **45**: 3449–59.

10 Leger S, Zwanenburg A, Pilz K, *et al.* A comparative study of machine learning methods for time-To-event survival data for radiomics risk modelling. *Sci Rep* 2017; **7**: 1–28.

11 Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. *Ann Surg Oncol* 2010; **17**: 1471–4.