EBioMedicine 48 (2019) 332–340

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Research paper

Tumor grading of soft tissue sarcomas using MRI-based radiomics

Jan C. Peeken [a,b,c,*], Matthew B. Spraker [d], Carolin Knebel [e], Hendrik Dapper [a], Daniela Pfeiffer [f], Michal Devecka [a], Ahmed Thamer [a], Mohamed A. Shouman [a], Armin Ott [g], Rüdiger von Eisenhart-Rothe [e], Fridtjof Nüsslin [a], Nina A. Mayr [d], Matthew J. Nyflot [d,h], Stephanie E. Combs [a,b,c]

[a] Department of Radiation Oncology, Klinikum Rechts der Isar, School of Medicine, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany
[b] Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Helmholtz Zentrum München, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany
[c] Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, Germany
[d] Department of Radiation Oncology, University of Washington, 1959 NE Pacific St, Box 356043, Seattle, WA 98195, United States of America
[e] Department of Orthopaedic Surgery, Klinikum Rechts der Isar, School of Medicine, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 München, Germany
[f] Department of Radiology, Klinikum Rechts der Isar, School of Medicine, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany
[g] Institute of Medical Informatics, Statistics and Epidemiology, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany
[h] Department of Radiology, University of Washington, Seattle, WA, United States of America

ARTICLE INFO

Article history:
Received 14 May 2019
Received in revised form 13 August 2019
Accepted 24 August 2019
Available online 12 September 2019

Keywords:
Soft tissue sarcoma
Radiomics
Tumor grading
MRI
Risk stratification
Biomarker

ABSTRACT

Background: Treatment decisions for multimodal therapy in soft tissue sarcoma (STS) patients greatly depend on the differentiation between low-grade and high-grade tumors. We developed MRI-based radiomics grading models for the differentiation between low-grade (G1) and high-grade (G2/G3) STS.
Methods: The study was registered at ClinicalTrials.gov (number NCT03798795). Contrast-enhanced T1-weighted fat saturated (T1FSGd), fat-saturated T2-weighted (T2FS) MRI sequences, and tumor grading following the French Federation of Cancer Centers Sarcoma Group obtained from pre-therapeutic biopsies were gathered from two independent retrospective patient cohorts. Volumes of interest were manually segmented. After preprocessing, 1394 radiomics features were extracted from each sequence. Features unstable in 21 independent multiple-segmentations were excluded. Least absolute shrinkage and selection operator models were developed using nested cross-validation on a training patient cohort (122 patients). The influence of ComBatHarmonization was assessed for correction of batch effects.
Findings: Three radiomic models based on T2FS, T1FSGd and a combined model achieved predictive performances with an area under the receiver operator characteristic curve (AUC) of 0.78, 0.69, and 0.76 on the independent validation set (103 patients), respectively. The T2FS-based model showed the best reproducibility. The radiomics model involving T1FSGd-based features achieved significant patient stratification. Combining the T2FS radiomic model into a nomogram with clinical staging improved prognostic performance and the clinical net benefit above clinical staging alone.
Interpretation: MRI-based radiomics tumor grading models effectively classify low-grade and high-grade soft tissue sarcomas.
Fund: The authors received support by the medical faculty of the Technical University of Munich and the German Cancer Consortium.

1. Introduction

Soft tissue sarcomas (STS) constitute rare malignant diseases [1]. In contrast to many other malignant tumors, treatment decisions strongly depend on pre-therapeutic tumor grading [2].

Tumor grading is commonly determined during histological workup of pre-therapeutic biopsies. Two distinct grading systems were developed by the National Cancer Institute and the French Federation of Cancer Centers Sarcoma Group (FNCLCC) [3,4]. In direct comparison, the FNCLCC system appeared to predict distant metastasis development

Abbreviations: 95%CI, 95% confidence interval; AUC, Area under the curve; C-index, Concordance index; FNCLCC, French Federation of Cancer Centers Sarcoma Group; GLCM, Gray level co-occurrence matrix; GLDM, Gray level dependence matrix; MCC, Maximum correlation coefficient; MRI, Magnetic resonance imaging; NPV, Negative predictive value; OR, odds ratio; OS, Overall survival; PPV, Positiv predictive value; ROC, receiver operator characteristic; STS, Soft tissue sarcomas; T1FSGd, Contrast-enhanced T1-weighted fat saturated MRI sequence; T2FS, Fat-saturated T2-weighted MRI sequence.

* Corresponding author at: Department of Radiation Oncology, Klinikum Rechts der Isar, School of Medicine, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany.
E-mail address: jan.peeken@tum.de (J.C. Peeken).

https://doi.org/10.1016/j.ebiom.2019.08.059

Research in context

Evidence before this study

Therapy decisions for patients with soft tissue sarcomas (STS) vastly depend on the differentiation between low-grade and high-grade STS. In contrast to low-grade STS, patients diagnosed with high-grade STS have an unfavorable prognosis and often suffer from the occurrence of distant metastases. As consequence, patients with high-grade STS receive multimodal therapy regiments including radiation therapy and/or chemotherapy. So far, tumor grading is defined by histological work up after invasive biopsies. Previous studies, could identify semantic magnetic resonance imaging (MRI) derived properties, such as contrast enhancement, necrosis or tumor heterogeneity, associated with higher tumor grading. First small single center studies were able to demonstrate an association of quantitative radiomics features with tumor grading, too.

Added value of this study

We have developed non-invasive tumor grading models for the differentiation between low-grade and high-grade sarcomas using MRI-based radiomics. All models were validated using an independent external patient cohort. The radiomic classifier based on contrast enhanced T1-weighted MRI sequences achieved significant patient risk stratification. Combining the T2-weighted sequence based radiomic model into a nomogram with clinical stages improved the clinical net benefit and prognostic performance above clinical staging alone.

Implications of all the available evidence

The study provides radiomic MRI-based models to non-invasively differentiate low-grade from high-grade sarcomas. The presented models may be used as a non-invasive grading estimate if the pathological workup does not yield a clear result, or the tumor exhibits inhomogeneous areas that are difficult to access via CT-targeted or open biopsy. Moreover, the proposed nomogram may be used for improved prognostic assesment prior to therapy.

and tumor mortality slightly better than the NCI system [5]. The most important differentiation is between G1 (referred to as low-grade) and G2 or G3 (referred to as high-grade) as it has direct therapeutic consequences in the primary setting. For instance, patients with high-grade STS regularly receive additional radiotherapy and/or chemotherapy in contrast to patients with low-grade STS [1,6,7].

As an alternative method for characterizing tumors, radiomics allows for large scale high-throughput analysis of imaging data revealing information beyond qualitative assessment by experts [8–10]. Such extracted radiomics features have been shown to predict different clinical endpoints such as patient prognosis or molecular aberrations [11,12]. Recently, first studies have indicated the potential of radiomics to predict tumor grading in multiple cancers such as neuroendocrine pancreatic tumors or gliomas [13,14]. For STS, two previous studies described the potential of radiomics to predict overall survival, distant disease progression, and response to neoadjuvant chemotherapy [15–18].

The scope of this work was to generate magnetic resonance imaging (MRI)-based radiomic grading models to differentiate low-grade from high-grade STS. We evaluated the influence of different MRI sequences to find the optimal prediction model. Finally, a radiomics nomogram was created for prognostic assessment. The propensity of patient risk stratification and the net benefit in clinical decision analysis was analyzed for the developed models.

2. Methods

2.1. Patients

The trial was registered at ClinicalTrials.gov (NCT03798795). All patients with histologically proven STS with known FNCLCC tumor grading determined by biopsy prior to therapy and availability of contrast-enhanced T1-weighted fat saturated (T1FSGd) as well as fat-saturated T2-weighted (T2FS) MRI sequences of pre-therapeutic MRI were included. Patient records were analyzed for patient demographics (see Table 1) and histological subtypes (see Supplemental Table 1) from two independent retrospective patients cohorts from the Technical University of Munich (TUM) (138 patients, one patient with two

Table 1
Patient demographics, tumor properties and outcome of patients included for model building.

Institution		TUM training	UW validation	*p*-Values[a]
Total patients		122 p	103 p	
Location	Extremity or trunk	115/122 p (94%)	99/103 p (96%)	1
	Abdomen/retroperitoneal	7/122 p (6%)	4/103 p (4%)	
Age		m 57 (sd: 16.7)	m 52.7 (sd:15.5)	.2
Gender	Female	57/122 p (47%)	47/103 p (46%)	1
	Male	65/122 p (53%)	56/103 p (54%)	
T-stage[b]	1	18/122 p (15%)	17/103 p (17%)	1
	2	104/122 p (85%)	86/103 p (83%)	
	a	11/122 p (9%)	1/103 p (1%)	1
	b	111/122 p (91%)	102/103 p (99%)	
M-stage[b]	0	117/122 p (96%)	102/103 p (99%)	1
	1	5/122 p (4%)	1/103 p (1%)	
N-stage[b]	0	120/122 p (98.0%)	103/103 p (100%)	1
	1	2/122 p (2%)	0/103 p (0%)	
Grading[c]	1	48/122 p (4.8%)	20/103 p (19.5%)	.05
	2	28/122 p (32.1%)	31/103 p (32.2%)	
	3	46/122 p (46.4%)	52/103 p (48.3%)	
Tumor volume		197.6 cc (sd: 391.4)	164.5 cc (sd: 428.8)	
AJCC-stage	IA	8/122 p (7%)	4/103 p (4%)	$p < .001$
	IB	40/122 p (33%)	16/103 p (16%)	
	IIA	8/122 p (7%)	13/103 p (13%)	
	IIB	5/122 p (4%)	23/103 p (22%)	
	III	57/122 p (47%)	47/103 p (46%)	
	IV	4/122 p (3%)	0/103 p (0%)	
Median OS		35.7 mo	44.7 mo	.2

Abbreviations: cc: cubic centimeters, m: median, mo: months, p: patients, TUM: Technical University of Munich, UW: University of Washington, sd: standard deviation.

[a] Categorial variables: Fisher' exact test (2 cohorts)) and continuous/Rank variables: Wilcoxon rank sum test (2 cohorts), log-rank test for survival, with bonferroni correction for multiple testing.

[b] Following AJCC staging system version 7 [36].

[c] According to French Federation of Cancer Centers Sarcoma Group (FFCCS) [5].

independent STS) and the University of Washington, Seattle (UW) (139 patients). Exclusion criteria were previous RT, Ewing sarcoma, primary bone sarcomas, and endoprosthesis-dependent MRI artifacts. Approval from the ethics committee was received in both institutions. Patients were treated after informed consent. Overall survival (OS) was calculated from the initial pathologic diagnosis to the time point of death or the time point of censoring.

2.2. Image acquisition and definition of volume of interests

Complete imaging studies (availability of both sequences of interest) were found in 122 patient in the TUM cohort and 103 patients in the UW cohort (see patient workflow in Supplemental Fig. 1). See Supplemental Tables 2 and 3 for MRI vendors and acquisition parameters.

Tumor segmentation was conducted manually using MIM software version 6.6 at UW by MS, MM and TC (MIM Software Inc., Cleveland, USA), Eclipse 13.0 (Varian Medical Systems, Palo Alto, USA), and iplan RT 4.1.2 (Brainlab, Munich, Germany) at TUM by JP. The volume of interest (VOI) was defined as the primary tumor excluding surrounding edematous changes. Multiple delineations were performed for 21 randomly selected patients by three radiation oncology residents (JP, TA, MS) in the TUM cohort (see Fig. 1) to compensate for operator-dependent bias. For comparison, dice coefficients (DC) were calculated using the DiceComputation module of 3D Slicer (3D Slicer, Version 4.8 stable release) [19].

2.3. Image preprocessing and radiomics feature extraction

Feature extraction, model building and statistical analyses were performed by JP and AO. In order to compensate for non-uniform intensity caused by field inhomogeneity, N4ITK MRI bias field correction was applied to each imaging study using Slicer3D implementation (3D Slicer, Version 4.8 stable release) [20]. The pyradiomics (version 2.1) implementation in python (version 3.6.4) was utilized for further preprocessing steps and radiomics feature extraction [21]. Intensity normalization redistributed the image at the mean with the standard deviation and a scale of 100. A fixed bin width of 10 was used for image discretization (a detailed description is provided in the supplemental material). Isotropic resampling to a voxel size of $3 \times 3 \times 3$ mm was performed using Bspline interpolation. All models were also calculated using a voxel size of $1 \times 1 \times 1$ mm showing inferior predictive performances.

Image reconstruction was performed applying wavelet decompositions filtering and Laplacian of Gaussian filtering with sigma values of 3.0, 4.0, and 6.0. Further on, local binary pattern-derived images were calculated using a level of one and two as well as the kurtosis image. 1394 features were extracted from filtered and original images in three dimensions including shape features, first-order features, and texture features. All extracted features are listed in Supplemental Table 4.

2.4. Feature reduction and predictive model building

Feature reduction, batch correction, modeling, and statistical analyses were performed in R (version 3.4.0, R core team, Vienna, Austria). All features susceptible to minor segmentation variances in the 21 patients that received multiple independent segmentations (intraclass correlation coefficient (3,1) of <0.8) were excluded.

For model building and unbiased performance evaluation on the training set, 10 iterations of five-fold nested cross validation were performed using the code published by Deist et al. built upon the "caret" package [9,22]. The total training set was split into five subgroups (outer folds). Each Subgroup was then split once more for five times (inner folds). Hyperparameters were optimized as part of the inner folds. The selected hyperparameters were then used for testing on the five outer folds. The total mean area under the receiver operator characteristic (ROC) curve (AUC) over all outer folds was calculated for model comparison (see Supplemental Fig. 2 for a graphical depiction). The

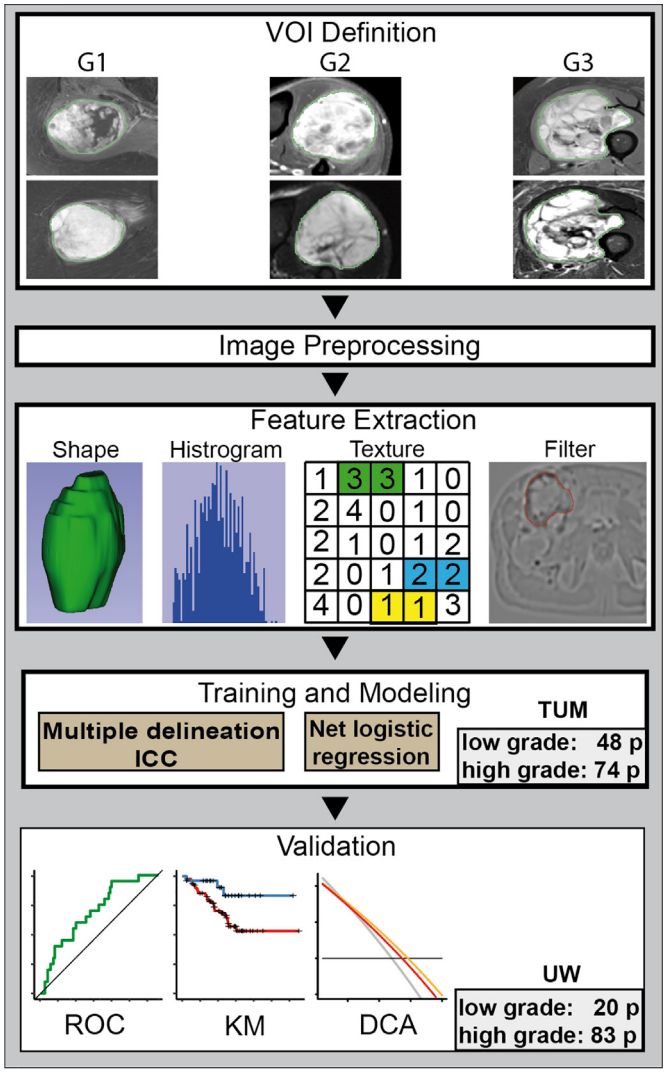


Fig. 1. The radiomics workflow. Abbreviations: DCA: decision curve analysis, ICC: intra class coefficient, KM: Kaplan Meier survival curve, ROC: receiver operator characteristic curve, TUM: Technical University of Munich, UW: University of Washington.

hyperparameter combination with the best mean performance was used to retrain a final model on the whole training set. Model performance with the optimal hyperparameter combination was assessed using cross validation as described above.

First, we compared the performance of seven machine learning techniques with inbuilt feature reduction (elastic net logistic regression, least absolute shrinkage and selection operator (LASSO), random forest, LogitBoost, decision tree, support vector machine and neural network). The influence of additional feature reduction by principal component analysis (PCA), resampling and reweighting were tested (see Supplemental Table 6) [22]. None of the techniques improved the performance further. We decided to apply the LASSO method without prior PCA as it showed good performance in cross validation, model simplicity and low tendency to overfitting (see supplemental material for method description and Supplemental Table 7 for selected hyperparameter values). Previous studies have demonstrated a competitive performance compared to other machine learning methods [22,23]. At the same time, it allows for a better interpretability due to the direct linear relationship between input features and outcome.

All models were trained to classify high-grade STS (G1) vs. low-grade STS (G2/G3). Three radiomic models *Radiomics-T1FSGd*, *Radiomic T2FS*, and *Radiomics-combined* were developed using the respective features as input. For comparison, logistic regression models of *Tumor-*

Volume alone, a *Clinical* model based upon "Age" and the TNM-staging system, and a combined model *Clinical-Volume-combined* were evaluated.

Finally, we trained elastic net regression models for the prediction of overall survival (OS) using five-fold cross validation using the "glmUtils" package. Model performance of all final models was finally determined on the external validation patient cohort (UW) to get an unbiased result.

2.5. Batch correction via ComBatHarmonization

Recently the ComBatHarmonization has been proposed as a method for the correction of batch effects among radiomics multicenter cohorts [24]. Its value to potentially improve reproducibility between different centers has been shown in multiple studies [25,26]. Based on the given feature distribution it estimates the additive and multiplicative batch effects using a maximum likelihood approach. We applied the R ComBat script (https://github.com/Jfortin1/ComBatHarmonization) correcting for MRI scanner models [27]. Further on, we evaluated the effect of using STS histology as a biological covariate.

2.6. Decision curve analysis

In order to analyze the clinical usefulness of the developed classifiers, decision curve analysis was conducted as described by Vickers et al. [https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/decision-curve-analysis] [28]. The net benefit is calculated by subtracting the proportion of false-positive patients from the proportion of true-positive patients, weighted by the relative harm of a false-negative and false-positive result. As a reference, the decision curves for treating all patients and treating no patients are displayed. A decision models shows a clinical benefit if it achieves larger net benefit values than both reference strategies.

2.7. Statistical analysis

ROC curves and respective AUC values were calculated and compared using the 1000-fold bootstrapping as implemented in the "fbroc" and "pROC" packages [29]. Patient stratification for survival was tested via Kaplan-Meier survival analysis using the "ggkm" package. Dichotomization was performed using the class output provided by the glmnet function inside of the "caret" package. Statistical significance was assessed using the log-rank test. Univariate logistic regression was used to assess correlation to tumor grading using the "survival" package. Calibration curves were generated using the "gbm" package. Bonferroni correction was performed in cases of multiple testing. A p-value below .05 was regarded as significant.

2.8. Data sharing statement

Due to patient privacy concerns the datasets are currently not publicly available. However, data may be obtained from the corresponding author on reasonable request and may be published online in the future after approval by the institutional ethics review boards.

3. Results

Fig. 1 shows the workflow of the study.

3.1. Patient characteristics and volume of interest definition

The overall distributions of tumor grading ($p = .005$) and histological subgroups ($p < .001$) between both cohorts were significantly different (see Table 1 and Supplemental Table 2, respectively). The training dataset consisted of 48/122 (39%) low-grade and 74/122 (61%) high-grade STS. With 20/103 (19%) low-grade STS and 83/103 (81%) high-grade STS the validation dataset was skewed towards high-grade tumors. The similarity between multiple segmentation was overall high with a mean dice similarity coefficient of 0.91 (standard deviation 0.069).

3.2. Training of the radiomic classifiers for tumor grading prediction

After radiomics feature reduction due to segmentation variability 1309, 1340, and 2609 features were used for the development of the *Radiomics-T1FSGd*, *Radiomics-T2FS* and *Radiomics-combined* models, respectively. During training of *Radiomics-T2FS*, *Radiomics-T1FSGd*, and *Radiomics-combined*, *24*, *22* and *13* features with non-zero coefficients were selected, respectively (see Supplemental Table 8 for selected features and feature importance ranking). Features were selected from various feature types and filtering methods. The most important T2FS-based features were first order intensity features based on wavelet filtering (wavelet.LHL_firstorder_Kurtosis and wavelet.HLH_firstorder_Mean), and the local binary pattern texture feature busyness (lbp.3D.k_ngtdm_Busyness). The T1FSGd based model was similarly dominated by wavelet.HLH_firstorder_Mean as for T2FS. In addition, wavelet decomposition based texture features such as "Maximum Correlation Coefficient" (MCC) computed on the "Gray level co-occurrence matrix" (GLCM) and "DependenceEntropy" derived from the "Gray Level Dependence Matrix" (GLDM) (wavelet.HLH_glcm_MCC and wavelet.HLH_gldm_DependenceEntropy) were among the three most important features.

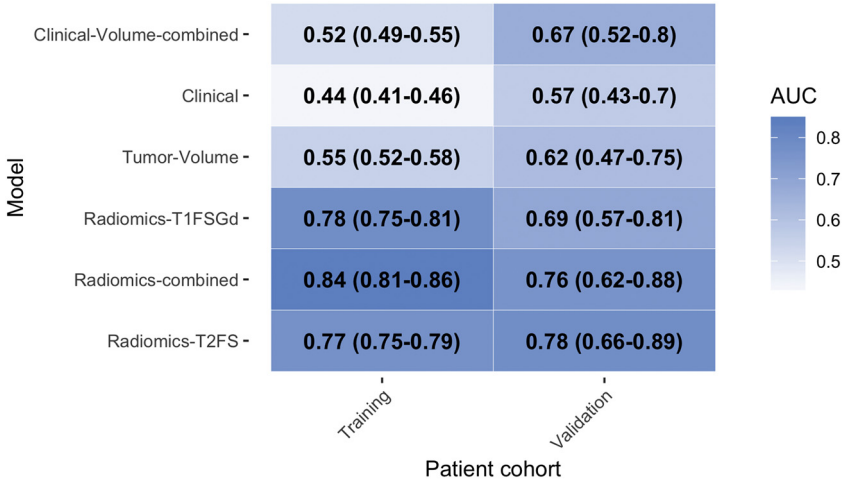
With AUC-values of 0.77 and 0.78 the models *Radiomics-T2FS* and *Radiomics-T1FSGd* showed a similar predictive performance ($p = .427$) on the training set (see Table 2 for AUC values and 95% confidence intervals (95%CI)). The *Radiomics-combined* model achieved a significantly better predictive capacity (AUC 0.84) than *Radiomics-T2FS* and *Radiomics-T1FSGd* ($p < .001$, respectively). The *Clinical* model (AUC: 0.55) and the *Tumor-Volume* model (AUC: 0.44) did not show a predictive capacity better than random. Combining clinical parameters and tumor volume as "*Clinical-Volume-combined*" model did not increase predictive performance further (AUC: 0.52). All three models performed significantly worse than all radiomic model s($p < .001$, for each comparison).

3.3. Independent validation of the developed classifiers

The predictive models were evaluated on the validation set after batch correction. The T2FS-based model showed good reproducibility with an AUC of 0.78 (absolute difference of +0.01) which was significantly better than *Clinical* but not *Tumor-Volume* ($p = .01$ and $p = .056$, respectively) (see Table 2 for AUC values and 95%CI, Fig. 2 for ROC curves, and Supplemental Fig. 3 for calibration curves). Validation of the *Radiomics-T1FSGd* model showed a larger drop in predictive performance with an absolute difference of −0.09 and an AUC of 0.69. The *Radiomics-combined* model showed a similar performance than the T2FS model (AUC 0.76, absolute difference −0.09) which was, however, not significantly better than *Clinical* or *Tumor-Volume* ($p = .056$ and $p = .165$, respectively). The observed difference in between the radiomic model did not reach statistical significance (*Radiomics-T1FSGd* vs. *Radiomics-T2FS*, $p = .192$, *Radiomics-T1FSGd* vs. *Radiomics-combined*, $p = .348$). The *Clinical* model and the *Tumor-Volume* model showed no predictive performance better than random. *Clinical-Volume-combined*, however, showed predictive performance significantly better than random with an AUC of 0.67.

In contrast to *Clinical*, *Tumor-Volume, Clinical-Volume-combined,* and *Radiomics-T1FSGd*, *Radiomics-T2FS* and *Radiomics-combined* were significantly associated with tumor grading in univariate logistic regression analysis in the independent validation cohort (odds ratio (OR) for a 10% increase in predictived probability: *Clinical* OR = 1.4 (95%CI 0.62-3.3) $p = 1.0$, *Tumor-Volume* OR = 1.5 (95%CI 0.9-6.1) $p = .35$, *Radiomics-T1FSGd* OR = 1.4 (95%CI 1.09-1.86) $p = .064$, *Radiomics-*

Table 2
Predictive performance metrics of the radiomic classifiers.



Area under the receiver operator characteristic curve (AUC) values for the differentiation of high-grade from low-grade soft tissue sarcomas. 95% confidence intervals are shown in parentheses.

T2FS OR 1.6 (95%CI 1.26-2.19) $p = .0021$, and *Radiomics-combined* OR = 2.1 (95%CI 1.39-3.30) $p = .0045$).

The *Radiomics-combined* model showed the highest accuracy (0.83) and sensitivity (0.90), but low specificity (0.50), for the prediction of high-grade STS compared to *Radiomics-T2FS* (accuracy: 0.78, sensitivity: 0.87, specificity: 0.40) and *Radiomics-T1FSGd* (accuracy: 0.78, sensitivity: 0.87, specificity: 0.40) (see Supplemental Table 9 for additional performance metrics) on the validation set.

3.4. Influence of batch correction

The influence of ComBatHarmonization on validation performance was different depending on the predictive model. The T1FSGd-based models *Radiomics-T1FSGd* improved with an absolute AUC difference of +0.08. *Radiomics-T2FS* and *Radiomics_combined* showed less marked increases in AUC of +0.03 and + 0.04, respectively. Application of histology as biological covariate did not increase predictive performance

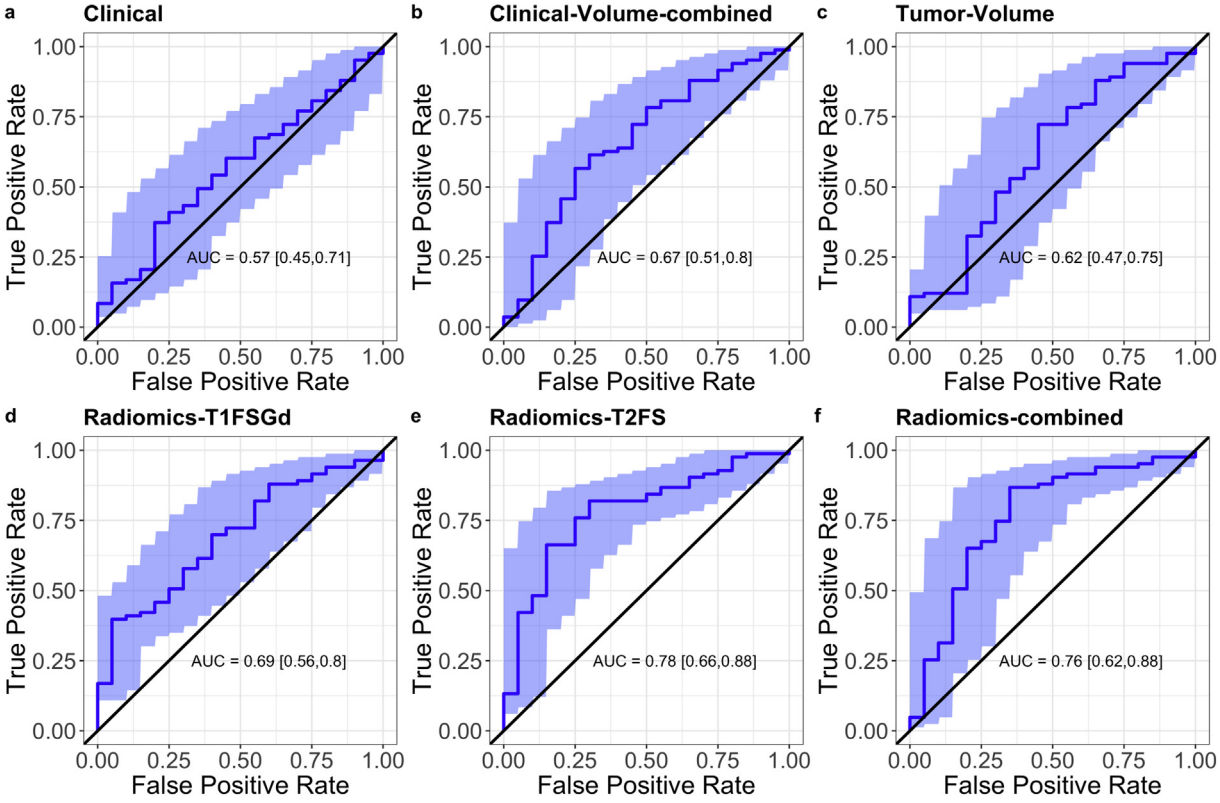


Fig. 2. Predictive performance of radiomics tumor grading models. Receiver operator characteristic curves (ROC) and the respective area under the curve (AUC) values depicting the performance of the prediction models *Clinical*, *Clinical-Volume-combined*, *Tumor-Volume*, *Radiomics-T1FSGd*, *Radiomics-T2FS*, and the *Radiomics-combined* on the validation cohort. The shaded blue areas depict the 95% confidence interval which is shown in parentheses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

profoundly rising the performance of *Radiomics-T1FSGd* to 0.70 while decreasing performances of *Radiomis-T2FS* and *Radiomics-combined* (AUC of 0.75 and 0.76, respectively) (see Supplemental Table 10).

3.5. Patient risk stratification

Tumor grading (high-grade vs. low-grade) significantly stratified patients for OS in the training and validation patient cohorts (p= 0.0006 and p = .045, respectively) (see Fig. 3 for Kaplan Meier survival curves and p-values). We used the classification of the developed radiomics grading models for dichotomization of the patient cohort into low-risk and high-risk patients. On the training set, all models achieved a separation of survival curves. The difference in survival was significantly different for the models *Radiomics-T1FSGd* (p = .0318), *Radiomics-T2FS* (p = .00328), and *Radiomics-combined* (p = .0204) but not for *Clinical-Volume* (p = .334). On the validation set, *Radiomics-T1FSGd* achieved a separation of survival curves with significant patient stratification (p = .0268). Interestingly, there was no significant separation of survival curves for the three other models.

3.6. Development of a clinical radiomics nomogram

Next, we analyzed if the proposed models may provide an incremental benefit above the existing clinical staging system. We generated multivariate nomograms by combining the *AJCC* staging system (7th edition) with the respective grading models. The model based on *Radiomics-T2FS* showed the best predictive performance for OS in the validation set with a concordance index (C-index) of 0.74 (95%CI 0.64–0.84) in comparison to *Radiomics-T1FSGd* (C-index: 0.71, 95%CI 0.61–0.81), *Radiomics-combined* (C-index: 0.72, 95%CI 0.62–0.83), and to the *AJCC* staging system alone (C-index 0.69, 95%CI 0.60–0.0.78). For comparison, tumor volume alone and inside of a similar multivariate model achieved a C-indics of 0.54 (95%CI 0.43–0.65) and 0.71 (95%CI 0.61-0.81), respectively. See Supplemental Table 11 for all training and testing performances.

For the best model based on *Radiomics-T2FS*, a nomogram was created (see Fig. 4a). With an AUC of 0.82 for 2-year survival, the model showed good discrimination and good calibration (Fig. 4b and c). Further on, it achieved significant patient risk stratification (p < .0001) (Fig. 4d). In decision curve analysis, the *Radiomics-T2FS nomogram* outperformed the "treat all" and "treat none" strategies, as well as the *AJCC* staging system and tumor grading alone in the threshold probability range between 0.2 and 0.55.

3.7. Prognostic radiomic prediction models show moderate performance

Finally, we trained radiomic prediction models directly for OS. The models based on the T1FSGd features, T2FS features, the combined feature set, and Tumor-Volume achieved predictive performances with C-indices of 0.55 (95%CI 0.45–0.65), 0.60 (95%CI 0.49–0.70), 0.60 (95%CI 0.50–0.69), and 0.54 (95% 0.43–0.65), respectively. Combining the *AJCC* staging system and the above mentioned models did not confer an incremental benefit (C-indices for T1FSGd: 0.67 (95%CI 0.57–0.77), T2FS: 0.70 (95%CI 0.59–0.80), combined features: 0.64 (95%CI 0.55–0.74), and Tumor-Volume: 0.71 (95%CI 0.61–0.81)).

4. Discussion

We have developed MRI-based radiomic models for tumor grading of STS focusing on the differentiation of high-grade STS from low-grade STS. A T2FS-sequence-based model showed a good reproducibility with the best performance on the independent validation patient set. A radiomic model involving T1FSGd-based features demonstrated a larger drop in predictive performance which was partly improved by ComBat harmonization. The Radiomics-T1FSGd model achieved significant patient risk stratification for OS. A nomogram combining *Radiomics-T2FS* and the *AJCC* clinical staging system achieved the best prognostic performance with significant patient stratification and a larger clinical net benefit than *AJCC* staging system alone. Radiomic models directly trained for the prediction of OS showed only moderate performtances.

Previous analyses demonstrated a correlation of qualitative MRI features such as peritumoral contrast enhancement with tumor grading [30]. In a recent small retrospective study encompassing 19 patients, diffusion weight MRI-based radiomics features were used to distinguish G2 and G3 STS [31]. A further study used T2FS sequence based radiomics to discriminate low-grade from high-grade STS in a small pilot study of 35 patients [32]. Although patient numbers were substantially low and

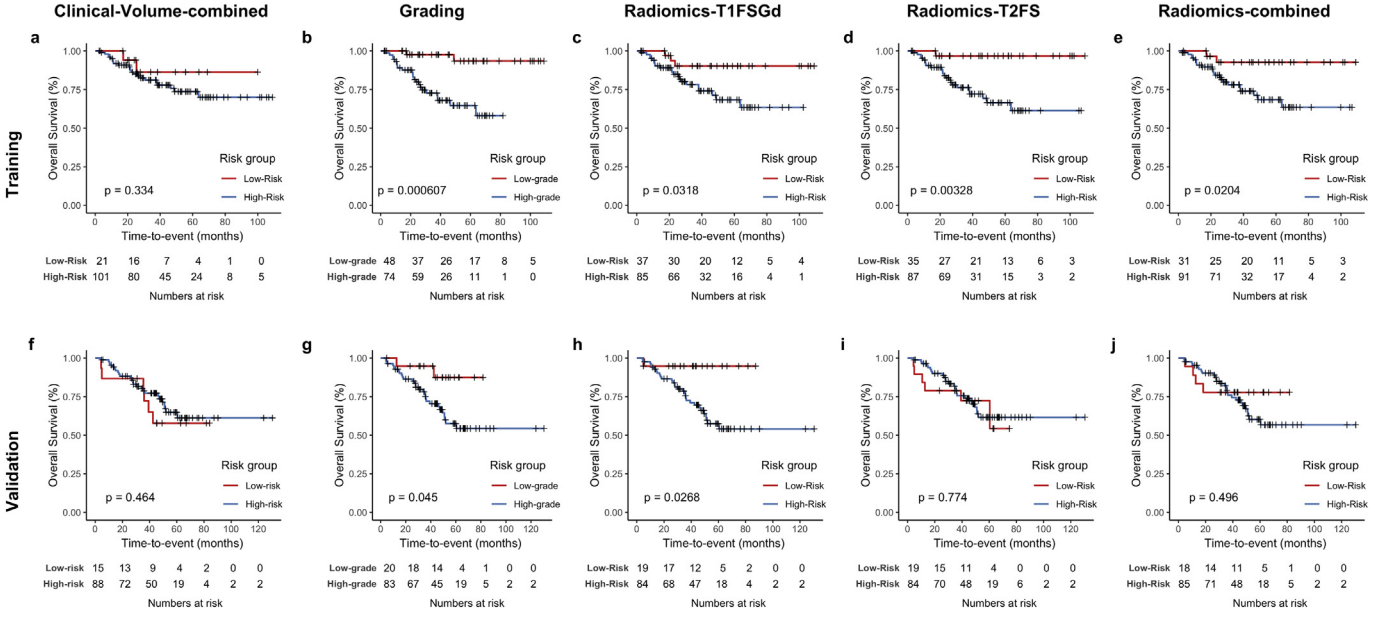


Fig. 3. Patient risk stratification. Kaplan Meier survival curves for patients' overall survival displaying patient stratification by the *Clinical-Volume-combined* model (a,f), tumor grading (low-grade vs. high-grade) (b,g), the *Radiomics-T1FSGd* model (c,h), the *Radiomics-T2FS* model (d,i) and the *Radiomics-combined* model (e,j) on the training (a,e) and validation (f,j) patient cohort.

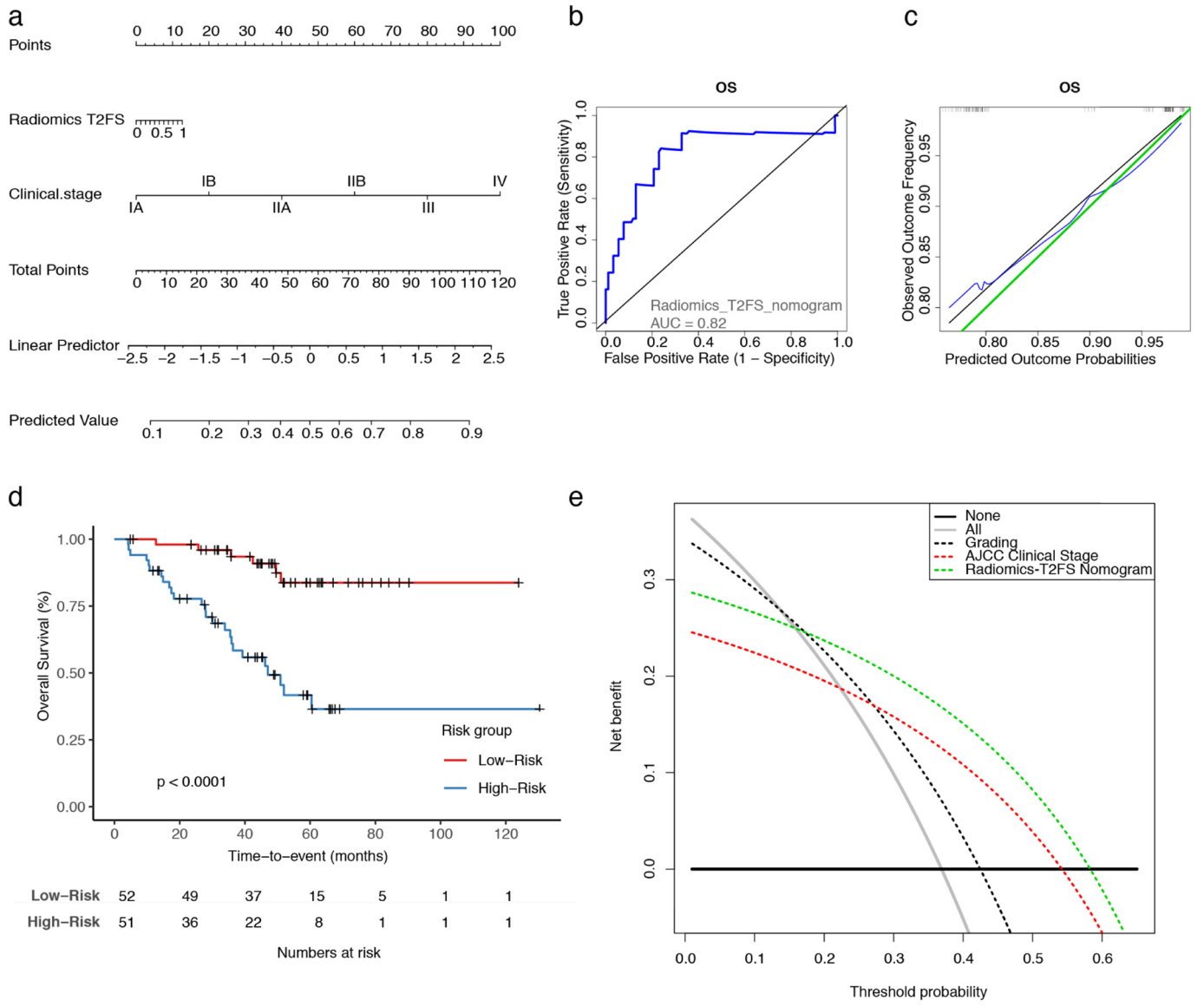


Fig. 4. A clinical radiomics nomogram. A multivariate nomogram combining the *Radiomics-T2FS* prediction model with the *AJCC* clinical staging system is illustrated (a). The receiver operator characteristic (ROC) curve and the representative area under the curve (AUC), as well as the calibration curve, each at two years, are shown (b,c). The Kaplan Meier survival curve for patients' overall survival displaying patient stratification by the proposed nomogram is depicted (d). Finally, decision durve analysis was performed comparing the net benefit by the *Radiomics-T2FS nomogram* with the *AJCC* clinical stage and tumor grading alone. The net benefit is calculated by subtracting the proportion of false-positive patients from the proportion of true-positive patients, weighted by the relative harm of a false-negative and false-positive result [28]. The threshold probability was calculated for death after five years. For reference, the two strategies "treat all" and "treat none" are displayed. A decision model shows a clinical benefit if the respective curve shows larger net benefit values than both reference strategies.

no external validation cohorts were used, the studies demonstrated the principal potential to correlate quantitative radiomic features to tumor grading. A further study could demonstrate the differentiation of G2 and G3 STS using computer tomography-based radiomics albeit with a low predictive performance with an AUC of 0.65 [18]. Although radiomics features seem to provide information regarding the differentiation of G2 or G3 STS, there is currently no clinical benefit for such a differentiation. Therefore, we focused on the clinically relevant differentiation of high-grade from low-grade STS.

An imaging-based prognostic classifier as developed in this study could be applied clinically in multiple scenarios. First, in the rare case that a biopsy is clinically difficult to perform due to the anatomical location, the radiomic classifier could be used instead to identify high-grade tumors as a basis for therapy decisions. Further practical applications may involve tumors that exhibit inhomogeneous areas that are difficult to access or have different access routes in CT-targeted or open biopsies. Secondly, if the pathological work-up does not yield a clear result the proposed models may be used as an additional biomarker. Thirdly, in radiotherapy planning the spatial distribution of relevant radiomics features could be used as a basis for dose painting following the radiomics target volume concept [9,33]. For instance, radiation dose could be escalated in high-grade subvolumes. Finally, as proposed in this study the MRI radiomic grading model may be combined with clinical staging providing complementary information and improving prognostic stratification.

In the current study we restricted the volumes of interest to the gross tumor volume excluding surrounding edematous changes in both sequences. By this, we hoped to ensure a better comparability of T2FS and T1FSGd based models. Secondly, inclusion of edematous changes may produce less reproducible segmentations and radiomic feature values could be influenced by the kind of the surrounding tissue. On the contrary, by using this approach we may have missed additional information such as measures for tissue infiltration.

Further on, we analyzed the helpfulness of radiomic models for prognostic assessment. Radiomic models directly trained to predict OS only showed moderate predictive performances. The *Radiomics-T1FSGd* grading model showed the propensity of significant patient risk stratification in the validation set in contrast to the models with the highest performance in predicting tumor grade (*Radiomics-T2FS* and *Radiomics-combined*). On the contrary, *Radiomics-T1FSGd* showed less improvement in prognostic performance in combination with the *AJCC* staging system than *Radiomics-T2FS* and *Radiomics-combined*. Considering that the T1FSGd-based model trained directly for OS showed worse predictive capacity than T2FS-based models contradicts the superior patient stratification of the *Radiomics-T1FSGd* grading model observed in this study.

Our models achieved accuracies for the detection of high-grade STS of 0.78 (*Radiomics-T1FSGd* and *Radiomics-T2FS*) to 0.83 (*Radiomics-combined*) with good positive predictive values (PPV) (0.86–0.88) on the expanse of negative predictive values (NPV). Thus, classifications of high-grade STS ("positive prediction") provided a more reliable results than the prediction of a low-grade STS. The number of false negative results (i.e. high-grade STS that were classified as low-grade STS) impaired the prognostic validation. The large observed difference between PPV and NPV, as well as the suboptimal patient stratification may be in part explained by the large prevalence of high-grade sarcomas in the validation set (81% of patients). Tumor grading itself achieved significant patient stratification in this skewed validation set only by a close margin ($p = .045$) making a significant patient stratification for surrogate markers, such as the proposed radiomic models, even more difficult. A future external TRIPOD type 4 validation with a more balanced patient data optimally reflecting the true tumor grading distribution may help to assess the impact of the proposed models in terms of classification and prognostic stratification [34].

There are several limitations to this work. In the current study, one could observe a drop in performance for the T1FSGd-based grading models. We applied nested cross validation for an optimally unbiased performance evaluation on the training set while trying to reduce overfitting. Still, reproducibility remains suboptimal with absolute differences in AUC of 0.09 and 0.08 for *Radiomics-T1FSGd* and *Radiomics-combined* between training and validation sets, respectively. There are potential explanations for this phenomenon. First, there was a large technical heterogeneity inside and between the two cohorts. In total, 20 MRI scanner types produced by four manufacturers were used across the two cohorts. Thirteen scanner types in the validation set were not present in the training set. The improved testing performance after performing batch correction with an absolute increase in AUC of up to 0.08 may be a sign of the important role of such batch-dependent effects. In contrast, *Radiomics-T2FS* showed good reproducibility even without batch correction. Secondly, STSs constitute one of the most heterogeneous malignant entities with over 100 histological subtypes [35]. In our study, a direct comparison of histological subtypes revealed a significant difference between training and testing cohort. Given the propensity of radiomics to predict tumor grading and molecular aberrations, the histological subtype may have a substantial influence on the radiomics phenotype. In particular, the amount of contrast enhancement of specific histologies may be an interfering factor, which may also explain that the T2FS-based model without contrast agent showed better reproducibility. In particular, the large proportion of liposarcomas in the training set (39%) may be an explanatory cause. Including histology as biological covariate into ComBat harmonization was not able to improve reproducibility. One solution to these problems may be a future prospective trial which would need to be sufficiently large to achieve a representative distribution of all histological subtypes. If future cohort sizes would achieve a critical number, even better performances may be possible by restricting model building to single histological subtypes.

To conclude, we developed MRI-based radiomic grading models differentiating high-grade from low-grade STS. External validation confirmed classification performance. A radiomic classifier using T2FS sequences appeared to be superior to a T1FSGd-based model predicting tumor grading. Integrating the *Radiomics-T2FS* model with the *AJCC* clincal staging system improved prognostic assessment and the clinical net benefit.

Funding

This work was funded in part by research support for JP within the KKF physician scientist program of the Medical Faculty of the Technical University of Munich (TUM) and Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich (to JP, SC). None of the sponsors was involved in study design, collection, analysis, and interpretation of data or writing of the manuscript. The corresponding author had the final responsibility for the decision to submit the manuscript for publication.

Acknowledgments

We sincerely thank Chapman TR and Macomber MW for segmentation of VOIs at UW.

Declaration of competing interests

The authors declare no potential conflicts of interest.

Author contributions

Jan C. Peeken: Design of the work, collection of data, data analysis, statistical analyses and writing of the manuscript.
Matthew B. Spraker: Collection of data and data analysis.
Carolin Knebel: Collection of data and writing of the manuscript.
Hendrik Dapper: Collection of data.
Daniela Pfeiffer: Data analysis.
Michal Devecka: Data analysis.
Ahmed Thamer: Collection of data, data analysis.
Mohamed A. Shouman: Collection of data, data analysis.
Armin Ott: Statistical analyses.
Rüdiger von Eisenhart-Rothe: Revision of the manuscript.
Fridtjof Nüsslin: Design of the work and revision of the manuscript.
Nina A. Mayr: Design of the work and revision of the manuscript.
Matthew J. Nyflot: Design of the work, literature search and revision of the manuscript.
Stephanie E. Combs: Design of the work and revision of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.08.059.

References

[1] Gutierrez JC, Perez EA, Franceschi D, Moffat FL, Livingstone AS, Koniaris LG. Outcomes for soft-tissue sarcoma in 8249 cases from a large state cancer registry. J Surg Res 2007;141:105–14.
[2] Amin MB, Edge S, Greene F, et al, editors. AJCC cancer staging manual. 8th ed. Springer International Publishing; 2017.
[3] Trojani M, Contesso G, Coindre JM, et al. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. Int J Cancer 1984;33:37–42.
[4] Costa J, Wesley RA, Glatstein E, Rosenberg SA. The grading of soft tissue sarcomas. Results of a clinicohistopathologic correlation in a series of 163 cases. Cancer 1984;53:530–41.
[5] Guillou L, Coindre JM, Bonichon F, et al. Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. J Clin Oncol 1997;15:350–62.

[6] Peeken JC, Knie C, Kessel KA, et al. Neoadjuvant image-guided helical intensity modulated radiotherapy of extremity sarcomas – a single center experience. Radiat Oncol 2019;14:4–11.
[7] Peeken JC, Goldberg T, Knie C, et al. Treatment-related features improve machine learning prediction of prognosis in soft tissue sarcoma patients. Strahlenther Onkol 2018;194:824–34.
[8] Peeken JC, Nüsslin F, Combs SE. "Radio-oncomics" - the potential of radiomics in radiation oncology. Strahlenther Onkol 2017;193:767–79.
[9] Peeken JC, Bernhofer M, Wiestler B, et al. Radiomics in radiooncology – challenging the medical physicist. Phys Med 2018;48:27–36.
[10] Peeken JC, Kessel KA, Nüsslin F, Braun AE, Combs SE. Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients. Strahlenther Onkol 2018;194:824–34.
[11] Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
[12] Rios Velazquez E, Parmar C, Liu Y, et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. Cancer Res 2017;77:3922–30.
[13] Pyka T, Gempt J, Hiob D, et al. Textural analysis of pre-therapeutic [18F]-FET-PET and its correlation with tumor grade and patient survival in high-grade gliomas. Eur J Nucl Med Mol Imaging 2016;43:133–41.
[14] Liang W, Yang P, Huang R, et al. A combined nomogram model to preoperatively predict histologic grade in pancreatic neuroendocrine tumors. Clin Cancer Res 2018. https://doi.org/10.1158/1078-0432.CCR-18-1305.
[15] Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol 2015;60:5471–96.
[16] Spraker MB, Wootton LS, Hippe DS, et al. MRI radiomic features are independently associated with overall survival in soft tissue sarcoma. Adv Radiat Oncol 2019;4:413–21.
[17] Crombé A, Périer C, Kind M, et al. T2-based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. J Magn Reson Imaging 2018:1–14.
[18] 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.
[19] Fedorov A, Beichel R, Kalphaty-Cramer J, et al. 3D slicers as an image computing platform for thw quantitative imaging network. Magn Reson Imaging 2012;30:1323–41.
[20] Tustison NJ, Gee JC. N4ITK: Nick's N3 ITK implementation for MRI Bias field correction. InsightJournal 2009:1–8.
[21] van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7.
[22] 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.
[23] 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.
[24] Steiger P, Sood R. How can radiomics be consistently applied across imagers and institutions? Radiology 2019;291:60–1.
[25] Lucia F, Visvikis D, Vallières M, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging 2018;46(4):864–77.
[26] Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 2019;291:53–9.
[27] Fortin J, Parker D, Tunç B, et al. NeuroImage harmonization of multi-site diffusion tensor imaging data. Neuroimage 2017;161:149–70.
[28] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2008;26:565–74.
[29] Pepe MS, Longton G, Janes H, Pepe MS, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. Stata J 2009;9:1–16.
[30] Crombé A, Marcellin P-J, Buy X, et al. Soft-tissue sarcomas: assessment of MRI features correlating with histologic grade and patient outcome. Radiology 2019;291(3):710–21.
[31] Corino VDA, Montin E, Messina A, et al. Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. J Magn Reson Imaging 2017:1–12.
[32] Zhang Y, Zhu Y, Shi X, et al. Soft tissue sarcomas: preoperative predictive histopathological grading based on radiomics of MRI. Acad Radiol 2018:1–7.
[33] Shiradkar R, Podder TK, Algohary A, Viswanath S, Ellis RJ, Madabhushi A. Radiomics based targeted radiotherapy planning (Rad-TRaP): a computational framework for prostate cancer treatment planning with MRI. Radiat Oncol 2016;11:148.
[34] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Eur Urol 2015;67:1142–51.
[35] Fletcher C, Bridge J, Hogendoorn P, Mertens F. WHO classification of tumours of soft tissue and bone. 4th ed.WHO; 2013 http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=70&codcch=4005 (accessed March 27, 2019).
[36] 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.