MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy

https://doi.org/10.1016/j.radonc.2021.08.023
Available online 4 October 2021
Accepted 27 August 2021
Received in revised form 15 August 2021
Received 26 April 2021
Article history:
The standard of care for high-grade soft-tissue sarcomas (STS) includes surgery, radiation therapy (RT) and/or chemotherapy (CTx). These treatment strategies achieve high local control rates but unfavorable overall survival (OS) and distant control [1-4]. RT can be delivered in the neoadjuvant or adjuvant setting. Compared to adjuvant RT, neoadjuvant RT offers several advantages including lower radiation doses, smaller target volumes, and reduced late toxicities [3,5]. The administration of chemotherapy remains more controversial. The phase-III EORTC62931 trial couldn’t demonstrate a survival benefit for adjuvant chemotherapy

Purpose: In high-grade soft-tissue sarcomas (STS) the standard of care encompasses multimodal therapy regimens. While there is a growing body of evidence for prognostic pretreatment radiomic models, we hypothesized that temporal changes in radiomic features following neoadjuvant treatment ("delta-radiomics") may be able to predict the pathologic complete response (pCR).

Methods: MRI scans (T1-weighted with fat-saturation and contrast-enhancement (T1FSGd) and T2-weighted with fat-saturation (T2FS)) of patients with STS of the extremities and trunk treated with neoadjuvant therapy were gathered from two independent institutions (training: 103, external testing: 53 patients). pCR was defined as <5% viable cells. After segmentation and preprocessing, 105 radiomic features were extracted. Delta-radiomic features were calculated by subtraction of features derived from MRI scans obtained before and after neoadjuvant therapy. After feature reduction, machine learning modeling was performed in 100 iterations of 3-fold nested cross-validation. Delta-radiomic models were compared with single timepoint models in the testing cohort.

Results: The combined delta-radiomic models achieved the best area under the receiver operating characteristic curve (AUC) of 0.75. Pre-therapeutic tumor volume was the best conventional predictor (AUC 0.70). The T2FS-based delta-radiomic model had the most balanced classification performance with a balanced accuracy of 0.69. Delta-radiomic models achieved better reproducibility than single timepoint radiomic models, RECIST or the peri-therapeutic volume change. Delta-radiomic models were significantly associated with survival in multivariate Cox regression.

Conclusion: This exploratory analysis demonstrated that MRI-based delta-radiomics improves prediction of pCR over tumor volume and RECIST. Delta-radiomics may one day function as a biomarker for personalized treatment adaptations.

© 2021 Elsevier B.V. All rights reserved. Radiotherapy and Oncology 164 (2021) 73–82

A R T I C L E   I N F O

Article history:
Received: 26 April 2021
Received in revised form: 15 August 2021
Accepted: 27 August 2021
Available online: 4 October 2021

Keywords:
Soft-tissue sarcoma
Delta radiomics
Neoadjuvant radiotherapy
Machine learning
Response prediction
MRI

A B S T R A C T

Purpose: In high-grade soft-tissue sarcomas (STS) the standard of care encompasses multimodal therapy regimens. While there is a growing body of evidence for prognostic pretreatment radiomic models, we hypothesized that temporal changes in radiomic features following neoadjuvant treatment (“delta-radiomics”) may be able to predict the pathologic complete response (pCR).

Methods: MRI scans (T1-weighted with fat-saturation and contrast-enhancement (T1FSGd) and T2-weighted with fat-saturation (T2FS)) of patients with STS of the extremities and trunk treated with neoadjuvant therapy were gathered from two independent institutions (training: 103, external testing: 53 patients). pCR was defined as <5% viable cells. After segmentation and preprocessing, 105 radiomic features were extracted. Delta-radiomic features were calculated by subtraction of features derived from MRI scans obtained before and after neoadjuvant therapy. After feature reduction, machine learning modeling was performed in 100 iterations of 3-fold nested cross-validation. Delta-radiomic models were compared with single timepoint models in the testing cohort.

Results: The combined delta-radiomic models achieved the best area under the receiver operating characteristic curve (AUC) of 0.75. Pre-therapeutic tumor volume was the best conventional predictor (AUC 0.70). The T2FS-based delta-radiomic model had the most balanced classification performance with a balanced accuracy of 0.69. Delta-radiomic models achieved better reproducibility than single timepoint radiomic models, RECIST or the peri-therapeutic volume change. Delta-radiomic models were significantly associated with survival in multivariate Cox regression.

Conclusion: This exploratory analysis demonstrated that MRI-based delta-radiomics improves prediction of pCR over tumor volume and RECIST. Delta-radiomics may one day function as a biomarker for personalized treatment adaptations.
Delta-radiomics in soft-tissue sarcomas

[6]. The ISG-STS1001 trial, however, showed a survival benefit after neoadjuvant CTx compared to a “histology-tailored” treatment approach [7].

The benefit of neoadjuvant treatment concepts lays in the possibility to assess treatment response. This information could then be used for individual therapy escalation [8]. As a potential biomarker in STS patients, the pathological complete response (pCR) is currently being used in prospective trials as a surrogate marker for patients’ outcomes [9]. A recent meta-analysis found a significant predictive value for OS [10].

Imaging constitutes an alternative tool to characterize tissue. As consequence, multiple authors have proposed quantitative imaging analysis (“radiomics”) as a potential novel method to assess treatment response [11,12]. Radiomics is defined as a high-throughput quantitative imaging analysis of imaging data [13]. Pre-defined features assessing the texture, intensity distribution, or shape of a volume of interest (VOI) are calculated and used as input for machine learning (ML) models [13-15]. Radiomics has been shown to predict various clinical and biological endpoints including pathological characteristics, prognosis, tumor progression, spatial infiltrations, and molecular aberrations in multiple cancer types [16-22]. In STS patients, there is growing evidence that radiomics can be successfully applied to predict OS, distant progression, and tumor grading using pretherapeutic imaging [23-29]. We hypothesize that temporal changes of radiomic features (“delta-radiomics”) obtained before and after neoadjuvant therapy may be able to predict treatment response [30,31]. Crombé et al. first demonstrated prediction of pCR after neoadjuvant CTx using a delta-radiomics approach in STS patients in a monocentric study [32].

We analyzed the potential of MRI-based delta-radiomics to predict pCR in STS patients that received neoadjuvant RT and/or CTx. Radiomic analysis was performed using the two MRI sequences “contrast-enhanced and fat-saturated T1-weighted” (T1FSGd) and “fat-saturated T2-weighted” (T2FS) obtained before and after neoadjuvant therapies. The results were compared to tumor volume changes, Response Evaluation Criteria in Solid Tumours (RECIST), and single timepoint radiomic models. All models were externally validated.

Material & methods

Patients

Two independent patient cohorts were retrospectively collected at the University of Washington/Seattle Cancer Care Alliance, referred to as “training cohort”, and the Technical University of Munich, referred to as “testing cohort”. The inclusion criteria were patients with STS of the extremities and trunk treated with neoadjuvant RT with or without CTx followed by tumor resection in curative intent. Exclusion criteria encompassed definitive, palliative, or postoperative RT, brachytherapy, other tumor locations, early abortion of RT (cut-off at 80% of planned total dose), osteosarcomas, Ewing sarcomas, rhabdomyosarcomas, endoprosthetic-dependent artifacts, missing pre- or post-therapeutic MRI scans, and incongruent image plane orientations between pre- and post-therapeutic MRI scans (see Supplemental Fig. 1 for a patient workflow). pCR values were obtained from all patients. If the information was missing, the surgical specimen was reassessed by board-certified pathologists at each institution (EC and KS) [33]. pCR was defined as less than 5% viable cells in the surgical specimen. The overall survival (OS) was calculated from initial pathologic diagnosis to the time point of death or the time point of censoring. Approval from the ethic committees was received (reference number 466/16 s). Informed consent was given before therapy. Data reporting follows the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations (Supplemental Material, Appendix) [34].

Image acquisition and segmentation

Each patient received pre-RT and post-RT MRI scans. Patients that received CTx (neoadjuvant in all cases) had an additional MRI before CTx administration. Table S1 describes acquisition parameters and scan planes. Tumor segmentation was performed manually (authors: MBS, JCP, VA; Supplemental Methods). To compensate for operator-dependent bias, multiple delineations were performed for 20 randomly selected patients on the pre-therapeutic MRI by three operators (authors: RA, MBS, JCP) in the training cohort (see Fig. 1). Dice similarity coefficients (DSC) were computed using 3D Slicer (DiceComputation module) [35].

Image preprocessing and radiomic feature extraction

See Supplemental Material for a detailed description of image preprocessing and radiomic feature extraction. In brief, preprocessing included bias field correction, intensity normalization, and isotropic resampling. 105 features per MR sequence were extracted using pyradiomics following the recommendations of the Imaging Biomarker Standardization Initiative (IBSI) [36,37]. Radiomic features included first-order, shape, and texture features (see Table S2 for a detailed listing). After CombatHarmonization [38-42], delta-radiomic features were calculated by absolute subtraction of post-therapeutic feature values from pre-therapeutic feature values \(x_{pre} - x_{post}\). For single timepoint models, radiomic features from pre- and post-RT MRIs were directly used as model input. For an exploratory analysis, radiomic feature were also extracted from post-Ctx MRIs if available.

Feature reduction

All feature reduction steps were performed using the training cohort. First, all features susceptible to variations in segmentation performed in the subset of patients with multiple segmentations were excluded. As a threshold, an intraclass correlation coefficient (ICC) (3,1) of 0.8 was used. The remaining features (T2FS: 72, T1FSGd: 103) were then used as input for the modeling pipeline including additional feature reduction and model training (Fig. 1). The feature reduction procedure was performed using 1000 bootstrap samples. For each bootstrap sample, two feature selection stages were applied. First, highly intercorrelated features defined by a Spearman correlation coefficient of greater than 0.8 were excluded. For identified feature pairs, the feature with the highest mean correlation to all remaining features was excluded. Second, the Boruta algorithm was applied to filter the most relevant features [43,44]. The features were ranked according to the frequency of their selection in the 1000 bootstrap runs. The final feature set was defined as the \(n\) top-ranking features with \(n\) defined as the median feature number selected over all bootstrap runs.

Modeling strategy

For ML model comparison and unbiased performance evaluation on the training set, 100 iterations of three-fold nested cross-validation were performed based on the code by Deist et al. built upon the “caret” package [45,46]. Three common ML techniques were compared to predict pCR: random forest (RF), elastic net regression (ENR), and LogitBoost [47-49]. After Synthetic Minority Oversampling Technique (SMOTE) for imbalance correction (see Supplemental Methods), hyperparameters were optimized using...
grid search as part of the inner folds. See Table S3 for hyperparameter tuning spaces. The selected hyperparameters were then used for testing on the five outer folds. The mean receiver operating characteristic (ROC) curve (AUC) over all outer folds was calculated for model comparison. The hyperparameter combination with the best mean performance was used to retrain a final model on the whole training set. The final models were externally validated on the testing cohort. 95% confidence intervals (95% CI) were estimated using 1000-fold bootstrapping.

For all developed models the complete pipeline was applied separately. In total, three different delta-radiomic models were developed: “Delta-T1FSGd” based on T1FSGd, “Delta-T2FS” based on T2FS, and “Delta-combined” combining both feature sets. For comparison, single timepoint radiomic models were developed using the pre-therapeutic MRI scans: “Pre-T1FSGd”, “Pre-T2FS”, and “Pre-combined”, or the post-therapeutic MRI scans: “Post-T1FSGd”, “Post-T2FS”, and “Post-combined”. Moreover, the value of the AJCC staging system (8th edition) “AJCC”, RECIST 1.1 (Supplemental Methods) [50], the tumor volume change “Delta-Volume”, the pre-therapeutic tumor volume “Pre-Volume”, and the post-therapeutic tumor volume “Post-Volume” were assessed. For quality assurance, the delta-radiomic models were retrained with permuted radiomic features. To assess the value of all models using a mixed cohort 100 iterations of three-fold nested cross-validation were applied as described above.

**Exploratory analysis of survival and transferability to chemotherapy**

Multivariate Cox proportional hazards regression was used to assess association with OS. The concordance index (C-index) was used to evaluate prognostic performance. Decision curve analysis was calculated for the multivariate models for OS at 3 years (see Supplemental Methods and caption of Fig. 3) [51,52].

We sought to evaluate if our developed models would function using delta-radiomic features calculated from MRIs obtained before and after neoadjuvant Ctx, too (see Supplemental Fig. S2 for an overview of MRI timepoints). As in the testing cohort no patient received Ctx, our models were tested on the subset of the training cohort that received Ctx. Chemotherapy-dependent delta-radiomic features we calculated by subtraction of radiomic feature calculated on the pre-therapeutic MRI and on the post-Ctx “interim” MRI (\(x_{\text{preCtx}} - x_{\text{postCtx}}\)).

**Statistical analysis**

Modeling and statistical analysis were performed using R (version 3.4.0, R core team, Vienna, Austria). Table S4 displays all R packages used. See Supplemental Methods for a description of outcome measures, feature importance, and calibration analysis. Delta-radiomic and multivariate models can be obtained from https://github.com/jacapan/Delta-Radiomics. Individual patient data may be available on reasonable request dependent on local ethics committee voting and in compliance with data protection rights.

**Results**

Patient demographics (training: 103, testing: 53 patients), staging groups, and RT doses were similar between the cohorts (Table 1). The frequency of neoadjuvant Ctx in addition to RT significantly differed between the training cohort 55% (\(n = 56\)) and the testing cohort 0% (\(p < 0.001\)). When Ctx was administered, it was always preceded RT and 91% (\(n = 51\)) of patients received an AIM (anthracycline + ifosfamide + mesna)-based regimen (Table S5). Ctx was a significant prognostic factor for OS (\(p = 0.03\)) in the training cohort. There was a trend towards an

![Fig. 1. The Delta-Radiomics Workflow. Abbreviations: Delta: delta-radiomics models, ENR: elastic net regression, ICC: intra-class correlation coefficient, pCR: pathological complete response, Pre: pre-radio(chemo)therapy models, Post: post-radio(chemo)therapy models, T1FSGd: contrast-enhanced T1-weighted fat saturated, T2FS: T2-weighted fat saturated.](image-url)
uneven histology distribution ($p = 0.087$, Table S6). pCR was achieved in 11% ($n = 11$) and 8% ($n = 5$) of patients in the training and testing cohort, respectively (predominantly pleomorphic sarcoma and myxoid liposarcoma, see Table S7). The similarity between multiple VOI delineations was rated with a DSC of 0.91 (standard deviation 0.035).

In nested cross-validation, the RF models achieved the best mean performance with a mean AUC of 0.73 (AUC ENR: 0.62, AUC LB 0.67) (Table S8a). When ranking the ML models for each fold by their AUC value, RF, ENR, and LB achieved mean ranks of 1.4, 2.55, and 2.05, respectively (Table S8b). Thus, the RF algorithm was chosen to train the final models.

In the training set, the predictive performances in AUC were comparable for delta-radiomic models, RECIST, and Delta-Volume (AUCs: 0.74–0.80) (Fig. 2: AUC-values, Fig. 3: ROC-curves, Supplemental Fig. S3: Calibration-curves). In the testing cohort, Delta-T1FSGd and Delta-combined showed better performances that were more similar to the training cohort performances than for Delta-T2FS with AUC values of 0.70 (95% CI: 0.43–0.92, AUC difference: −0.05), 0.75 (95% CI: 0.56–0.93, AUC difference: −0.05), and 0.85 (95% CI: 0.33–0.92, AUC difference: −0.09), respectively. RECIST (AUC 0.6, 95% CI: 0.33–0.86), Delta-Volume (AUC 0.43, 95% CI: 0.06–0.82), and AJCC (AUC 0.51, 95% CI: 0.25–0.75) achieved predictive performances closer to random. After permutation of radiomic features no predictive performance above random was achieved in both cohorts (AUC-training: 0.38, 0.52, 0.59; AUC-testing: 0.45, 0.48, 0.47 for Delta-T1FSGd, Delta-T2FS, and Delta-combined, respectively).

Next, we evaluated the predictive performance of single timepoint models (Table 3: AUC-values, Supplemental Fig. S4: ROC curves). The T1FSGd-based models achieved better predictive performances than the T2FS-based models in both single timepoints in the training set. In the external validation cohort, however, the only single timepoint models with relevant predictive performance were Post-combined and Pre-Volume with AUC-values of 0.68 (95% CI: 0.44–0.91, AUC difference: −0.06) and 0.70 (95% CI: 0.48–0.9, AUC difference: +0.08).

Due to the imbalanced dataset, accuracy did not produce reliable metrics with high values for non-discriminative predictors. The best performances were achieved by Delta-T2FS (Matthew’s correlation coefficient 0.48, balanced accuracy 0.69, F1-score: 0.5) (Table S9 displays all remaining models). Due to the imbalance

### Table 1

<table>
<thead>
<tr>
<th>Institution</th>
<th>Training (UW)</th>
<th>Testing (TUM)</th>
<th>p-value$^1$</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accrual time</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Patients</strong></td>
<td>2008–2017</td>
<td>2010–2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Patients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1FSGd sequence</td>
<td>102 p (98%)</td>
<td>53 p (90%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2FS sequence</td>
<td>98 p (96%)</td>
<td>49 p (83%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both sequences</td>
<td>96 p (94%)</td>
<td>43 p (73%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recurrent</strong></td>
<td>0 p</td>
<td>5 p (8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>m 54 (19–86)</td>
<td>m 57 (22–87)</td>
<td>0.331</td>
<td>1</td>
</tr>
<tr>
<td>male</td>
<td>38 (37%)</td>
<td>24 p (41%)</td>
<td>0.737</td>
<td>1</td>
</tr>
<tr>
<td><strong>T-stage</strong>$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6 p (6%)</td>
<td>1 p (2%)</td>
<td>0.431</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>35 p (34%)</td>
<td>24 p (41%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>35 p (34%)</td>
<td>23 p (39%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>26 p (25%)</td>
<td>11 p (19%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N-stage</strong>$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>102 p (100%)</td>
<td>58 p (98%)</td>
<td>0.366</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0 p (0%)</td>
<td>1 p (2%)</td>
<td>0.074</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>22 p (22%)</td>
<td>5 p (8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>39 p (38%)</td>
<td>23 p (39%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>41 p (40%)</td>
<td>31 p (57%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AJCC-Stage</strong>$^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA</td>
<td>2 p (2%)</td>
<td>0 p (0%)</td>
<td>0.16</td>
<td>1</td>
</tr>
<tr>
<td>IB</td>
<td>20 p (20%)</td>
<td>5 p (8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>4 p (4%)</td>
<td>1 p (2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIIA</td>
<td>28 p (27%)</td>
<td>22 p (37%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIIB</td>
<td>45 p (44%)</td>
<td>31 p (53%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>3 p (3%)</td>
<td>0 p (0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper Extremities</td>
<td>17 p (17%)</td>
<td>7 p (12%)</td>
<td>0.169</td>
<td>1</td>
</tr>
<tr>
<td>Lower Extremities</td>
<td>75 p (74%)</td>
<td>51 p (87%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trunk</td>
<td>10 p (10%)</td>
<td>1 p (2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prognosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Median OS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Therapy information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Margin-status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>positive</td>
<td>20 p (20%)</td>
<td>7 p (12%)</td>
<td>0.026</td>
<td>0.38</td>
</tr>
<tr>
<td>negative</td>
<td>81 p (80%)</td>
<td>48 p (81%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Dose</td>
<td>m 50 Gy (42–60 Gy)</td>
<td>m 50 Gy (50–56 Gy)</td>
<td>0.303</td>
<td>1</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>56 p (55%)</td>
<td>0 p (0%)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Viable cells after neoadjuvant therapy</td>
<td>m 40% (0–100%)</td>
<td>m 30% (0–100%)</td>
<td>0.150</td>
<td>1</td>
</tr>
<tr>
<td>pCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>positive</td>
<td>11 (11%)</td>
<td>5 (8%)</td>
<td>0.787</td>
<td>1</td>
</tr>
<tr>
<td>negative</td>
<td>91 (89%)</td>
<td>54 (92%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** *: p-value < 0.05, pCR: pathological complete response, AJCC: American Joint Committee on Cancer staging system, m: median, p: patients, r: range, RT: radiation therapy.

1 Wilcoxon rank-sum test for continuous and ordinal variables, Fisher’s exact test for nominal variables, log-rank test for comparison of survival times. Corrected for multiple testing by Bonferroni correction ("p-value adjusted").

2 Following AJCC staging manual 8th edition [61].

3 According to the French Federation of Cancer Centers Sarcoma Group (FNCLCC).
and small test set, calibration curves were “weak” for all models (Supplemental Fig. S3). The Hosmer-Lemeshow test was significant for all models reflecting the poor calibration of all models [53].

The final feature numbers used as input for modeling were 16, 15, and 31 features for Delta-T1FSGd, Delta-T2FS, and Delta-combined. Feature importance was assessed for the Delta-radiomics models (Table S10). For Delta-T1FSGd, the three most important features included two texture features “Busyness” and “SizeZoneNonUniformityNormalized”, as well as the shape feature “Flatness”. Delta-T2FS was also dominated by “Flatness” and two texture features “SizeZoneNonUniformity” and “ZoneEntropy”. In the combined model, the important features from the single modality models also scored highest.

In an exploratory analysis we tested the association of the delta-radiomic models, Pre-Volume, and pCR with OS in the combined patient cohort corrected for AJCC and age as known prognostic variables. All delta-radiomic models, but not pCR and Pre-Volume, were significantly associated with OS in multivariate Cox regression (Table S11). In the training set, Delta-T2FS, Delta-combined, and pCR were significantly associated with OS, whereas in the test set none of the predictors, including age and AJCC, were associated with OS. The C-index of the clinical baseline model (AJCC+ age) of 0.68 in the training set could be improved by adding pCR (0.73), and Delta-T2FS (0.69) (Table S12). In the test set, however, only Delta-combined (0.69) and Delta-T1FSGd (0.71) improved the C-index of the clinical model (0.68). Likewise, both models achieved a net clinical benefit above the clinical model and the pCR-based multivariate model in decision curve analysis in the test set (Fig. 4). Exemplary patient cases are displayed in Fig. 5.

To determine the transferability to neoadjuvant chemotherapy delta-radiomics, we tested our delta-radiomic models using delta-radiomic features derived from MRIs obtained before and after neoadjuvant CTx on the subset of the training cohort that received CTx. Delta-T1FSGd, Delta-T2FS, and Delta-combined achieved AUC values of 0.90 (0.80–0.98), 0.88 (0.69–0.99), and 0.91 (0.74–1.00), respectively.

Because the two cohorts were found to have different patient characteristics, especially in use of chemotherapy and tumor histology, we performed a secondary analysis for all models using a mixed patient cohort. After 100 iterations of 3-fold nested cross-validation on the complete mixed cohort, the delta-radiomic models achieved the best performances (Delta-combined: AUC 0.79 (95% CI: 0.78–0.80), Delta-T2FS AUC 0.73 (95% CI: 0.72–0.75), Delta-T1FSGd 0.73 (95% CI: 0.72–0.74) compared to volume-based metrics (RECIST AUC 0.64 (95% CI: 0.63–0.66), Delta-Volume AUC 0.62 (95% CI: 0.6–0.64), Pre-Volume (0.66 (95% CI: 0.64–0.67)) or single time point radiomic models (best model: Post-combined AUC 0.64 (95% CI: 0.63–0.65)) (Supplemental Fig. S5). The delta-radiomics model performance was also comparable to the primary analysis (maximal testing AUC 0.75).

Discussion

We could demonstrate delta-radiomics-based response assessment in patients receiving neoadjuvant therapy. Delta-radiomic models achieved better reproducibility than single timepoint radiomic models, RECIST, or the peri-therapeutic volume change. The combined delta-radiomic model and T2FS-based model achieved the best predictive performance in terms of AUC. The best classification performance in terms of balanced accuracy was achieved by the T2FS-based model. Pre-therapeutic tumor volume was the best single timepoint predictor. While association with pCR was the primary hypothesis, the delta-radiomic models were significantly associated with OS in multivariate Cox regression models in a combined cohort.

While there is a growing body of evidence suggesting association between pre-treatment radiomics and outcomes in STS, data on delta-radiomics remains scarce. In a small 30-patient study, delta-radiomic analysis of diffusion-weighted MR imaging improved prediction of the treatment effect score (response threshold 50%) in internal cross validation [54]. Crombé et al. analyzed the value of T2-weighted sequence-based delta-radiomics for pCR predictions in STS patients after neoadjuvant CTx in a monocentric cohort of 65 patients [55]. The final model achieved an AUC of 0.63 within a 15-patient holdout set. Lin et al. conducted a similar study in osteosarcoma patients using CT imaging. A performance with an AUC of 0.82 in an internal testing cohort was demonstrated [56]. Our delta-radiomic models had comparable performances with AUC-values ranging from 0.68 to 0.75. Since there were significant differences in histologies, imaging modalities, distribution of outcome variables, and treatment regimens, the performances within the above mentioned studies cannot be directly compared. One advantage of our study is that an independent test cohort was used providing a TRIPOD type III validation [34].

TRIPOD type III validation was achieved through the use of independent training and testing cohorts for our primary analysis [34]. However, as both cohorts were found to have significant differences in clinical features, histologies, and treatment regimens (e.g. chemotherapy), we performed a post-hoc secondary analysis with a mixed cohort and 3-fold nested cross validation to evaluate the effects of the cohort construction. Again, the delta-radiomic models outperformed volume-based metrics such as RECIST or single timepoint radiomic-models. Even though differences in the val-
idation method mean the results cannot be directly compared, the performances in the mixed cohort were similar compared to that achieved on the primary training set.

In this study we used pCR as a surrogate marker for patient outcome instead of directly predicting OS, due to the small patient cohorts. pCR is often used in prospective trials as a surrogate marker reducing the need for long follow up and large patient numbers [9]. Ultimately, a response marker needs to prove its usefulness for patients’ survival. We addressed this fact by performing exploratory multivariate Cox regression analyses. Interestingly, we could demonstrate an association of the delta-radiomic models with OS in the combined and training cohorts. pCR itself was only significant in the training cohort. With its low patient and event number, none of the prognostic factors, including age and AJCC, were significant in the test cohort. In the literature, multiple retrospective studies showed contradicting results regarding the prognostic value of pCR [57,58]. The previously mentioned meta-analysis encompassing 1663 predominantly retrospectively assessed patients demonstrated significant association of OS independent of the treatment modality [10]. Regarding the small patient numbers, we therefore see pCR as a valid endpoint for this exploratory work. Future larger studies should assess direct predictions of OS.

In many ways, delta-radiomics is a concept similar to assessing treatment response via changes in tumor volume by the RECIST criteria [50]. In contrast to the delta radiomic models, RECIST, Delta-Volume, and Post-Volume did not show similar predictive performances in the test set. Similar observations have been previously described [11,32]. Pre-Volume, however, appeared to be a stable predictor for pCR. The finding that radiomic features had better predictive performance than volume alone is of interest because radiomic features have been criticized as overly correlated to the VOI volume [59]. Delta-radiomic features may capture radiation-induced biologic changes occurring within the STS such as infarction, necrosis, fibrosis, or hyalinization [60].

In other cancer types, pCR plays an increasing role in treatment personalization. In breast cancer, certain pre-stratified patients without pCR after neoadjuvant Ctx receive additional adjuvant Ctx [8]. Future directions may harness imaging-based response assessment prior to surgery to identify those patients with STS who may benefit from additional neoadjuvant therapeutic escalation.

We also evaluated the transferability of the developed models to peri-CT delta-radiomic features. A high predictive performance could be observed. These results may be overly optimistic as all patients receiving CTx were in the training set. Still, this exploratory analysis suggests that the transfer of radiomic models to different cytotoxic therapies should be investigated.

Pathological response in clinical practice is evaluated on representative samples of the STS. This procedure constitutes a compromise between accurate response estimation and time expenditure [33]. As consequence, the determination of viable cells may be prone to a certain sampling error. By using a binarized endpoint, this risk may be reduced as minor numerical deviations far from the cut-point do not affect the result. Still, this uncertainty in the endpoint constitutes a limitation of this approach. Moreover, other pathological measures such as hyalinization were not available for

**Fig. 3.** Receiver operating characteristic curves (ROC) of the delta-radiomic models in the testing patient cohort. The shaded area represents the 95% confidence interval. The circles represent the cut-points applied for classification (median). Abbreviations: AJCC: American Joint Committee on Cancer staging system 8th edition, AUC: area under the ROC curve, Delta: delta-radiomics models, RECIST: Response Evaluation Criteria in Solid Tumors, T1FSGd: contrast-enhanced T1-weighted fat saturated, T2FS: T2-weighted fat saturated.
all patients among both institutions, but constitute interesting alternative response markers [60].

This analysis bears several additional limitations. First, the retrospective nature of this study may be a reason for potential bias in data selection [61]. Second, the heterogeneity in treatment regimens impaired therapy-specific modeling potentially reducing performance. Third, patient numbers were overall low, especially in the test set. Large standard deviations made direct comparisons between models difficult and impeded interpretability of multivariate models, especially in the context of the imbalanced outcome measure. For instance, in the test set none of the known prognostic factors was significantly associated with OS. Still, this study remains the largest reported delta-radiomic analysis in STS patients. Fourth, technical variations in image acquisition between cohorts and between the two timepoints used for delta-radiomic feature calculation may have hindered better performances and generalizability. Finally, the optimal approach to calculate deltarradiomic features is unknown. We decided for absolute subtraction of feature values as performed in previous studies, but the assessment of the relative change may be an alternative method. A prospective trial, with pre-defined acquisition protocols optimized for feature reproducibility, may enable better response evaluation.

To conclude, we could demonstrate prediction of pathological complete response using the delta-radiomics principle in STS patients.

Fig. 4. Decision curve analysis of delta-radiomics multivariate models. Decision curve analysis was performed comparing the net benefit of the delta-radiomic multivariate models with the clinical model (AJCC+age) and the pCR multivariate model in the testing set. 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-positive and false-negative result [52]. The threshold probability was calculated for death after three years. The two extreme strategies “treat all” and “treat none” are displayed as a reference. A decision model shows a clinical benefit if the decision curve shows larger net benefit than both reference strategies. In the panel right panel in the second row “Clinical” and “pCR_Clinical” are overlapping. Abbreviations: AJCC: American Joint Committee on Cancer staging system 8th edition, AUC: area under the ROC curve, Delta: delta-radiomics models, pCR: pathological complete response, T1FSGd: contrast-enhanced T1-weighted fat saturated, T2FS: T2-weighted fat saturated.
patients. Delta-radiomic models achieved better reproducibility than single timepoint radiomic models, RECIST, or the peri-therapeutic volume change and was associated with overall survival. The models also functioned in patients using MRIs obtained before and after chemotherapy. We conclude, that delta-radiomic features may capture radiation-induced biological changes and may function as a treatment response biomarker.

Funding

This work was funded by an ESTRO mobility grant and a physician-scientist program of the Helmholtz Zentrum München (JCP).

Declaration of Competing Interest

DH received research funding from GE Healthcare, Philips Healthcare, Canon Medical Systems USA, and Siemens Healthineers. PL reports, within the submitted work, grants/sponsored research agreements from radiomics SA. PL has minority shares in the company radiomics SA. He is co-inventor of two issued patents with royalties on radiomics (PCT/2014/050248, PCT/2014/050728) licensed to Radiomics SA, one non-patented invention (softwares) licensed to Radiomics SA and two non-issues, non-licensed patents on Deep Learning-Radiomics (n° 17123818, PCT/2020/050794). PL confirms that none of the above entities or funding was involved in the preparation of this paper. HW has minority shares in the company radiomics SA. The other authors declare no potential conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2021.08.023.

References

Delta-radiomics in soft-tissue sarcomas


82