A CT-based radiomics classification model for the prediction of histological type and tumour grade in retroperitoneal sarcoma (RADSARC-R): a retrospective multicohort analysis

Amani Arthur, Matthew R Orton*, Robby Emsley, Sharon Vit, Christian Kelly-Morland, Dirk Strauss, Jason Lunn, Simon Doran, Hafida Lmalem, Axelle Nzokirantevye, Saskia Litiere, Sylvie Bonvalot, Rick Haas, Alessandro Gronchi, Dirk Van Gestel, Anne Ducassou, Chandrajit P Raut, Pierre Meeus, Mateusz Spalek, Matthew Hatton, Cecile Le Pechoux, Khin Thway, Cyril Fisher, Robin Jones, Paul H Huang†, Christina Messiou†*

Summary

Background Retroperitoneal sarcomas are tumours with a poor prognosis. Upfront characterisation of the tumour is difficult, and under-grading is common. Radiomics has the potential to non-invasively characterise the so-called radiological phenotype of tumours. We aimed to develop and independently validate a CT-based radiomics classification model for the prediction of histological type and grade in retroperitoneal leiomyosarcoma and liposarcoma.

Methods A retrospective discovery cohort was collated at our centre (Royal Marsden Hospital, London, UK) and an independent validation cohort comprising patients recruited in the phase 3 STRASS study of neoadjuvant radiotherapy in retroperitoneal sarcoma. Patients aged older than 18 years with confirmed primary leiomyosarcoma or liposarcoma proceeding to surgical resection with available contrast-enhanced CT scans were included. Using the discovery dataset, a CT-based radiomics workflow was developed, including manual delineation, sub-segmentation, feature extraction, and predictive model building. Separate probabilistic classifiers for the prediction of histological type and low versus intermediate or high grade tumour types were built and tested. Independent validation was then performed. The primary objective of the study was to develop radiomic classification models for the prediction of retroperitoneal leiomyosarcoma and liposarcoma type and histological grade.

Findings 170 patients recruited between Oct 30, 2016, and Dec 23, 2020, were eligible in the discovery cohort and 89 patients recruited between Jan 18, 2012, and April 10, 2017, were eligible in the validation cohort. In the discovery cohort, the median age was 63 years (range 27–89), with 83 (49%) female and 87 (51%) male patients. In the validation cohort, median age was 59 years (range 33–77), with 46 (52%) female and 43 (48%) male patients. The highest performing model for the prediction of histological type had an area under the receiver operator curve (AUROC) of 0·928 on validation, based on a feature set of radiomics and approximate radiomic volume fraction. The highest performing model for the prediction of histological grade had an AUROC of 0·882 on validation, based on a radiomics feature set.

Interpretation Our validated radiomics model can predict the histological type and grade of retroperitoneal sarcomas with excellent performance. This could have important implications for improving diagnosis and risk stratification in retroperitoneal sarcomas.

Funding Wellcome Trust, European Organisation for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group, the National Institutes for Health, and the National Institute for Health and Care Research Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and The Institute of Cancer Research.

Lancet Oncol 2023; 24: 1277–86

*Contributed equally as co-first authors

†Contributed equally as co-senior authors

The Institute of Cancer Research, London, UK (A Arthur MRCPCH, M R Orton PhD, R Emsley PgC, S Vit MSc, J Lunn MSci, S Doran PhD, K Thway MD, Prof R Jones MRCP, P H Huang PhD, Prof C Messiou MD); **The Royal Marsden NHS Foundation Trust, London, UK** (C Kelly-Morland FRCR, D Strauss FRCS, K Thway, Prof R Jones, Prof C Messiou); **The European Organisation for Research and Treatment of Cancer, Brussels, Belgium** (H Lmalem MSc, A Nzokirantevye MSc, S Litiere PhD); **Institut Curie, Hopital de Paris, Paris, France** (Prof S Bonvalot PhD); **The Netherlands Cancer Institute (Antoni Van Leeuwenhoekziekenhuis), Amsterdam, Netherlands** (Prof R Haas PhD); **Department of Surgery, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy** (A Gronchi FSSO); **Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium** (D Van Gestel PhD); **Centre Hospitalier Universitaire de Toulouse, Toulouse, France** (A Ducassou MD); **Institut Claudius Regaud, Toulouse, France** (A Ducassou); **Institut Universitaire du Cancer de Toulouse Oncopole, Toulouse, France** (A Ducassou); **Brigham and Women's Hospital, Boston, MA, USA** (Prof C P Raut MD); **Dana-Farber Cancer Institute, Boston, MA, USA** (Prof C P Raut); **Harvard Medical School, Boston, MA, USA** (Prof C P Raut); **Centre Leon Berard, Lyon, France** (P Meeus MD); **Maria Sklodowska-Curie National**

Introduction

Retroperitoneal sarcomas are large and complex tumours that account for 12–15% of all soft tissue sarcomas and their prognosis is poorer than that of extremity sarcomas.[1,2] Liposarcoma and leiomyosarcoma are the most common retroperitoneal sarcoma histologies. Compared with superficial lesions, retroperitoneal sarcomas are more challenging to obtain a biopsy sample for and are prone to sampling bias.[3] Clinical trials have explored neoadjuvant therapy, including the phase 3 STRASS (European Organisation for Research and Treatment of Cancer 62092) trial, which assessed neoadjuvant radiotherapy in retroperitoneal sarcoma.[4]

Radiomics is used in oncological imaging to extract and mine variables from medical images and non-invasively quantify the global radiological phenotype of tumours.[5–7] However, successful clinical translation is elusive.[8,9] Progress has been hindered by the limited generalisability of data, variations in methods, and absence of independent validation cohorts.[10,11]

Research Institute of Oncology, Warsaw, Poland (M Spalek PhD); **Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK** (M Hatton FRCR); **Gustave Roussy, Villejuif, France** (C Le Pechoux MD); **University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (Prof C Fisher DSc)

Correspondence to: Prof Christina Messiou, The Royal Marsden NHS Foundation Trust, London SW3 6JJ, UK
christina.messiou@rmh.nhs.uk

Research in context

Evidence before this study

Radiomics has become an important concept in oncological imaging to quantify the characteristics of a tumour in a non-invasive and global manner. We searched PubMed for articles in English from Oct 20, 2021, to Jan 5, 2023, using the following search terms: "retroperitoneal sarcoma", "soft tissue sarcoma", "radiomics in soft tissue sarcoma", and "imaging in retroperitoneal sarcoma". All radiomics studies were examined, with a focus on independently validated studies. Some retrospective radiomics studies supported the use of radiomics in the prediction of histological grade for soft tissue sarcoma. However, patients with retroperitoneal sarcoma were under-represented, and most studies did not have independent validation. The two largest validated studies had area under receiver operator curve (AUROC) values of 0·75 and 0·8 for the prediction of grade. Few patients with retroperitoneal sarcoma were included in these studies and models were MRI-derived (standard-of-care imaging for retroperitoneal sarcoma is CT), making results challenging to extrapolate. No radiomics studies to date have approached the prediction of histological type.

Added value of this study

To the best of our knowledge, RADSARC-R is the largest cohort analysed by radiomics for patients with retroperitoneal soft tissue sarcoma and the only one validated in an external independent cohort. The study incorporates repeatability analysis and a novel artificial intelligence pipeline for radiomic feature selection to identify robust and more interpretable features to maximise direct clinical applicability. The model produced good results in the prediction of the two most common histological types (liposarcoma and leiomyosarcoma) with an AUROC of 0·928 and the differentiation between low grade and intermediate or high grade tumours with an AUROC of 0·882. The use of multi-centre heterogeneous data, particularly in the validation cohort, aligns this study with real-world data and increases the reliability and potential generalisability of our models.

Implications of all the available evidence

Patients with retroperitoneal sarcoma continue to have a poor prognosis, poorer than that of sarcoma of other anatomical sites. The need for neoadjuvant therapy is currently guided by doing an invasive biopsy that can lead to the under-grading of tumours in up to 68% of patients. The clinical tools available for the upfront determination of histological type and grade urgently need to be improved to allow for improved risk stratification and management. Our radiomics model to predict histological subtype and grade presents a conceptual advance and an opportunity for a change in personalised care for patients with retroperitoneal sarcoma.

Although more than 50 soft tissue sarcoma radiomics studies have been completed, few include retroperitoneal sarcomas, and the majority use single-centre datasets without independent validation.[5] The limited interpretation of the quantitative radiological phenotype in retroperitoneal sarcomas and its association with tumour biology is a missed opportunity.

This study, radiomics in sarcoma of the retroperitoneum (RADSARC-R), aims to develop and validate a CT-based radiomics model for histological type classification and prediction of grade in retroperitoneal liposarcoma and leiomyosarcoma. We have developed a predictive model for histology and grade that was validated in an external independent cohort[4] with the intended goal of eventual clinical translation and use by health-care professionals.

Methods

Study design and participants

This was a retrospective cohort study that applied radiomics analysis retrospectively to a discovery cohort followed by the validation of findings in an independent multi-institution cohort. The discovery cohort comprised patients with primary retroperitoneal liposarcoma or leiomyosarcoma undergoing surgery at the Royal Marsden Hospital (London, UK) between Oct 30, 2016, and Dec 23, 2020, who were retrospectively identified from a prospectively maintained database. The inclusion criteria were: (1) histologically confirmed retroperitoneal liposarcoma or leiomyosarcoma; (2) older than 18 years of age; (3) primary and unifocal disease; (4) baseline venous phase contrast-enhanced CT; (5) CT scan images including the entire tumour volume without artifacts; and (6) minimum clinical dataset required for radiomic model development available. The exclusion criteria were: (1) other histological types; (2) baseline or venous phase contrast-enhanced CT scans unavailable; (3) image artifacts; and (4) missing clinical data. For patients who had received neoadjuvant therapy, baseline scans and histology were used. Patient sex was defined as per the electronic patient records and no ethnicity data were collected.

For the validation cohort, patients with primary retroperitoneal liposarcoma and leiomyosarcoma recruited as part of the STRASS trial (between Jan 18, 2012, and April 10, 2017) were included.[4] Contributing sites were The Royal Marsden Hospital, London, UK; Institut Curie, Hopital de Paris, Paris, France; Gustave Roussy, Villejuif, France; The Netherlands Cancer Institute (Antoni Van Leeuwenhoekziekenhuis), Amsterdam, The Netherlands; Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; Centre Hospitalier Universitaire de Toulouse, Toulouse, France; and Dana-Farber Cancer Institute, Boston, MA, USA. Patient and scan selection criteria matched those of the discovery cohort. Patients from the Royal Marsden Hospital in the discovery cohort

who were also enrolled into STRASS were not included in the validation cohort.

Ethical approval was obtained for this retrospective study (NHS REC16/EE/0213) and a data transfer agreement allowed the inclusion of the validation cohort from the STRASS trial. A summary of the study workflow is shown in the appendix (p 17).

Image analysis

In the discovery cohort, the CT scan images were pseudonymised and transferred to the eXtensible Neuroimaging Archive Toolkit (XNAT) platform for image curation and segmentation.[12,13] Imaging data were in the Digital Imaging and Communications in Medicine format. In the validation cohort, pseudonymised CT scan images from the European Organisation for Research and Treatment of Cancer repository were uploaded onto XNAT. CT scan variables are listed in the appendix (pp 7–8).

The tumour was manually delineated on all slices using XNAT contouring tools producing 3D regions of interest. Three independent users completed whole lesion segmentations (experienced sarcoma radiologist CK-M, clinical fellow AA, and senior research radiographer RE). Segmentations completed by RE were reviewed by AA and an independent senior sarcoma radiologist (CM).

A semiautomated sub-segmentation tool was used to obtain radiomic volume fraction (RVF) features for four sub-regions associated with low, middle, high, and very high Hounsfield units. Sub-segmentations were obtained semi-automatically using the algorithm described in the appendix (p 1), which makes use of patient-specific Hounsfield unit thresholds in conjunction with morphological operations with the aim of generating sub-segmentations that are similar to those a human user would create manually. Because of challenges in automatically identifying the visually apparent high Hounsfield unit regions in some tumours, guide regions of interest were drawn on a single slice of each contiguous high Hounsfield unit sub-region.

An alternative approach for computing the sub-region volume fractions was developed to reduce user subjectivity. The RVF feature values obtained from the algorithm were used to derive fixed Hounsfield unit thresholds, from which approximate RVF (ARVF) estimates were computed by simple image thresholding. This procedure is detailed in the appendix (pp 2–3, 14–15) and applied to the discovery dataset only to ensure that the estimates of predictive performance obtained using the validation dataset are not biased to user subjectivity. The thresholds thus derived were –50, 19, and 228 Hounsfield units (low is <–50, middle is –50 to <19, high is 19 to <228, very high is ≥228).

Radiomic features

A summary of the algorithm workflow is shown in the appendix (p 16). Radiomic features were computed using pyradiomics version 3.0.1 (Imaging Biomarker Standardisation Initiative compliant algorithm),[14] yielding 105 features from three feature groups: 14 shape, 18 first order, and 73 texture; full details are in the appendix (pp 6–7). The radiomic feature groups were combined with the RVF and ARVF features to give three feature sets as outlined in the appendix (p 9). Models were built using all three feature sets and the best performing model was used in the validation set. CT imaging protocols had variations in slice thickness and pixel spacing so all images were resampled to a 1×1×5 mm voxel size using trilinear interpolation (the most frequently acquired slice thickness was 5 mm in 72 [42%] of 170 scans; appendix pp 6–7). Because of the non-isotropic voxel size, the texture features were obtained using a 2D neighbourhood, and the image values were quantised in steps of 25 Hounsfield units before texture feature computation. The features were augmented with the RVF and ARVF feature groups. For each of these feature groups, the four volume fraction features were linearly dependent (since they should sum up to 1), so the middle Hounsfield units volume feature was removed, leaving the other three volume features as a linearly independent set. A subset of radiomic features were log-transformed as described in the appendix (p 3).

Feature reproducibility was assessed by repeat segmentation of 30 scans in the discovery dataset that were selected at random (20 liposarcoma and ten leiomyosarcoma). Segmentations were completed by a further independent senior research radiographer (SV) and senior sarcoma radiologist (CM) who were masked to the initial segmentations. The intraclass correlation coefficients (two-way random effects, absolute agreement, single rater, and measurement) were computed for all features, and features where this coefficient was less than 0·75 were rejected.[15]

Machine learning pipeline

Separate probabilistic binary classifiers were built using the Python sklearn toolbox (version 0.24.2) to predict the tumour type (liposarcoma *vs* leiomyosarcoma) and tumour grade (low grade [1] *vs* intermediate or high grade [>1] and low or intermediate grade [<3] *vs* high grade [3]).

A recently developed machine learning pipeline was used, which is designed to discover models that are easier to interpret[16] (appendix pp 3–5, 16). The pipeline is based on a nested cross-validation structure, where the outer cross-validation provides performance estimates (ie, area under the receiver operator curve [AUROC]), and the inner cross-validation is used for tuning variable optimisation. The pipeline was applied to the discovery set to obtain cross-validated performance estimates, and the so-called model generation sub-routine was applied to the entire discovery set to obtain a single model that was tested in the validation dataset. Further details are in the appendix (pp 3–5).[16] The pipeline included a step that removed correlated features by comparing pairwise

See **Online** for appendix

For the **eXtensible Neuroimaging Archive Toolkit platform** see www.xnat.org

feature correlations (Spearman correlation) with a threshold, and this threshold was optimised, as described in the appendix (pp 3–5).

Statistical analysis

Analyses were performed in Python version 3.6.8 for the machine learning pipeline and Python version 3.9.13 for the calibration statistics. This was a feasibility study and sample size calculations were not performed. The models were derived from a pipeline that includes variable selection, therefore conventionally computed CIs and p values would not be valid and are not available for model feature selection. Continuous variables were expressed as median and IQR, whereas categorical variables were expressed as count and percentages. The diagnostic performance of radiomics models for predicting histological type and grade was evaluated against assessment of the surgical specimen by expert sarcoma pathologists (KT, with 14 years experience, and CF, with 41 years experience) using accuracy, sensitivity, specificity, and positive and negative predictive values. The discriminative ability of the models was assessed using receiver operating characteristic (ROC) curves, from which the AUROC was computed. p values for the null hypothesis that the AUROC is equal to 0·5 (no discrimination) were computed using Wilcoxon rank statistics. ROC metrics were computed using cross-validation in the discovery dataset. An analysis of errors was performed using the validation dataset that was independent of the development dataset and not used during model development. Errors were identified as cases where the tumour type or grade identified from histological examination of the surgical specimen did not agree with the type or grade predicted by the model. Model performance was also assessed using Hosmer–Lemeshow score calibration statistics. For AUROC, p values were two-sided, and for the Hosmer–Lemeshow score, p values were one-sided and a cutoff value of 0·05 was used for statistical significance. A Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis checklist is provided in the appendix (pp 12–13).

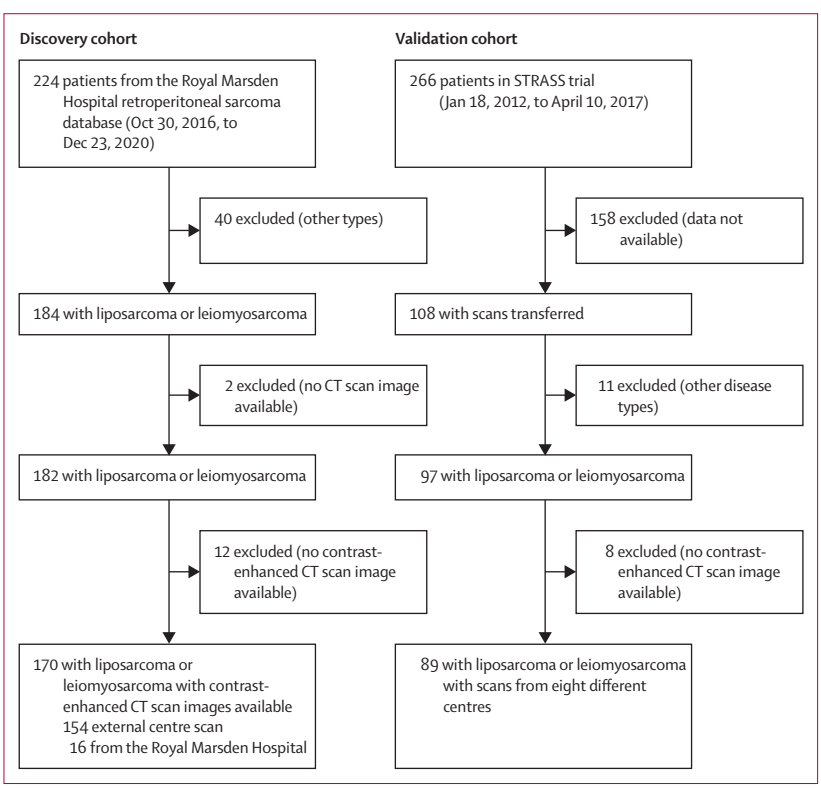

Figure 1: **Study profiles of the discovery and validation cohorts showing the reasons for exclusion and final cohort numbers**

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Patient inclusion is summarised in figure 1. 170 patients were included in the discovery cohort. The median age was 63 years (range 27–89) with 83 (49%) female patients and 87 (51%) male patients. 117 (69%) patients had liposarcoma and 53 (31%) patients had leiomyosarcoma. 38 (22%) had grade 1 tumours, 74 (44%) had grade 2 tumours, and 58 (34%) had grade 3 tumours (table 1; appendix p 18). For 59 (35%) of 170 patients in the discovery cohort, the initial reporting radiologist was unable to offer a diagnosis for histology; 86 (74%) of 117 were correctly diagnosed as liposarcoma, 23 (43%) of 53 were correctly diagnosed as leiomyosarcoma, and two (2%) of 117 patients with liposarcoma were incorrectly diagnosed as having leiomyosarcoma.

120 (71%) of the 170 patients did not have grade available from the diagnostic biopsy records as reported by two experienced sarcoma pathologists (KT and CF) at the time of diagnosis. Where available (50 [29%] of 170 patients), the comparison of core biopsy and surgical specimen grade showed that 22 (44%) of 50 were correctly graded and 28 (56%) of 50 were incorrect (27 [96%] of 28 were assigned a higher grade than originally reported [six patients upgraded from grade 1 to 2; three patients were assigned a higher grade than originally reported from grade 1 to 3; and 18 patients upgraded from grade 2 to 3] and one [4%] was assigned a lower grade than originally reported [from grade 3 to 2]), reported by KT and CF.

89 eligible patients from the STRASS trial, from eight centres, were included in the validation cohort (figure 1). The median age was 59 years (range 33–77), with 46 (52%) female patients and 43 (48%) male patients. 76 (85%) patients had liposarcoma and 13 (15%) patients had leiomyosarcoma. 33 (37%) patients had grade 1 tumours, 33 (37%) had grade 2 tumours, and seven (8%) had grade 3 tumours. 16 (18%) patients had no grade

recorded. The patients without grade recorded were included for the validation of histology; however, they were excluded from the validation of grade (table 1; appendix p 18).

The manual delineation process of CT scan images from both the discovery and validation cohorts showed some liposarcoma with visually discernible lower Hounsfield unit areas, probably corresponding to areas of abnormal fat; however, the majority (152 [89%] of 170 scans in the discovery cohort and 59 [66%] of 89 scans in the validation cohort) showed little radiographic distinction between leiomyosarcoma and liposarcoma (appendix p 19). The semiautomated sub-segmentation was also successfully applied to scans from both cohorts (appendix p 19).

For the discrimination of leiomyosarcoma from liposarcoma, the model derived from the discovery dataset showed excellent cross-validated performance for all three feature sets evaluated (radiomics, radiomics plus RVF, and radiomics plus ARVF), with AUROCs ranging between 0·912 and 0·944 (figure 2A; appendix p 10). The highest AUROC was attained using radiomics plus ARVF; for this feature set, the ARVF and texture feature group combination was consistently selected across all feature correlation thresholds, substantially more frequently than the other feature groups, such as shape and first-order statistics, and more than 95% of the cross-validation splits for the optimum threshold of 0·62 (figure 2B; appendix pp 10, 20).

The logistic regression coefficients for the final model for histology prediction are detailed in table 2. Much of the predictive power of the final model was found to come from two features with the largest coefficients, indicating that patients with larger Hounsfield unit volume fractions or with larger values of gldm_SmallDependenceLowGrayLevelEmphasis were more likely to have leiomyosarcoma. In contrast, patients with very high Hounsfield unit volume fractions were more likely to have liposarcoma, and similarly for the three other texture features, glszm_SmallAreaEmphasis, ngtdm_Strength (log), and also for gldm_

	Discovery cohort (N=170)	Validation cohort (N=89)
Median age, range	63 (27–89)	59 (33–77)
Sex		
Female	83 (49%)	46 (52%)
Male	87 (51%)	43 (48%)
Eastern Cooperative Oncology Group (discovery cohort) and WHO score (validation cohort)		
0	103 (61%)	78 (88%)
1	65 (38%)	11 (12%)
2	2 (1%)	0
Histological type		
Liposarcoma	117 (69%)	76 (85%)
Leiomyosarcoma	53 (31%)	13 (15%)
Grade		
1	38 (22%)	33 (37%)
2	74 (44%)	33 (37%)
3	58 (34%)	7 (8%)
Unknown*	0	16 (18%)
Treatment		
Neoadjuvant radiotherapy and surgery	3 (2%)	46 (52%)
Neoadjuvant chemotherapy and surgery	2 (1%)	0
Surgery only	165 (97%)	43 (48%)

Data are n (%), unless otherwise indicated. *When tumour grade was missing, these cases were excluded for the purposes of testing the predictive model for grade.

Table 1: **Full baseline clinicopathological details for the discovery and validation cohorts**

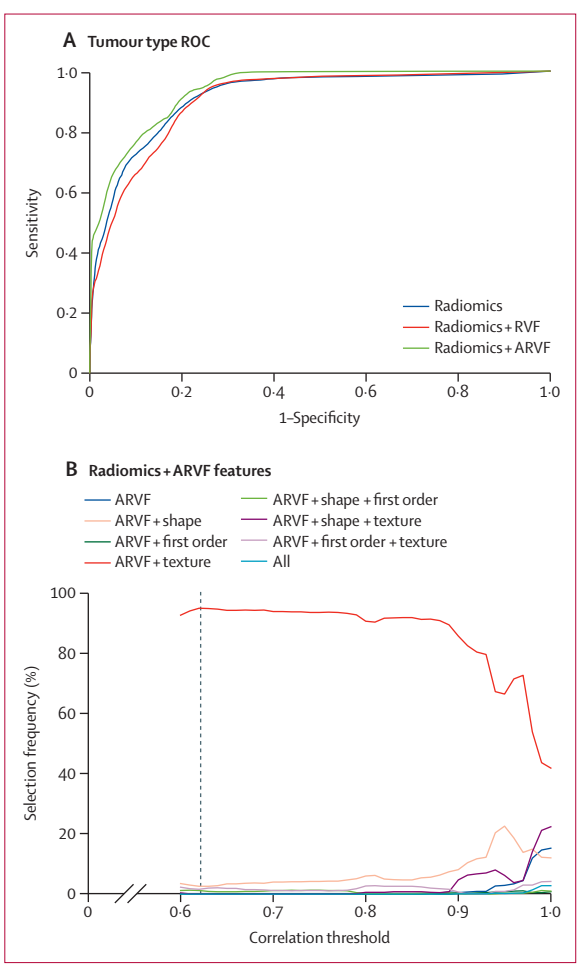

Figure 2: **The performance and feature selection for the final tumour type prediction model developed from the discovery dataset**
(A) ROC of the three feature combinations tested with radiomics (blue), radiomics plus RVF (red), and radiomics plus ARVF (green). Radiomics plus ARVF had the highest area under the ROC. (B) The selection frequency of features across cross-validation splits for the radiomics plus ARVF combination, as a function of the threshold used in the hierarchical correlation feature reduction step (0·62). ARVF=approximate radiomic volume fraction. ROC=receiver operating characteristic. RVF=radiomic volume fraction.

	Coefficient	Frequency (%)
High Hounsfield units approximate volume fraction	4·782	95·2%
gldm_SmallDependenceLowGrayLevelEmphasis	2·411	96·2%
Very high approximate Hounsfield units volume fraction	−0·662	99·7%
glszm_SmallAreaEmphasis	−0·373	96·0%
ngtdm_Strength (log)	−0·208	91·0%
gldm_DependenceNonUniformityNormalized (log)	−0·129	89·6%

Features are ordered on coefficient magnitude, and the second column gives the frequency that each feature appeared in the models obtained from the different cross-validation splits using the optimum correlation thresholds of 0·62.

***Table 2*: Features (derived from pyradiomics version 3.0.1) and logistic regression coefficients for the final histological type prediction model that was developed from the discovery dataset**

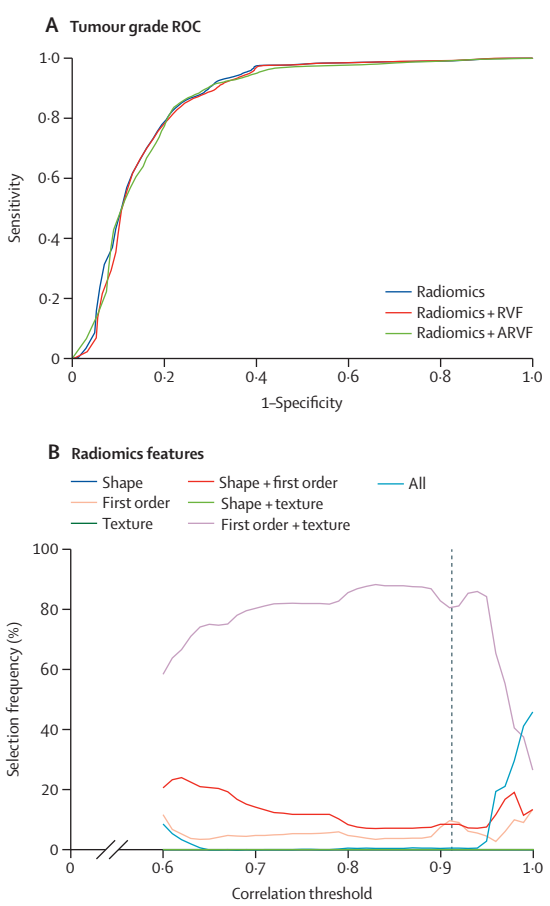

***Figure 3*: The performance and feature selection for the final tumour grade prediction model developed from the discovery dataset**

(A) ROC of the three feature combinations tested with radiomics (blue), radiomics plus RVF (red), and radiomics plus ARVF (green). Radiomics had the highest area under the ROC. (B) The selection frequency of features across cross-validation splits for the radiomics feature set, as a function of the threshold used in the hierarchical correlation feature reduction step (0·93). ARVF=approximate radiomic volume fraction. ROC=receiver operating characteristic. RVF=radiomic volume fraction.

DependenceNonUniformityNormalized (log). Four of six of these features were present in 95·2–99·7% of cross-validation splits, implying the discovery process to be stable in these data for this model.

For the prediction of low grade versus intermediate or high grade tumours, the model developed from the discovery dataset had good AUROCs in the range of 0·857–0·863 for the three feature sets (figure 3A; appendix p 10). The highest AUROC was attained in the radiomics only feature set. The feature group frequency had a selection of first-order statistics and texture features for most feature correlation thresholds (figure 3B; appendix pp 10, 21) including at the optimum threshold of 0·93. The radiomics only feature set was selected as the final model for prediction of tumour grade.

The final model for predicting histological grade included nine features from the first-order and texture groups in the final feature set (table 3). The firstorder_90Percentile feature had the largest coefficient magnitude and a selection frequency of 98·1%, suggesting that much of the predictive power of this model came from this feature with larger values associated with higher tumour grades.

In addition, we evaluated the ability of the model to discriminate between low (grade 1) or intermediate (grade 2) versus high (grade 3) tumours. The model developed from the discovery dataset had acceptable AUROCs in the range of 0·714–0·733 (appendix pp 11, 22) for the three feature sets. The radiomics plus ARVF feature set yielded the highest AUROC.

With the best performance for histology classification, the radiomics plus ARVF model was evaluated in the STRASS validation cohort. This analysis yielded an AUROC of 0·928 (p<0·0001; figure 4A; appendix p 10). Other measures of discrimination for the chosen model in this cohort were an accuracy of 0·843, a sensitivity of

	Coefficient	Frequency (%)
firstorder_90Percentile	1·811	98·1%
glszm_ZoneVariance (log)	0·412	82·3%
gldm_DependenceNonUniformityNormalized (log)	0·369	85·3%
glcm_ClusterShade (log)	−0·192	80·6%
firstorder_Kurtosis (log)	−0·141	71·8%
ngtdm_Strength (log)	−0·136	73·6%
firstorder_RootMeanSquared	−0·093	88·7%
glszm_LargeAreaHighGrayLevelEmphasis (log)	0·070	49·2%
firstorder_InterquartileRange	0·016	50·0%

Features are ordered on coefficient magnitude, and the second column gives the frequency that each feature appeared in the models obtained from the different cross-validation splits using the optimum correlation thresholds of 0·93.

***Table 3*: Features (derived from pyradiomics version 3.0.1) and logistic regression coefficients for the tumour grade prediction model that was developed from the discovery cohort**

0·923, a specificity of 0·829, a positive predictive value of 0·480, and a negative predictive value of 0·984 (calculated at the optimal Youden index threshold). The Hosmer–Lemeshow statistic for calibration was 5·34 (p=0·72), indicating a non-significant test of poor calibration. The corresponding highest performing model for the prediction of grade, the radiomics model, was selected for validation. This yielded an AUROC of 0·882 (p<0·0001; figure 4B; appendix p 10). Other measures of discrimination for the chosen model in this cohort were an accuracy of 0·823, a sensitivity of 0·800, a specificity of 0·848, a positive predictive value of 0·865, and a negative predictive value of 0·778. The Hosmer–Lemeshow statistic for calibration was 63·5 (p<0·0001), indicating a significant test of poor calibration. It was not possible to validate the radiomics plus ARVF model for discriminating low or intermediate grade tumours versus high grade tumours because of the small number of grade 3 tumours in STRASS.

Discussion

Our radiomics models successfully predicted histopathological type and grade; they differentiated liposarcoma from leiomyosarcoma with an AUROC of 0·928 and predicted grade with an AUROC of 0·882 on validation. To the best of our knowledge, this is the largest retroperitoneal sarcoma cohort analysed by radiomics and the only one validated in an independent cohort. Furthermore, we have undertaken rigorous cross-validation and used sub-segmentation and feature selection techniques to enrich for interpretable features.

Arguably, liposarcoma can display distinguishable semantic radiographic features if abnormal fat is present, indicating a well differentiated lipomatous component. However, in the absence of abnormal fat, liposarcoma and leiomyosarcoma can be indiscernible when using conventional CT; therefore, tissue sampling is needed for diagnosis and grading. Expert centres have concluded that no radiological criteria are sufficient to anticipate a specific diagnosis of a retroperitoneal sarcoma except well differentiated liposarcoma,[17] and in this study, the reporting radiologist was not able to offer a diagnosis in 35% of patients and was only able to correctly diagnose 73% of liposarcoma and 43% of leiomyosarcoma. Our model accuracy of 0·843 (ie, 84·3%) has notable potential as a tool to support differential diagnosis. Because of intratumoural heterogeneity, histopathological grade can vary greatly between tumour regions, resulting in a risk of under-grading and making suboptimal therapeutic decisions. In leiomyosarcoma, under-grading by core biopsy can be up to 68%.[3] In this study, 56% of tumours were incorrectly graded on biopsy compared with the surgical specimen. The accuracy of this model of 0·823 (ie, 82·3%) for predicting grade is therefore also promising. Radiomics offers a consistent method not influenced by radiologist interpretation, and one that might be used by non-expert users. Few radiomics studies have focused on retroperitoneal sarcoma or CT-derived radiomics in soft tissue sarcomas. Given the central role of tumour grading in patient risk stratification and treatment planning, previous radiomic studies in sarcoma have sought to evaluate the predictive capability of radiomics models for grade; however, they have generally used MRI.[18–24] The primary drawback of most of these studies is the absence of independent external validation and under-representation of retroperitoneal sarcoma, possibly because MRI is not routine practice for retroperitoneal sarcomas. Novel to the existing literature, our study addresses the absence of a radiomics model specific to retroperitoneal sarcoma for the prediction of tumour grade, and has good and consistent performance in an independent cohort with an AUROC of 0·882 on validation and accuracy of 0·823. In addition, subset analyses from the STRASS trial have suggested a possible role for neoadjuvant radiotherapy in low grade liposarcoma tumours.[4] Our study focuses on distinguishing low from intermediate and high-grade tumours, and the identification of low-grade liposarcoma might help to inform presurgical treatment planning. Furthermore, to our knowledge, for the first time, our model enables the classification of histological type in

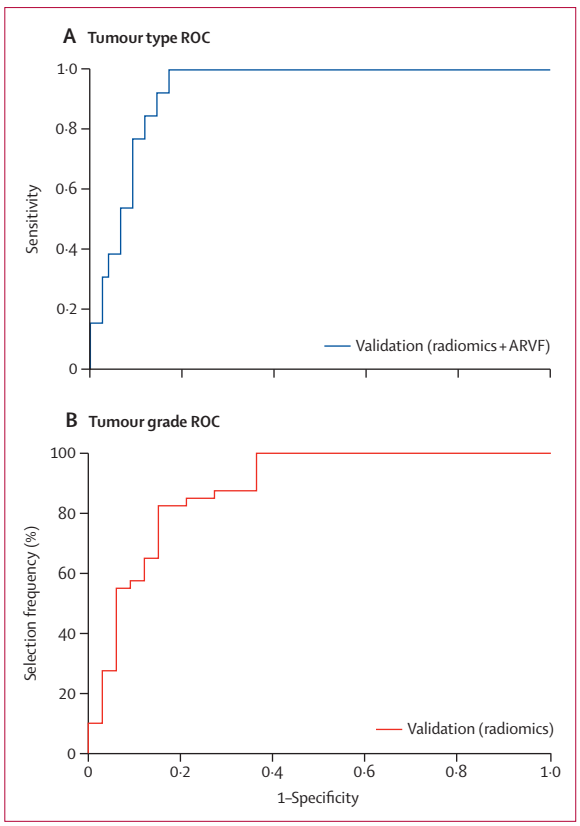

Figure 4: **ROCs for tumour type and tumour grade following independent external validation**
In the graphs, the blue line shows the validation plus ARVF model, and the red line shows radiomics alone model. ARVF=approximate radiomic volume fraction. ROC=receiver operating characteristic.

leiomyosarcoma and liposarcoma, which can frequently be challenging to discern with conventional imaging. This finding could be a useful aide to tissue sampling with unclear histological type morphology and negative ancillary tests, and to inform immediate care when tissue sampling is not possible or successful.

Most patients in the discovery cohort had baseline CT scans acquired in their local centres and protocols were therefore not standardised. It is important that radiomics models are designed to be applicable to the real-world setting that includes heterogeneous data. Our model has excellent performance using heterogeneous and real-world data. Our validation cohort was also compiled from different centres, and the similar performance observed in the discovery and validation datasets corroborates the generalisability of the radiomic feature sets selected by our models. It is also worth noting that management (surgery alone) was almost uniform in the discovery cohort despite variability in tumour grade between patients. This finding reflects the retrospective nature of the data collection, which was during a period when neoadjuvant therapy was not often an option for patients with retroperitoneal sarcoma.

The current consensus is that, although the selection of stable validated features is crucial for the development of a clinically useable radiomics model, interpretability enables the end user to maximise its potential.[25] We applied a recently developed pipeline designed to improve model interpretability compared with standard radiomics pipelines.[16] Furthermore, our novel sub-segmentation method attempted to address the often heterogeneous radiological phenotypes of retroperitoneal sarcoma seen on CT and our method using fixed thresholds for sub-region segmentation makes it more likely to be reproducible.[26]

Our model for predicting histology includes ARVF and texture features, which suggests that none of the shape and first-order features are useful in predicting histology. The ARVF features are high Hounsfield unit and very high Hounsfield unit fractions (calcification). Notably, the high Hounsfield unit fraction relationship with leiomyosarcoma and very high Hounsfield unit (calcification) relationship with liposarcoma mirrors what is currently accepted as the morphological phenotypes of these histologies. The final model for the prediction of grade is based on a radiomics-only feature set, and one that comprises first-order and textural features. This finding suggests differences in grade are attributable to lesion characteristics captured by the pixel Hounsfield unit values or image texture, or both, and, importantly, characteristics that are harder to distinguish via radiographic features (eg, shape). The 90Percentile feature relates to the brightness of the brightest 10% of pixels and higher tumour grades are associated with higher values. Notably, a key feature in the model for predicting histology is the ARVF high Hounsfield unit volume fraction, and thus predictions of both tumour type and grade are dominated by the brightest pixels in the tumour. Finally, our radiomics model shows that using CT images, distinct tumour features can still be derived to aid the prediction of histology and grade. Our choice of CT increases the model applicability with no additional imaging required.

We acknowledge the limitations of this study. Its retrospective nature leaves it susceptible to potential bias. Given the rarity of soft tissue sarcomas, retrospective cohorts make studies such as ours possible, where valuable findings can be consolidated and validated through prospective studies. Limited by the low incidence of other retroperitoneal histologies, we only included liposarcoma and leiomyosarcoma and acknowledge that our model is not generalisable to all retroperitoneal sarcomas. There is an imbalance between the number of patients with leiomyosarcoma and liposarcoma within our discovery cohort, which is reflective of the real-world differences in the incidence of these histologies. However, our assessment was that this imbalance was insufficient to require additional resampling or weighting techniques, which are generally recommended for an imbalance greater than 1:10. In addition, because of the small numbers of patients with grade 3 tumours in the validation dataset, we could not validate our model for low or intermediate versus high grade tumours. Clinically, these tumours were more likely to represent dedifferentiated liposarcoma and leiomyosarcoma, which are substantially harder to distinguish with conventional imaging and carry a higher risk of under-grading by biopsy. This can be explored prospectively to further improve our model and better inform clinical practice. Furthermore, the small dataset limits the number of extractable radiomic features. However, this study surpasses most radiomic studies within this tumour type and anatomical site. The intensive internal cross-validation performed further improves confidence in the final feature sets selected and their stability. Because of the limited dataset, harmonisation techniques could not be applied. This risks extra noise on some affected radiomic features, resulting in a lower predictive power and rejection by the model pipeline.[27] Correct harmonisation might have improved their predictive capabilities and therefore we accept the risk of rejecting potentially useful features. The low calibration statistic values for the model predicting grade suggest that a larger dataset would be necessary to obtain a well calibrated model.

To conclude, our study provides a foundation for the further development and prospective validation of these models. These models could be further developed to address the intricate complexities of intermediate and higher grade retroperitoneal sarcomas and for exploring the value of combining other radiological features or clinical data to the predictive performance of the models. The availability of high-quality STRASS trial data is an opportunity to explore a prognostic model for clinical

outcome and this is in development. Finally, by developing this model further, it could harmonise entry into future prospective clinical trials and standardise patient staging and prognostication.

Contributors

AA, PHH, and CM were responsible for conceptualisation and funding acquisition. AA was responsible for project management, and PHH and CM supervised the study. AA, DS, SD, HL, AN, SL, SB, RH, AG, DVG, AD, CPR, PM, MS, MH, CLP, and RJ were involved in data curation. AA, MRO, and CM had access to and verified all data. AA, MRO, PHH, and CM were responsible for the formal analysis. AA, MRO, RE, SV, CK-M, JL, SD, KT, CF, PHH, and CM conducted the methods of the study. JL and SD were involved in the pseudo-anonymisation and transfer of scans onto the eXtensible Neuroimaging Archive Toolkit platform. HL, AN, and SL were project administrators for the STRASS trial data. AA, MRO, PHH, and CM were responsible for the literature search and writing the original draft. All authors critically appraised the study concept and design, reviewed and approved the manuscript, and accept responsibility for the decision to submit for publication.

Declaration of interests

AA reports funding from the Wellcome Trust paid to their institution. MRO, SD, and CM report funding paid to their institutions from the National Institute for Health and Care Research Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and The Institute of Cancer Research, London. MRO reports funding from the Royal Marsden Cancer Charity to their institution. SD reports funding to their institution from Cancer Research UK. RH declares a previous role as the President of the Connective Tissue Oncology Society (in 2021). AG reports grants from PharmaMar and Nanobiotic; consulting fees from Novartis, Pfizer, Bayer, Lilly, PharmaMar, SpringWorks, and Boehringer Ingelheim; and honoraria from Deciphera. CPR reports royalties from UpTo Date. DVG reports grants from Elekta, IntraOp, and Orfit to their institution; consulting fees from Sanofi, Takeda, Novartis, and Merck; payment or honoraria from Elekta for lectures and support for attending meetings or travel; payment from Sanofi for expert testimony; and also declares a role in the Belgian College of Oncology and a research partnership with Elekta, Siemens, IntraOp and, MIM software, and a commercial partnership with Orfit, VisionRT, Philips, and Precisionxray. CLP reports payment from AstraZeneca to their institution for lectures; support for attending meetings and travel by Janseen and Ose Immunotherapeutics; and participation on advisory boards for AstraZeneca, Bristol Myers Squibbs and Varian. RJ reports grants from MSD and GSK and consulting fees from Adaptimmune, Astex, Athenex, Bayer, Boehringer Ingelheim, Blueprint, Clinigen, Eisai, Epizyme, Daichii, Deciphera, Immunedesign, Immunicum, Karma Oncology, Lilly, Merck, Mundipharma, PharmaMar, SpringWorks, Synox, Tracon, and UpTo Date. CM reports funding from European Organisation for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group to their institution; payment or honoraria from TeleMedicine Clinic academy and GE Healthcare; and support for meetings from International Cancer Imaging Society. CM also declares a role as trustee of the British Institution of Radiology and is a Fellow of the International Cancer Imaging Society. All other authors declare no competing interests.

Data sharing

De-identified individual participant data, including a data dictionary defining each field within the dataset, the algorithm, and images can be shared upon reasonable request to the corresponding author (CM). These data will be available beginning 9 months and ending 36 months after Article publication. Access to these data is controlled by the Research and Development Division at the Royal Marsden Hospital. Access can be obtained by contacting the corresponding author (CM) and will require a protocol review and data access agreement with the Royal Marsden Hospital. For access to data provided by the European Organisation for Research and Treatment of Cancer, please contact the corresponding author for further details.

Acknowledgments

We acknowledge funding from the Wellcome Trust, the National Institutes for Health, European Organisation for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group, and the European Organisation for Research and Treatment of Cancer, and thank them for their invaluable data contribution. This work represents independent research funded by the National Institute for Health and Care Research Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and The Institute of Cancer Research, London. The views expressed are those of the authors and not necessarily those of the National Institutes of Health Research or the Department of Health and Social Care. Finally, we acknowledge the patients included in this study, without whom this work would not have been possible.

References

1 Dangoor A, Seddon B, Gerrand C, Grimer R, Whelan J, Judson I. UK guidelines for the management of soft tissue sarcomas. *Clin Sarcoma Res* 2016; **6:** 20.

2 Stojadinovic A, Yeh A, Brennan MF. Completely resected recurrent soft tissue sarcoma: primary anatomic site governs outcomes. *J Am Coll Surg* 2002; **194:** 436–47.

3 Schneider N, Strauss DC, Smith MJ, et al. The adequacy of core biopsy in the assessment of smooth muscle neoplasms of soft tissues: implications for treatment and prognosis. *Am J Surg Pathol* 2017; **41:** 923–31.

4 Bonvalot S, Gronchi A, Le Péchoux C, et al. Preoperative radiotherapy plus surgery versus surgery alone for patients with primary retroperitoneal sarcoma (EORTC-62092: STRASS): a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol* 2020; **21:** 1366–77.

5 Crombé A, Fadli D, Italiano A, Saut O, Buy X, Kind M. Systematic review of sarcomas radiomics studies: bridging the gap between concepts and clinical applications? *Eur J Radiol* 2020; **132:** 109283.

6 Shur JD, Doran SJ, Kumar S, et al. Radiomics in oncology: a practical guide. *Radiographics* 2021; **41:** 1717–32.

7 Arthur A, Johnston EW, Winfield JM, et al. Virtual biopsy in soft tissue sarcoma. How close are we? *Front Oncol* 2022; **12:** 892620.

8 Huang EP, O'Connor JPB, McShane LM, et al. Criteria for the translation of radiomics into clinically useful tests. *Nat Rev Clin Oncol* 2023; **20:** 69–82.

9 Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. *Lancet Digit Health* 2021; **3:** e195–203.

10 Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017; **14:** 749–62.

11 Napel S, Mu W, Jardim-Perassi BV, Aerts HJWL, Gillies RJ. Quantitative imaging of cancer in the postgenomic era: Radio(geno) mics, deep learning, and habitats. *Cancer* 2018; **124:** 4633–49.

12 Doran SJ, Al Sa'd M, Petts JA, et al. Integrating the OHIF Viewer into XNAT: achievements, challenges and prospects for quantitative imaging studies. *Tomography* 2022; **8:** 497–512.

13 Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 2007; **5:** 11–34.

14 Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020; **295:** 328–38.

15 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; **15:** 155–63.

16 Orton MR, Hann E, Doran SJ, et al. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx Renal study. *Cancer Imaging* 2023; **23:** 76.

17 Morosi C, Stacchiotti S, Marchianò A, et al. Correlation between radiological assessment and histopathological diagnosis in retroperitoneal tumors: analysis of 291 consecutive patients at a tertiary reference sarcoma center. *Eur J Surg Oncol* 2014; **40:** 1662–70.

18 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* 2018; **47:** 829–40.

19 Navarro F, Dapper H, Asadpour R, et al. Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging. *Cancers (Basel)* 2021; **13:** 2866.

20 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.

21 Wang H, Chen H, Duan S, Hao D, Liu J. Radiomics and machine learning with multiparametric preoperative MRI may accurately predict the histopathological grades of soft tissue sarcomas. *J Magn Reson Imaging* 2020; **51:** 791–97.

22 Xu W, Hao D, Hou F, Zhang D, Wang H. Soft tissue sarcoma: preoperative MRI-based radiomics and machine learning may be accurate predictors of histopathologic grade. *AJR Am J Roentgenol* 2020; **215:** 963–69.

23 Yan R, Hao D, Li J, et al. Magnetic resonance imaging-based radiomics nomogram for prediction of the histopathological grade of soft tissue sarcomas: a two-center study. *J Magn Reson Imaging* 2021; **53:** 1683–96.

24 Zhang Y, Zhu Y, Shi X, et al. Soft tissue sarcomas: preoperative predictive histopathological grading based on radiomics of MRI. *Acad Radiol* 2019; **26:** 1262–68.

25 Tomaszewski MR, Gillies RJ. The biological meaning of radiomic features. *Radiology* 2021; **299:** E256.

26 Messiou C, Morosi C. Imaging in retroperitoneal soft tissue sarcoma. *J Surg Oncol* 2018; **117:** 25–32.

27 Bishop CM. Pattern recognition and machine learning. New York, NY: Springer, 2006.