Strahlenther Onkol (2018) 194:824–834
https://doi.org/10.1007/s00066-018-1294-2

ORIGINAL ARTICLE

Treatment-related features improve machine learning prediction of prognosis in soft tissue sarcoma patients

Jan C. Peeken[1,7] · Tatyana Goldberg[2] · Christoph Knie[1] · Basil Komboz[2] · Michael Bernhofer[3] · Francesco Pasa[4,5] · Kerstin A. Kessel[1,6,7] · Pouya D. Tafti[2] · Burkhard Rost[3] · Fridtjof Nüsslin[1] · Andreas E. Braun[2] · Stephanie E. Combs[1,6,7]

Received: 5 October 2017 / Accepted: 5 March 2018 / Published online: 20 March 2018

Abstract

Background and purpose Current prognostic models for soft tissue sarcoma (STS) patients are solely based on staging information. Treatment-related data have not been included to date. Including such information, however, could help to improve these models.

Materials and methods A single-center retrospective cohort of 136 STS patients treated with radiotherapy (RT) was analyzed for patients' characteristics, staging information, and treatment-related data. Therapeutic imaging studies and pathology reports of neoadjuvantly treated patients were analyzed for signs of response. Random forest machine learning-based models were used to predict patients' death and disease progression at 2 years. Pre-treatment and treatment models were compared.

Results The prognostic models achieved high performances. Using treatment features improved the overall performance for all three classification types: prediction of death, and of local and systemic progression (area under the receiver operatoring characteristic curve (AUC) of 0.87, 0.88, and 0.84, respectively). Overall, RT-related features, such as the planning target volume and total dose, had preeminent importance for prognostic performance. Therapy response features were selected for prediction of disease progression.

Conclusions A machine learning-based prognostic model combining known prognostic factors with treatment- and response-related information showed high accuracy for individualized risk assessment. This model could be used for adjustments of follow-up procedures.

Keywords Biomarker · Precision medicine · Prognostic model · Random forest · Decision support systems

Both authors contributed equally: Jan C. Peeken, Tatyana Goldberg.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00066-018-1294-2) contains supplementary material, which is available to authorized users.

✉ Jan C. Peeken
jan.peeken@tum.de

[1] Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany

[2] Allianz SE, Königinstraße 28, 80802 Munich, Germany

[3] Department for Bioinformatics and Computational Biology, Informatik 12, Technical University of Munich (TUM), Boltzmannstraße 3, 85748 Garching, Germany

[4] Department of Computer Science, Informatik 9, Technical University of Munich (TUM), Boltzmannstraße 3, 85748 Garching, Germany

[5] Chair of Biomedical Physics, Department of Physics, Technical University of Munich (TUM), James-Franck-Straße 1, 85748 Garching, Germany

[6] Institute of Innovative Radiotherapy (iRT), Department of Radiation Sciences (DRS), Helmholtz Zentrum München, Ingolstaedter Landstraße 1, 85764 Neuherberg, Germany

[7] Partner Site Munich, Deutsches Konsortium für Translationale Krebsforschung (DKTK), Munich, Germany

Therapieinformationen verbessern auf maschinellem Lernen basierende prognostische Einschätzungen für Patienten mit Weichteilsarkomen

Zusammenfassung

Hintergrund und Zielsetzung Aktuelle prognostische Modelle für Patienten mit Weichteilsarkomen basieren primär auf Staginginformationen. Therapieinformationen werden dabei nicht berücksichtigt. Die Berücksichtigung solcher Daten könnte die Vorhersage verbessern.

Material und Methoden Für eine retrospektive, monozentrische, strahlentherapeutisch behandelte Kohorte mit 136 Weichteilsarkompatienten wurden Patientencharakteristika, Staging und therapieassoziierte Daten erhoben. Potenzielle mit dem Therapieansprechen assoziierte Informationen von neoadjuvant behandelten Patienten wurden aus therapeutischen Magnetresonanztomographie(MRT)-Datensätzen und pathologischen Befunden erhoben. Auf Basis dieser Informationen wurden Random-Forest-Modelle für die Vorhersage des 2-Jahres-Überlebens bzw. des Progresses generiert. Prätherapie- und Therapiemodelle wurden verglichen.

Ergebnisse Die prognostischen Modelle zeigten insgesamt eine gute Vorhersagekraft. Die Hinzunahme von Therapieinformationen konnte die Vorhersageeffizienz der 3 Klassifikationen verbessern: Vorhersage des Versterbens sowie des lokalen und systemischen Progresses („area under the receiver operating or characteristic curve“ [AUC] je 0,87, 0,88 und 0,84). Strahlentherapieassoziierte Informationen wie das Planungszielvolumen und die Gesamtdosis hatten einen großen Einfluss auf die Vorhersagekraft. Mit dem Therapieansprechen assoziierte Informationen wurden für die Vorhersage des Progresses selektiert und zeigten so eine mögliche prognostische Bedeutung.

Schlussfolgerung Auf maschinellem Lernen basierende prognostische Modelle zeigten eine hohe Genauigkeit für die Vorhersage des Überlebens und Krankheitsprogresses durch Einschluss von Informationen zur Therapie und zum Therapieansprechen. Diese Modelle könnten für die individuelle Risikoabschätzung in der Nachsorge verwendet werden.

Schlüsselwörter Biomarker · Präzisionsmedizin · Prognostisches Modell · Random Forest · Entscheidungsunterstützungssystem

Introduction

Soft tissue sarcomas (STS) constitute a histologically heterogeneous group of malignancies of mesenchymal origin. They constitute about 1% of all malignant disorders, with a reported incidence rate of 37.8 cases per 1,000,000 persons per year [1]. The median survival of a large cohort of 8249 patients treated between 1981 and 2004 was 25 months [1].

In large retrospective analyses, prognostic markers have been identified for local and distant progression as well as survival, including resection margin status, age, histological subgroups, histological grading, recurrent disease, tumor size, nodal status, distant metastasis, tumor depth, and the tumor's location [2–6]. Tumor size appears to be significantly correlated with progression and survival above the formerly used thresholds in the 7th edition of the American Joint Committee on Cancer (AJCC) TNM staging system of 5 cm in the greatest dimension [6, 7].

Nowadays, stage-adapted multimodal therapy regimens involving limb-sparing surgery, radiation therapy, or chemotherapy constitutes the standard of care for young and elderly patients [8].

A well-established nomogram on the basis of 12-year survival data was created at the Memorial Sloan Kettering Cancer Center based on a multivariate Cox proportional hazard model including age, tumor depth, tumor location, histological grade, and tumor size [9]. In an independent external validation, the model achieved a concordance index of 0.76 (comparable to the area under the receiver operating characteristic [ROC] curve [AUC]) [10].

In a recent publication, the performance of the AJCC stage was criticized to be non-sufficient, since it does not represent tumor site, histology, and size in an appropriate manner. The authors propose to include further prognostic variables in the above-mentioned categories for model improvement [5]. For optimal patient risk assessment, multiple independent models were necessary to represent distinct patient subgroups (e. g., tumor site-specific). The novel 8th version of the AJCC staging system starts addressing this issue by implementing four separate staging groups as head and neck, abdomen/thoracic visceral organs, extremities/trunk, and retroperitoneum [11]. The clinical utility, however, still needs to be proven as done for other entities [12].

The human cognitive capacity of estimating prognosis appears to be limited to five features per classification [13]. In recent years, machine learning approaches have emerged as an alternative means for model generation. Due to the fundamental improvements in handling complex and large datasets, machine learning-based clinical decision support systems combining multidimensional data may lead the way to future precision medicine [14].

In this work, we focused on the generation of a machine learning-based model for the classification of patient survival as well as of local and distant progression of patients with STS. Besides known prognostic clinical and pathological features, treatment modalities and treatment-related features were also included into the model. The resulting models were tested on an internal validation cohort and the contribution of single features was tested. We thus demonstrate how machine learning can be used for personalized prognostic assessment based on different types of input data.

Methods

Study design and patient cohort

In total, 136 STS patients treated with radiotherapy (RT) from 2007 to 2014 at our institution were included into this study. The overall survival, calculated from the end of RT to the time point of death or the time point of censoring, was 19.9 months (95% confidence interval 17–22.8 months) with 53 reported deaths. Patients with sarcomas of the head and neck were excluded due to low patient numbers.

As binary endpoints for the machine learning models, a 2-year (2y) threshold was chosen to balance the groups and minimize the number of censored patients. Censored patients were excluded prior to any modeling. The 2y survival was 58% (38 deaths, 53 alive, and 45 censored). The 2y local and systemic progression was 42 and 57%, respectively (local: 31 progress, 43 no progress, 62 censored; systemic: 54 progress, 41 no progress, 41 censored).

All clinical and molecular data were collected in the Munich Innovative Radiotherapy (MIRO) database. This study was approved by the ethical committee of the Technical University of Munich (reference number 466/16).

Patient characteristics and features included for model building

Prognostic features

Patient records were assessed for age, TNM staging according to the 7th edition of the AJCC staging system, tumor site, histological grading, and histological subtype (Table 1). In total, nine variables were used as input features for pre-treatment model generation. If the TNM stage was documented with the suffix x, the respective information was excluded. Missing data are indicated in Table 1.

Treatment- and response-related features

Treatment-related information including resection status, radiotherapy type, total delivered dose, single delivered dose, boost technique, planning target volume (PTV), chemotherapy, and surgery was assessed. Combined, thirteen input features were used for treatment models (Table 2). Generally, patients received RT with neoadjuvant, adjuvant, additive, definitive, or palliative intent. RT was delivered using a helical tomotherapy (129 patients, 94.9%), intensity-modulated radiation therapy (2 patients, 1.5%), or three-dimensional conformal RT (5 patients, 4.7%). Pre-treatment gross tumor volume (GTV) and the volume of the primary tumor (PT) were contoured on magnetic resonance imaging (MRI) studies using T2-weighted and fat-saturated T1-weighted contrast-enhanced sequences. GTV was defined as PT plus surrounding edematous changes. If MRI was not available, contouring was conducted on planning computer tomography (CT) datasets. For patients who received neoadjuvant RT, the PT after RT and its relative change compared to the PT pre-RT were assessed. Post-treatment PTs were contoured on MRI using the same sequences at the first follow-up 6 weeks after the end of RT. The percentage of viable cells after neoadjuvant RT was determined in analogy to the European Organization for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group (EORTC-STBSG) recommendations for pathological examination on a continuous scale [15].

Prediction method

Random forest is the algorithm that implements an ensemble of decision trees constructed from randomly selected features and training data points [16]. Due to its high predictive power, short training periods, good interpretability, and the ability to handle noisy and incomplete data, random forest has been the algorithm of choice for many machine learning-based prediction tasks [16–20]. In this study, the WEKA v38 [21] implementation of random forest was used to develop altogether six prediction models: (i) treatment model for 2y survival, (ii) pre-treatment model for 2y survival, (iii) treatment model for 2y systemic progression, (iv) pre-treatment model for 2y systemic progression, (v) treatment model for 2y local progression, and (vi) pre-treatment model for 2y local progression.

Cross-validation

For all testing purposes, the data set for each of the six prediction models was randomly split into five subsets. For each of the five subsets, four sets were used for training the model (optimizing free parameters, i. e., the number of trees and the feature set) in a nested five-fold cross-validation

Table 1 Prognostic features for pre-therapy models

Feature	Distribution within patient cohort
Age at diagnosis, median (±SD)	56.8 y (±22.8 y)
T-stage, (%)	
T1	34 p (25.0%)
T2	95 p (69.9%)
Not classified	7 p (5.1%)
T-stage, suffix (%)	
T-a (superficial tumor), (%)	7 p (5.1%)
T-b (deep tumor), (%)	99 p (72.8%)
Data not applicable/available	30 p (22.1%)
N-stage, (%)	
N0	94 p (69.1%)
N1	8 p (5.9%)
Nx	34 p (25.0%)
M-stage, (%)	
M0	90 p (66.2%)
M1	23 p (16.9%)
Mx	23 p (16.9%)
Recurrent disease, (%)	23 p (16.9%)
Tumor site, (%)	
Upper extremities	12 p (8.8%)
Lower extremities	57 p (41.9%)
Pelvis	25 p (18.4%)
Thorax	19 p (14.0%)
Abdomen	10 p (7.4%)
Retroperitoneum	13 p (9.6%)
Histological grading[a]*, (%)*	
G1	3 p (2.2%)
G2	49 p (36.0%)
G3	55 p (40.4%)
G4 (Ewing sarcoma)	19 p (13.9%)
No classification	10p (7.4%)
Histological subtype, (%)	
Liposarcoma	26 p (19.1%)
Synovial sarcoma	9 p (0.7%)
Spindle cell sarcoma	4 p (0.7%)
Myxofibrosarcoma	19 p (14.0%)
Ewing sarcoma	22 p (16.2%)
Leiomyosarcoma	11 p (8.1%)
Rhabdomyosarcoma	4 p (2.9%)
Angiosarcoma	3 p (2.2%)
Pleomorphic sarcoma	30 p (22.1%)
Other[b]	4 p (2.9%)
No classification	4 p (5.8%)

Nine prognostic features (highlighted in italics) were used for development of pre-therapeutic prediction models for 2y survival, and 2y systemic, and 2y local progression. For each feature, its sub-categories and the number of patients within the corresponding sub-category in this study are listed. T-, N- and M-stage respond to categories of the TNM classification of malignant tumours according to the 7th edition of the AJCC staging system.

p patients, *SD* standard deviation, *y* years

[a] French Federation of Cancer Centers Sarcoma Group: Ewing sarcoma equals grade 4

[b] Epithelioid sarcoma, myxosarcoma, alveolar soft part sarcoma

Table 2 Treatment- and response-related features for treatment models

Features	Distribution within patient cohort
Resection status, (%)	
R0	48 p (40.0%)
R1/R2	16 p (13.4%)
Rx	6 p (5.0%)
Missing data	39 p (32.3%)
Radiotherapy type	
Neoadjuvant	71 p (52.2%)
Adjuvant	28 p (20.6%)
Additive	15 p (11.0%)
Definitive	15 p (11.0%)
Palliative	7 p (5.2%)
Total delivered dose, median (±SD)	50 Gy (±8.7 Gy)
Single delivered dose, median (±SD)	1.8 Gy (±0.4 Gy)
Boost	
No boost	63 p (46.3%)
Simultaneous boost	68 p (50.0%)
Sequential boost	5 p (3.7%)
PTV, median (±SD)	134 p: 1910.2 ml (±2238.3 ml)
GTV, median (±SD)	102 p: 554.6 ml (±1619.34 ml)
PT pre-RT, median (±SD)	73 p: 258.9 ml (±604.9 ml)
PT post-RT, median (±SD)	68 p: 263.40 ml (±735.3 ml)
Relative decrease in PT after RT, median (±SD)	68 p: –14.4% (±51.7%)
Viable cells after neoadjuvant RT, median (±SD)	50 p: 50% (±32.2%)
Chemotherapy	
Yes	89 p (65.4%)
No	47 p (34.6%)
Surgery	
Yes	119 p (87.5%)
No	17 p (12.5%)

In addition to nine prognostic features, treatment prediction models for 2y survival, 2y systemic, and 2y local progression were developed using 13 treatment and response-related features (highlighted in italics). For each feature, its sub-categories and the number of patients within the corresponding sub-category in this study are listed. For continuous features the number patients with available data and the median and standard deviation are shown

CTV clinical target volume, *GTV* gross tumor volume, *ml* milliliter, *p* patients, *PT* primary tumor, *PTV* planning target volume, *RT* radiation therapy, *SD* standard deviation, *2y* 2 year

[22], and one for testing its predictive power. These subsets were then rotated such that each subset was used for testing exactly once. The average over five subsets used for testing provided the final results. No information from the test split was used during the training phase.

Feature selection

In this study, features for training of three treatment and three pre-treatment models were selected within the cross-validation setting described above. At every new rotation, the following iterative protocol was used: The process was started with all 22 features for the treatment models and 9 features for the pre-treatment models. Their predictive importance was determined by applying the WEKA's Relief Attribute Evaluation function [23]. After removing the feature with the smallest importance, the performance of the model was assessed (in terms of AUC). In case of performance increase, the feature was excluded from the set of best performing features and the remaining features were evaluated. The backward selection was stopped as soon as no further increase in AUC could be observed.

Performance evaluation

The performance was assessed on test sets, as described above, using standard measures. AUC, averaged over five rounds of training and testing, served as a single performance estimator. The corresponding true positive rate (TPR) and false positive rate (FPR) were defined by:

$$\mathrm{TPR} = \frac{TP}{TP+FN} \quad \mathrm{FPR} = \frac{FP}{FP+TN} \tag{1}$$

The overall two-state accuracy (referred to as Q_2) was used:

$$Q_2 = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

Finally, class-specific values were compiled:

$$\begin{aligned} \mathrm{Precision}_{\mathrm{dead/progress}} &= \frac{TP}{TP+FP} \\ \mathrm{Precision}_{\mathrm{alive/no\ progress}} &= \frac{TN}{TN+FN} \end{aligned} \tag{3}$$

$$\begin{aligned} \mathrm{Recall}_{\mathrm{dead/progress}} &= \frac{TP}{TP+FN} \\ \mathrm{Recall}_{\mathrm{alive/no\ progress}} &= \frac{TN}{TN+FP} \end{aligned} \tag{4}$$

where TP are the true positives and FP are the false positives. Similarly, TN are the true negatives and FN are the false negatives. The levels of precision (also called "positive predictive value") and recall (also called "sensitivity") were monitored as a function of the reliability score (RS) of the prediction, which ranged between 0% (non-reliable) and 100% (most reliable). For positive predictions (patient's death and disease progression), RS was computed by multiplying the random forest probability score by 100 and for negative predictions it was subtracted from 100.

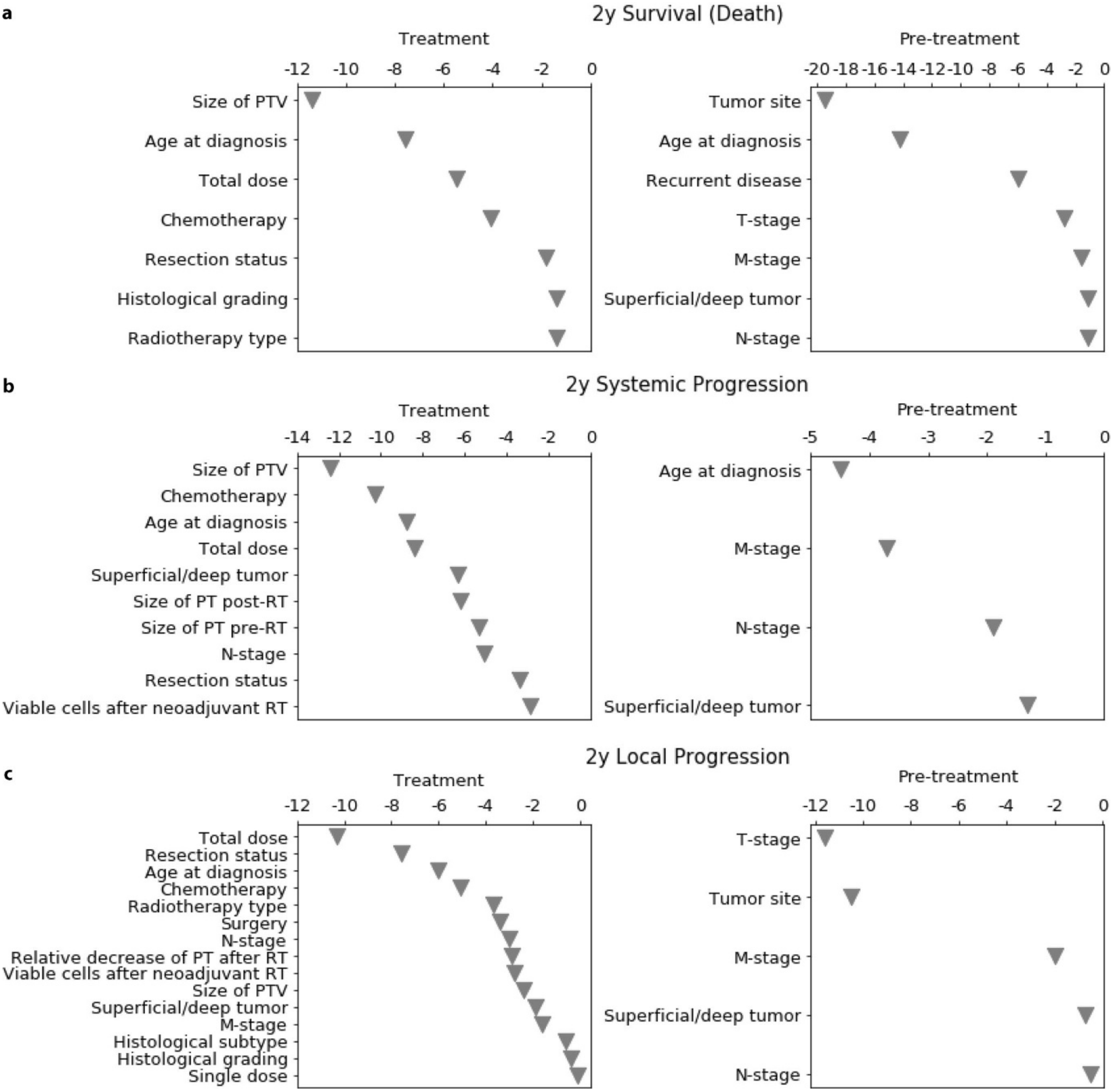

Fig. 1 Importance assessment of the input features. **a** 7 features were selected to be most beneficial for the predictive performance of the treatment and 7 distinct features for the pre-treatment 2y survival model; **b** 10 and 4 features for the treatment and pre-treatment 2y systemic disease progression models, respectively; and **c** 15 and 5 features for the treatment and pre-treatment 2y local disease progression models, respectively. The X-axis shows the absolute decrease in AUC (Methods) if the respective feature was removed from both training and test sets of the random forest model. In this step, no model parameter optimization was performed. All values reported here are averages over a cross-validated test set. Note: differences in AUC were multiplied by 100 for a better readability. *AUC* area under the receiver operating characteristic curve, *PTV* planning target volume, *PT* primary tumor, *RT* radiation therapy, *2y* 2 year

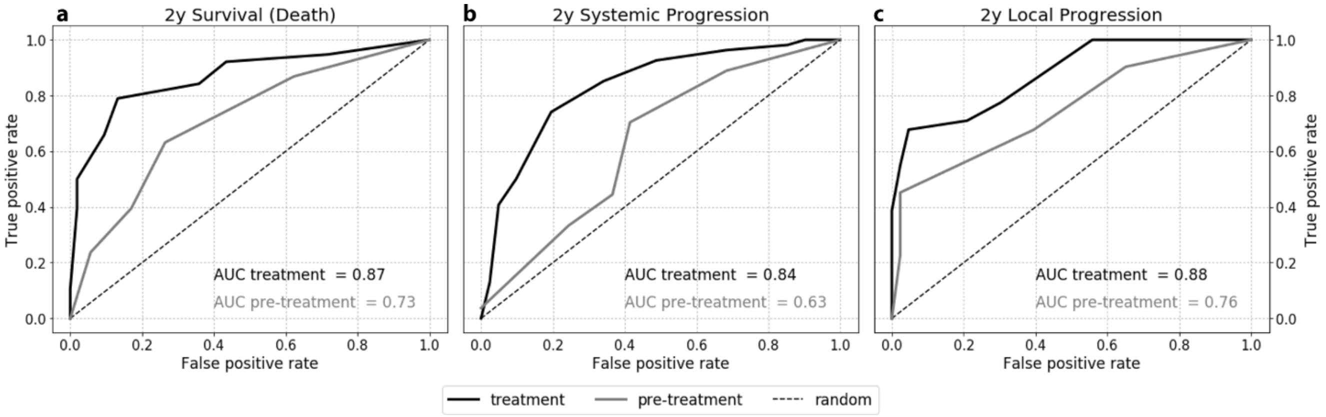

Fig. 2 High performance of the prediction model in cross-validation. All values are based on cross-validated averages of models for the prediction of **a** patients' death at 2y, **b** systemic disease progression at 2y, and **c** local disease progression at 2y. Each subfigure shows ROC curves and AUC performance values for both treatment and pre-treatment models (*black* and *gray lines*, respectively) and their comparison to a random classifier (*dashed line*) with AUC of 0.5. *AUC* area under the ROC curve, *ROC* receiver operating characteristic, *2y* 2 year

Statistics

Kaplan–Meier survival curves were generated following classification of patients from the cross-validated test sets. Significant separation of survival curves was tested using the log rank test. A *p*-value of 0.05 or lower was regarded as statistically significant. All analyses were performed using the Graphpad Prism version 5.0c (GraphPad Software Inc., La Jolla, CA, USA).

Results

In total, six different models were developed for 2y patient survival, 2y systemic disease progression, and 2y local disease progression. Pre-treatment models were generated using nine features listed in Table 1, while treatment models used the additional 13 features listed in Table 2.

Few features most predictive

Our backward selection algorithm yielded at most 15 features to be informative for each of the six prediction models developed (Fig. 1). The success of all models was dominated by the first few features.

For pre-treatment models, the prognostic feature "tumor site" appeared to be among the most discriminative features with an impact of nearly 0.20 AUC in predicting death and nearly 0.11 in predicting local progression. Age at diagnosis was another strongly discriminative feature with an AUC difference up to 0.14 in predicting death and nearly 0.05 in predicting systemic progression. TNM staging information and tumor depth were identified as additional discriminative features for all three pre-treatment models, with an impact of at most 0.04 in absolute AUC.

Treatment models were, in contrast, predominantly based on treatment information. While age at diagnosis appeared among the impactful features (AUC decrease up to 0.09) for all three models, total delivered dose and chemotherapy were among them as well, with an AUC change up to 0.11 (Fig. 1c, treatment model). PTV was another most distinguishable feature for the prediction of death and systemic progression, with an AUC change up to 0.12. The influence on model performance of these treatment-related features compared to the pre-treatment information underlines their strong impact.

From the features assessed after neoadjuvant RT, size of the PT after RT was selected into the predictive model for systemic progression, its relative decrease after RT for the prediction of local progression, and the percentage of viable cells after RT was selected into both models with an AUC change up to 0.06.

Treatment features improved model performance

Random forest prediction models estimate the probability of patients' death and systemic and local disease progression. Through a simple threshold, this probability gives a binary prediction (e.g., death at (RS)> 50% and alive at RS $\leq$ 50%). At these thresholds, the overall two-state accuracy for the prediction models was Q_2 > 84% for therapeutic and Q_2 > 67% for pre-therapeutic 2y survival models, Q_2 > 77% for therapeutic and Q_2 > 60% for pre-therapeutic 2y systemic progression models, and Q_2 > 84% for therapeutic and Q_2 > 77% for pre-therapeutic 2y local progression models (Supplementary Table S1). By iterating over the whole spectrum of thresholds, the ROC curves and the AUC values for the six models were also computed (Fig. 2, Supplementary Table S2). The prediction of death at 2 years using the pre-treatment survival model achieved an AUC

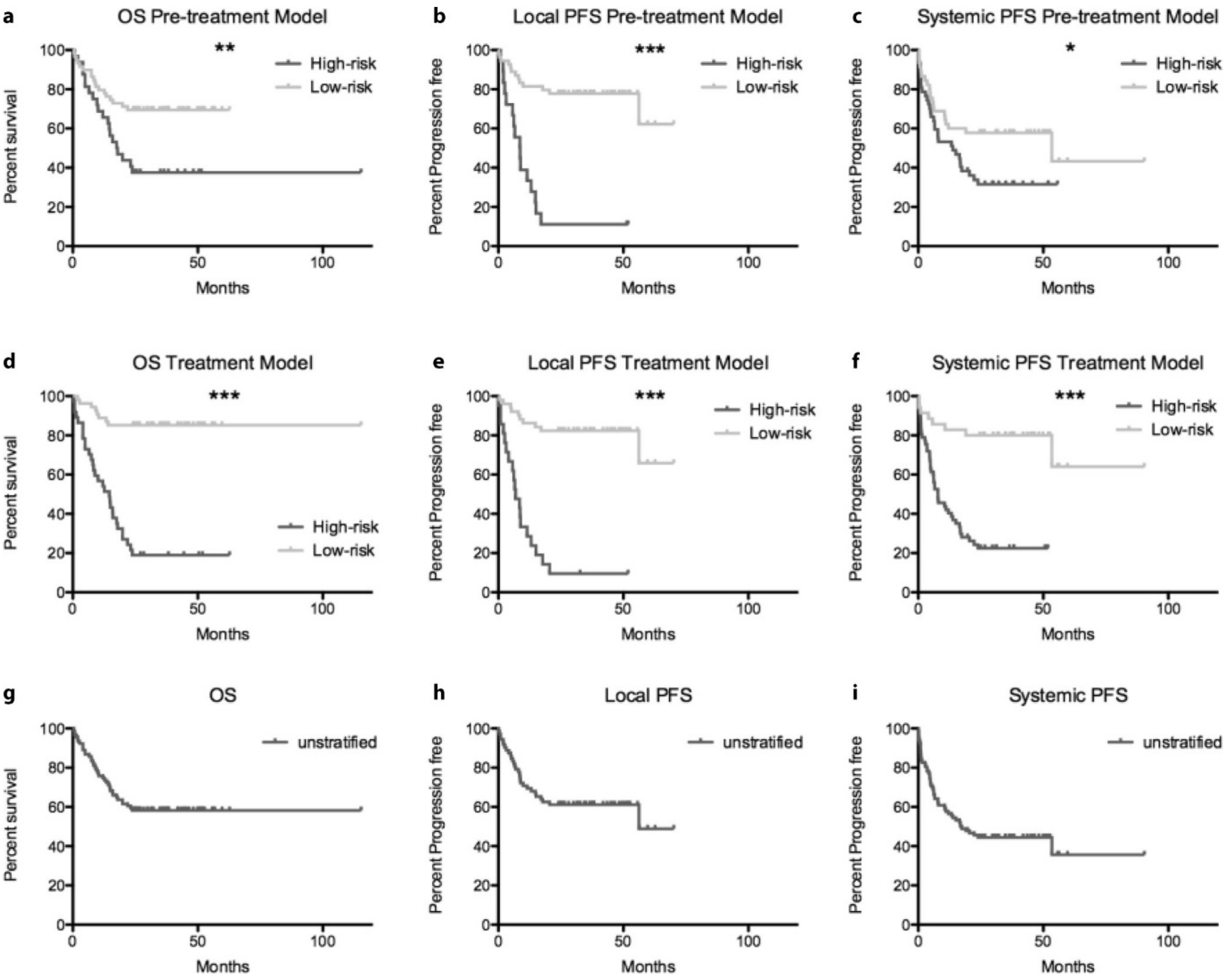

Fig. 3 Prognostic models significantly differentiate high-risk from low-risk patients. All patients from the cross-validated test sets of six prediction models were divided into two groups following the classification by the predictive models. Kaplan–Meier survival curves were generated for the pre-treatment models (**a**, **b**, **c**) and treatment models (**d**, **e**, **f**) predicting OS, local PFS, and systemic PFS. Unstratified Kaplan–Meier curves are shown for OS, local PFS, and systemic PFS (**g**, **h**, **i**). Significant separation was determined by log rank test, * $p > 0.05$, ** $p > 0.01$, *** $p > 0.0001$. *OS* overall survival, *PFS* progression-free survival

of 0.73 (random classifier: AUC 0.5). Adding in therapeutic features improved the performance to an AUC of 0.87 using the treatment model. It appeared that the difference in performance was similar between treatment (AUC of 0.88) and pre-treatment (AUC of 0.76) models for 2y local disease progression. This was, however, different for the treatment and pre-treatment models for 2y systemic disease progression, as they performed with a larger difference in AUC (post: 0.84 and pre: 0.63). In order to compare the performance with existing prognostic models, we analyzed the predictive performance of the 7th edition of the AJCC staging system, as all STS were staged according to this [7]. Compared to treatment models, the staging system achieved overall lower predictive performances, with AUCs of 0.58 for death at 2y, 0.68 for 2y local progression, and 0.66 for 2y systemic progression.

Finally, risk group stratification of patients in the validation was tested following classification by the predictive models. All models were able to significantly split patients into a high-risk and a low-risk group (see Fig. 3 for Kaplan–Meier curves).

Thus, predicting systemic disease progression at 2y with prognostic features alone appears to be the most difficult task. Nevertheless, prognostic features were strong indicators for death and disease progression, and enable high prediction performance. A combination including treatment-related features, however, considerably increased the performance of prediction models further.

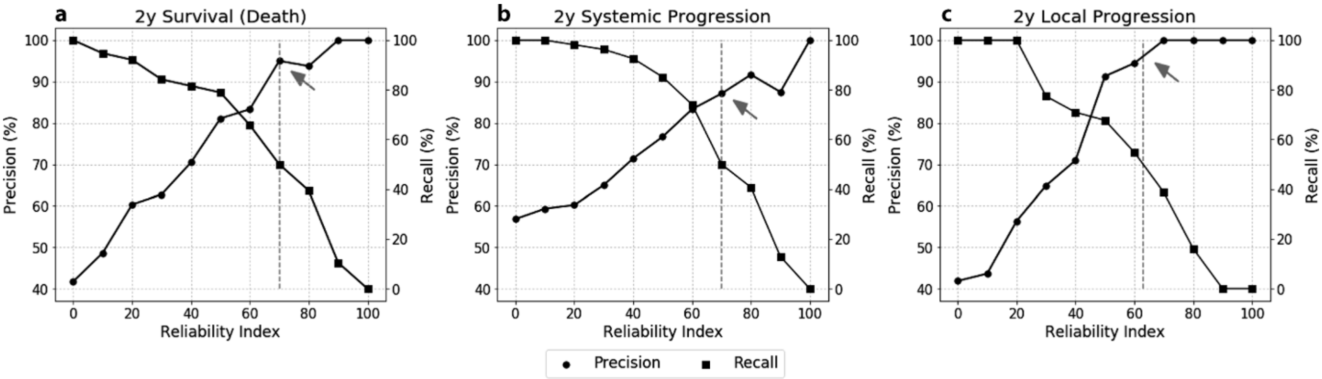

Fig. 4 Predictions with a higher reliability score are better. Precision/recall curves above a given threshold in reliability score (RS), ranging from 0 = unreliable to 100 = most reliable, are shown for random forest-based predictions of **a** patients' death at 2y, **b** disease systemic progression at 2y, and **c** disease local progression at 2y. All models shown here are treatment models. Death is predicted for 50% of patients at RSs above 70 (*gray dashed cross-line*); for these, precision is above 95% (*gray arrow*). Similar values are achieved for disease progression models: half of patients with systemic disease progression are predicted with RS above 70 and precision >87%, and half of patients with local disease progression are predicted with RS above 63 and precision >95%

Reliability score provides more confidence in predictions

Each binary decision of the prediction models comes with an RS (Methods) that measures the level of confidence of a particular prediction. The value of RS correlates strongly with the accuracy of the model (Fig. 4) and thus enables users to focus on more reliable predictions or knowingly broadcast the less reliable ones. By making a choice for an RS threshold, any user can read off Fig. 4 what to expect from the choice. For instance, when choosing 50% predicted patients with the status of death using the treatment survival model, nearly 95% of the predictions were correct (Fig. 4a, gray arrow). These predictions were obtained at high RS70 (Fig. 4a, gray dashed cross-line). A similar level of precision of 95% was also achieved for more than half of the predictions for local disease progression with the treatment model, but at RS ≥63 (Fig. 4c). For systemic disease progression, 50% of best predictions of the treatment model were correct in over 87% of all cases at RS ≥70 (Fig. 4b). Performance estimates of pre-treatment models are shown in Supplementary Figure S1.

Discussion

To the best of our best knowledge, the prognostic models presented in this work are the first machine learning-based methods combining pre-treatment and treatment-related information for STS patients. Random forests were selected as underlying classification approaches due to their short training times, easy interoperability of results, and the ability to deal with missing and incomplete data.

The pre-treatment random forest models achieved predictive performances with overall AUCs of 0.73, 0.63, and 0.76 for predicting patients' death, systemic disease progression, and local disease progression, respectively, in an internal validation cohort. Selected features were known prognostic factors such as age, TNM stage, and tumor site as clinically expectable [2, 4]. With AUCs of 0.87, 0.84, and 0.88, respectively, the treatment models achieved a higher prediction performance for patient's death as well as for systemic and local disease progression. All models allowed for significant risk group stratification. The feature importance analyses gave insights into how much a single treatment-related feature contributed to the improved performance of the treatment models. Therapy-related features such as PTV, chemotherapy, and total delivered dose had the strongest influence on prediction performance, with a reduction in AUC of up to 0.12 when excluded from all three treatment models. Pathological and imaging features assessed after neoadjuvant RT were selected for disease progression prediction, demonstrating a potential for response assessment.

Hence, we demonstrated that by integrating treatment-related information into predictive models, a superior performance is achievable. Established prognostic models, such as the AJCC staging system or existing nomograms, are currently used for treatment decisions [9, 11]. In patients treated with postoperative RT, our treatment models could be used accordingly if the planned RT doses were used. In contrast, for neoadjuvantly treated patients, the model requires information that becomes available only after therapy (such as margin status). In this case, our models could be used to adjust follow-up procedures with a high accuracy. For instance, patients with a high risk of disease progression or death could be monitored more closely after therapy (e. g., shorter imaging intervals for a longer period after therapy or adjusted imaging modalities). Moreover, depending on the prediction of local or systemic progression, the follow-up procedure could be adjusted accordingly. This might lead to a better distribution of resources inside the medical system, with earlier diagnosis of progression and improved treatment options. As a result, patients with a low risk of recurrence could be spared unnecessary procedures.

Recently novel predictive biomarkers have been proposed. For instance, Valliere et al. demonstrated the feasibility of quantitative high throughput analysis of MRIs ("radiomics") to predict distant progression in STS [24]. A comprehensive genomic characterization of STS by the Cancer Genome Atlas Research Network may lead to novel molecular determinators of prognosis [25].

It should be noted that the prognostic models described in this work were generated based on a single-center retrospective cohort of 136 patients. In contrast to prospective data, this dataset is thus vulnerable to selection- and information bias [26]. Moreover, we chose to predict dichotomized endpoints at 2y, which lead to exclusion of censored patients constituting a further source for potential bias. Due to the heterogeneity of STS, certain subgroups were represented with only very few patients. However, for safe appliance of a prognostic model to other STS patients, comparable patient characteristics as well as treatment plans are essential. In this context, seldom irradiated G1-graded sarcomas, R2 resection status, and rare histological subtypes (including synovial sarcoma, spindle cell sarcoma, angiosarcoma, rhabdomyosarcoma, epithelioid sarcoma, myxosarcoma) were insufficiently covered in the underlying study population. In addition, only 4 patients were younger than 10 years old. Therefore, this model cannot be reliably applied to pediatric STS patients. Patients with sarcomas of the head and neck were excluded beforehand and are thus not covered by the model. Despite the large benefits of a random forest model, a larger number of patients is necessary for model training and testing. Due to limited overall patient numbers, this may lead to a certain instability in prediction performance (e.g., causing large standard errors) and the feature backward selection. The latter could explain inconsistencies in feature selection between pretherapeutic and therapeutic models.

The prognostic models were thoroughly evaluated on an internal patient group using the standard cross-validation techniques. To guarantee effectiveness for a distinct patient group, however, an external validation step is necessary. Our models were published online (www.predictcancer.org) to enable external validation in independent patient cohorts.

Conclusion

This is the first study presenting six machine learning-based prognostic models on the basis of known prognostic factors and treatment-related information, which are, to our best knowledge, the first ones of their kind. RT-related features such as PTV and total delivered dose had preeminent importance for model performances. Each prediction is accompanied by a reliability score, which allows users to focus on most accurate predictions. These machine learning-based models constitute a first step towards future decision-support systems leading the way to more personalized medicine. In a next step, genomic or proteomic data from biopsies, resected tissues, or liquid biopsies must be included, and radiomics information quantifying texture, shape, and intensity of an individual tumor could enhance prediction of local or systemic disease progression [3, 4, 13, 27–32]. Cross-institutional large cohorts with standardized data will be required for testing and optimization of these machine-learning methods, as is currently being done by, e.g., the EUROCAT consortium [30].

Funding Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich (to JP, KK, SC), Alexander von Humboldt Foundation through German Federal Ministry for Education and Research, and the Bavarian Competence Network for Technical and Scientific High Performance Computing (to MB and BR), Allianz (to TG, BK, PT and AB).

Compliance with ethical guidelines

Conflict of interest J.C. Peeken, T. Goldberg, C. Knie, B. Komboz, M. Bernhofer, F. Pasa, K.A. Kessel, P.D. Tafti, B. Rost, F. Nüsslin, A.E. Braun, and S.E. Combs declare that they have no competing interests.

Ethical standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

1. Gutierrez JC, Perez EA, Franceschi D et al (2007) Outcomes for soft-tissue sarcoma in 8249 cases from a Large State Cancer Registry. J Surg Res 141:105–114. https://doi.org/10.1016/j.jss.2007.02.026
2. Zagars GK, Ballo MT, Pisters PWT et al (2003) Prognostic factors for patients with localized soft-tissue sarcoma treated with conservation surgery and radiation therapy: an analysis of 1225 patients. Cancer 97:2530–2543. https://doi.org/10.1002/cncr.11365
3. Pisters PW, Leung DH, Woodruff J, Shi W, Brennan MF (1996) Analysis of prognostic factors in 1,041 patients with localized soft tissue sarcomas of the extremities. J Clin Oncol 14:1679–1689. https://doi.org/10.1200/JCO.1996.14.5.1679

4. Ramanathan RC, A'Hern R, Fisher C, Thomas JM (1999) Modified staging system for extremity soft tissue sarcomas. Ann Surg Oncol 6:57–69
5. Maki RG, Moraco N, Antonescu CR et al (2013) Toward better soft tissue sarcoma staging: building on american joint committee on cancer staging systems versions 6 and 7. Ann Surg Oncol 20:3377–3383. https://doi.org/10.1245/s10434-013-3052-0
6. Suit HD, Mankin HJ, Wood WC et al (1988) Treatment of the patient with stage M0 soft tissue sarcoma. J Clin Oncol 6:854–862. https://doi.org/10.1200/JCO.1988.6.5.854
7. Edge SB, Compton CC (2010) The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 17:1471–1474. https://doi.org/10.1245/s10434-010-0985-4
8. Andrä C, Klein A, Dürr HR et al (2017) External-beam radiation therapy combined with limb-sparing surgery in elderly patients (>70 years) with primary soft tissue sarcomas of the extremities. Strahlenther Onkol 193:604–611. https://doi.org/10.1007/s00066-017-1109-x
9. Kattan MW, Leung DHY, Brennan MF (2002) Postoperative nomogram for 12-year sarcoma-specific death. J Clin Oncol 20:791–796. https://doi.org/10.1200/JCO.2002.20.3.791
10. Eilber FC, Brennan MF, Eilber FR et al (2004) Validation of the postoperative nomogram for 12-year sarcoma-specific mortality. Cancer 101:2270–2275. https://doi.org/10.1002/cncr.20570
11. Amin MB, Edge S, Greene F et al (eds) (2017) AJCC cancer staging manual, 8th edn. Springer, Cham
12. Tufman A, Kahnert K, Kauffmann-Guerrero D et al (2017) Clinical relevance of the M1b and M1c descriptors from the proposed TNM 8 classification of lung cancer. Strahlenther Onkol 193:392–401. https://doi.org/10.1007/s00066-017-1118-9
13. Abernethy AP, Etheredge LM, Ganz PA et al (2010) Rapid-learning system for cancer care. J Clin Oncol 28:4268–4274. https://doi.org/10.1200/JCO.2010.28.5478
14. Lambin P, van Stiphout RGPM, Starmans MHW et al (2013) Predicting outcomes in radiation oncology–multifactorial decision support systems. Nat Rev Clin Oncol 10:27–40. https://doi.org/10.1038/nrclinonc.2012.196
15. Wardelmann E, Haas RL, Bovée JVMG et al (2016) Evaluation of response after neoadjuvant treatment in soft tissue sarcomas; the European Organization for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group (EORTC-STBSG) recommendations for pathological examination and reporting. Eur J Cancer 53(021):84–95. https://doi.org/10.1016/j.ejca.2015.09.021
16. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
17. Zimmer VA, Glocker B, Hahner N et al (2017) Learning and combining image neighborhoods using random forests for neonatal brain disease classification. Med Image Anal 42:189–199. https://doi.org/10.1016/j.media.2017.08.004
18. Chen T, Cao Y, Zhang Y et al (2013) Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med. https://doi.org/10.1155/2013/298183
19. Liu M, Xu X, Tao Y, Wang X (2017) An improved random forest method based on RELIEFF for medical diagnosis. IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), pp 44–49
20. Kotti M, Duffell LD, Faisal AA, McGregor AH (2017) Detecting knee osteoarthritis and its discriminating parameters using random forests. Med Eng Phys 43:19–29. https://doi.org/10.1016/j.medengphy.2017.02.004
21. Rastgoo M, Lemaître G, More O et al (2016) Classification of melanoma lesions using sparse coded features and random forests. Proceedings Volume 9785, Medical Imaging 2016: Computer-Aided Diagnosis; 97850C. https://doi.org/10.1117/12.2216973
22. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21:631–643. https://doi.org/10.1093/bioinformatics/bti033
23. Eibe F, Hal MA, Ian H (2016) Witten the WEKA workbench. Online appendix for "data mining: practical machine learning tools and techniques". Morgan Kaufmann, Burlington.
24. Vallières M, Freeman CR, Skamene SR, El Naqa I (2015) 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 60:5471–5496. https://doi.org/10.1088/0031-9155/60/14/5471
25. Abeshouse A, Adebamowo C, Adebamowo SN, Akbani R, Akeredolu T, Ally A et al (2017) Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. Cell 171(8):950–965.e2. https://doi.org/10.1016/j.cell.2017.10.014
26. Sica GT (2006) Bias in research studies. Radiology 238:780–789. https://doi.org/10.1148/radiol.2383041109
27. Wulfkuhle JD, Liotta LA, Petricoin EF (2003) Proteomic applications for the early detection of cancer. Nat Rev Cancer 3:267–275. https://doi.org/10.1038/nrc1043
28. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696. https://doi.org/10.1038/nrg2841
29. Vallieres M, Kumar A, Sultanem K, El Naqa I (2013) FDG-PET image-derived features can determine HPV status in head-and-neck cancer. Int J Radiat Oncol Biol Phys 87:467. https://doi.org/10.1016/j.ijrobp.2013.06.1236
30. Deist TM, Jochems A, van Soest J et al (2017) Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004
31. Peeken JC, Nüsslin F, Combs SE (2017) Radio-oncomics. Strahlenther Onkol. https://doi.org/10.1007/s00066-017-1175-0
32. Wichmann H, Güttler A, Bache M et al (2014) Inverse prognostic impact of ErbB2 mRNA and protein expression level in tumors of soft tissue sarcoma patients. Strahlenther Onkol 190:912–918. https://doi.org/10.1007/s00066-014-0655-8