



CLINICAL ARTICLE

A Scoring System for Predicting Neoadjuvant Chemotherapy Response in Primary High-Grade Bone Sarcomas: A Multicenter Study

Fangzhou He, MD¹ , Lu Xie, MD¹, Xin Sun, MD¹, Jie Xu, MD¹, Yuan Li, MD², Rong Liu, MD², Kunkun Sun, MD³, Danhua Shen, MD³, Jin Gu, MD⁴, Tao Ji, MD¹, Wei Guo, PhD, MD¹ 

Department of ¹Musculoskeletal Tumor Center, ²Radiology and ³Pathology, Peking University People's Hospital and ⁴Department of Surgical Oncology, Peking University Shougang Hospital, Beijing, China

Objective: Currently, there is a lack of good clinical tools for evaluating the effect of chemotherapy preoperatively on primary high-grade bone sarcomas. Our goal was to investigate the predictive value of the clinical findings and establish a scoring system to predict chemotherapy response.

Methods: We conducted a retrospective multicenter cohort study and reviewed 322 patients with primary high-grade bone sarcomas. Patients who routinely received neoadjuvant chemotherapy and underwent primary tumor resection with an assessment of tumor necrosis rate (TNR) were enrolled in this study. The medical records of patients were collected from November 1, 2011, to March 1, 2018, at Peking University People's Hospital (PKUPH) and Peking University Shougang Hospital (PKUSH). The mean age of the patients was 16.2 years (range 3–52 years), of whom 65.5% were male. The clinical data collected before and after neoadjuvant chemotherapy included the degree of pain, laboratory inspection, X-ray, CT, contrast-enhanced magnetic resonance (MR), and positron emission tomography-computed tomography (PET-CT). Several machine learning models, including logistic regression, decision trees, support vector machines, and neural networks, were used to classify the chemotherapy responses. Area under the curve (AUC) of the scoring system to predict chemotherapy response is the primary outcome measure.

Results: For patients without events, a minimum follow-up of 24 months was achieved. The median follow-up time was 43.3 months, and it ranged from 24 to 84 months. The 5 years progression-free survival (PFS) of the included patients was 54.1%. The 5 years PFS rate was 39.7% for poor responders and 74.9% for good responders. Features such as longest diameter reduction ratio (up to three points), clear bone boundary formation (up to two points), tumor necrosis measured by magnetic resonance (up to two points), maximum standard uptake value (SUV_{max}) decrease (up to three points), and significant alkaline phosphatase decrease (up to 1 point) were identified as significant predictors of good histological response and constituted the scoring system. A score ≥ 4 predicts a good response to chemotherapy. The scoring system based on the above factors performed well, achieving an AUC of 0.893. For nonmeasurable lesions (classified by the revised Response Evaluation Criteria in Solid Tumors [RECIST 1.1]), the AUC was 0.901.

Conclusion: We first devised a well-performing comprehensive scoring system to predict the response to neoadjuvant chemotherapy in primary high-grade bone sarcomas.

Key words: Bone sarcomas; Chemotherapy response; Diagnosis support; Machine learning; Scoring system

Address for correspondence Wei Guo, PhD, MD, Musculoskeletal Tumor Center, Peking University People's Hospital, No. 11 Xizhimen South Street, Beijing 100044, China Email: bonetumor@163.com

Fangzhou He and Lu Xie authors contributed equally to this work.

Received 14 March 2021; accepted 25 July 2022



Introduction

The gold standard for evaluating the chemotherapy response of bone sarcomas is still pathologic examination.¹ However, the tumor necrosis rate (TNR) can be obtained only postoperatively; it cannot be used for preoperative evaluation of a patient's chemotherapy response. Nevertheless, preoperative evaluation of chemotherapy response is extremely valuable because it can be used to: (i) determine whether preoperative chemotherapy should be terminated or whether the chemotherapy regimen should be changed during neoadjuvant chemotherapy; (ii) assess the risk of local recurrence for limb-salvage surgery; and (iii) evaluate the therapeutic responses of unresectable tumors.² The significance of this study is to establish a preoperative model for predicting the TNR.

In 1979, the World Health Organization (WHO)³ proposed the concept of objective response, which was defined as 50% reduction in the product of perpendicular diameters of tumors. Clinical evaluation concepts of complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD) have become the standard criteria used to evaluate the efficacy of new agents in almost all solid tumors. Currently, the revised Response Evaluation Criteria in Solid Tumors (RECIST 1.1) is the most commonly used criterion for predicting histological response.⁴ However, existing problems have been raised regarding the scope of the revised RECIST. First, the intraosseous tumor volume changes only minimally with treatment. Therefore, the change in the longest diameter of the intraosseous tumor cannot represent the histological response of the tumor to chemotherapy. Second, intraosseous lesions and sclerotic lesions are thought of as being nonmeasurable, and only lytic bone lesions or mixed lytic-sclerotic lesions with soft tissue components ≥ 1 cm are evaluated to determine anatomic changes after therapy.⁴ Third, the inherent deficiency of this assessment method is that it does not consider morphological or metabolic changes in the treated target lesions.

Current researches on chemotherapy response assessment involves many fields. For sarcomas, alkaline phosphatase (ALP) and lactate dehydrogenase (LDH) in osteosarcoma,^{5,6} as well as LDH in Ewing's sarcoma,⁷ have been identified and reported as prognostic serum markers. In radiology, increased calcium deposition within the neoplastic bone and the presence of a calcified shell surrounding the tumor are considered to represent progressive healing and a good histological response.⁸ Contrast-enhanced MRI⁹ can help oncologists identify the necrotic fraction of sarcoma. The addition of gadopentetate dimeglumine to the static T1-weighted sequence enabled the separation of unenhanced necrosis and hemorrhage from enhanced inflammation and viable tumors.⁹ In recent years, some studies have revealed that decreased fluorine-18-fluorodeoxyglucose (18F-FDG) uptake after chemotherapy or radiotherapy is associated with a good pathological response.^{10,11} In positron emission tomography (PET), the response criteria in solid tumors (PERCIST 1.0)¹¹ requires a 30% decline in the standard

uptake value (SUV) for a response. At present, no method for chemotherapy response assessment considers all these factors. A meta-analysis¹² suggested that PET-CT, the most expensive imaging examination, had an AUC of 0.72–0.81 for predictive performance, which was still lower than clinical need. Therefore, a more comprehensive method is urgently needed.

Given all these clinical findings, we must consider how to analyze these data most effectively. Machine learning (ML) techniques are being increasingly widely used in different medical sciences to determine the most effective variables and predict outcomes, and experimental results have shown that these models not only improve the classification accuracy but also have strong universality.^{13,14} ML voting mechanisms can help establish a scoring system to predict the histological response to chemotherapy.

The purpose of this study was to: (i) establish a practical multi-factor scoring system for predicting the chemotherapy response to neoadjuvant chemotherapy in primary malignant bone sarcomas; and (ii) compare the prediction performance of different types of machine learning models.

Methods

Patient Selection Criteria

We retrospectively collected and reviewed the medical records of patients with primary high-grade bone sarcomas who received neoadjuvant chemotherapy from November 1, 2011, to March 1, 2018, at Peking University People's Hospital (PKUPH) and Peking University Shougang Hospital (PKUSH).

Patients were potentially included when they met the following criteria: (i) patients with histologically confirmed primary high-grade bone sarcomas; (ii) patients who routinely received neoadjuvant chemotherapy; (iii) patients who underwent primary tumor resection with an assessment of TNR according to the method reported by Rosen *et al.*;¹ and (iv) patients who had complete pre- and post-neoadjuvant chemotherapy imaging, including X-ray, CT, and contrast-enhanced MRI of the tumors as well as chest CT (with each layer ≤ 5 mm), bone scan, and either a PET-CT of the tumor site or whole-body PET-CT instead of the aforementioned 3 imaging modalities.

The exclusion criteria were as follows: (i) patients whose follow-up information or postoperative chemotherapy evaluation was unavailable; and (ii) patients with other fatal diseases. From the initial 358 eligible patients, those whose follow-up information or evaluation after postoperative chemotherapy was unavailable were excluded ($n = 36$); thus, 322 patients were included (270 from PKUPH and 52 from PKUSH).

Neoadjuvant Chemotherapy

Patients were treated according to PKUPH's chemoprotocols. Four-drug regimen¹⁵ (high-dose methotrexate [MTX], doxorubicin [ADM], cisplatin [CDP] and ifosfamide [IFO]) for

osteosarcoma and undifferentiated high-grade pleomorphic sarcoma (UPS). Five-drug regimen¹⁶ (vincristine [V], doxorubicin [A], cyclophosphamide [C], ifosfamide [I] and etoposide [E]) for Ewing sarcoma (ES).

Clinical and Imaging Data Collection

The clinical data collected before and after neoadjuvant chemotherapy included patient characteristics, disease features, clinical manifestations, images, and laboratory inspection.

Tumor Necrosis Rate

In this study, all pathology slides were reviewed by two senior pathologists who evaluated all surgical resection specimens but were blinded to the patient's clinical status. TNR was graded using tumor histopathological response grading¹ (Huvos classification), in which grade I was 0% to 49%, grade II was 50% to 89%, grade III was 90% to 99%, and grade IV was 100% necrosis. The response was classified as “poor” for grades I and II and “good” for grades III and IV.

Pain

The degree of pain was assessed by the visual analogue scale (VAS), which consists of a straight line ranging from 0–10 that represents pain intensity. Pain relief was defined as a lower VAS after the last cycle of chemotherapy than before the first cycle of chemotherapy.

Imaging

Images were obtained within 2 weeks prior to initiation of chemotherapy and 2 weeks after the last cycle of chemotherapy. The time interval between post-neoadjuvant chemotherapy imaging and tumor resection was less than 2 weeks. Two radiologists who were blinded to the study information and outcomes independently evaluated the images and recorded changes in the size of the target lesions pre- and post-therapy (Fig. 1). The longest axial diameters of the extraosseous tumor component were measured in extremities. For pelvic sarcomas or spinal sarcomas, the longest diameters of the tumor were measured. All measurements were measured on contrast-enhanced T1-weighted sequence and reported as the means.

Some specific phenomena were defined. Clear bone boundary formation was defined as the development of a fully sclerotic rim of lesions on X-ray or CT^{8,17} (Fig. 1). Increased CT density was defined as a 50% increase in the CT value (Hounsfield unit, Hu) over baseline after treatment,¹⁷ which reflected the degree of bone healing inside the tumor. Tumor necrosis measured by MR was described as a distinct nonenhanced solid area emerging within the lesion after chemotherapy (when no such area had existed before chemotherapy). If a significant area of necrosis was already observed on MR before chemotherapy, this item was no longer evaluated and assigned 0 point. The necrotic area shows a high signal on fat-saturated T2-weighted MR images and a lack of central enhancement

on postcontrast subtraction MR images and can be surrounded by irregular rim enhancement^{9,18} (Fig. 1). If the two radiologists reported different results, they would discuss together to form a consensus. In the case where no consensus was reached, a third expert was consulted.

PET-CT images were evaluated by the Department of Nuclear Medicine of the two hospitals. After image reconstruction, FDG uptake was measured as SUV_{max}. Protocols for image acquisition are described in Appendix S1.

Laboratory Inspection

ALP decrease ratio was calculated as $\frac{ALP_{pre} - ALP_{post}}{ALP_{pre}}$. ALP_{pre} refers to the ALP level before the first cycle of chemotherapy. ALP_{post} refers to the ALP level after the last cycle of chemotherapy. The calculation of LDH decrease ratio is the same as above. The factor ALP was only evaluated in patients with osteosarcoma.

The VAS, ALP, LDH, and SUV_{max} values were extracted from the official reports of the examinations.

Statistical Analysis

Figure S1 shows a flowchart of the model development and validation procedure. The full dataset was randomly split into a training set (75%) and a testing set (25%) used to evaluate and compare the performances of competing models. We adopted a supervised learning approach. Several classification models, including ML methods (logistic regression [LR], decision tree [DT], support vector machine [SVM], and neural network [NN]), were used to classify the histological response into two groups: “good” and “poor.” The metrics of prediction accuracy and the area under the receiver operating characteristic (ROC) curve (AUC) are useful for evaluating model performances.¹⁹ These metrics were calculated and evaluated by executing the trained models on the testing set.

LR analysis was used to determine the significant factors involved in the decision support system. The tolerance criteria for the multivariate analysis were 0.05 of α risk for admission and 0.1 of α risk for rejection. Stepwise selection in both directions was performed under these criteria. Here, locally weighted scatter-plot smoother (LOESS)²⁰ curves were examined to identify the cut points of continuous variables associated with notable changes in the risk of the outcome. The coefficient of the LR formula rounded off to the nearest integer can be assigned as the weight of the variables in a scoring system.²¹ The scores of each predictive factor were summed to produce a total score. The optimal threshold of the score to predict good response is the point closest to the top-left corner of the ROC plot with a perfect balance between sensitivity and specificity. After establishing the scoring system, it underwent performance evaluations similar to the other ML algorithms. This scoring system further underwent calibration evaluation with the testing set to evaluate the agreement between the predicted and observed risks.

Many factors, such as first-line chemotherapy, second-line treatments, surgery and malignant tumor grade, can

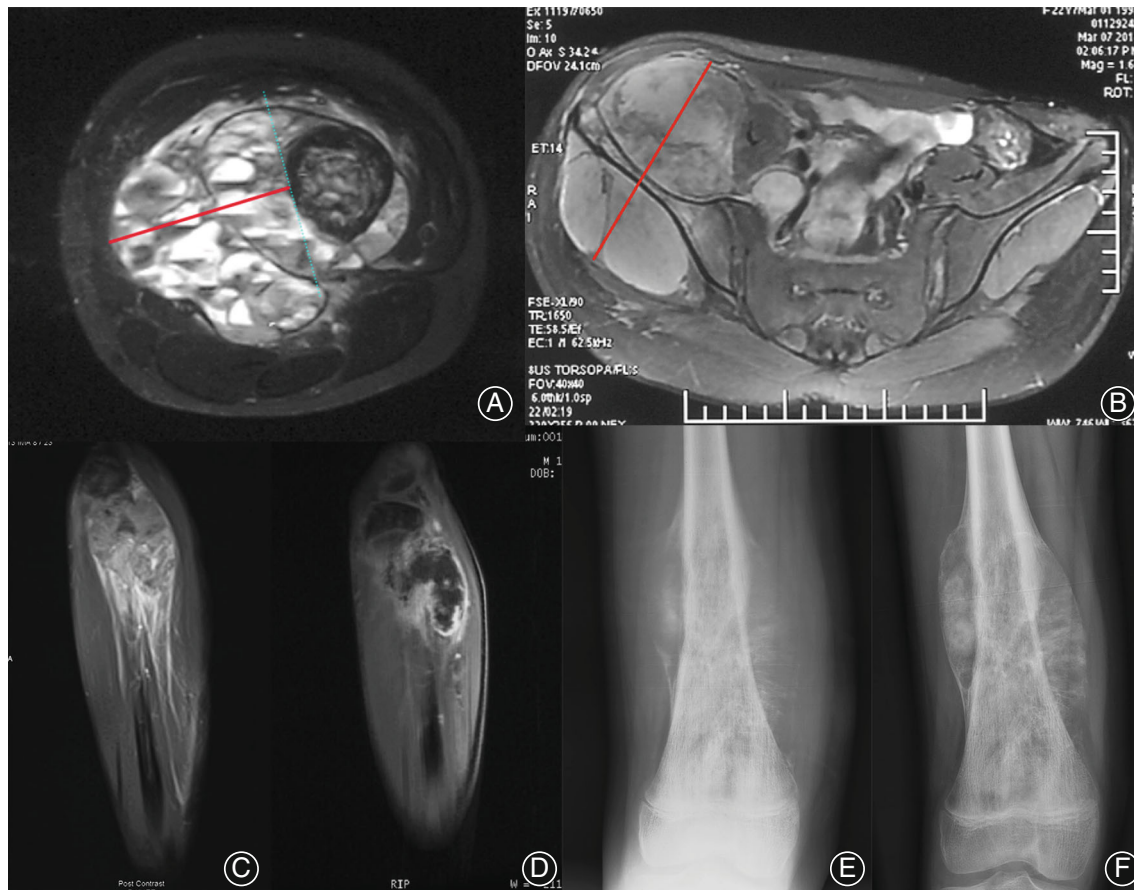


Fig. 1 Imaging assessment methods involved in this study. (A) Axial MR image of a distal femoral osteosarcoma. The red line indicates the longest axial diameter of the extraosseous tumor component at the distal femur. (B) Axial MR image of a pelvic osteosarcoma. The red line indicates the longest diameter of the pelvic osteosarcoma. (C, D) Pre- and post-contrast-enhanced MR images of proximal tibial osteosarcoma. Lack of central enhancement on postcontrast-subtraction MR images surrounded by irregular rim enhancement (D). (E, F) Pre- and post-contrast radiograph of a distal femoral osteosarcoma. Bone healing shows a smooth and clear edge on the image (F).

influence the overall survival of individual patients, while in this study, progression-free survival (PFS) was not influenced by second-line drug treatments or radiotherapy. For this purpose, PFS was selected as the endpoint of survival data. All data-mining tasks in this study were performed using R version 3.6.0.

Ethics Approval

Informed consent was obtained from all participants and/or their legal guardian(s). This study was approved by the ethical committee of Peking University People's Hospital (No. 2018PHB170-01) and Peking University Shougang Hospital (No. IRBK-2020-060-02).

Results

Clinical Characteristics and External Consistency

Of the 322 patients, 242 patients were randomly assigned to the training set, and 80 were assigned to the testing set. No

significant differences between the two sets were noted regarding the analyzed factors (Table 1). For patients without events, a minimum follow-up of 24 months was achieved. The median follow-up time was 43.3 months, and it ranged from 24 to 84 months. The 5 years PFS of the included patients was 54.1%. The 5 years PFS rate was 39.7% for poor responders and 74.9% for good responders (Logrank $P < 0.001$).

ML Models

The final predictive performance of each algorithm and the ROC curves of the ML methods validated on the testing set are shown in Fig. 2. The predictive performances of LR, SVM, and NN were similar (AUC = 0.899, 0.907, 0.893, respectively; LR vs SVM, $P = 0.5873$; LR vs NN, $P = 0.4291$. DeLong's test). Moreover, they outperformed DT (AUC of DT = 0.841; LR vs DT, $P = 0.0149$). Therefore, we assume that LR can satisfactorily reflect the overall contribution of variables. Because the coefficients of an LR formula can be

TABLE 1 Clinical characteristics and external consistency

Variable	Training set (mean ± SD or count)	Testing set (mean ± SD or count)	P value of no difference
Age	16.31 ± 7.87 year	15.79 ± 7.74 year	0.600
Gender			
Male	161	50	0.511
Female	81	30	
Location			
Lower extremity	175	62	0.643
Upper extremity	28	11	
Pelvis	31	6	
Spine	7	1	
Other(rib)	1	0	
Histological response			
Good	101	32	0.785
Poor	141	48	
Pathology			
Osteosarcoma	205	73	0.316
Ewing sarcoma	33	7	
UPS	4	0	
Initial Enneking stage			
IIA	10	4	0.332
IIB	182	64	
IIIB	50	12	
Pain relief			
Relieved	227	76	0.904
Not relieved	15	4	
Longest diameter reduction ratio	14.81 ± 61.13%	11.17 ± 69.56%	0.677
Clear bone boundary formation			
Yes	88	25	0.406
No	154	55	
Increased CT density			
Yes	86	30	0.751
No	156	50	
Tumor necrosis measured by MR			
Yes	50	21	0.296
No	192	59	
SUV _{max} decrease ratio	33.11 ± 36.32%	44.92 ± 28.81%	0.157
ALP decrease ratio	22.71 ± 45.91%	3.90 ± 130.13%	0.209
LDH decrease ratio	8.057 ± 57.09%	9.49 ± 47.20%	0.828

Notes: No significant differences between the training set and the testing set were noted regarding the analyzed factors. P value, t-test for numeric variables, chi-squared test for categorical variables, and Fisher's exact test for contingency table.; Abbreviations: ALP, alkaline phosphatase; LDH, lactate dehydrogenase; MR, magnetic resonance; SUV_{max}, maximum standard uptake value; UPS, undifferentiated high-grade pleomorphic sarcoma.

easily extracted, interpreted, and shared, we further performed LR variate analysis. The results of the best performing parameters and the tuning process can be found in the Appendix S2.

Scoring System

Weight of the Factors

Regarding the LR features, the longest diameter reduction ratio, clear bone boundary formation, tumor necrosis measured by MR, SUV_{max} decrease ratio, and ALP decrease ratio were identified as significant predictors of good histological response (Table S1).

Three of the five variables in the LR result were continuous. The cutoff values of the longest diameter reduction ratio were specified as -10%, 20%, and 50%, while the cutoff points of the SUV_{max} decrease ratio were 30% and 60% (Fig. S2). Due

to the results of the LR univariate analysis ($P = 0.0121$), we additionally used a post-chemotherapy SUV_{max} cutoff value of 2.5 to discriminate histological response. An ALP reduction to the normal range after chemotherapy could be considered significant (LR univariate analysis $P = 0.0480$). For patients whose ALP levels were too high to decline to a normal level, a decrease in ALP of more than 70% suggests a probability of a good histological response above 50% (Fig. S2). After relabeling the records, the scoring system was developed on the training set. Factors such as the longest diameter reduction ratio and SUV_{max} decrease had the highest weights, followed by clear bone boundary formation and tumor necrosis measured by MR. The categorical criteria and weights of the factors are shown in Table 2.

Predictive Performance

This weight-based scoring system is named Peking University Score (PKU Score). The ROC threshold point for the

PKU score is 3.5, which means that scores of 4 and higher indicate a good response. Subsequently, the AUC, overall

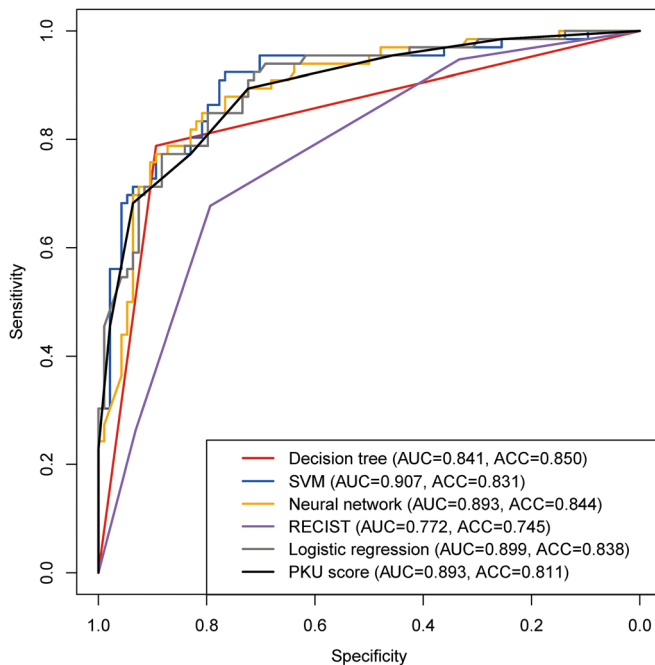


Fig. 2 The receiver operating characteristic (ROC) curves of the machine learning methods. Peking University (PKU) score and RECIST 1.1 were validated on the testing set, and the PKU score had a higher AUC than RECIST 1.1 (0.893 vs 0.772, $P < 0.001$). SVM, support vector machine; AUC, area under the curve; ACC, accuracy.

prediction accuracy, specificity and sensitivity of the scoring system on the testing set were 0.893, 0.836, 0.774 and 0.811, respectively. The PKU score had a higher AUC than RECIST 1.1 (0.893 vs 0.772, $P < 0.001$). The predictive performance of each individual score is shown in Table S2. The logistic calibration of the PKU score is close to the ideal line (Fig. 3). The Dxy and Brier scores of this scoring system are 0.797 and 0.122, respectively, suggesting a good match. Harrell's unreliability index $U = -0.004$ ($P = 0.526$) indicates that the model is not overfitted.

Clinical Evaluation Generated from PKU Score

A survival analysis of PFS was performed to connect the scores to the oncology outcome data. The post hoc log-rank analysis showed that patients with scores of 0, 1–3, or more had differences in disease progression (Fig. 4). Therefore, 0 points was classified as an unfavorable response (5 years PFS was 30.4%), 1–3 points was classified as a moderate response (5 years PFS was 43.5%), and four points or more was classified as a favorable response (5 years PFS was 74.6%; post hoc log-rank test: unfavorable vs moderate, $P = 0.017$; moderate vs favorable, $P < 0.001$). Although the purpose of this scoring system was not to assess survival time but to predict tumor pathological response, the results corroborated the correlation between survival outcome and chemotherapy response. Comparing the baseline characteristics of different PKU score groups, there was no significant difference in the initial Enneking stage (chi-square test P value 0.213, Table 3).

TABLE 2 Peking University score

Variable	Logistic regression coefficient	P value	Scores
Clear bone boundary formation			
No	Reference	—	0
Yes	1.902	<0.001	2
Tumor necrosis measured by MR			
No	Reference	—	0
Yes	1.962	<0.001	2
Longest diameter reduction ratio			
$\leq -10\%$	Reference	—	0
$> -10\%$ and $\leq 20\%$	0.823	0.170	1
$> 20\%$ and $\leq 50\%$	1.732	0.00290	2
$> 50\%$	2.801	<0.001	3
Significant ALP decrease ^a			
No	Reference	—	0
$> 70\%$ or decrease to normal level	0.826	0.0408	1
SUV _{max} decrease			
$\leq 30\%$ and post-chemotherapy SUV _{max} > 2.5	Reference	—	0
$> 30\%$, $\leq 60\%$ and post-chemotherapy SUV _{max} > 2.5	0.59	0.295	1
$> 60\%$ or post-chemotherapy SUV _{max} ≤ 2.5	3.181	<0.001	3

Notes: The coefficient of the logistic regression formula rounded off to the nearest integer can be assigned as the weight of the variables in a scoring system. The scores of each predictive factor were summed to produce a total score: zero points, unfavorable response; 1–3 points, moderate response; and ≥ 4 points, favorable response.; Abbreviations: ALP, alkaline phosphatase; MR, magnetic resonance; SUV_{max}, maximum standard uptake value.; ^aALP was only evaluated in patients with osteosarcoma.

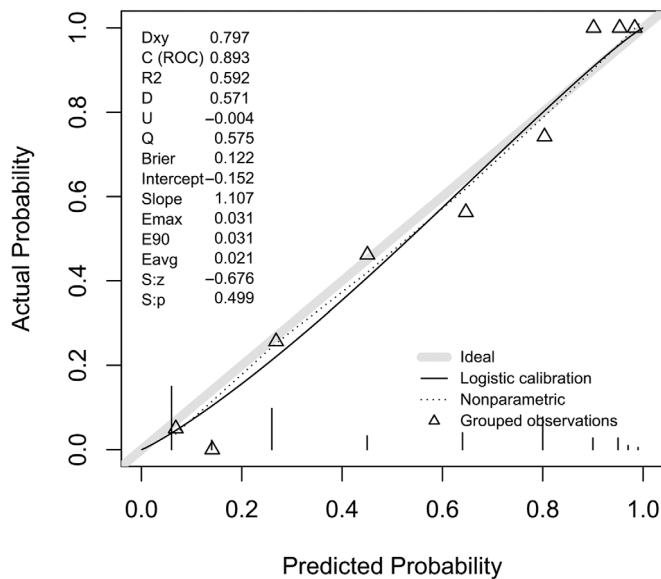


Fig. 3 The calibration plot evaluates the agreement between predicted and observed risks. The predicted probability using the testing set data is plotted against the observed probability, and the deviation from the ideal line indicates the difference between the predicted and observed risks. Somers' Dxy rank correlation measures how two pairs of variables are connected ($-1 =$ all pairs disagree, $1 =$ all pairs agree). The Brier score measures the total difference between the event and the forecast probability of that event as an average squared difference ($0 =$ perfect forecaster, $1 =$ perfect misforecaster). The Dxy and Brier scores of this scoring system are 0.797 and 0.122, respectively, suggesting a good match. Harrell's unreliability index $U = -0.004$ ($P = 0.526$) indicates that the model is not overfitted. ROC, receiver operating characteristic.

Traditional Nonmeasurable Lesions

A total of 38 nonmeasurable tumors (classified by RECIST 1.1) included 13 intraosseous-only, 24 sclerotic-only tumors, and one combined tumor. The variant analysis results showed that clear bone boundary formation and SUV_{max} decrease were associated with a good response (P values of 0.0492 and <0.001 , respectively). The prediction accuracy and the AUC of the PKU score for nonmeasurable lesions were 0.842 and 0.901, respectively.

Discussion

Main Findings of the Study

We found it available to use a weight-based scoring system to preoperatively assess the response of primary malignant bone tumors to chemotherapy. Features such as longest diameter reduction ratio, clear bone boundary formation, tumor necrosis measured by magnetic resonance, SUV_{max} decrease, and significant alkaline phosphatase decrease were identified as significant predictors of good histological response and constituted the scoring system.

Limitations of Current Tools

Much effort has been directed toward the development of standardized and reproducible methods for evaluating the response of tumors over the past several decades.²²⁻²⁴ Currently, the most commonly used criteria are the revised RECIST guidelines (version 1.1).⁴ These anatomic criteria focus predominantly on unidimensional physical measurement of solid tumors. These criteria provide highly concordant response assessments compared with bidimensional measurements and are also more reproducible. RECIST 1.1 guarantees reliability, but it is not integrated with clinical presentation and new imaging techniques. In addition, RECIST 1.1 is beneficial but not for intraosseous or sclerotic bone sarcomas.⁴ Therefore, a comprehensive evaluation system that can predict the treatment response of bone sarcomas is urgently needed. Based on the RECIST 1.1, PERCIST, and MDA criteria, we focused on patients with lesions that originated in bones who received neoadjuvant chemotherapy. We reviewed the relevant studies^{5,12,25,26} and evaluated a number of morphologic and metabolic parameters to validate the clinical prediction and standardized methods of measuring lesions.

Based on clinical experience, we generally think that patients who respond well to chemotherapy will achieve significant pain relief, but this study shows that the pain of patients with poor response will also be reduced. Pain relief depends on receiving chemotherapy rather than chemotherapy response.

The Longest Axial Diameters of the Extraosseous Tumor Component

In primary sarcomas of long bone, in most cases, the longitudinal intramedullary extent never decreases with chemotherapy, while the axial dimension of the extraosseous tumor component shows reduction after chemotherapy.^{27,28} Therefore, the longest axial diameters of the extraosseous tumor component were measured in the extremities. For pelvic sarcomas or spinal sarcomas, because the shape of the bone is irregular, the longest diameters of the tumor were measured. The mean pre-treatment longest axial diameter of the extraosseous tumor component was 3.1 cm (range, 0.48–15 cm). There were 14 cases of tumors with longest diameter more than 8 cm, of which nine cases (64%) had a necrosis rate $<90\%$ and five cases (36%) had a necrosis rate greater than or equal to 90%. There were 308 cases of tumors with longest diameter less than or equal to 8 cm, of which 180 cases (58%) had a necrosis rate $<90\%$ and 128 cases (42%) had a necrosis rate greater than or equal to 90%. The chi-square test showed no significant difference in the necrosis rate between the two groups ($P = 0.875$). We did not observe a higher rate of necrosis in large tumors. In this study, we observed some contradictions that although the majority of tumors that had a good response to chemotherapy showed a decrease in volume (91%, 121/133), there were also tumors that had a good response to chemotherapy that increased in size by more than 10% (6.8%, 9/133).

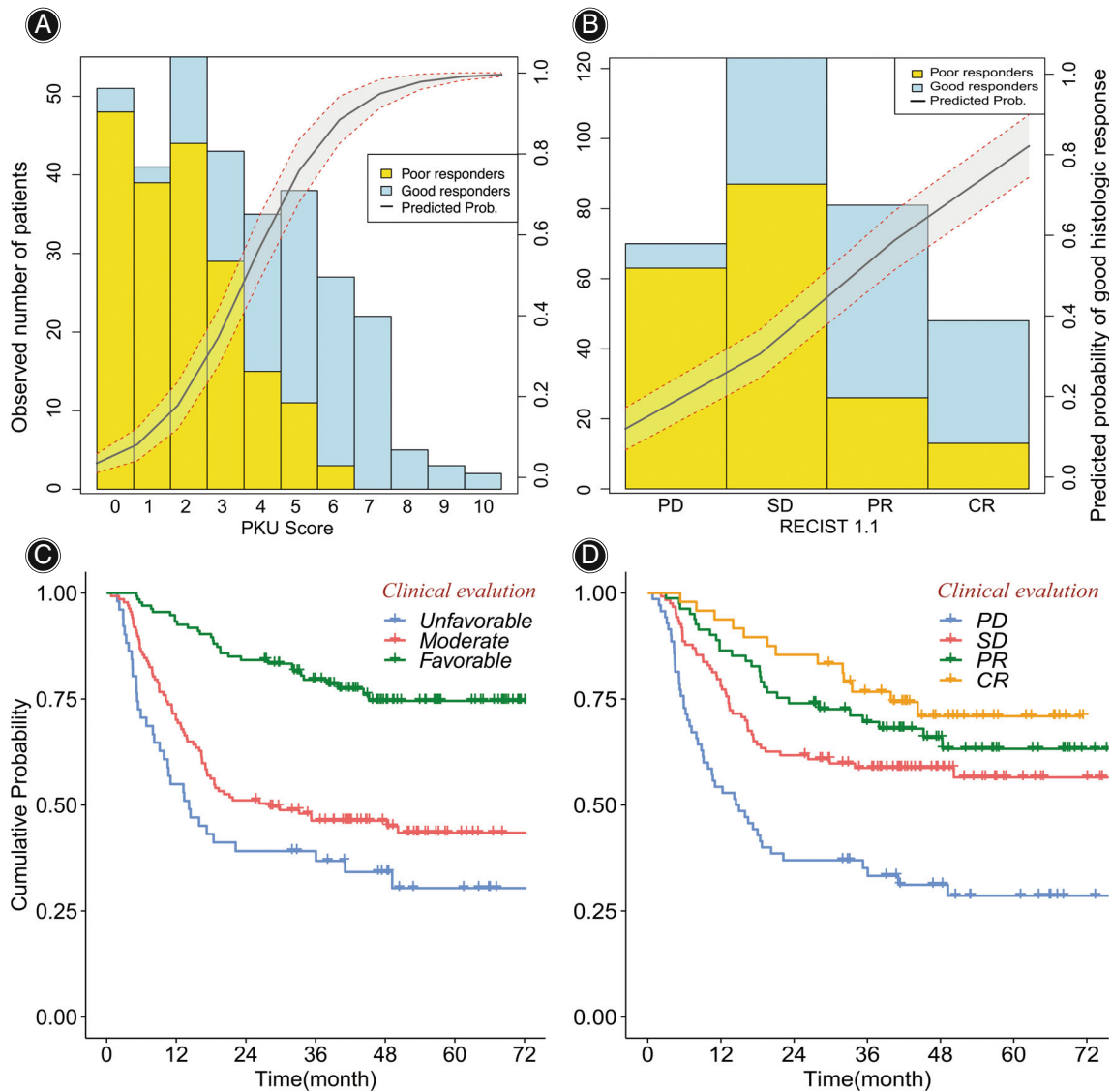


Fig. 4 Distribution and survival analysis of patients grouped by Peking University (PKU) score or RECIST 1.1. (A, B) Bar plot of histological response grouped by PKU score (A) or RECIST 1.1 criteria (B). The probability of a good chemotherapy response increases with PKU score, forming a steep S-shaped curve with better discrimination than RECIST 1.1. (C, D) Kaplan–Meier estimate of progression-free survival (PFS) of the entire cohort by clinical evaluation generated from PKU score (C) or RECIST 1.1 (D). Post hoc log-rank analysis showed that patients with scores of 0, 1–3, or higher had differences in disease progression (post hoc log-rank test: unfavorable versus moderate, $p = 0.017$; moderate versus favorable, $P < 0.001$). In addition, it was difficult to distinguish PFS from disease status evaluated by RECIST 1.1 (complete response (CR) vs partial response (PR), $P = 0.334$; PR vs stable disease (SD), $P = 0.160$; SD vs progressive disease (PD), $P < 0.001$).

This phenomenon may be related to the metabolic activity of tumor necrosis.

Clear Bone Boundary Formation

The typical pattern of response to chemotherapy in osteosarcoma is the increase in bone matrix production leading to diffuse calcification of the tumor mass. However, increased

CT density of the lesion sometimes fails to distinguish between neoplastic osteogenesis and bone healing: bone healing shows smooth and clear edges on an X-ray image, while the edges of neoplastic osteogenesis are irregular. Neoplastic osteogenesis means that osteosarcoma cells are still active. Therefore, although both increased CT density and clear bone boundary formation were significant in the

TABLE 3 Baseline characteristics of different PKU score groups

Initial Enneking stage	PKU score groups		
	Unfavorable response (0)	Moderate response (1–3)	Favorable response (≥4)
IIA	2	5	7
IIB	34	100	113
IIIB	15	26	20

Abbreviation: PKU, Peking University.

univariate analysis, clear bone boundary formation performed better in the multivariate analysis. The interobserver agreement rate between the two radiologists for evaluating clear bone boundary formation was 97.5% (314/322).

Tumor Necrosis Measured by MR

MR is the most sensitive technique for detecting marrow-based lesions. However, MRI does not provide sufficient information on the degree of tumor viability, the clinical parameter determining tumor response and prognosis.²⁹ This led to the investigation of dynamic MRI. Because the application of dynamic contrast-enhanced MRI is far less than regular static contrast-enhanced MRI, static contrast-enhanced MRI is still our most important clinical examination. Hao *et al.*³⁰ studied chemotherapy-induced MRI changes, suggesting that changes in tumor volume, peritumoral edema, boundaries around the extraosseous component, and hemorrhage were associated with chemotherapy. Amit *et al.*³¹ reported that the nonenhanced area on contrast-enhanced MRI had a high correlation with histological necrosis. In our experience, although the overall accuracy of tumor necrosis measured by MR in predicting response to chemotherapy is not satisfactory (66.5%, 214/322), it is a factor of low sensitivity (36.1%, 48/133) but high specificity (87.8%, 166/189). This means that tumors with a good response to chemotherapy may not be distinguishable by this method, but tumors with this imaging phenomenon are very likely to be necrotic. Such high specificity can also contribute to a comprehensive evaluation system. The interobserver agreement rate between the two radiologists for this factor was 96.9% (312/322).

Positron Emission Tomography-Computed Tomography

PERCIST 1.0¹¹ were introduced in 2009 as guidelines for systematic and structured assessment of response to therapy with fluorine 18 fluorodeoxyglucose (FDG) PET in patients with cancer, with suggested application in clinical trials and, potentially, in the clinical practice of PET reporting. We adopted the cutoff points of the SUV_{max} decrease ratio of 30% and 60% in our LOESS curve. The cutoff point of 30% was also recommended by PERCIST. Due to the results of the LR analysis (univariate regression $P = 0.0121$), we additionally used a post-chemotherapy SUV_{max} cutoff value of 2.5 to discriminate histological response. The same cutoff value was also adopted by previous studies.^{10,12,32}

Alkaline Phosphatase

ALP has been shown to be produced directly by human osteosarcoma cells, and its level can be increased in patients with osteosarcoma.³³ Given the disagreement regarding ALP as a prognostic indicator for osteosarcoma in previous studies, recent studies have reassessed the enzyme as a tumor marker and prognostic predictor, and these rigorous studies have reaffirmed the predictive value of elevated ALP for poor prognosis in patients with osteosarcoma.^{5,34–36} In our study, ALP was a predictive factor of good response, but the weight was lower than that of other factors. ALP is generally not considered a prognostic factor for ES and UPS. In this study, there were 44 cases of these two types of tumors, and 42 out of the 44 cases had an ALP score of zero. Therefore, in patients without osteosarcoma, we did not evaluate ALP, and their ALP scores were adjusted to zero.

Traditional Nonmeasurable Lesions

Our scoring system also performed well in assessing non-measurable lesions (classified by RECIST 1.1). The longest diameter reduction ratio of intraosseous lesions could not be evaluated and was assigned 0 points. The contrast-enhanced MR image of sclerotic osteosarcoma exhibited a low signal; therefore, even when necrosis was present, it was difficult to identify from MR images. Of all the factors, only clear bone boundary formation and SUV_{max} decrease elicited a good response. This finding indicates that PET-CT is an essential tool when evaluating the chemotherapy response to sclerotic osteosarcoma.

Heterogeneity

Another potential problem is that the chemotherapy-related imaging changes between osteosarcoma and Ewing sarcoma are not identical. We used 278 patients with osteosarcoma as a data set and trained the prediction model, which resulted in a score of 2 for clear bone boundary formation, 2 for tumor necrosis measured by MR, 3 for longest diameter reduction ratio, 1 for a significant reduction in ALP, and 4 for a decrease in SUV_{max} . SUV_{max} has a slightly higher in the PKU score (for osteosarcoma dataset). The performance parameters of the PKU score for predicting the response to chemotherapy in osteosarcoma, such as the cutoff value, AUC, overall accuracy, specificity, and sensitivity, were 3.5,

0.8873, 0.8165468, 0.8373, and 0.7857, respectively. Weighting and prediction performance were observed to be similar between the system for the osteosarcoma dataset and the overall dataset. This study included 40 cases of Ewing sarcoma, and the model of Ewing sarcoma was trained using the same procedure, which resulted in a score of 2 for clear bone boundary formation, 1 for tumor necrosis measured by MR, 4 for the longest diameter reduction ratio, and 3 for a decrease in SUV_{max} . The change in the longest axial diameter of the extraosseous tumor component seemed to be weighted the most. However, because the sample size is small, barely meeting the minimum requirement of 10 times the sample size of the variables, the system is not stable enough. The performance parameters of the PKU score for predicting the response to chemotherapy in Ewing sarcoma, such as the cutoff value, AUC, overall accuracy, specificity, and sensitivity, were 3.5, 0.8008, 0.75, 0.8095238, and 0.6842, respectively. The performance parameters in the Ewing sarcoma dataset were slightly lower than those in the OS dataset but still higher or similar to those reported in other studies (AUC was 0.807 in Saleh *et al.*'s study³⁷ on dynamically enhanced MRI; AUC was 0.750 in El-Hennawy *et al.*'s study³⁸ on FDG PET).

Comparison between ML Models

Logistic regression is a linear model for classification. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. SVMs are a set of supervised learning methods used for classification. The main advantage of SVMs is still effective in cases where number of dimensions is greater than the number of samples. A neural network is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. The main advantages of NN is capability to learn non-linear models. Some advantages of DT are: (i) Using a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained; (ii) requiring little data preparation. The disadvantages of DT include: (i) can be unstable because small variations in the data might result in a completely different tree being generated; and (ii) can create over-complex trees that do not generalize the data well. This is called overfitting.

The predictive performances of LR, SVM, and NN were similar (AUC = 0.899, 0.907, 0.893, respectively; LR vs SVM, $P = 0.5873$; LR vs NN, $P = 0.4291$. DeLong's test.). Moreover, they outperformed DT (AUC of DT = 0.841; LR vs DT, $P = 0.0149$). The results show that the scoring system have a good predictive performance and linear fit. There is no clear advantage to more complex algorithms (SVM and NN).

Limitations

However, our study still has several limitations. First, it is a retrospective study with inevitable selection biases. Second, due to the rarity of bone sarcomas and the requirement for

complete PET-CT examinations, the number of patients was relatively small. Our ongoing prospective study will validate the predictive values of this system in a large number of patients. In future studies, we hope that larger prospective multicenter studies could be developed to test and promote the application of the prediction scoring system.

Conclusion

We first devised a comprehensive scoring system to predict the response to neoadjuvant chemotherapy preoperatively in primary high-grade bone sarcomas. This new scoring system considers all the clinical exam results and performed well in estimating patient chemotherapy response. Our proposed system can also assess the response of traditional non-measurable lesions (classified by RECIST 1.1).

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 82072970) and Medical Service and Security Capacity Enhancement Project (No. 2199000730).

Supporting Information

Additional Supporting Information may be found in the online version of this article on the publisher's web-site:

Appendix S1 Supporting Information

Fig. S1 Flowchart of the model development and validation procedure. The full dataset was split into a model training set (75%) and a testing set (25%) that were used for evaluating and comparing the performances of competing models. Random sampling was performed as an attempt to balance the class distributions of the two levels of histological responses. For each selected technique, we tested a series of values for the tuning process with the optimal parameters determined based on the model performance. The model training set was further randomly divided into 10 subsets. The holdout method was repeated 10 times with different datasets (10-fold cross-validation (CV)). The error estimations from three repeated 10-fold CVs were taken and averaged to give the final error estimation of the model. The coefficients of the factors in the logistic regression were then converted into scoring weights. Discrimination and calibration evaluation of the Peking University (PKU) score were performed on the testing set.

Fig. S2 Locally weighted scatter-plot smoother (LOESS) curve of the continuous variables. LOESS curves were examined to identify the cutoff points of continuous variables associated with notable changes in the probability of a good response.

TABLE S1 Logistic regression on the training set.

TABLE S2 Predictive performance table of the Peking University (PKU) score for the full set.

References

1. Rosen G, Caparros B, Huvos AG, Kosloff C, Nirenberg A, Cacavio A, et al. Preoperative chemotherapy for osteogenic sarcoma: selection of postoperative adjuvant chemotherapy based on the response of the primary tumor to preoperative chemotherapy. *Cancer*. 1982;49(6):1221–30. [https://doi.org/10.1002/1097-0142\(19820315\)49:6<1221::aid-cnrcr2820490625>3.0.co;2-e](https://doi.org/10.1002/1097-0142(19820315)49:6<1221::aid-cnrcr2820490625>3.0.co;2-e)
2. Isakoff MS, Bielack SS, Meltzer P, Gorlick R. Osteosarcoma: current treatment and a collaborative pathway to success. *J Clin Oncol*. 2015;33(27):3029–35. <https://doi.org/10.1200/JCO.2014.59.4895>
3. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976;38(1):388–94. [https://doi.org/10.1002/1097-0142\(197607\)38:1<388::aid-cnrcr2820380156>3.0.co;2-a](https://doi.org/10.1002/1097-0142(197607)38:1<388::aid-cnrcr2820380156>3.0.co;2-a)
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz L, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–47. <https://doi.org/10.1016/j.ejca.2008.10.026>
5. Ren HY, Sun LL, Li HY, Ye ZM. Prognostic significance of serum alkaline phosphatase level in osteosarcoma: a meta-analysis of published data. *Biomed Res Int*. 2015;2015:160835. <https://doi.org/10.1155/2015/160835>
6. Meyers PA, Heller G, Healey J, Huvos A, Lane J, Marcove R, et al. Chemotherapy for nonmetastatic osteogenic sarcoma: the memorial Sloan-Kettering experience. *J Clin Oncol*. 1992;10(1):5–15. <https://doi.org/10.1200/JCO.1992.10.1.5>
7. Cotterill SJ, Ahrens S, Paulussen M, Jürgens H, Voüte P, Gadner H, et al. Prognostic factors in Ewing's tumor of bone: analysis of 975 patients from the European intergroup cooperative Ewing's sarcoma study group. *J Clin Oncol*. 2000;18(17):3108–14. <https://doi.org/10.1200/JCO.2000.18.17.3108>
8. Shirkhoda A, Jaffe N, Wallace S, Ayala A, Lindell MM, Zomoza J. Computed tomography of osteosarcoma after intraarterial chemotherapy. *Am J Roentgenol*. 1985;144(1):95–9. <https://doi.org/10.2214/ajr.144.1.95>
9. de Baere T, Vanel D, Shapeero LG, Charpentier A, Terrier P, di Paola M. Osteosarcoma after chemotherapy: evaluation with contrast material-enhanced subtraction MR imaging. *Radiology*. 1992;185(2):587–92. <https://doi.org/10.1148/radiology.185.2.1410378>
10. Hamada K, Tomita Y, Inoue A, Fujimoto T, Hashimoto N, Myoui A, et al. Evaluation of chemotherapy response in osteosarcoma with FDG-PET. *Ann Nucl Med*. 2009;23(1):89–95. <https://doi.org/10.1007/s12149-008-0213-5>
11. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50(Suppl 1):122S–50S. <https://doi.org/10.2967/jnumed.108.057307>
12. Li H, Zhao H, Wang B, Xiaojin W, Zhiyu W, Shuier Z, et al. 18F-FDG positron emission tomography for the assessment of histological response to neoadjuvant chemotherapy in osteosarcomas: a meta-analysis. *Surg Oncol*. 2012;21(4):e165–70. <https://doi.org/10.1016/j.suronc.2012.07.002>
13. Xiong XL, Zhang RX, Bi Y, Zhou WH, Yu Y, Zhu DL. Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese adults. *Curr Med Sci*. 2019;39(4):582–8. <https://doi.org/10.1007/s11596-019-2077-4>
14. Bucholz M, Ding X, Wang H, Glass DH, Wang H, Prasad G, et al. A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Syst Appl*. 2019;130:157–71. <https://doi.org/10.1016/j.eswa.2019.04.022>
15. Xu J, Xie L, Guo W. Neoadjuvant chemotherapy followed by delayed surgery: is it necessary for all patients with nonmetastatic high-grade pelvic osteosarcoma? *Clin Orthop Relat Res*. 2018;476(11):2177–86. <https://doi.org/10.1097/CORR.0000000000000387>
16. Guo W, Tang X, Tang S, Yang Y. Preliminary report of combination chemotherapy including arsenic trioxide for stage III osteosarcoma and Ewing sarcoma. *Zhonghua Wai Ke Za Zhi*. 2006;44(12):805–8.
17. Hamaoka T, Madewell JE, Podoloff DA, Hortobagyi GN, Ueno NT. Bone imaging in metastatic breast cancer. *J Clin Oncol*. 2004;22(14):2942–53. <https://doi.org/10.1200/JCO.2004.08.181>
18. He F, Qin L, Bao Q, et al. Pre-operative chemotherapy response assessed by contrast-enhanced MRI can predict the prognosis of Enneking surgical margins in patients with osteosarcoma. *J Orthop Res*. 2019;37(1):258–64. <https://doi.org/10.1002/jor.24143>
19. Carter JV, Pan J, Rai SN, Galanduk S. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. *Surgery*. 2016;159(6):1638–45. <https://doi.org/10.1016/j.surg.2015.12.029>
20. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*. 1979;74(368):829–36. <https://doi.org/10.1080/01621459.1979.10481038>
21. Zhang Z, Zhang H, Khanal MK. Development of scoring system for risk stratification in clinical medicine: a step-by-step tutorial. *Ann Transl Med*. 2017;5(21):436. <https://doi.org/10.21037/atm.2017.08.22>
22. Pinker K, Riedl C, Weber WA. Evaluating tumor response with FDG PET: updates on PERCIST, comparison with EORTC criteria and clues to future developments. *Eur J Nucl Med Mol Imaging*. 2017;44(Suppl 1):55–66. <https://doi.org/10.1007/s00259-017-3687-3>
23. Gupta S, Kang HC, Sun J, Matrana MR, Tannir NM, Choi H. Tumor response criteria for assessing pazopanib first-line therapy in patients with metastatic renal cell carcinoma (mRCC). *Abdom Radiol*. 2020;45(6):1872–82. <https://doi.org/10.1007/s00261-019-02354-z>
24. Ashikyan O, Bradshaw SB, Dettori NJ, Hwang H, Chhabra A. Conventional and advanced MR imaging insights of synovial sarcoma. *Clin Imaging*. 2021;76:149–55. <https://doi.org/10.1016/j.clinimag.2021.02.010>
25. Kelleher FC, O'Sullivan H. Monocytes, macrophages, and osteoclasts in osteosarcoma. *J Adolesc Young Adult Oncol*. 2017;6(3):396–405. <https://doi.org/10.1089/jayao.2016.0078>
26. Ashikyan O, Bradshaw SB, Dettori NJ, Hwang H, Chhabra A. Conventional and advanced MR imaging insights of synovial sarcoma. *Clin Imaging*. 2021;76:149–55. <https://doi.org/10.1016/j.clinimag.2021.02.010>
27. Saifuddin A, Sharif B, Gerrand C, Whelan J. The current status of MRI in the pre-operative assessment of intramedullary conventional appendicular osteosarcoma. *Skeletal Radiol*. 2019;48(4):503–16. <https://doi.org/10.1007/s00256-018-3079-1>
28. van der Woude HJ, Bloem JL, Hogendoorn PCW. Preoperative evaluation and monitoring chemotherapy in patients with high-grade osteogenic and Ewing's sarcoma: review of current imaging modalities. *Skeletal Radiol*. 1998;27(2):57–71. <https://doi.org/10.1007/s002560050339>
29. Errani C, Kreshak J, Ruggieri P, Alberghini M, Picci P, Vanel D. Imaging of bone tumors for the musculoskeletal oncologic surgeon. *Eur J Radiol*. 2013;82(12):2083–91. <https://doi.org/10.1016/j.ejrad.2011.11.034>
30. Hao Y, An R, Xue Y, et al. Prognostic value of tumoral and peritumoral magnetic resonance parameters in osteosarcoma patients for monitoring chemotherapy response. *Eur Radiol*. 2021;31(5):3518–29. <https://doi.org/10.1007/s00330-020-07338-y>
31. Amit P, Malhotra A, Kumar R, Kumar L, Patro DK, Elangovan S. Evaluation of static and dynamic MRI for assessing response of bone sarcomas to preoperative chemotherapy: correlation with histological necrosis. *Indian J Radiol Imaging*. 2015;25(3):269–75. <https://doi.org/10.4103/0971-3026.161452>
32. Raciborska A, Bilska K, Drabko K, Michalak E, Chaber R, Pogorzala M, et al. Response to chemotherapy estimates by FDG PET is an important prognostic factor in patients with Ewing sarcoma. *Clin Transl Oncol*. 2016;18(2):189–95. <https://doi.org/10.1007/s12094-015-1351-6>
33. Agustina H, Asyifa I, Aziz A, Hernowo BS. The role of osteocalcin and alkaline phosphatase immunohistochemistry in osteosarcoma diagnosis. *Patholog Res Int*. 2018;2018:6346409. <https://doi.org/10.1155/2018/6346409>
34. Kim SH, Shin KH, Moon SH, Jang J, Kim HS, Suh J, et al. Reassessment of alkaline phosphatase as serum tumor marker with high specificity in osteosarcoma. *Cancer Med*. 2017;6(6):1311–22. <https://doi.org/10.1002/cam4.1022>
35. Hao H, Chen L, Huang D, Ge J, Qiu Y, Hao L. Meta-analysis of alkaline phosphatase and prognosis for osteosarcoma. *Eur J Cancer Care*. 2017;26(5):e12536. <https://doi.org/10.1111/ecc.12536>
36. de Rodrigues LCS, Holmes KE, Thompson V, Piskun C, Lana S, Newton M, et al. Osteosarcoma tissues and cell lines from patients with differing serum alkaline phosphatase concentrations display minimal differences in gene expression patterns. *Vet Comp Oncol*. 2016;14(2):e58–69. <https://doi.org/10.1111/vco.12132>
37. Saleh MM, Abdelrahman TM, Madney Y, Mohamed G, Shokry AM, Moustafa AF. Multiparametric MRI with diffusion-weighted imaging in predicting response to chemotherapy in cases of osteosarcoma and Ewing's sarcoma. *Br J Radiol*. 2020;93(1115):20200257. <https://doi.org/10.1259/bjr.20200257>
38. El-Hennawy G, Moustafa H, Omar W, et al. Different (18) F-FDG PET parameters for the prediction of histological response to neoadjuvant chemotherapy in pediatric Ewing sarcoma family of tumors. *Pediatr Blood Cancer*. 2020;67(11):e28605. <https://doi.org/10.1002/psc.28605>