European Radiology (2019) 29:5441–5451
https://doi.org/10.1007/s00330-019-06082-2

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Convolutional neural network to predict the local recurrence of giant cell tumor of bone after curettage based on pre-surgery magnetic resonance images

Yifeng He[1,2] · Jiapan Guo[3] · Xiaoyi Ding[2] · Peter M. A. van Ooijen[3] · Yaping Zhang[1] · An Chen[1] · Matthijs Oudkerk[3] · Xueqian Xie[1]

Received: 20 November 2018 / Revised: 24 January 2019 / Accepted: 7 February 2019 / Published online: 11 March 2019

Abstract

Objective To predict the local recurrence of giant cell bone tumors (GCTB) on MR features and the clinical characteristics after curettage using a deep convolutional neural network (CNN).

Methods MR images were collected from 56 patients with histopathologically confirmed GCTB after curettage who were followed up for 5.8 years (range, 2.0 to 9.5 years). The inception v3 CNN architecture was fine-tuned by two categories of the MR datasets (recurrent and non-recurrent GCTB) obtained through data augmentation and was validated using fourfold cross-validation to evaluate its generalization ability. Twenty-eight cases (50%) were chosen as the training dataset for the CNN and four radiologists, while the remaining 28 cases (50%) were used as the test dataset. A binary logistic regression model was established to predict recurrent GCTB by combining the CNN prediction and patient features (age and tumor location). Accuracy and sensitivity were used to evaluate the prediction performance.

Results When comparing the CNN, CNN regression, and radiologists, the accuracies of the CNN and CNN regression models were 75.5% (95% CI 55.1 to 89.3%) and 78.6% (59.0 to 91.7%), respectively, which were higher than the 64.3% (44.1 to 81.4%) accuracy of the radiologists. The sensitivities were 85.7% (42.1 to 99.6%) and 87.5% (47.3 to 99.7%), respectively, which were higher than the 58.3% (27.7 to 84.8%) sensitivity of the radiologists ($p < 0.05$).

Conclusion The CNN has the potential to predict recurrent GCTB after curettage. A binary regression model combined with patient characteristics improves its prediction accuracy.

Key Points

- *Convolutional neural network (CNN) can be trained successfully on a limited number of pre-surgery MR images, by fine-tuning a pre-trained CNN architecture.*
- *CNN has an accuracy of 75.5% to predict post-surgery recurrence of giant cell tumors of bone, which surpasses the 64.3% accuracy of human observation.*
- *A binary logistic regression model combining CNN prediction rate, patient age, and tumor location improves the accuracy to predict post-surgery recurrence of giant cell bone tumors to 78.6%.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06082-2) contains supplementary material, which is available to authorized users.

✉ Xueqian Xie
xiexueqian@hotmail.com

[1] Radiology Department, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, HaiNing Rd.100, Shanghai 200080, China

[2] Radiology Department, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine, RuiJin No.2 Rd.197, Shanghai 200025, China

[3] University Medical Center Groningen, Center for Medical Imaging – North East Netherlands, University of Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands

Keywords Artificial intelligence · Magnetic resonance imaging · Giant cell tumor of bone · Prognosis

Abbreviations

CNN	Convolutional neural network
GCTB	Giant cell tumor of bone
PMMA	Polymethylmethacrylate
RMSProp	Root mean square prop
t-SNE	t-Distributed stochastic neighbor embedding

Giant cell tumor of bone (GCTB) represents 20% of benign bone tumors, 75% of which are located around the knee joint [1]. Although benign, GCTB has a tendency to local invasion [2]. Treatment is mainly curettage which is a less invasive surgical option than wide resection [3]. However, the tumor recurrence rate after curettage is reported to be high (13–65%) [4, 5].

The risk factors for recurrent GCTB include patient age, tumor location, and extension [4–10]. Apart from patient characteristics, the involvement of marrow or regional soft tissue is difficult to evaluate before surgery [11–14]. Medical imaging captures intratumoral heterogeneity. The heterogeneity is associated with the prognosis [15, 16]. Cystic changes and irregular margins on MR were reported as influential factors for recurrent GCTB [17, 18].

Deep learning dramatically improves the performance in analyzing images and many other domains [19–22]. The convolutional neural network (CNN) is a class of deep, feed-forward neural networks, wherein the information moves in a forward direction from the input to the output nodes. CNN algorithms have been applied successfully to medical imaging, such as detection of breast cancer [23], differentiation of liver masses [24], and classification of interstitial lung disease [25]. However, training a CNN from scratch requires a large dataset to optimize the numerous parameters.

Accumulating image characteristics obtained in a large generic dataset and fine-tuning it with a small specific dataset, also known as transfer learning, are a potentially effective way to mitigate data requirements in machine learning [26, 27]. Li et al employed 132 frames of wireless capsule endoscopy image with bleeding signs as the training dataset to achieve an accuracy of 98.62% to identify gastrointestinal bleeding [28]. Choi et al used 278 fundus photographs in 3 retinal disease categories as the training dataset and reached an accuracy of 72.8% in the test dataset [29]. Considering the situation that the clinical data of some uncommon diseases are difficult to obtain, transfer learning may help to analyze the limited amount of data. We intend to predict the local recurrence of GCTB using a pre-trained deep CNN architecture fine-tuned with MR images. Further improvement will be achieved by establishing a regression model combining CNN prediction and patient characteristics.

Materials and methods

Patients

From January 2005 to July 2015, 60 patients who underwent MR imaging and lesion curettage for GCTB located in the proximal tibia or distal femur and had histopathologically confirmed results were identified. To exclude the impact of the increased complexity of surgical treatment, four patients with pathological fractures were excluded. Finally, 56 patients were included for further analysis. This retrospective clinical study was approved by the institutional review board and the requirement for written informed consent was waived.

MR imaging

All patients were examined using a 1.5-T MR (Signa, GE Healthcare) with dedicated extremity coils. No contrast media were given. Axial, sagittal, and coronal images were obtained on T1-weighted (T_1W) and T2-weighted (T_2W) sequences. Fat-suppressed T_2W imaging was conducted on the sagittal or coronal plane. The field of view varied depending on the tumor size and adjacent region. The slice thickness and interslice gap were 5 mm and 0.5 mm, respectively. The image resolution ranged from 192 × 256 to 256 × 256 pixels.

Surgical curettage

All patients were treated with curettage by senior orthopedic surgeons. The neoplastic tissue was removed with a curette through a wide cortical window. The remainder of the tumor in the cavity was eliminated with a high-speed burr. Phenol was applied in the borders of the cavity with cotton-tipped applicators and then was neutralized with alcohol in 31 cases; the remaining 25 cases were treated without additional adjuvant. Finally, the tumor cavity was

carefully packed with a polymethylmethacrylate (PMMA) filling.

Follow-up and recurrence

All patients were instructed to have an MR examination whenever they were symptomatic (abnormal pain and swelling); otherwise, they would be re-examined by X-ray or MR annually. The patients were monitored for at least 2 years, because most GCTB recurrence happens within 2 years [30, 31]. The extension of the radiolucent zone on radiography or high signal intensity around PMMA on T_2W imaging after bone cement filling was considered local recurrence [32].

Grouping outline

In machine learning, validation method means fine-tuning the neural network architecture, so as to reach a robust classifier with good generalization ability, while test method evaluates the prediction accuracy of the model [33]. To assess the performance of the CNN model and compare between the CNN and radiologists, we used two different methods to split the dataset, which consisted of images of T_1W and T_2W sequences. The first method followed fourfold cross-validation, which is widely used in machine learning to assess its generalization ability. The total sample of 56 cases was randomly partitioned into equal sizes of 4 subsets. Each subset included 14 cases. No overlap was found among these four datasets. In four consecutive deep learning procedures, three subsets (42 cases) were used each time to train the model, and one subset (14 cases) was used to test the model.

The second approach was a random 50–50% split of the dataset into the training and test datasets, instead of cross-validation, because humans will remember what they had learned. The training dataset was utilized to train the CNN model as well as radiologists. Thus, 56 patients were divided equally into the training and test groups. The training dataset (11 recurrent and 17 non-recurrent) was used to fine-tune the CNN and to train radiologists. The test dataset (12 recurrent and 16 non-recurrent) was used to validate the accuracy of the CNN and radiologists.

Data preparation and image processing

We categorized the pre-surgery MR images into two classes: class A represented the recurrent tumor, and class B represented the non-recurrent tumor. Our data contained sets of images corresponding to the same lesion but from multiple viewpoints and imaging sequences, i.e., T_1W and T_2W images of the axial, sagittal, and coronal planes. Blurry or unclear images without details of tumor texture were deleted.

We converted the images into the grayscale Joint Photographic Experts Group (JPEG) format and resized each image to a resolution of 299×299 pixels to make it compatible with the original inception v3 network architecture.

Before fine-tuning the CNN, we first automatically augmented the images to increase the number of training images [34]. Data augmentation was achieved by performing geometric transformations to train a robust model that was invariant to such transformations. Each image was randomly rotated between 0 and 359° and distorted with the aspect ratio ranging from 0.75 to 1.33. Additionally, the margin region was cropped with the central fraction = 0.875, and the image was flipped vertically or horizontally with a probability of 0.5. Through the aforementioned process, 100 to 200 images were generated from one image. Subsequently, the generated series of images were randomly shuffled.

Image augmentation generated 216,900 images to train the CNN model. From the fourfold cross-validation method, 155,200, 117,600, 134,200, and 142,000 images were used for the four training datasets, respectively. The corresponding test datasets consisted of 259, 394, 365, and 376 images, respectively.

When comparing the performance of the CNN and radiologists, the training dataset for the CNN consisted of 118,200 images generated by image augmentation. The test dataset comprised 741 images.

Training algorithm and environment

We employed transfer learning in inception v3 CNN architecture (Online Fig. 1, Online Table 1), which was pre-trained by the ImageNet dataset, which is a large database designed for visual object recognition including 1.28 million images over 1000 generic object classes [34]. We reduced the final classifier layer from the original 1000 categories to two and subsequently fine-tuned the parameters using backpropagation across all layers with our training dataset (Fig. 1). T-Distributed stochastic neighbor embedding (t-SNE) for the fully connected layer was used to visualize the distribution of logits.

We performed deep learning using a computer with a graphics processing unit (GTX1080, nVidia) and 64GB random-access memory. The source code of inception v3 architecture is available in the online appendix. A programming tool (Python 3.6.1) and an integrated development environment (PyCharm 2017.2.4, JetBrains) were used to develop

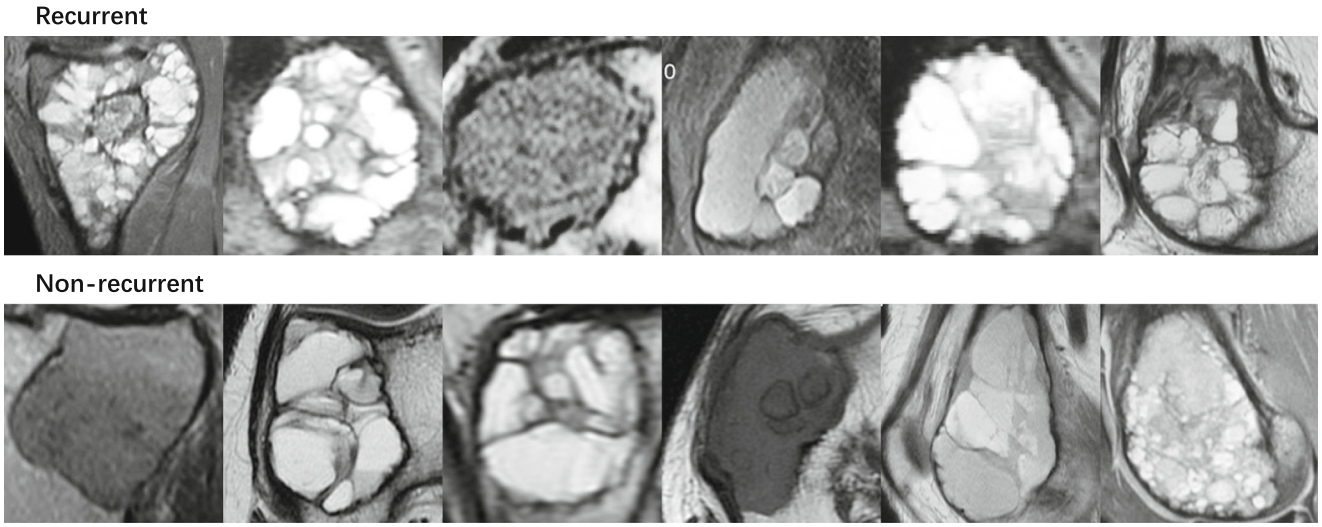

Fig. 1 Representative magnetic resonance images between recurrent and non-recurrent cases in the training group. The tumors visually show an irregular margin in the recurrent category

codes based on an open-source deep learning framework (TensorFlow 1.3, Google).

Inference algorithm

Given a patient examined by MR resulting in multiple images, CNN inference for each image was binary (recurrent or non-recurrent), indicating whether an image had recurrent or non-recurrent feature learned from the training dataset. By counting all the feasible images from each patient, the CNN program outputs a probability (0 to 100%), suggesting tumor recurrence as follows:

$$p = \frac{r}{n}$$

where p is the predicted recurrent probability of GCTB, r is the number of images with recurrent feature, and n is the total image number.

Training and test of radiologists

After reviewing the current publications on risk factors and image features associated with GCTB recurrence [4, 6, 9, 17, 18, 32], four musculoskeletal radiologists of at least 12 years' experience studied MR images in the training dataset (28 cases) for image features predicting recurrent GCTB, including cystic change and irregular margin [17, 18]. Next, each radiologist independently determined the probability of recurrent GCTB (0 to 100%) for the patients in the test dataset, with knowledge of the predilection of young patients (≤ 29 years [18]) and GCTB located in the proximal tibia. The average predicted probability was calculated as the mean from four radiologists.

Statistics

To model recurrent probability into a binary outcome, a predicted probability > 50% is considered recurrent while ≤ 50% is considered non-recurrent [35]. The inter-observer agreement among the four radiologists was accessed by the intraclass correlation coefficient (ICC). Using the fourfold cross-validation method in this study, the prediction accuracy was calculated as the mean of four test subsets using the follow-up results as the reference.

A binary logistic regression model was established to predict local recurrence of GCTB, combining CNN prediction (recurrent percentage) and patient characteristics (age and tumor location), which have been shown to be influential factors for GCTB recurrence [4, 10, 31]. A similar model combining radiologist prediction and patient characteristics was also established. The formula for each regression model was established.

Prediction performance for recurrent GCTB by the CNN and radiologists was indicated as accuracy, sensitivity, specificity, and positive and negative predictive values, as well as by receiver operating characteristic (ROC) analysis with an area under the ROC curve (AUC). Multivariate ROC analysis of DeLong's method [36] based on the logistic regression model was performed to compare the prediction performances among the CNN, CNN combined with the regression model, and radiologists. Zombie plots were used to visualize the sensitivity and specificity [37].

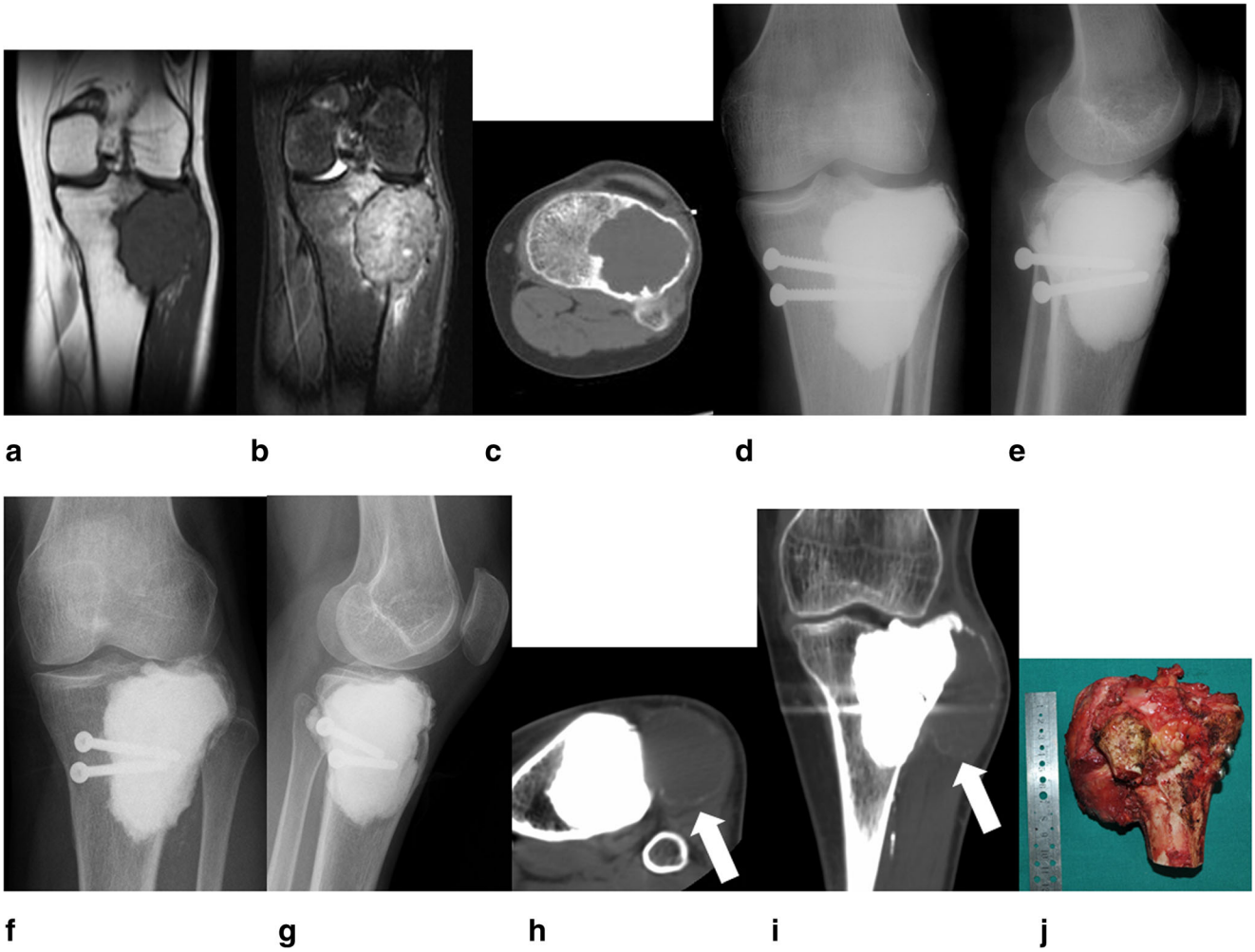

Fig. 2 A 24-year-old woman with histopathologically confirmed giant cell tumor of bone. (**a**, **b**) Coronal MR images show low signal intensity and high signal intensity on T_1WI and fat-suppressed T_2WI, respectively. (**c**) The main CT feature is an osteolytic lesion with osteosclerosis in the proximal tibia. Compared with (**d**, **e**) X-ray examination after operation, (**f**, **g**) X-ray images indicate no sign of recurrence after 8 years of follow-up. (**h**, **i**) But axial and sagittal CT images reveal recurrent soft mass (arrow) in the ninth year after follow-up. (**j**) The surgical specimen confirms recurrent soft mass

A p value smaller than 0.05 was regarded as statistically significant. Statistical analyses were performed using a software package (MedCalc 13.0, MedCalc Software).

Results

Patient and image characteristics

Fifty-six GCTB patients (27 men and 29 women; median age, 29 years [range, 18–64 years]) with 56 tumors were enrolled. Localization of the tumor was in 30 (53.6%) the proximal tibia and in 26 (46.4%) the distal femur. The mean follow-up period was 5.8 years (range, 2.0 to 9.5 years). Of the 56 patients, 26 were followed up by X-ray, 13 by MR, and 13 by both X-ray and MR, and 4 were followed up clinically, with a result that 23 (41.1%) tumors were diagnosed as recurrent. Twenty of the 23 cases (87.0%) recurred within 2 years (mean 1.28 years), two cases (8.7%) in the fifth year, and one (4.3%) in the eighth year. Of the 23 patients with recurrent GCTB, 14 were identified by X-ray, 6 by MR, and 3 by both X-ray and MR. A representative recurrent case is shown in Fig. 2.

CNN and CNN regression model

In the first method of data splitting, we experimented with fourfold cross-validation. Next, we used t-SNE to visualize the logits distribution at the fully connected layer during the iteration procedure, in which homogeneously distributed logits became two clusters (Online Fig. 2).

Table 1 Confusion matrix of the CNN, CNN regression model, and radiologists to predict recurrent giant cell tumor of bone on the test dataset

		CNN		CNN regression model		Radiologist 1		Radiologist 2		Radiologist 3		Radiologist 4		Radiologist average	
True label		CNN prediction by fourfold cross-validation													
	Recur	14	9	16	7										
	Non-recur	4	29	3	30										
		CNN vs radiologists													
	Recur	6	6	7	5	6	6	8	4	6	6	10	2	7	5
	Non-recur	1	15	1	15	3	13	6	10	7	9	3	13	4	12
		Recur	Non-recur	Recur	Non-recur	Recur	Non-recur	Recur	Non-recur	Recur	Non-recur	Recur	Non-recur	Recur	Non-recur
		Predicted label													

CNN convolutional neural network

The accuracy of the CNN to predict the prognosis of GCTB (recurrent or non-recurrent) after intralesional curettage was 76.8% (95% confidence interval [CI] 63.6 to 87.0%) for all the cases in the test datasets (Tables 1 and 2). An example of a non-recurrent case is shown in Fig. 3. A formula was established based on the CNN regression model as the following:

$$p = \frac{1}{1 + e^{-(-2.791+1.327a+0.820b+3.701c)}}$$

where p, a, b, and c represent the overall predicted probability, age (0 indicated ≤ 29 years, 1 indicated > 29 years), location (0 indicated the distal femur, 1 indicated the proximal tibia), and CNN predicted probability, respectively. The CNN regression model increased the accuracy to 82.1% (69.6 to 91.1%). The CNN showed a sensitivity and specificity of 77.8% (52.4 to 93.6%) and 76.3% (59.8 to 88.6%), respectively, to determine recurrent GCTB (Tables 1 and 2). However, the CNN regression model showed a higher sensitivity and specificity of 84.2% (60.4 to 96.6%) and 81.1% (64.8 to 92.0%), respectively. The details of the binary logistic regression are shown in Table 3.

CNN vs radiologists

In the second method of data splitting, we divided the dataset equally into training and test datasets and compared the performance between the CNN and radiologists. For fair competition, the CNN and radiologists learned and were tested on the same dataset. The ICC for four radiologists was 0.729 (95% CI 0.518 to 0.862). A formula was established based on the CNN regression model:

$$p = \frac{1}{1 + e^{-(-2.069+1.069a+0.828b+3.207c)}}$$

where p, a, b, and c represent the overall predicted probability, age (0 indicated ≤ 29 years, 1 indicated > 29 years), location (0 indicated the distal femur, 1 indicated the proximal tibia), and CNN predicted probability, respectively. The CNN and CNN regression model successfully predicted 75.5% (95% CI 55.1 to 89.3%) and 78.6% (59.0 to 91.7%), respectively, of the cases in the test dataset (Table 1). On the other hand, the accuracy values of the four radiologists were 64.3% (44.1 to 81.4%) (Table 2). The CNN and CNN regression model showed sensitivities of 85.7% (42.1 to 99.6%) and 87.5% (47.3 to 99.7%), respectively, to determine recurrent GCTB, greater than 58.3% (27.7 to 84.8%) by radiologists ($p < 0.05$). However, the CNN and CNN regression model showed specificities of 71.4% (47.8 to

Table 2 Diagnostic performance of the CNN, CNN regression model, and radiologists to predict recurrent giant cell tumor of bone on the test dataset using the follow-up results as a reference

	CNN	CNN regression model	Radiologist average
CNN prediction by fourfold cross-validation			
Sensitivity	77.8% [14/18] (52.4 to 93.6%)	84.2% [16/19] (60.4 to 96.6%)	
Specificity	76.3% [29/38] (59.8 to 88.6%)	81.1% [30/37] (64.8 to 92.0%)	
PPV	60.8% [14/23] (45.5 to 74.3%)	69.6% [16/23] (53.3 to 82.0%)	
NPV	87.9% [29/33] (75.0 to 94.6%)	90.9% [30/33] (77.8 to 96.6%)	
Accuracy	76.8% [43/56] (63.6 to 87.0%)	82.1% [46/56] (69.6 to 91.1%)	
CNN vs radiologists			
Sensitivity	85.7% [6/7] (42.1 to 99.6%)	87.5% [7/8] (47.3 to 99.7%)	58.3% [7/12] (27.7 to 84.8%)
Specificity	71.4% [15/21] (47.8 to 88.7%)	75.0% [15/20] (50.9 to 91.3%)	75.0% [12/16] (47.6 to 92.7%)
PPV	50.0% [6/12] (32.3 to 67.2%)	58.3% [7/12] (38.5 to 75.6%)	63.6% [7/11] (39.7 to 82.3%)
NPV	93.8% [15/16] (70.6 to 98.9%)	93.8% [15/16] (70.6 to 98.9%)	70.6% [12/17] (53.7 to 83.2%)
Accuracy	75.0% [21/28] (55.1 to 89.3%)	78.6% [22/28] (59.0 to 91.7%)	64.3% [18/28] (44.1 to 81.4%)

CNN convolutional neural network, *PPV* positive predictive value, *NPV* negative predictive value

88.7%) and 75.0% (50.9 to 91.3%), respectively, similar to 75.0% (47.6 to 92.7%) by radiologists ($p > 0.05$).

ROC analysis

The AUC of the receiver operating characteristic of the CNN regression model was 0.832 (95% CI 0.708 to 0.919), significantly greater than 0.781 (0.650 to 0.880) of the CNN ($p = 0.012$). When comparing the performance between CNN and radiologists, the AUC values of the CNN, CNN regression model, radiologist average, and radiologist regression model were 0.688 (0.486 to 0.848), 0.797 (0.603 to 0.924), 0.753 (0.554 to 0.895), and 0.776 (0.590 to 0.911), respectively. The CNN regression model showed a greater AUC values than the CNN and radiologists ($p < 0.05$). However, the radiologist regression model did not significantly improve the AUC compared with the radiologists ($p > 0.05$). ROCs and zombie plots are shown in Fig. 4.

Discussion

In this study, we analyzed pre-surgery MR images of GCTB using a CNN, which detects image features not perceptible to human eyes. We established a logistic regression model integrated with tumor location, patient age, and CNN prediction to forecast recurrent GCTB. In previous studies, proximal tibia and younger age were influential factors for GCTB recurrence [17, 18]. Our results showed that the CNN had a prediction accuracy of 75.5%, higher than 64.3% by human observation. However, CNN regression model combined with age and lesion location further improved the accuracy to 78.6%, which helped most in differentiating recurrent from non-recurrent tumors.

Evaluation of the recurrent risk of GCTB may play an important role to help guide treatment for patients. Important alternatives for GCTB treatment are en bloc resection, curettage, with the potential addition of high-speed burr, phenol, and liquid nitrogen [1]. Wide resection is associated with a lower recurrence rate, but weakens adjacent joints [38]. Curettage may spare the adjacent joints but is associated with a higher likelihood of recurrence. The goal would be to find a reliable prediction model to choose the relevant method.

GCTB comprises of mononuclear stromal cells and characteristic multinucleated giant cells that exhibit osteoclastic activity [32]. The typical appearance is an osteolytic lesion with a well-defined but non-sclerotic margin that is eccentric and extends near the articular surface. GCTB may have invasive features, including cortical expansion or destruction. Chen et al established a multivariate logistic regression model combining tumor size, number, and margins of the soft tissue extension [8]. This model significantly associated with the recurrence of GCTB. But the observation of soft tissue extension is subjective, which heavily relies on observer experience and is influenced by inter-observer variability. A method based on CNN may provide an objective and repeatable way to predict the recurrence of GCTB.

In this study, phenol was applied in the borders of the cavity in 31 cases. The remaining 25 cases were treated without additional adjuvant. A publication showed that the application of phenol was not correlated with local recurrence after intralesional curettage [18]. Furthermore,

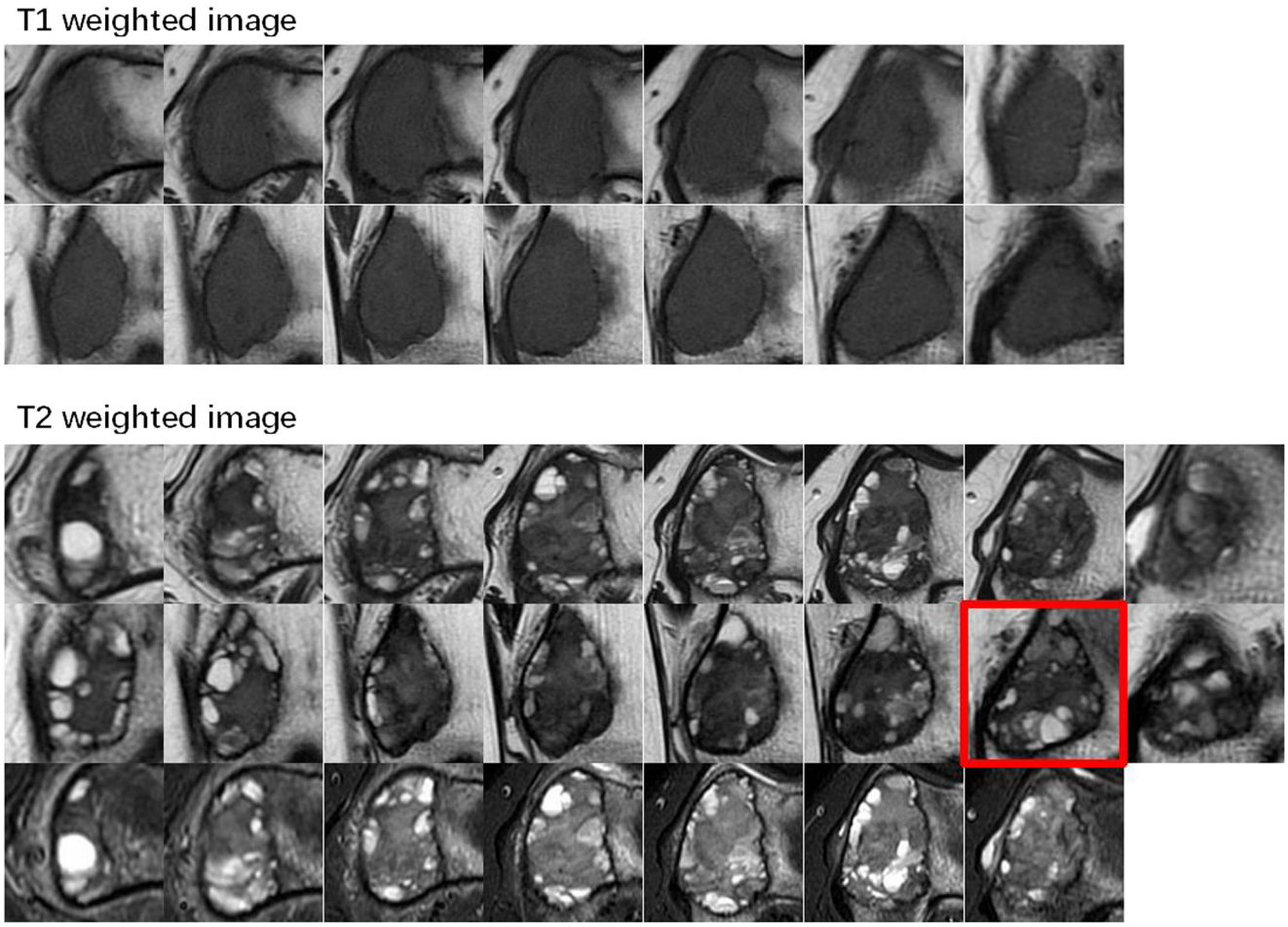

Fig. 3 A representative case from the test dataset female, 29 years old. A giant cell tumor was found in the distal femur bone. Four radiologists read the images and predicted recurrent probabilities of 60%, 60%, 80%, and 60%, because the cystic sign was observed. The average human prediction was 65%. However, the CNN identified only one image of 37 with recurrent feature (the red box). The recurrent probability of CNN was 1/37 = 2.7%. Finally, this patient's follow-up result was non-recurrent after 2 years

Table 3 Binary logistic regression model for the prediction of recurrent giant cell tumor of bone

	B	Standard error	Wald	*p*	Exp(*B*)
CNN regression model by fourfold cross-validation					
Age	1.327	0.699	3.600	0.048	3.768
Location	0.820	0.668	1.506	0.220	2.271
CNN	3.701	1.118	10.960	0.001	40.798
Constant	− 2.791	0.820	11.589	0.001	0.061
CNN regression model trained by 28 cases					
Age	1.069	0.942	1.286	0.257	2.911
Location	0.828	0.928	0.796	0.372	2.289
CNN	3.207	1.684	3.626	0.057	24.708
Constant	− 2.069	1.043	3.933	0.047	0.126

B represents the regression coefficient. Wald represents the Wald chi-square statistic, which tests the unique contribution of each predictor in the context of the other predictors. Exp(*B*) indicates the odds ratio predicted by the model

CNN convolutional neural network

we randomly divided the cases into the training and test datasets. The utilization of phenol may not influence the likelihood of recurrence. We followed the patients for 5.8 years and found 87% recurrent cases within 2 years. Abat et al followed 97 cases of GCTB patients for 12 years after surgery and found 84% recurrences within 1 year [7]. Because most recurrence occurred within 2 years, recurrence after 2 years is exceptional [1]. The shorter follow-up period may not lead to underestimation of recurrent rate.

There are limitations to this study. First, although local relapse after 2 years was considered exceptional [31], we could not exclude the possibility of recurrence after evaluation in this study. A longer follow-up period would provide more information to strengthen the conclusion. Second, the sample size was restricted. We used transfer learning and image augmentation. We expect that increasing the training dataset would improve the accuracy and robustness of deep learning in the future.

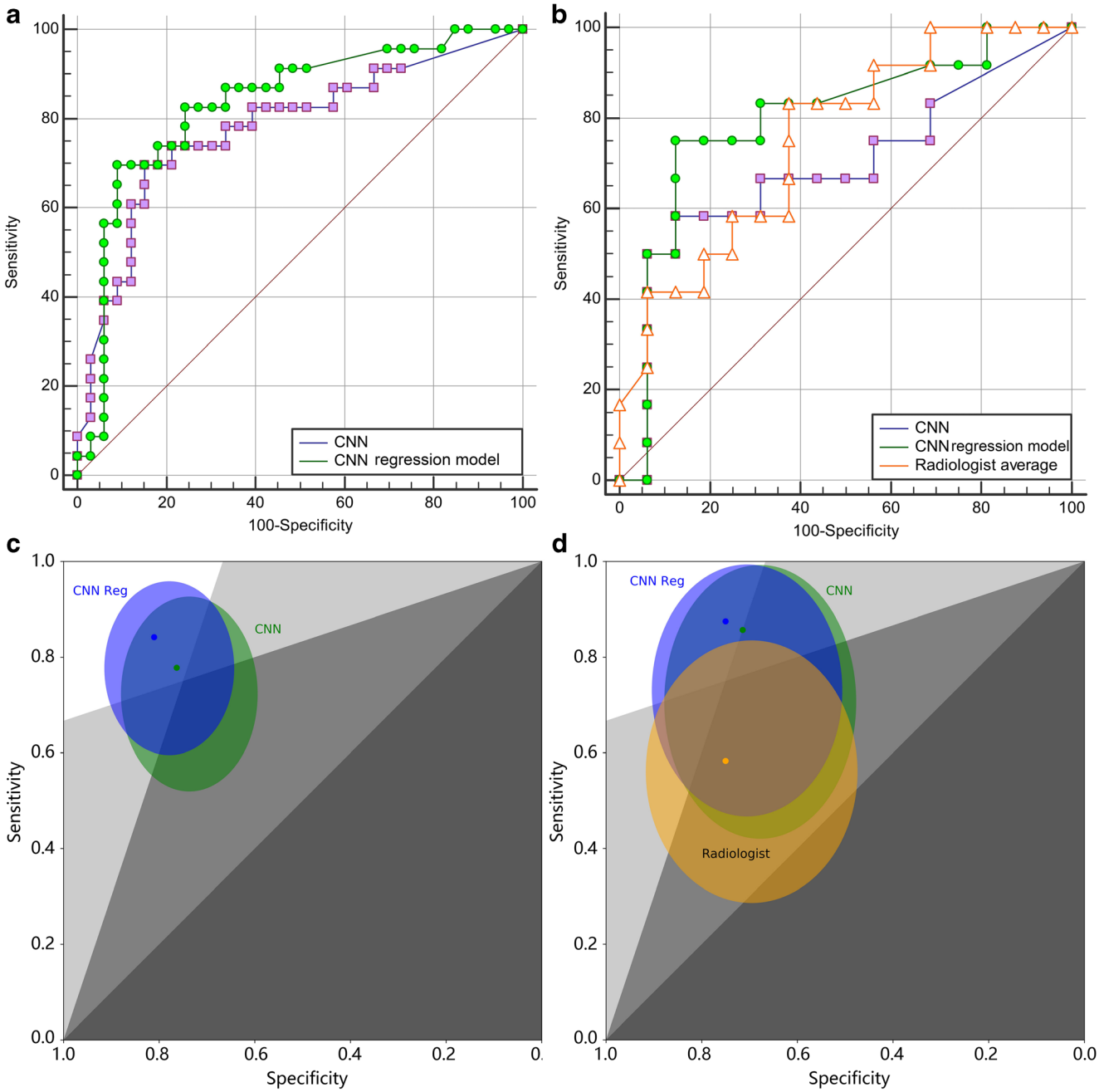

Fig. 4 **a** Comparison of receiver operating characteristic curves among the CNN and CNN regression model to predict recurrent giant cell tumor of bone. CNN convolutional neural network. **b** Comparison of receiver operating characteristic curves among the CNN, CNN regression model, and radiologist average to predict recurrent giant cell tumor of bone on the test dataset trained by 28 cases. CNN convolutional neural network. **c** A zombie plot for the four-way cross-validation shows that the 95% confidence region for the CNN with regression is located slightly further toward the upper left-hand corner than the plain CNN. However, the two confidence regions overlap considerably, making it unlikely that there is a credible difference between them. **d** A zombie plot compares the average radiologist scores with those of the CNN + regression. The plain regression also shows little credible difference in the confidence regions for the two CNNs. Although the radiologist region also has considerable overlap with the two CNNs, there is still considerable non-overlap. The large confidence regions shown here and large areas of overlap are a symptom of the very small sample sizes in this study

Third, the patients were recruited in one center, and the image acquisition was performed using one MR equipment. A multicentric study on MR equipment from different vendors may improve the performance of CNN and further enhance the conclusion.

Conclusion

In summary, the transfer learning-based CNN might be helpful for the prediction of local recurrence of GCTB after curettage. The combination of CNN prediction and patient features

improves its prediction performance. Deep learning surpasses human observation to predict recurrent GCTB. Clinical consequences might be to influence the initial choice of surgical technique or decision of a closer follow-up in patients that are predefined as at a high risk for recurrence.

Acknowledgements We thank Dr. Xiaochun Yuan, Dr. Qin Chen, Dr. Gehong Yao, and Dr. Lin Zhang in Shanghai General Hospital for analyzing the images.

Funding This study was sponsored by the National Natural Science Foundation of China (project no. 81471662), Ministry of Science and Technology of China (2016YFE0103000), Science and Technology Commission of Shanghai Municipality (16411968500 and 16410722300), Shanghai Jiao Tong University (ZH2018ZDB10), Shanghai Jiao Tong University School of Medicine - Gaofeng Clinical Medicine Grant Support (20181814), and Clinical Research Innovation Plan of Shanghai General Hospital (CTCCR-2018B04).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Xueqian Xie, MD PhD, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the co-authors, Dr. Jiapan Guo, PhD, from the University Medical Center Groningen, The Netherlands, kindly provided statistical advice and IT supports for this manuscript.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- Retrospective
- Diagnostic study
- Performed at one institution

References

1. van der Heijden L, Dijkstra PD, van de Sande MA et al (2014) The clinical approach toward giant cell tumor of bone. Oncologist 19: 550–561
2. Ghert MA, Rizzo M, Harrelson JM, Scully SP (2002) Giant-cell tumor of the appendicular skeleton. Clin Orthop Relat Res 400: 201–210
3. van der Heijden L, Dijkstra PD, Campanacci DA, Gibbons CL, van de Sande MA (2013) Giant cell tumor with pathologic fracture: should we curette or resect? Clin Orthop Relat Res 471:820–829
4. Klenke FM, Wenger DE, Inwards CY, Rose PS, Sim FH (2011) Giant cell tumor of bone: risk factors for recurrence. Clin Orthop Relat Res 469:591–599
5. Arbeitsgemeinschaft Knochentumoren, Becker WT, Dohle J et al (2008) Local recurrence of giant cell tumor of bone after intralesional treatment with and without adjuvant therapy. J Bone Joint Surg Am 90:1060–1067
6. He Y, Zhang J, Ding X (2017) Prognosis of local recurrence in giant cell tumour of bone: what can we do? Radiol Med 122:505–519
7. Abat F, Almenara M, Peiro A, Trullols L, Bague S, Gracia I (2015) Giant cell tumour of bone: a series of 97 cases with a mean follow-up of 12 years. Rev Esp Cir Ortop Traumatol 59:59–65
8. Chen L, Ding XY, Wang CS et al (2014) In-depth analysis of local recurrence of giant cell tumour of bone with soft tissue extension after intralesional curettage. Radiol Med 119:861–870
9. Teixeira LE, Vilela JC, Miranda RH, Gomes AH, Costa FA, de Faria VC (2014) Giant cell tumors of bone: nonsurgical factors associated with local recurrence. Acta Orthop Traumatol Turc 48: 136–140
10. Siddiqui MA, Seng C, Tan MH (2014) Risk factors for recurrence of giant cell tumours of bone. J Orthop Surg (Hong Kong) 22:108–110
11. Turcotte RE, Wunder JS, Isler MH et al (2002) Giant cell tumor of long bone: a Canadian Sarcoma Group study. Clin Orthop Relat Res 397:248–258
12. Gouin F, Dumaine V, French Sarcoma and Bone Tumor Study Groups GSF-GETO (2013) Local recurrence after curettage treatment of giant cell tumors in peripheral bones: retrospective study by the GSF-GETO (French Sarcoma and Bone Tumor Study Groups). Orthop Traumatol Surg Res 99:S313–S318
13. Klenke FM, Wenger DE, Inwards CY, Rose PS, Sim FH (2011) Recurrent giant cell tumor of long bones: analysis of surgical management. Clin Orthop Relat Res 469:1181–1187
14. Wang H, Wan N, Hu Y (2012) Giant cell tumour of bone: a new evaluating system is necessary. Int Orthop 36:2521–2527
15. Sanduleanu S, Woodruff HC, de Jong EEC et al (2018) Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score. Radiother Oncol 127:349–360
16. Huang YQ, Liang CH, He L et al (2016) Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol 34:2157–2164
17. He Y, Wang J, Rui W et al (2018) Retrospective investigation of "paint brush borders" sign in association with local recurrence of giant cell tumor of bone after intralesional curettage. J Bone Oncol 10:41–48
18. He Y, Wang J, Zhang J, Yuan F, Ding X (2017) A prospective study on predicting local recurrence of giant cell tumour of bone by evaluating preoperative imaging features of the tumour around the knee joint. Radiol Med 122:546–555
19. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436–444
20. Mnih V, Kavukcuoglu K, Silver D et al (2015) Human-level control through deep reinforcement learning. Nature 518:529–533
21. Silver D, Huang A, Maddison CJ et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489
22. Silver D, Schrittwieser J, Simonyan K et al (2017) Mastering the game of Go without human knowledge. Nature 550:354–359
23. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A (2017) Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 52:434–440
24. Yasaka K, Akai H, Abe O, Kiryu S (2017) Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 286:887–896

25. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S (2016) Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 35:1207–1216
26. Banerjee I, Crawley A, Bhethanabotla M, Daldrup-Link HE, Rubin DL (2018) Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma. Comput Med Imaging Graph 65:167–175
27. Chartrand G, Cheng PM, Vorontsov E et al (2017) Deep learning: a primer for radiologists. Radiographics 37:2113–2131
28. Li XL, Zhang H, Zhang XL, Liu H, Xie GT (2017) Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images. Conf Proc IEEE Eng Med Biol Soc 2017:1994–1997
29. Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH (2017) Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. PLoS One 12:e0187336
30. Lausten GS, Jensen PK, Schiodt T, Lund B (1996) Local recurrences in giant cell tumour of bone. Long-term follow up of 31 cases. Int Orthop 20:172–176
31. Kivioja AH, Blomqvist C, Hietaniemi K et al (2008) Cement is recommended in intralesional surgery of giant cell tumors: a Scandinavian Sarcoma Group study of 294 patients followed for a median time of 5 years. Acta Orthop 79:86–93
32. Chakarun CJ, Forrester DM, Gottsegen CJ, Patel DB, White EA, Matcuk GR Jr (2013) Giant cell tumor of bone: review, mimics, and new developments in treatment. Radiographics 33:197–211
33. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R. Springer, Berlin
34. Esteva A, Kuprel B, Novoa RA et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542: 115–118
35. Hohmann E, Wetzler MJ, D'Agostino RB Jr (2017) Research pearls: the significance of statistics and perils of pooling. Part 2: predictive modeling. Arthroscopy 33:1423–1432
36. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
37. Richardson ML (2016) The Zombie plot: a simple graphic method for visualizing the efficacy of a diagnostic test. AJR Am J Roentgenol 207:W43–W52
38. Li D, Zhang J, Li Y et al (2016) Surgery methods and soft tissue extension are the potential risk factors of local recurrence in giant cell tumor of bone. World J Surg Oncol 14:114