An updated systematic review of radiomics in osteosarcoma: utilizing CLAIM to adapt the increasing trend of deep learning application in radiomics

Jingyu Zhong1†, Yangfan Hu1†, Guangcheng Zhang2, Yue Xing1, Defang Ding1, Xiang Ge1, Zhen Pan3, Qingcheng Yang3, Qian Yin4, Huizhen Zhang4, Huan Zhang5* and Weiwu Yao1*

Abstract

Objective: To update the systematic review of radiomics in osteosarcoma.

Methods: PubMed, Embase, Web of Science, China National Knowledge Infrastructure, and Wanfang Data were searched to identify articles on osteosarcoma radiomics until May 15, 2022. The studies were assessed by Radiomics Quality Score (RQS), Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement, Checklist for Artificial Intelligence in Medical Imaging (CLAIM), and modified Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool. The evidence supporting radiomics application for osteosarcoma was rated according to meta-analysis results.

Results: Twenty-nine articles were included. The average of the ideal percentage of RQS, the TRIPOD adherence rate and the CLAIM adherence rate were 29.2%, 59.2%, and 63.7%, respectively. RQS identified a radiomics-specific issue of phantom study. TRIPOD addressed deficiency in blindness of assessment. CLAIM and TRIPOD both pointed out shortness in missing data handling and sample size or power calculation. CLAIM identified extra disadvantages in data de-identification and failure analysis. External validation and open science were emphasized by all the above three tools. The risk of bias and applicability concerns were mainly related to the index test. The meta-analysis of radiomics predicting neoadjuvant chemotherapy response by MRI presented a diagnostic odds ratio (95% confidence interval) of 28.83 (10.27–80.95) on testing datasets and was rated as weak evidence.

Conclusions: The quality of osteosarcoma radiomics studies is insufficient. More investigation is needed before using radiomics to optimize osteosarcoma treatment. CLAIM is recommended to guide the design and reporting of radiomics research.

Keywords: Osteosarcoma, Radiomics, Machine learning, Quality improvement, Systematic review

Key points

• The MRI-radiomics in predicting neoadjuvant chemotherapy response is supported by weak evidence.
• The quality of osteosarcoma radiomics studies has been improved recent two years.
• CLAIM can adapt the increasing trend of deep learning application in radiomics.
Introduction

Osteosarcoma is the most common primary high-grade sarcoma of the skeleton, in which the tumor cells produce neoplastic bone [1]. Imaging is the key examination in the work-up of osteosarcoma management, from diagnosis, staging, treatment evaluation, to follow-up [1–3]. The diagnosis of osteosarcoma generally starts with X-ray radiography and is followed by CT for further evaluation. A contrast-enhanced MRI scan is useful in diagnosis completion and soft tissue involvement assessment and is usually the last step before biopsy of local disease. A chest CT scan is substantial for lung metastases detection. For patients with pathologically confirmed osteosarcoma, a whole-body PET examination has been recommended for initial staging rather than a bone scintigraphy nowadays. The treatment evaluation and follow-up imaging commonly include local CT and MRI scans and chest CT scan. In most cases, the current imaging approach with physical, laboratory, and histopathological examinations can guide clinicians to an appropriate curative plan, but there remain difficulties in differential diagnosis of osteosarcoma subtypes, prediction of response to treatment, and prognosis concerns including survival, recurrence, and lung metastasis [2, 3]. Radiomics, utilizing a plethora of strategies for extracting underlying information from medical images, has been used to overcome such challenges [4–7]. Radiomics models have been deemed as a promising approach for addressing clinical problems related to osteosarcoma patients, especially for predicting their response to neoadjuvant chemotherapy (NAC) [8].

Our preliminary search suggested that radiomics studies in osteosarcoma patients have doubled since the publication of the previous review [8], indicating necessity for updates on this rapidly developing field. It is unclear whether the radiomics study quality has improved in recent years. Next, the study quality and risk of bias of radiomics research on osteosarcoma have been only assessed by the Radiomics Quality Score (RQS) [7] and the modified Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool [9]. An additional evaluation using the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist [10] has been recommended to identify several significant items for reporting transparency of radiomics studies [11–14]. Further, RQS and TRIPOD may not be totally suitable for current radiomics studies, since recently developed deep radiomics applies convolutional neural networks to analyze these extracted features [15–18]. The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [19] has been demonstrated as a useful tool to improve design and reporting of deep...
learning research [20, 21]. It is potentially better for the evaluation of current radiomics studies with increasing application of deep learning. Finally, the level of evidence supporting the radiomics application in osteosarcoma has not been evaluated yet [22]. It is of importance to provide an overall evidence strength rating before translating radiomics into clinical practice [23, 24]. Therefore, we hypothesized that the publication of the previous review could improve the radiomics study quality in osteosarcoma, and that CLAIM is a better tool for current radiomics studies.

The aim of the present study is to provide an updated systematic review of radiomics in osteosarcoma with quality assessment and evidence-level rating and find out whether CLAIM can better identify disadvantages in current radiomics studies.

**Methods**

**Protocol and registration**

The updating of this systematic review was decided according to a three-step decision framework [25] and was conducted in the style of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [26]. The review protocol (CRD42020175383) and updating information are present in Additional file 1: Note S1. The PRISMA checklist for current systematic review and meta-analysis is present as Additional file 2.

**Literature search and selection**

An up-to-date literature search was performed via PubMed, Embase, Web of Science, China National Knowledge Infrastructure, and Wanfang Data until May 15, 2022, by two reviewers each with 4 years’ experience in radiology and radiomics research. Disagreements were solved by a review group consisting of radiologists, orthopedists, and pathologists with different levels of experience. All primary research assessing the role of radiomics in osteosarcoma treatment for diagnostic, prognostic, or predictive purposes was considered eligible for the current review. No publication period restrictions were applied, while only articles in English, Japanese, Chinese, German or French were available. The two reviewers screened the titles and abstracts after the removal of duplications and obtained the full-texts and their supplementary materials. The same reviewers determined their eligibility according to the inclusion and exclusion criteria. Other potentially eligible articles were identified...

| Study Characteristics | Data |
|-----------------------|------|
| Sample size, mean ± standard deviation, median (range) | 86.6 ± 45.8, 81 (17–191) |
| Journal type, n (%) | N = 29 |
| Imaging | 13 (44.8) |
| Non-imaging | 16 (55.2) |
| First authorship, n (%) | N = 29 |
| Radiologist | 19 (65.5) |
| Non-radiologist | 10 (34.5) |
| Imaging modality, n (%) | N = 29 |
| CT | 9 (31.0) |
| MRI | 14 (48.3) |
| PET | 6 (20.7) |
| Biomarker, n (%) | N = 33 |
| Diagnostic | 3 (9.1) |
| Predictive | 18 (54.5) |
| Prognostic | 12 (36.4) |
| Model type, n (%) | N = 33 |
| Type 1a: Developed model validated with exactly the same data | 8 (24.2) |
| Type 1b: Developed model validated with resampling data | 8 (24.2) |
| Type 2a: Developed model validated with randomly splitting data | 12 (36.4) |
| Type 2b: Developed model validated with non-randomly splitting data | 1 (3.0) |
| Type 3: Developed model validated with separate data | 4 (12.1) |
| Type 4: Validation only | 0 (0.0) |

There were 33 radiomics models identified in 29 included studies. The model type was determined according to criteria in TRIPOD statement. TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
from the reference lists of relevant articles and reviews. For uncertainties, the review group was consulted. The search and selection strategy is shown in Additional file 1: Note S2.

**Data extraction and quality assessment**

We used a data collection sheet for bibliographical information, study characteristics, radiomics considerations, and model metrics (Additional file 1: Table S1) [8]. The eligible studies employed RQS [7], TRIPOD [10], CLAIM [19], and QUADAS-2 tools [9] (Additional file 1: Tables S2–S5). The RQS is a consensus list composed of sixteen items for methodological issues specific to radiomics studies and is later summarized to six key domains [12–14]. The TRIPOD statement provides a checklist consisting of thirty-seven items in twenty-two criteria aiming to promote transparency of prediction model studies and is recommended for identifying room of improvement in radiomics studies [12–14]. The CLAIM includes forty-two items in seven topics that should be viewed as a best practice to guide presentation of AI research [19]. The CLAIM has seldomly been employed for quality assessment of radiomics studies [20, 21]. However, we assumed that CLAIM is suitable for radiomics studies evaluation, as radiomics is a subset of AI application in medical imaging [15–18]. The QUADAS-2 tool was tailored to our review by modifying the signaling questions [8]. Two reviewers independently extracted the data and evaluated the studies. Disagreements were solved by discussion with the review group. Topics discussed are recorded in Additional file 1: Note S3.

**Data synthesis and analysis**

The statistical analysis was performed with R language version 4.1.3 within RStudio version 1.4.1106 [27]. The RQS rating, the ideal percentage of RQS, and adherence rates of RQS, TRIPOD and CLIAM were calculated. In case a score of at least one point for each item was obtained without minus points, it was considered to have basic adherence, as those have been reported [12–14]. For example, if the item of validation in RQS obtained 2–5 points, it was considered as basic adherent, while...

---

**Table 2**

| Imaging in osteosarcoma and radiomics study topics. Imaging examination is a routine in diagnosis and treatment decision in osteosarcoma. Radiomics has shown potential in personal precision medicine in this process. The study topics and number of radiomics studies in osteosarcoma with imaging modality were summarized. Note two studies built a prediction model for the response to NAC and a prognostic model for survival, respectively; and two studies built one prediction model for response to NAC and a prognostic model for survival, respectively. This resulted in 33 radiomics models in 29 included studies. OS osteosarcoma, ES Ewing sarcoma, CS chondrosarcoma.

| Imaging examination | Diagnosis |
|--------------------|-----------|
| OS vs ES | 1 MRI |
| OS vs ES & CS | 1 CT |
| Metastatic vs non-metastatic lung nodules | 1 CT |
| **Staging imaging** | **Predictive** |
| Metastasis at diagnosis | 2 MRI |
| Response to neoadjuvant chemotherapy | 2 CT |
| **Pathology** | **Prognostic** |
| Survival | 3 CT |
| Recurrence | 1 CT |
| Metastasis | 1 MRI |
| Imaging | 2 MRI |
| Follow up | Follow up | Follow up | Follow up |
it was regarded as without basic adherence when it was rated as -5 points. The QUADAS-2 assessment result was summarized. Pearson correlation test was used for the correlation analysis between the ideal percentage of RQS, the TRIPOD adherence rate and the CLAIM adherence rate. Subgroup analysis was performed to compare the ideal percentage of RQS, the TRIPOD adherence rate and the CLAIM adherence rate by journal type, first authorship, imaging modality, and publication period. A two-tailed \( p \) value < 0.05 indicated statistical significance, unless otherwise specified. Post hoc multiple comparisons were adjusted using the Bonferroni method. The detailed data analysis method is described in Additional file 1: Note S4.

The meta-analysis was performed using Stata software version 15.1 [28]. In the current systematic review, the role of MRI-driven radiomics in prediction of osteosarcoma patients’ response to NAC was addressed repeatedly. To present the true performance of the radiomics model, corresponding meta-analysis was conducted based on results of testing datasets. The two-by-two tables were directly extracted from the articles or reconstructed based on available data. The diagnostic odds ratio (DOR) with 95% confidence interval (CI) and corresponding \( p \) value were calculated using the random-effect model. Sensitivity, specificity, positive and negative likelihood ratio were estimated. A summary receiver operating characteristic (SROC) curve was drawn. The Cochran’s Q test and the Higgins \( I^2 \) test were used for heterogeneity assessment. The funnel plot was drawn with Egger’s test and Begg’s test, and the Deeks’ funnel plot was constructed with Deeks’ funnel plot asymmetry test for publication bias. A two-tailed \( p \) value > 0.10 indicated a low publication bias. The trim and fill method was employed to estimate the number of missing studies. The level of evidence supporting the clinical application of radiomics in osteosarcoma was rated based on results of meta-analysis (Additional file 1: Table S6) [22].

**Results**

**Literature search**

The search yielded 251 records in total, in which 142 remained after removing duplicates. After screening the titles and abstracts, full-texts of 47 articles were retrieved and reviewed. Ultimately, 29 articles were included in this systematic review [29–57] (Fig. 1). No additional eligible study was identified through hand search of relevant reviews and reference lists of eligible articles.

**Study characteristics**

Table 1 and Fig. 2 summarized the characteristics of 33 models described in the 29 included studies. The sample size of studies ranged from 17 to 191 patients with a median of 81 patients. More than a half of studies were
published on non-imaging journals (55.2%), while the majority of first authorship belonged to radiologists (65.5%). The most utilized imaging modality was MRI (48.3%). Almost half of the models aimed to predict the response to NAC (48.5%), followed by prognostic models for survival (18.2%) and those for recurrence or metastasis (9.1%). Most models were validated within the same data with or without resampling (48.5%), while a limited number of models were externally validated (12.1%). The detailed characteristics of studies are present in Additional file 1: Tables S7–S10.

### Study quality

Figure 3 summarized the results of study quality evaluation. Table 2 showed that the median (range) of RQS for current osteosarcoma radiomics studies was 10 (3–18), with a percentage of the ideal score of 29.2% (305/1044) and the adherence rate of 44.6% (207/464). Tables 3 and 4 presented that the TRIPOD and CLAIM adherence rates were 59.2% (481/812) and 63.7% (961/1508), respectively. The risk of bias and applicability concerns were mainly related to the index test. The individual assessment for each study is present in Additional file 1: Tables S11–S14.

RQS addressed a radiomics-specific issue of phantom study (0.0%) and the deficiency in cut-off analysis (0.0%) and cost-effectiveness analysis (0.0%). TRIPOD emphasized the shortness in reporting title (6.9%), blindness of assessment for outcomes and predictors (10.3%; 13.7%), and stating study objective in introduction (24.1%). Both RQS and CLAIM indicated a low percentage of comparing the model with the benchmark (27.6%; 27.6%), while both TRIPOD and CLAIM pointed out the disadvantages in sample size or power calculation (10.3%; 13.7%), and missing data handling (20.7%; 20.7%). CLAIM identified extra lacking in reporting in data de-identification (10.3%), stating study hypothesis in introduction (13.8%), and failure analysis (17.2%). All the above three tools emphasized the validation (25/145, 17.2%; 32/64, 50.0%; 16/29, 55.2%) and open science or additional information (10/116, 8.6%; 6/58, 10.3%; 6/87, 6.9%). The correlation between RQS and TRIPOD ($r = 0.7498, p < 0.001$) was moderate, while that between TRIPOD and CLAIM

| Table 2 | RQS rating of included studies |
|---------|-------------------------------|
| **16 items according to 6 key domains** | Range | Median (range) | Percentage of ideal score, n (%) | Adherence rate, n (%) |
| **Total 16 items** | –8–36 | 10 (3–18) | 305/1044 (29.2) | 207/464 (44.6) |
| **Domain 1: protocol quality and stability in image and segmentation** | 0–5 | 2 (0–3) | 50/145 (34.5) | 50/116 (43.1) |
| Protocol quality | 0–2 | 1 (0–1) | 22/58 (37.9) | 22/29 (75.9) |
| Multiple segmentations | 0–1 | 1 (0–1) | 20/29 (69.0) | 20/29 (69.0) |
| Test–retest | 0–1 | 0 (0–1) | 8/29 (27.6) | 8/29 (27.6) |
| Phantom study | 0–1 | 0 (0–0) | 0/29 (0.0) | 0/29 (0.0) |
| **Domain 2: feature selection and validation** | –8 to 8 | 5 (–8 to 8) | 94/232 (40.5) | 49/58 (84.5) |
| Feature reduction or adjustment of multiple testing | –3 to 3 | 3 (–3–3) | 69/87 (79.3) | 26/29 (89.7) |
| Validation | –5 to 5 | 2 (–5 to 5) | 25/145 (17.2) | 23/29 (79.3) |
| **Domain 3: biologic/clinical validation and utility** | 0–6 | 2 (0–5) | 69/174 (39.7) | 61/116 (52.6) |
| Non-radiomics features | 0–1 | 1 (0–1) | 18/29 (62.1) | 18/29 (62.1) |
| Biologic correlations | 0–1 | 1 (0–1) | 27/29 (93.1) | 27/29 (93.1) |
| Comparison to “gold standard” | 0–2 | 0 (0 to 2) | 16/58 (27.6) | 8/29 (27.6) |
| Potential clinical utility | 0–2 | 0 (0–1) | 8/58 (13.8) | 8/29 (27.6) |
| **Domain 4: model performance index** | 0 to 5 | 2 (1–4) | 61/145 (42.1) | 35/87 (40.2) |
| Cut-off analysis | 0–1 | 0 (0–0) | 0/29 (0.0) | 0/29 (0.0) |
| Discrimination statistics | 0–2 | 2 (1–2) | 49/58 (84.5) | 29/29 (100.0) |
| Calibration statistics | 0–2 | 0 (0–2) | 12/58 (20.7) | 6/29 (20.7) |
| **Domain 5: high level of evidence** | 0–8 | 0 (0–7) | 21/232 (9.1) | 3/58 (5.2) |
| Prospective study | 0–7 | 0 (0–7) | 21/203 (10.3) | 3/29 (10.3) |
| Cost-effectiveness analysis | 0–1 | 0 (0–0) | 0/29 (0.0) | 0/29 (0.0) |
| **Domain 6: open science and data** | 0–4 | 0 (0–2) | 10/116 (8.6) | 9/29 (31.0) |

The ideal score was described as score and percentage of score to ideal score for each item. In the cases where a score of one point per item was obtained, the study was considered to have basic adherence to each item. The adherence rate was calculated as proportion of the number of articles with basic adherence to number of total articles.

RQS Radiomics Quality Score
Table 3  TRIPOD adherence of included studies

| 37 Selected items in 22 criteria according to 7 sections (N = 29) | Study, n (%) |
|---------------------------------------------------------------|--------------|
| Overall (excluding items 5c, 11, 14b, 10c, 10e, 12, 13, 17, and 19a) | 481/812 (59.2) |
| **Section 1: Title and Abstract** | |
| 1. Title—identify developing/validating a model, target population, and the outcome | 18/58 (31.0) |
| 2. Abstract—provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions | 16/29 (55.2) |
| **Section 2: Introduction** | |
| 3a. Background—Explain the medical context and rationale for developing/validating the model | 36/58 (62.1) |
| 3b. Objective—Specify the objectives, including whether the study describes the development/validation of the model or both | 7/29 (24.1) |
| **Section 3: Methods** | |
| 4a. Source of data—describe the study design or source of data (randomized trial, cohort, or registry data) | 29/29 (100.0) |
| 4b. Source of data—specify the key dates | 29/29 (100.0) |
| 5a. Participants—specify key elements of the study setting including number and location of centers | 29/29 (100.0) |
| 5b. Participants—describe eligibility criteria for participants (inclusion and exclusion criteria) | 22/29 (75.9) |
| 5c. Participants—give details of treatment received, if relevant (N = 25) | 16/25 (64.0) |
| 6a. Outcome—clearly define the outcome, including how and when assessed | 27/29 (93.1) |
| 6b. Outcome—report any actions to blind assessment of the outcome | 3/29 (10.3) |
| 7a. Predictors—clearly define all predictors, including how and when assessed | 10/29 (34.5) |
| 7b. Predictors—report any actions to blind assessment of predictors for the outcome and other predictors | 4/29 (13.8) |
| 8. Sample size—explain how the study size was arrived at | 3/29 (10.3) |
| 9. Missing data—describe how missing data were handled with details of any imputation method | 6/29 (20.7) |
| 10a. Statistical analysis methods—describe how predictors were handled | 29/29 (100.0) |
| 10b. Statistical analysis methods—specify type of model, all model-building procedures (any predictor selection), and method for internal validation | 21/29 (72.4) |
| 10d. Statistical analysis methods—specify all measures used to assess model performance and if relevant, to compare multiple models (discrimination and calibration) | 6/29 (20.7) |
| 11. Risk groups—provide details on how risk groups were created, if done (N=0) | n/a |
| **Section 4: Results** | |
| 13a. Participants—describe the flow of participants, including the number of participants with and without the outcome. A diagram may be helpful | 16/29 (55.2) |
| 13b. Participants—describe the characteristics of the participants, including the number of participants with missing data for predictors and outcome | 26/29 (89.7) |
| 14a. Model development—specify the number of participants and outcome events in each analysis | 23/29 (79.3) |
| 14b. Model development—report the unadjusted association between each candidate predictor and outcome, if done (N = 5) | 4/5 (80.0) |
| 15a. Model specification—present the full prediction model to allow predictions for individuals (regression coefficients, intercept) | 21/29 (72.4) |
| 15b. Model specification—explain how to use the prediction model (nomogram, calculator, etc.) | 11/29 (37.9) |
| 16. Model performance—report performance measures (with confidence intervals) for the prediction model | 20/29 (69.0) |
| **Section 5: Discussion** | |
| 18. Limitations—Discuss any limitations of the study | 28/29 (96.6) |
| 19b. Interpretation—Give an overall interpretation of the results | 29/29 (100.0) |
| 20. Implications—Discuss the potential clinical use of the model and implications for future research | 29/29 (100.0) |
| **Section 6: Other information** | |
| 21. Supplementary information—provide information about the availability of supplementary resources, such as study | 0/29 (0.0) |
| 22. Funding—give the source of funding and the role of the funders for the present study | 6/29 (20.7) |
| **Section 7: Validation for Model type 2a, 2b, 3, and 4 (N = 16)** | |
| 10c. Statistical analysis methods—describe how the predictions were calculated | 15/16 (93.8) |
| 10e. Statistical analysis methods—describe any model updating (recalibration), if done (N = 0) | n/a |
| 12. Development versus validation—Identify any differences from the development data in setting, eligibility criteria, outcome, and predictors | 10/16 (62.5) |
| 13c. Participants (for validation)—show a comparison with the development data of the distribution of important variables | 2/16 (12.5) |
| 17. Model updating—report the results from any model updating, if done (N=0) | n/a |
| 19a. Interpretation (for validation)—discuss the results with reference to performance in the development data and any other validation data | 5/16 (31.3) |
(r = 0.9004, p < 0.001) and that between RQS and CLAIM (r = 0.8158, p < 0.001) were high (Additional file 1: Fig. S1).

Figure 4 presents results of study quality evaluation with impact factor, sample size, and publication year. We compared the quality of studies published before and after the previous review and found that the ideal percentage of RQS (22.7% vs. 33.8%, p = 0.020), the TRIPOD adherence rate (53.6% vs. 63.4%, p = 0.026), and the CLAIM adherence rate (56.1% vs. 69.1%, p = 0.007) have all been improved (Additional file 1: Table S15 and Additional file 1: Fig. S2). Subgroup analysis also found that imaging modalities utilized in studies have influence on TRIPOD and CLAIM adherence rates (p = 0.002, p = 0.004). The journal type and first authorship did not significantly influence study quality (both p > 0.05).

### Meta-analysis

The meta-analysis of radiomics predicting NAC response by MRI presented a DOR of 28.83 (95% CI 10.27–80.95) on testing datasets of 115 osteosarcoma patients in total [35, 39, 55, 57] (Fig. 5). The corresponding metrics indicates a dramatic performance (Additional file 1: Figs. S3–S7). The Cochran’s Q test (Q = 5.18, p = 0.160) and Higgins I-square statistic (I² = 42.04%) indicated that the heterogeneity was moderate. The funnel plot with Egger’s test (p = 0.035) and Begg’s test (p = 0.089) and the Deeks’ funnel plot with Deeks’ asymmetry test (p = 0.069) revealed that the likelihood of publication bias was high (Additional file 1: Figs. S8–S9). The trim and fill analysis estimated that two studies were missing (Additional file 1: Fig. S10). However, the adjusted DOR was 20.53 (95% CI 7.80–54.06; p < 0.001). The level of evidence supporting the application of radiomics in predicting NAC response by MRI is rated as weak (Table 5). All meta-analyzed data are presented in Additional file 1: Table S16.

### Discussion

We provided an updated systematic review on osteosarcoma radiomics. Although the overall methodological and reporting quality of included studies was still suboptimal, it has improved after the publication of the previous review. The evidence supporting MRI-driven radiomics to predict NAC response in osteosarcoma has been rated as weak based on meta-analysis of testing data. CLAIM has shown unique ability in capturing deficiency in radiomics studies with deep learning.

In the previous review, the most frequently investigated question was whether radiomics could predict the NAC response [8], and it is still the most attractive topic nowadays in osteosarcoma radiomics [29–32, 35, 38–44, 49, 53, 55, 57]. The current review identified two studies each for differential diagnosis [37, 54], for metastasis at diagnosis [46, 47], and for early recurrence [33, 34] of osteosarcoma, while none of the previous twelve studies touched upon these topics. These achievements cover the routine for osteosarcoma and have potential in aiding clinicians to improve their treatment decision. MRI is currently the most frequently utilized imaging modality, and CT has exceeded PET to become the second. In terms of MRI techniques, T1 mapping and dynamic contrast-enhanced MRI have been introduced into osteosarcoma radiomics studies [31, 55]. However, whether these advanced techniques allow radiomics to better answer the clinical questions has not been fully investigated. Although most of studies segmented ROIs manually, two studies and one study, respectively, employed the region growing method based on the threshold of SUV [40, 42] and a deep learning nnU-Net [57] for automatic segmentation. These approaches may liberate radiologists from time-consuming segmentation workloads and potentially make osteosarcoma radiomics an automatic pipeline for clinical use. In addition to segmentation, deep learning models have been compared with radiomics models and showed better performance in predicting NAC response and metastasis [41, 42], and the performance of radiomics models improved when incorporating deep learning features [50]. The application of deep learning has not been detected by the previous review, but currently more studies used deep learning to further mine information in images. More studies tested their model using datasets from other institutions [33–35, 45] or splitting testing datasets [39, 44, 46–48, 50–57] to show the true performance of their models, while none of the studies in the previous review has been externally validated. The improvements in validation settings allow us to meta-analyze the performance of radiomics for prediction of NAC response based on testing datasets. The pooled DOR is lower than that in the previous review (28.83 vs 43.68), but result of the present review is more robust and interpretable [23]. We only included MRI-driven radiomics models which have been evaluated on testing datasets [35, 39, 55, 57], while the previous meta-analysis was carried out based on any imaging modality or dataset.
| Table 4 | CLAIM adherence of included studies |
|---|---|
| CLAIM items (N = 29) | Study, n (%) |
| Overall (excluding item 27) | 961/1508 (63.7) |
| Section 1: Title and Abstract | 53/58 (91.4) |
| 1. Title or abstract—Identification as a study of AI methodology | 29/29 (100.0) |
| 2. Abstract—Structured summary of study design, methods, results, and conclusions | 24/29 (82.8) |
| Section 2: Introduction | 55/87 (63.2) |
| 3. Background—Scientific and clinical background, including the intended use and clinical role of the AI approach | 29/29 (100.0) |
| 4a. Study objective | 22/29 (75.9) |
| 4b. Study hypothesis | 4/29 (13.8) |
| Section 3: Methods | 700/1044 (67.0) |
| 5. Study design—Prospective or retrospective study | 29/29 (100.0) |
| 6. Study design—Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial | 29/29 (100.0) |
| 7a. Data—Data source | 29/29 (100.0) |
| 7b. Data—Data collection institutions | 29/29 (100.0) |
| 7c. Data—Imaging equipment vendors | 25/29 (86.2) |
| 7d. Data—Image acquisition parameters | 22/29 (75.9) |
| 7e. Data—Institutional review board approval | 28/29 (96.6) |
| 7f. Data—Participant consent | 24/29 (82.8) |
| 8. Data—Eligibility criteria | 22/29 (75.9) |
| 9. Data—Data pre-processing steps | 20/29 (69.0) |
| 10. Data—Selection of data subsets (segmentation of ROI in radiomics studies) | 26/29 (89.7) |
| 11. Data—Definitions of data elements, with references to Common Data Elements | 29/29 (100.0) |
| 12. Data—De-identification methods | 3/29 (10.3) |
| 13. Data—How missing data were handled | 6/29 (20.7) |
| 14. Ground truth—Definition of ground truth reference standard, in sufficient detail to allow replication | 27/29 (93.1) |
| 15a. Ground truth—Rationale for choosing the reference standard (if alternatives exist) | 0/29 (0.0) |
| 15b. Ground truth—Definitive ground truth | 29/29 (100.0) |
| 16. Ground truth—Manual image annotation | 17/29 (58.6) |
| 17. Ground truth—Image annotation tools and software | 10/29 (34.5) |
| 18. Ground truth—Measurement of inter- and intra-rater variability; methods to mitigate variability and/or resolve discrepancies | 9/29 (31.0) |
| 19a. Data Partitions—Intended sample size and how it was determined | 29/29 (100.0) |
| 19b. Data Partitions—Provided power calculation | 4/29 (13.8) |
| 19c. Data Partitions—Distinct study participants | 23/29 (79.3) |
| 20. Data Partitions—How data were assigned to partitions; specify proportions | 22/29 (75.9) |
| 21. Data Partitions—Level at which partitions are disjoint (e.g., image, study, patient, institution) | 22/29 (75.9) |
| 22a. Model—Provided reproducible model description | 21/29 (72.4) |
| 22b. Model—Provided source code | 0/29 (0.0) |
| 23. Model—Software libraries, frameworks, and packages | 20/29 (69.0) |
| 24. Model—Initialization of model parameters (e.g., randomization, transfer learning) | 23/29 (79.3) |
| 25. Training—Details of training approach, including data augmentation, hyperparameters, number of models trained | 16/29 (55.2) |
| 26. Training—Method of selecting the final model | 21/29 (72.4) |
| 27. Training—Ensembling techniques, if applicable (N = 14) | 8/14 (57.1) |
| 28. Evaluation—Metrics of model performance | 29/29 (100.0) |
| 29. Evaluation—Statistical measures of significance and uncertainty (e.g., confidence intervals) | 20/29 (69.0) |
| 30. Evaluation—Robustness or sensitivity analysis | 10/29 (34.5) |
| 31. Evaluation—Methods for explainability or interpretability (e.g., saliency maps), and how they were validated | 11/29 (37.9) |
| 32. Evaluation—Validation or testing on external data | 16/29 (55.2) |
| Section 4: Results | 90/174 (51.7) |
| 33. Data—Flow of participants or cases, using a diagram to indicate inclusion and exclusion | 16/29 (55.2) |
| 34. Data—Demographic and clinical characteristics of cases in each partition | 25/29 (86.2) |
Study quality has improved since the publication of the previous review. However, the overall study quality is suboptimal. RQS and TRIPOD have identified disadvantages in phantom study, cut-off analysis, cost-effectiveness analysis, blindness of assessment, sample size calculation, and missing data handling, which have
been repeatedly addressed [8, 12–14]. The previous review only employed RQS for quality assessment. RQS was a specialized tool proposed to help the radiomics community assess the quality and value of a radiomics study. However, RQS was tailored on hand-crafted features. As deep learning is gaining momentum, the current version of RQS may not capture the strengths and weaknesses of deep learning radiomics studies correctly [58]. TRIPOD is a similar example that aimed to promote transparency reporting of diagnostic accuracy model studies and has been recommended to identify room for improvements in radiomics studies [11]. Nevertheless, the current version of TRIPOD may not capture some unique challenges with machine learning or AI application [59]. In contrast, CLAIM captured unique shortness in our review, such as data de-identification and failure analysis. CLAIM has been employed as a useful tool for quality evaluation in deep learning studies [20, 21], and our review demonstrated the feasibility of CLAIM in radiomics studies. We further confirmed that CLAIM can serve as a better review and study design guideline in radiomics studies. CLAIM may guide the update of TRIPOD and RQS, because it not only includes general reporting criteria, but also allows extra distinction of unique shortness in deep learning. CLAIM may even replace RQS and TRIPOD, considering the overlapping items and high correlation between these tools. The researchers are still reticent in publishing the RQS and TRIPOD for their radiomics studies [58]. Only one study in our review included RQS, TRIPOD and CLAIM as supplementary materials [57].

Our review has several limitations. First, our review focused on osteosarcoma radiomics studies. The conclusion should be interpreted with caution when expanded to other diseases. However, it provided insights for the design and reporting radiomics studies. Second, our study only included AI research applying the radiomics approach, but overlooked those conducted with only deep learning for segmentation [60–62] or modeling [63]. However, the secondary aim of our study is to find out whether CLAIM can better identify disadvantages in radiomics studies than the currently recommended RQS and TRIPOD. As CLAIM is suitable for both radiomics and deep learning radiomics studies, future review is encouraged to carry out without the restriction of radiomics. Third, it has not been investigated in our review how to weigh each item in CLAIM. The previous reviews have created subitems for some evaluations [20] and weighed them as equal [21]. We treated each subitem as equal, but it is necessary to find out whether it is appropriate. Fourth, we did not employ more specific tools for evaluation, because they are not suitable or are currently under development [59, 64–68]. The review may benefit from the increasing study reporting guidelines for clinical studies using AI in healthcare, because they pay extra attention to additional factors which do not neatly conform to traditional reporting guidelines, especially details relating to technical algorithm development [69].

| Study       | Good | Poor |
|-------------|------|------|
|             | TP   | FP   | FN   | TN   |
| Zhong 2022  | 9    | 7    | 2    | 25   |
| Zhang 2021  | 10   | 1    | 2    | 8    |
| Dufau 2019  | 10   | 1    | 0    | 6    |
| Chen 2021   | 8    | 2    | 3    | 21   |

**Diagnostic odds ratio [95% CI]**

- **Zhong 2022**: 16.07 [2.80 - 92.15]
- **Zhang 2021**: 40.00 [3.05 - 524.83]
- **Dufau 2019**: 91.00 [3.20 - 1000.00]
- **Chen 2021**: 28.00 [3.92 - 199.94]
- **COMBINED**: 28.83 [10.27 - 80.95]

Q = 5.18, df = 3.00, p = 0.16
I² = 42.04 [0.00 - 100.00]

**Fig. 5** Forest plots of diagnostic odds ratios. The performance of radiomics in prediction of NAC response in osteosarcoma patients based on testing datasets. TP pathological good responders predicted as good responders, FP pathological poor responders predicted as good responders, FN pathological good responders predicted as poor responders, TN pathological poor responders predicted as poor responders.
The prediction performance of radiomics for NAC response in osteosarcoma patients

| Clinical question | MRI-driven radiomics prediction model for NAC response in osteosarcoma patients |
|-------------------|--------------------------------------------------------------------------------|
| Number of studies | 4                                                                               |
| Good responder/sample size | 44/115                                                                          |

Pooled analysis

- DOR (95% CI): 28.83 (10.27–80.95)
- p value for DOR: p < 0.001
- Sensitivity (95% CI): 0.84 (0.70–0.92)
- Specificity (95% CI): 0.85 (0.74–0.91)
- PLR (95% CI): 5.43 (3.11–9.49)
- NLR (95% CI): 0.19 (0.09–0.37)
- AUC (95% CI): 0.91 (0.88–0.93)
- Heterogeneity: Higgins 𝑖^2 test, 𝑖^2 = 42.04%
- Publication bias: Cochran’s Q test, Q = 5.18, p = 0.160
- Trim and fill method

- Number of missing studies: 2
- Adjusted DOR (95% CI): 20.53 (7.80–54.06)
- p value for adjusted DOR: p < 0.001
- Level of Evidence: Weak

**Clinical question MRI-driven radiomics prediction model for NAC response in osteosarcoma patients**

**Number of studies**: 4

**Good responder/sample size**: 44/115

**Pooled analysis**

- **DOR (95% CI)**: 28.83 (10.27–80.95)
- **p value for DOR**: p < 0.001
- **Sensitivity (95% CI)**: 0.84 (0.70–0.92)
- **Specificity (95% CI)**: 0.85 (0.74–0.91)
- **PLR (95% CI)**: 5.43 (3.11–9.49)
- **NLR (95% CI)**: 0.19 (0.09–0.37)
- **AUC (95% CI)**: 0.91 (0.88–0.93)

**Heterogeneity**

- **Higgins 𝑖^2 test**: 𝑖^2 = 42.04%
- **Cochran’s Q test**: Q = 5.18, p = 0.160

**Publication bias**

- **Egger’s test**: p = 0.035
- **Begg’s test**: p = 0.069
- **Deeks’ test**: p = 0.069

**Trim and fill method**

- **Number of missing studies**: 2
- **Adjusted DOR (95% CI)**: 20.53 (7.80–54.06)
- **p value for adjusted DOR**: p < 0.001
- **Level of Evidence**: Weak

**Table 5** The prediction performance of radiomics for NAC response in osteosarcoma patients

Abbreviations

- AI: Artificial intelligence
- CI: Confidence interval
- CLAIM: Checklist for Artificial Intelligence in Medical Imaging
- DOR: Diagnostic odds ratio
- NAC: Neoadjuvant chemotherapy
- QUADAS-2: Modified Quality Assessment of Diagnostic Accuracy Studies
- RQS: Radiomics Quality Score
- SROC: Summary receiver operating characteristic
- TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

**Supplementary Information**

The online version contains supplementary material available at [https://doi.org/10.1186/s13244-022-01277-6](https://doi.org/10.1186/s13244-022-01277-6)

**Additional file 1** Note S1: Review protocol. Note S2: Search strategy and study selection. Note S3: Consensus reached during data extraction and quality assessment. Note S4: Data synthesis and analysis methods. Note S5: Table S1: PICOT of included studies. Table S2: Radiomics methodology issue of included studies. Table S3: Model presentation and performance metrics of included studies. Table S4: Table S5: QUADAS-2 tool for risk of bias and concern on application. Table S6: Category of five levels of supporting evidence of meta-analyses. Table S7: Study characteristics of included studies. Table S8: PICOT of included studies. Table S9: Radiomics methodological issue of included studies. Table S10: Model presentation and performance metrics of included studies. Table S11: RQS rating per study. Table S12: TRIPOD adherence per study. Table S13: CLAIM adherence per study. Table S14: QUADAS-2 assessment per study. Table S15: Subgroup analysis of study quality according to study characteristics. Table S16: Model metrics of studies included in meta-analysis. Figure S1: Forrest plot of pooled positive likelihood ratio. Figure S2: Forrest plot of pooled negative likelihood ratio. Figure S3: Forrest plot of pooled specificity. Figure S4: Forrest plot of pooled sensitivity. Figure S5: Forrest plot of pooled positive likelihood ratio. Figure S6: Forrest plot of pooled negative likelihood ratio. Figure S7: SROC curve of the model performance. Figure S8: Funnel plot of studies included in meta-analysis. Figure S9: Deeks’ funnel plot of studies included in meta-analysis. Figure S10: Trim and fill analysis of studies included in meta-analysis.

**Acknowledgments**

The authors would like to express their gratitude to Dr. Jia Geng and Dr. Leping Si for their contribution on the previous version of this systematic review and meta-analysis, and Dr. Shiqi Mao for his suggestions on data visualization.

**Author contributions**

JYZ and YYH performed the literature search, data extraction, and quality assessment. JYZ performed meta-analyses, visualized data, and drafted the original version of manuscript. All authors read and approved the final manuscript.

**Funding**

This study has received funding by Yangfan Project of Science and Technology Commission of Shanghai Municipality (22YF1442400), Shanghai Science and Technology Commission Science and Technology Innovation Action Plan Innovation Development Field (18411953000), Medicine and Engineering Combination Project of Shanghai Jiao Tong University (YG2019ZDB09), and Research Fund of Tongren Hospital, Shanghai Jiao Tong University School of Medicine (TRKYRC-XX202204, 2020TRY1(JL)B06, 2020TRY1(JC)C07, TRGG202101, TRYJ2021JC06). They played no role in the study design, data collection or analysis, decision to publish, or manuscript preparation.

**Availability of data and materials**

All data generated or analyzed during this study are included in this published article and its supplementary information files.

**Declarations**

**Ethics approval and consent to participate**

Not applicable.
Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1 Department of Imaging, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, No. 1111 Xianxia Road, Shanghai 200336, China. 2 Department of Sports Medicine, Shanghai Jiao Tong University Affiliated Sixth People’s Hospital, No. 600 Yishan Road, Shanghai 200233, China. 3 Department of Pathology, Shanghai Jiao Tong University Affiliated Sixth People’s Hospital, No. 600 Yishan Road, Shanghai 200233, China. 4 Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197 Rujin 2nd Road, Shanghai 200025, China.

Received: 6 June 2022  Accepted: 24 July 2022
Published online: 20 August 2022

References
1. WHO Classification of Tumours Edition Board (2020) World Health organization classification of tumours: WHO classification of tumours of soft tissue and bone, 5th edn. IARC Press, Lyon
2. Strauss SJ, Frezza AM, Abecassis N et al Guidelines Committee, EURACAN, GENTURIS and ERN PaedCan (2021) Bone sarcomas: ESMO-EURACAN-GENTURIS-ERN PaedCan clinical practice guideline for diagnosis, treatment and follow-up. Ann Oncol 32(12):1520–1536. https://doi.org/10.1093/annonc/mdab189
3. National Comprehensive Cancer Network (2021) NCCN clinical practice guidelines in oncology: bone cancer, version 2. 2022. https://www.nccn.org/professionals/physician_gls/pdf/bone.pdf. Accessed 8 Oct 2021
4. Lambin P, Rios-Velázquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48(4):441–446. https://doi.org/10.1016/j.ejca.2011.11.036
5. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278(2):563–577. https://doi.org/10.1148/radiol.2015151169
6. O’Connor JP, Aboagye EO, Adams JE et al (2017) Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 14:169–186. https://doi.org/10.1038/nrclinonc.2016.162
7. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14(12):749–762. https://doi.org/10.1038/nrclinonc.2017.141
8. Zhong J, Hu Y, Si L et al (2021) A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. Eur Radiol 31(3):1526–1535. https://doi.org/10.1007/s00330-020-07221-w
9. Whiting PF, Rutjes AW, Westwood ME et al QUADAS-2 Group (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
10. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 162(1):55–63. https://doi.org/10.7326/MII4-0697
11. Park SH (2022) Guidelines for the successful conduct and reporting of systematic review and meta-analysis of diagnostic test accuracy studies. Korean J Radiol 23(3):295–297. https://doi.org/10.3348/kjr.20210963
12. Park JE, Kim D, Kim HS et al (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 30(1):523–536. https://doi.org/10.1007/s00330-019-06350-z
13. Won SY, Park YW, Park M, Ahn SS, Kim J, Lee SK (2020) Quality reporting of radiomics analysis in mild cognitive impairment and Alzheimer’s disease: a roadmap for moving forward. Korean J Radiol 21(2):1345–1354. https://doi.org/10.3348/kjr.2020.0715
14. Park CJ, Park YW, Ahn SS et al (2022) Quality of radiomics research on brain metastasis: a roadmap to promote clinical translation. Korean J Radiol 23(1):77–88. https://doi.org/10.3348/kjr.20210421
15. Bi WL, Hosny A, Schabath MB et al (2019) Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 69:127–157. https://doi.org/10.3332/caac.21552
16. Shur JD, Doran SJ, Kumar S et al (2021) Radiomics in oncology: a practical guide. Radiographics 41(6):1717–1732. https://doi.org/10.1148/rg.2021200010
17. Cheng PM, Montagon E, Yamashita R et al (2021) Deep learning an update for radiologists. Radiographics 41(5):1427–1445. https://doi.org/10.1148/rg.2021200021
18. Marti-Bonmati L, Koh DM, Riklund K et al (2022) Considerations for artificial intelligence clinical impact in oncologic imaging: an AHAII position paper. Insights Imaging 13:189. https://doi.org/10.1186/ s13442-022-01226-9
19. Morgan J, Moy L, Kahn CE J (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2(2):e200029. https://doi.org/10.1148/rai.2020200029
20. O’Shea RJ, Sharkey AR, Cook GJR, Goh V (2021) Systematic review of research design and reporting of imaging studies applying convolutional neural networks for radiological cancer diagnosis. Eur Radiol 31(10):7969–7983. https://doi.org/10.1007/s00330-021-08781-2
21. Si L, Zhong J, Huo J, et al (2022) Deep learning in knee imaging: a systematic review using a checklist for artificial intelligence in medical imaging (CLAIM). Eur Radiol 32(2):1353–1361. https://doi.org/10.1007/s00330-021-08190-4
22. Dang Y, Hou Y (2021) The prognostic value of late gadolinium enhancement in heart diseases: an umbrella review of meta-analyses of observational studies. Eur Radiol 31(7):4528–4537. https://doi.org/10.1007/s00330-020-07437-w
23. Gitto S, Cuocolo R, Albano D et al (2021) CT and MRI radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies. Insights Imaging 12(11):68. https://doi.org/10.1186/s13442-021-01008-3
24. Crombé A, Fadli D, Italiano A, Saut Q, Bux Y, Kind M (2020) Systematic review of sarcomas radiomics studies: bridging the gap between concepts and clinical applications? Eur J Radiol 132:109283. https://doi.org/10.1016/j.ejrad.2020.109283
25. Garner P, Hopewell S, Chandler J et al Panel for updating guidance for systematic reviews (PUGs) (2016) When and how to update systematic reviews: consensus and checklist. BMJ 354:i3507. https://doi.org/10.1136/bmj.i3507
26. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
27. Mangiafico SS (2016) Summary and analysis of extension program evaluation in R, version 1.1.90. http://rcompanion.org/handbook/. Accessed May 2022
28. Cochrane screening and diagnostic test methods group (2022) Cochrane handbook for systematic reviews of diagnostic test accuracy, version 2. https://training.cochrane.org/handbook-diagnostic-test-accuracy. Accessed May 2022
29. Baidya Kayal E, Kandasamy D, Khare K, Bakhshi S, Sharma R, Mehndiratta A (2019) Intravoxel incoherent motion (IVIM) for response assessment in patients with osteosarcoma undergoing neoadjuvant chemotherapy. Eur J Radiol 119:108635. https://doi.org/10.1016/j.ejrad.2019.08.004
30. Baidya Kayal E, Kandasamy D, Khare K, Bakhshi S, Sharma R, Mehndiratta A (2021) Texture analysis for chemotherapy response evaluation in osteosarcoma using MRI imaging. NMR Biomed 34(2):e4426. https://doi.org/10.1002/nbm.4426
31. Baidya Kayal E, Sharma N, Sharma R, Bakhshi S, Kandasamy D, Mehndiratta A (2022) T1 mapping as a surrogate marker of chemotherapy response evaluation in patients with osteosarcoma. Eur J Radiol 148:110170. https://doi.org/10.1016/j.ejrad.2022.110170
32. Bailly C, Leforestier R, Campion L et al (2017) Prognostic value of FDG-PET indices for the assessment of histological response to neoadjuvant chemotherapy and outcome in pediatric patients with Ewing sarcoma and osteosarcoma. PLoS One 12(8):e0183841. https://doi.org/10.1371/journal.pone.0183841
33. Chen H, Liu J, Cheng Z et al (2020) Value of radiomics nomogram based on T1WI for pretreatment prediction of relapse within 1 year in osteosarcoma: a multicenter study. Chin J Radiol 54(9):874–881. https://doi.org/10.3760/cma.j.issn.112149-20200512-00675 (in Chinese)

34. Chen H, Liu J, Cheng Z et al (2020) Development and external validation of an MRI-based radiomics nomogram for pretreatment prediction for early relapse in osteosarcoma: a retrospective multicenter study. Eur J Radiol 129:109066. https://doi.org/10.1016/j.ejrad.2020.109066

35. Chen H, Zhang X, Wang X et al (2021) MRI-based radiomics signature for pretreatment prediction of pathological response to neoadjuvant chemotherapy in osteosarcoma: a multicenter study. Eur Radiol 31(10):7913–7924. https://doi.org/10.1007/s00330-021-07748-6

36. Cho YJ, Kim WS, Choi YH et al (2019) Computerized texture analysis of pelvic osteosarcoma patients with osteosarcoma: differentiation of pelvic osteosarcoma and ewing sarcoma using radiomics analysis based on T2-weighted images and contrast-enhanced T1-weighted images. Biomed Res Int 2020:978603. https://doi.org/10.1155/2020/978603

37. Dai Y, Yin P, Mao N et al (2020) Differentiation of pelvic osteosarcoma and ewing sarcoma using radiomics analysis based on T2-weighted images and contrast-enhanced T1-weighted images. Biomed Res Int 2020:978603. https://doi.org/10.1155/2020/978603

38. Djuričić GJ, Ahammer H, Rajković S (2022) Directionally sensitive fractal machine learning approaches with PCA. Contrast Media Mol Imaging 2022:3515080. https://doi.org/10.1155/2019/3515080

39. Dufau J, Bouhamama A, Leporq B et al (2019) Prediction of chemotherapy response in primary osteosarcoma using the machine learning technique on radiomic data. Bull Cancer 106(11):983–999. https://doi.org/10.1016/j.bulcan.2019.07.005 (in French)

40. Jeong SY, Kim W, Byun BH et al (2019) Prediction of chemotherapy response of osteosarcoma using baseline 18F-FDG textural features machine learning approach with PCA. Contrast Media Mol Imaging 2019:3515080. https://doi.org/10.1155/2019/3515080

41. Kim BC, Kim J, Kim K et al (2021) Preliminary radiogenomic evidence for neoadjuvant chemotherapy in osteosarcoma: a multicenter study. Eur Radiol 31(10):7913–7924. https://doi.org/10.1007/s00330-021-07748-6

42. Cho YJ, Kim WS, Choi YH et al (2019) Computerized texture analysis of pelvic osteosarcoma patients with osteosarcoma: differentiation of pelvic osteosarcoma and ewing sarcoma using radiomics analysis based on T2-weighted images and contrast-enhanced T1-weighted images. Biomed Res Int 2020:978603. https://doi.org/10.1155/2020/978603

43. Lee SK, Jee WH, Jung CK et al (2020) Prediction of poor responders to neoadjuvant chemotherapy in high-grade osteosarcoma. Cancer Imaging 20(1):7. https://doi.org/10.1186/s41591-021-02030-9

44. Lee SK, Jee WH, Jung CK et al (2020) Prediction of poor responders to neoadjuvant chemotherapy in high-grade osteosarcoma. Cancer Imaging 20(1):7. https://doi.org/10.1186/s41591-021-02030-9

45. Liu J, Lian T, Chen H et al (2021) Pretreatment prediction of relapse risk in patients with osteosarcoma using radiomics nomogram based on CT: a retrospective multicenter study. Biomed Res Int 2021:6674471. https://doi.org/10.1155/2021/6674471

46. Luo Z, Li J, Liao Y, Li R, Shen X, Chen W (2022) Radiomics analysis of multiparametric MRI for prediction of synchronous lung metastases in osteosarcoma. Front Oncol 12:802234. https://doi.org/10.3389/fonc.2022.802234

47. Pereira HM, Leite Duarte ME, Ribeiro Damasceno I, de Oliveira Moura Santos LA, Nogueira-Barbosa MH (2021) Machine learning-based CT radiomics features for the prediction of pulmonary metastasis in osteosarcoma. Br J Radiol 94(1142):20210391. https://doi.org/10.1259/bjr/20210391

48. Shen H, Kim W, Byun BH et al (2019) Metastasis risk prediction model in osteosarcoma using metabolomic imaging phenotypes: a multivariable radiomics model. PLoS One 14(11):e0225242. https://doi.org/10.1371/journal.pone.0225242

49. Song H, Jiao Y, Wei W et al (2019) Can pretreatment 18F-FDG PET tumor texture features predict the outcomes of osteosarcoma treated by neoadjuvant chemotherapy? Eur Radiol 29(7):3945–3954. https://doi.org/10.1007/s00330-019-06074-2

50. Yan Y, Zhang L, Gao Q et al (2021) Machine learning-based radiomics nomogram with dynamic contrast-enhanced MRI of the osteosarcoma for evaluation of efficacy of neoadjuvant chemotherapy. Front Oncol 11:758921. https://doi.org/10.3389/fonc.2021.758921

51. Zhao S, Su Y, Duan J et al (2019) Radiomics signature extracted from diffusion-weighted magnetic resonance imaging predicts outcomes in osteosarcoma. J Bone Oncol 19:100263. https://doi.org/10.1016/j.jbo.2019.100263

52. Zhong J, Zhang C, Hu Y et al (2022) Automated prediction of the neoadjuvant chemotherapy response in osteosarcoma with deep learning and an MRI-based radiomics nomogram. Eur Radiol. https://doi.org/10.1007/s00330-022-08735-1

53. Guo J, Vaidyanathan A, Deprez L et al (2022) A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev 42(1):426–440. https://doi.org/10.1002/med.21846

54. Collins GS, Dhiman P, Andaut Navarro CL et al (2021) Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 11(7):e048008. https://doi.org/10.1136/bmjopen-2020-048008

55. Huang L, Xia W, Zhang B, Qiu B, Gao X (2017) MSFCN-multiple supervised fully convolutional networks for the osteosarcoma segmentation of CT images. Comput Methods Progr Biomed 143:67–74. https://doi.org/10.1016/j.cmpb.2017.02.013

56. Zhang R, Huang L, Xia W, Zhang B, Qiu B, Gao X (2018) Multiple supervised residual network for osteosarcoma segmentation in CT images. Comput Med Imaging Graph 63:1–8. https://doi.org/10.1016/j.compmedimag.2018.01.006

57. Wu J, Yang S, Gou F et al (2022) Intelligent segmentation medical assistance system for MRI images of osteosarcoma in developing countries. Comput Math Methods Med 2022:770583. https://doi.org/10.1155/2022/770583

58. Huang B, Wang J, Sun M et al (2020) Feasibility of multi-parametric magnetic resonance imaging combined with machine learning in the assessment of necrosis of osteosarcoma after neoadjuvant chemotherapy: a preliminary study. BMC Cancer 20(1):322. https://doi.org/10.1186/s12885-020-06625-1

59. Soonderajah V, Ashrafian H, Rose S et al (2021) A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 27(10):1663–1665. https://doi.org/10.1038/s41591-021-01517-0

60. Vasely B, Nagendran M, Campbell B et al DECIDE-AI expert group (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 28(5):924–933. https://doi.org/10.1038/s41591-022-01772-9

61. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, SPIRIT-AI and CONSORT-AI Consensus Group (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26(9):1351–1363. https://doi.org/10.1038/s41591-020-1037-7

62. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group (2020) Reporting guidelines for clinical trial...
reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26(9):1364–1374. https://doi.org/10.1038/s41591-020-1034-x

68. Sounderajah V, Ashrafian H, Aggarwal R et al (2020) Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. Nat Med 26(6):807–808. https://doi.org/10.1038/s41591-020-0941-1

69. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ (2021) Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform 28(1):e100385. https://doi.org/10.1136/bmjhci-2021-100385

70. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295(2):328–338. https://doi.org/10.1148/radiol.2020191145

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.