



A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation

Jingyu Zhong¹ · Yangfan Hu² · Liping Si¹ · Geng Jia² · Yue Xing² · Huan Zhang³ · Weiwu Yao¹

Received: 25 May 2020 / Revised: 12 July 2020 / Accepted: 21 August 2020 / Published online: 2 September 2020
© European Society of Radiology 2020

Abstract

Objectives To assess the methodological quality and risk of bias in radiomics studies investigating diagnosis, therapy response, and survival of patients with osteosarcoma.

Methods In this systematic review, literatures on radiomics in osteosarcoma were included and assessed for methodological quality through the radiomics quality score (RQS). The risk of bias and concern of application was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool. A meta-analysis of studies focusing on predicting osteosarcoma response to neoadjuvant chemotherapy was performed.

Results Twelve radiomics studies exploring osteosarcoma were identified, and five were included in meta-analysis. The RQS reached an average of 20.4% (6.92 of 36) with good inter-rater agreement (ICC 0.95, 95% CI 0.85-0.99). Four studies validated results with an internal dataset, none of which used external dataset; one study was prospectively designed, and another one shared part of the dataset. The risk of bias and concern of application were mainly related to index test aspect. The meta-analysis showed a diagnostic odds ratio of 43.68 (95%CI 13.5-141.31) for predicting response to neoadjuvant chemotherapy with high heterogeneity and low methodological quality.

Conclusions The overall scientific quality of included studies is insufficient; however, radiomics remains a promising technology for predicting treatment response, which might guide therapeutic decision-making and related to prognosis. Improvements in study design, validation, and open science needs to be made to demonstrate the generalizability of findings and to achieve clinical applications. Widespread application of RQS, pre-trained RQS scoring procedure, and modification of RQS in response to clinical needs are necessary.

Key Points

- Limited radiomics studies were established in osteosarcoma with mean RQS of 20.4%, commonly due to unvalidated results, retrospective study design, and absence of open science.
- Meta-analysis of radiomics studies predicting osteosarcoma response to neoadjuvant chemotherapy showed high diagnostic odds ratio 43.68, while high heterogeneity and low methodological quality were the main concerns.
- A previously trained data extraction instrument allowed reaching moderate inter-rater agreement in RQS applications, while RQS still needs improvement to become a wide adaptive tool in reviews of radiomics studies, in routine self-check before manuscript submitting and in study design.

Jingyu Zhong and Yangfan Hu contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00330-020-07221-w>) contains supplementary material, which is available to authorized users.

✉ Huan Zhang
huanzhangy@163.com

✉ Weiwu Yao
yaoweiwuhuan@163.com

¹ Department of Imaging, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, No. 1111 Xianxia Road, Changning District, Shanghai 200050, China

² Department of Radiology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, No. 600 Yishan Road, Xuhui District, Shanghai 200233, China

³ Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin 2nd Road, Huangpu District, Shanghai 200025, China

Keywords Osteosarcoma · Machine learning · Quality improvement · Neoadjuvant therapy · Systematic review

Abbreviations

CI	Confidence intervals
DOR	Diagnostic odds ratio
HSROC	Hierarchical summary receiver operating characteristic
ICC	Correlation coefficient
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-analysis
PROSPERO	International Prospective Register Of Systematic Reviews
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RQS	Radiomics quality score
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Introduction

Osteosarcoma is the most common primary malignant osseous sarcoma with most cases developing in children and adolescents [1]. Radiologic examinations are useful tools in osteosarcoma diagnosis [2–5]. Osteosarcoma is often detected on plain radiograph with a contrast-enhanced MRI scan as the next step in the diagnostic work-up; a chest CT scan is essential for lung metastases detection; PET examination or a bone scan is generally recommended for initial staging in osteosarcoma patients [1].

For treatment considerations, chemotherapy has been considered as essential of high-grade osteosarcoma [5]. Surgery of the primary tumor after chemotherapy is a conventional approach [6]; and the histologic response to neoadjuvant chemotherapy evaluated based on tumor necrosis of excision specimen [7] is crucial for treatment strategy and is related to prognosis of patients [8]. Although aggressive treatment plans improve prognosis of patients who were likely to exhibit poor survival, not all patients benefit from these approaches. In clinical settings, expert radiologists may provide informative reports for clinicians to decide treatment strategy [2–4], and if patients could be stratified by radiologic examinations, personalized medicine strategy may be realized [9]. However, imaging interpretation relies largely on radiologists; therefore, reports vary due to uncontrollable subjective factors.

Radiomics, a bunch of strategies extracting quantitative, minable, high-dimensional data from medical images, is capable for generating imaging biomarkers which may not be visible to naked eye [9–12]. Quantitative, reader independent imaging biomarkers could support clinical decision and increase diagnostic, predictive, and prognostic accuracy [13].

In recent years, extensive research using radiomic methods and even artificial intelligence tried and succeeded in linking radiologic image to lesion characterization, treatment response, and patient outcome; nonetheless, translation into clinical practice has not yet realized [14]. For radiomics to cross the translational gap between an exploratory research method and a valued addition to precision medicine workflows, challenges including technical and biological validity and regulatory and ethical problems as well as cost-effectiveness still need to be overcome, in which this process radiomics quality score (RQS) may be employed as a useful tool (Fig. 1) [9, 15, 16].

Furthermore, no previous study has been undertaken a systematic research on radiomics in osteosarcoma. The factors affecting the performance of radiomics in osteosarcoma should be identified to further improve its clinical translation. Thus, the aim of our study was to establish whether the methodological quality of studies published on radiomics in osteosarcoma for multiple purposes poses barriers to effective clinical application. A meta-analysis of the radiomics utility in prediction of neoadjuvant chemotherapy response to osteosarcoma was performed to evaluate its ability of proposed models to answer this clinically relevant question.

Materials and methods

Protocol and registry

This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) statement [17]. A review protocol was drafted [18], and has been submitted to the International Prospective Register Of Systematic Reviews (PROSPERO).

Literature search and study selection

The structured search via PubMed, Embase, and Web of Science was performed until 30 Apr 2020 by two reviewers both with 2 years of experience in radiology, independently. Disagreements were resolved by discussion or with the help of a third reviewer with 4 years of experience. This review included primary research assessing the role of radiomics in patients with osteosarcoma for diagnostic, prognostic, or predictive purpose. Two reviewers selected potential studies by screening titles and abstracts. Articles that met inclusion criteria were obtained in full. The full text was determined for further eligibility by two same reviewers. In the case of uncertainties, a third reviewer was consulted to reach final consensus. The reference lists of included studies were

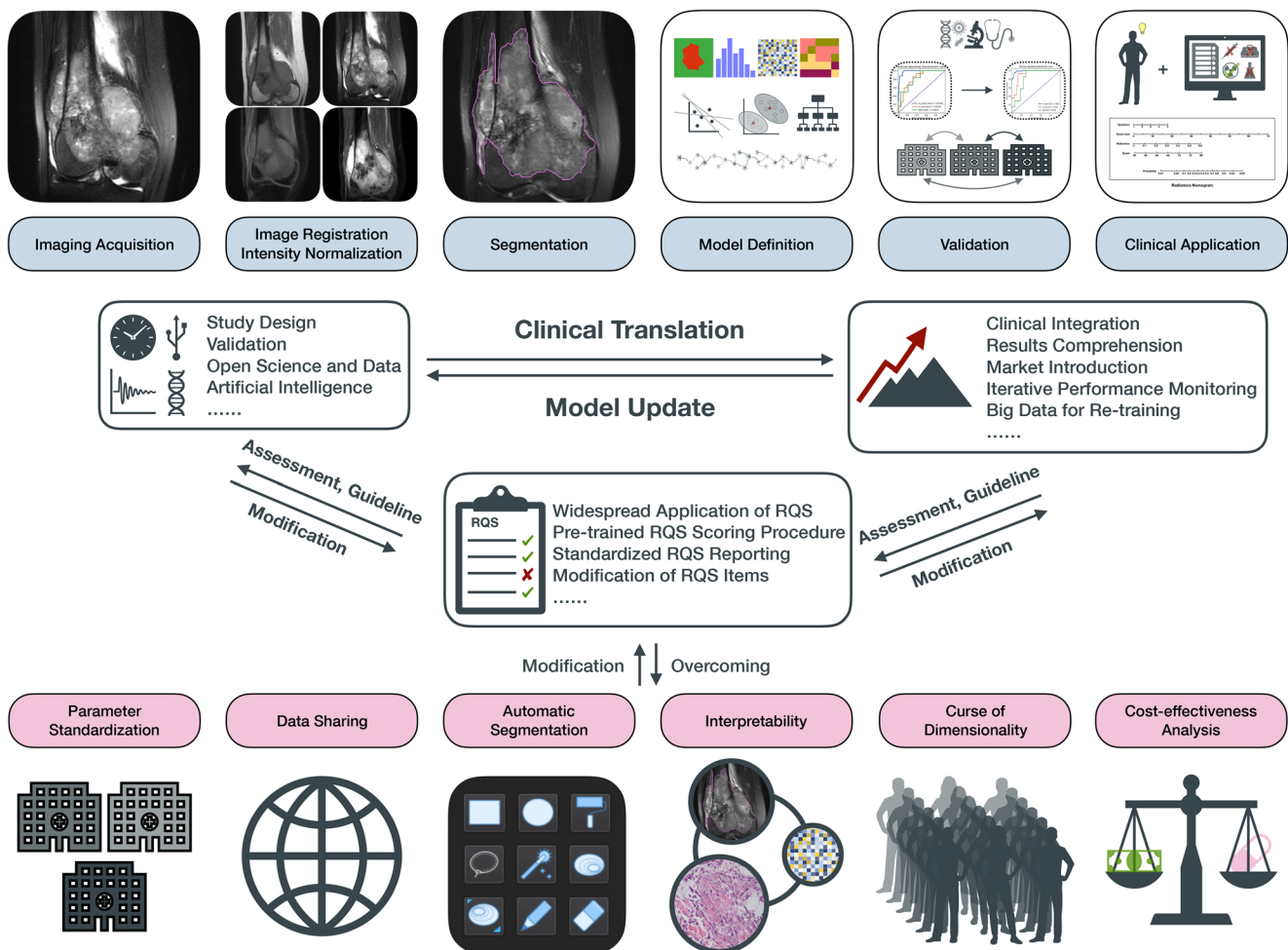


Fig. 1 The radiomics research and role of RQS. A typical radiomics workflow includes image and acquisition and post-processing; manual semi-automatic, or automatic segmentation; model definition using classical machine learning algorithms or deep learning method; external and prospective validation; and finally, clinical application. RQS is a useful tool to assess the methodological quality of this workflow and further

reflecting challenges and insufficiencies in radiomics studies, such as lack of prospective design, absence of external validation, and unwillingness to share data. On the other hand, modification of RQS is deemed to be necessary, either according to other predictive model reporting checklists or in response to actual practical needs

screened for additional, potentially eligible articles. Detailed search strategies and selection criteria can be found in [supplementary materials](#).

Data extraction and quality assessment

The eligible articles were assessed by the RQS for methodological quality [9] and by the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool for the risk of bias and concern of application [19]. The RQS was a recently accepted tool to measure the methodological rigor of radiomics workflow [20]. The RQS checklist is described in Table S1 [9]. The assessment interrogates 16 components and rated resulting with a minimum score with – 8 to 0 defined as 0% and a maximum score with 36 points defined as 100%. The QUADAS-2 tool was employed for presenting bias in patient selection, index test, reference standard, and flow and timing. The tool was tailored to

our research question through signaling questions for risks of bias specific to current study [19].

We developed a data collection instrument for study data, RQS, and QUADAS-2 score based on previous articles [9, 19, 21]. Two reviewers independently extracted study data into the instrument from two randomly chosen articles that fully met the inclusion criteria to test and adjust the tool. Disagreements were discussed with the third reviewer in order to achieve a shared appropriate understanding of each parameter. The data collection instrument is described in Table S2. Two same reviewers then measured and rated each study independently and recorded data for further analysis.

Data synthesis and analysis

Statistical analysis was performed with SPSS and R language using raters package, while the meta-analysis was performed

with Stata using the *metan*, *midas*, and *metandi* packages [20–22]. The summed RQS rating per study was calculated and the average rating of all raters is reported. Inter-rater agreement for single items of the RQS was calculated with modified Fleiss kappa statistic, while the interclass correlation coefficient (ICC) was determined to describe inter-rater agreement for the summed RQS [20, 21].

In current systemic review, the response prediction of osteosarcoma to neoadjuvant chemotherapy was addressed repeatedly; therefore, these studies were included in the meta-analysis. Two-by-two tables were extracted, if documented, or reconstructed based on published data. Sensitivity, specificity, positive and negative likelihood ratio, and diagnostic odds ratio (DOR) and their 95% confidence intervals (95% CIs) were calculated as effect size. A hierarchical summary receiver operating characteristic (HSROC) curve was plotted to show the diagnostic accuracy.

For heterogeneity assessment, Cochran's *Q* test and Higgins inconsistency index (*I*²) test were used to estimate the heterogeneity among the studies included in the meta-analysis. HSROC curve was drawn to visually assess the difference between the 95% confidence region and prediction region. A funnel plot and Deeks funnel plot were constructed to visually assess the risk of publication bias, and Egger's test and Deeks funnel plot asymmetry test were performed. The trim and fill method was used to estimate the number of missing studies. Detailed statistical methods were described in the [supplementary materials](#).

Results

Literature search

The search strategy yielded 30 studies from PubMed, 21 from Embase and 24 from Web of Science. After exclusion of 32 duplicates, 43 unique records of titles and abstracts were screened. Among these, fourteen were selected for possible inclusion and their full text retrieved. Review of the full text resulted in ultimate inclusion of twelve articles in the systematic review [23–34]. No additional study was included by hand search of their reference lists. Five studies [25, 26, 28, 29, 31] that interrogated response to neoadjuvant chemotherapy were included into meta-analysis (Fig. 2).

Study characteristics

Tables S3 and S4 summarize aims and characteristics of included studies. Five studies investigated treatment response prediction, three interrogated survival prediction, and two attempted to answer both clinical questions by radiomics method, while the remaining two studies explored stratification of metastatic risk, and differentiation of benign and malignant pulmonary nodules in osteosarcoma patients,

respectively. In terms of used modalities, seven studies used metabolic imaging methods, including PET and advanced MRI sequences. In terms of applied MRI sequence, one used conventional MRI sequence and contrast-enhanced T1-weighted imaging; three used advanced MRI sequence, two with DWI and one with IVIM, respectively.

Quality assessment

The twelve studies reached a mean \pm standard deviation RQS of 6.92 ± 6.00 , median 5, and range – 5 to 16. The average percentage RQS was 20.4% with a maximum of 44.4%. Average RQS rating per component and inter-rater agreement are presented in Table 1.

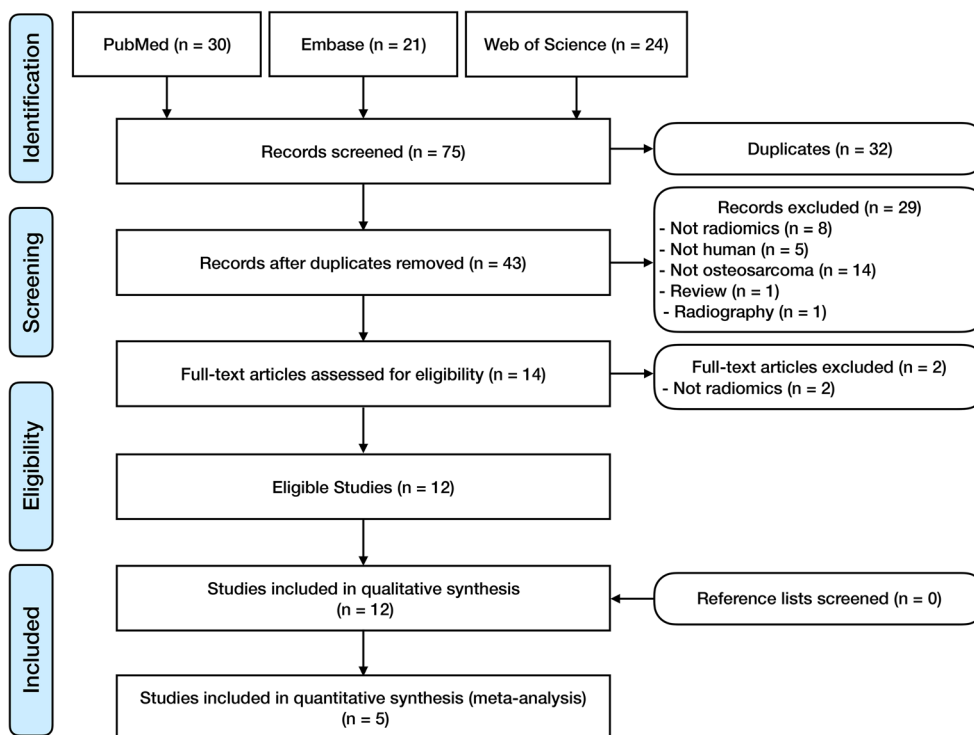
Most of all studies reported well-documented image acquisition protocols; however, eleven studies relied on prospectively acquired data, and only one included plan for radiomics analysis in its prospective study protocol. Ten studies acquired images using the same equipment, while two study included images from three CT scanners. However, none of them performed a phantom study. Multiple segmentation was conducted in seven studies, in which six were performed by two or more readers, and the remaining one identified tumor using the region-growing algorithm and then confirmed by a physician. Three studies conducted imaging at multiple time points and extracted their radiomics features respectively.

Twelve studies in this review included a total sample size of 964 patients. These studies extracted between 10 and 474 features from 16 to 191 patients, in which one study investigated 42 pulmonary nodules from sixteen patients. The ratio between features and patients ranged from 3.1 times more patients than features to 3.2 times more features than patients. Feature reduction and adjustment was performed in ten studies, in which eight underwent multiple testing. Five studies combined radiomics with clinical information or human objective assessment of image. Validation of radiomics signatures on internal datasets was performed in five of the studies; none of them employed external datasets. For model assessment, discrimination statistics results were usually provided, while calibration statistics results were less mentioned, and none of the study performed cutoff analysis.

Concerning biological validation and clinical utility, most studies compared their model with gold standard. The correlation between tumor biology and radiomics features were discussed in three to provide a more holistic model. Only two studies evaluated whether the model was sufficiently robust for clinical practice by decision curve analysis, but cost-effectiveness analysis was performed in none. Surprisingly, one study made its data partially available to the public.

Risk of bias and applicability concerns were assessed by QUADAS-2 and summarized in Fig. 3. Most included studies were regarded as having a moderate risk of bias. Risk of bias and application concerns relating to index testing were

Fig. 2 Flow diagram of the study selection process for this systematic review and meta-analysis



frequently observed. Some studies did not provide enough observations per predictor variable to produce reasonably

stable estimates. Feature reduction and adjustment process were not described in detail to allow replication.


Table 1 Average rating and inter-rater agreement per component of RQS

No.	RQS scoring item	Range	Average	Inter-rater agreement	
				S* or ICC	95% CI
1	Image protocol quality	0 to 2	0.92	0.62	0.60–0.64
2	Multiple segmentations	0 to 1	0.58	0.82	0.80–0.84
3	Phantom study on all scanners	0 to 1	0.00	1.00	1.00–1.00
4	Imaging at multiple time points	0 to 1	0.25	0.75	0.73–0.77
5	Feature reduction or adjustment for multiple testing	–3 to 3	2.00	0.75	0.73–0.77
6	Multivariable analysis with non-radiomics features	0 to 1	0.42	1.00	1.00–1.00
7	Detect and discuss biological correlates	0 to 1	0.25	0.75	0.73–0.77
8	Cut-off analyses	0 to 1	0.00	1.00	1.00–1.00
9	Discrimination statistics	0 to 2	1.50	0.54	0.52–0.55
10	Calibration statistics	0 to 2	0.25	0.80	0.78–0.82
11	Prospective study registered in a trial database	0 to 7	0.58	1.00	1.00–1.00
12	Validation	–5 to 5	–2.08	1.00	1.00–1.00
13	Comparison to ‘gold standard’	0 to 2	1.83	0.62	0.60–0.64
14	Potential clinical utility	0 to 2	0.33	1.00	1.00–1.00
15	Cost-effectiveness analysis	0 to 1	0.00	1.00	1.00–1.00
16	Open science and data	0 to 4	0.08	1.00	1.00–1.00
Total points: –8 to 0 = 0%, 36 = 100%		0% to 100%	6.92 = 20.4%	0.95	0.85–0.99

CI confidence interval, ICC interclass correlation coefficient, RQS Radiomics Quality Score, S* Fleiss Kappa statistics.

Fig. 3 Quality assessment of included studies by QUADAS-2 tool. The authors’ judgments for each domain of each included study were reviewed. The proportion of included studies that indicated low, unclear, or high risk and applicability concerns is shown in green, yellow, and red, respectively

	Risk of bias				Applicability concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Bailly 2017	+	+	+	?	+	+	+
Cho 2019	?	+	+	-	?	+	+
Dufau 2019	+	-	+	?	+	-	+
Jeong 2019	+	+	+	+	+	?	+
Kayal 2019	+	-	-	+	+	-	?
Lee 2020	+	?	+	-	+	?	+
Lin 2020	+	+	+	+	+	+	+
Sheen 2019	+	-	?	+	+	-	+
Song 2019	+	-	+	+	+	?	+
Wu 2018	+	+	+	+	+	+	+
Xu 2019	+	+	+	?	+	?	+
Zhao 2019	+	+	+	+	+	+	+



The reproducibility of the RQS and QUADAS-2 was calculated. The ICC for the RQS was 0.95 (95% CI 0.85–0.99). Moderate agreement was achieved in evaluating image protocol, discrimination statistics, and gold standard; substantial or almost perfect agreement was reached for the remaining elements of the RQS. Absolute agreement of the seven indicator questions of the QUADAS-2 ranged from 66.7 to 91.7%. RQS score and QUADAS-2 assessment per study per element are presented in Tables S5 and S6.

Prediction of response to chemotherapy

Since only one of the five included studies had a validation dataset, meta-analysis was performed only in the training dataset with a sample size of 328 patients. Individual selected studies showed high DOR for predicting response to neoadjuvant chemotherapy, ranging from 25.46 to 470.59, and the pooled DOR was 43.68 (95% CI 13.50–141.31; Fig. 4). Furthermore, the pooled sensitivity and specificity were 86% (95% CI 65–95%) and 88% (95% CI 79–94%), respectively (Fig. S1). The pooled positive likelihood ratio and negative

likelihood ratio were 7.16 (95% CI 3.96–12.94) and 0.16 (95% CI 0.06–0.43), respectively (Fig. S1). The AUC was 0.91 (95% CI 0.89–0.94), which indicates a high diagnostic performance (Fig. S2).

Cochran’s *Q* test implied that heterogeneity was present ($Q = 10.137, p = 0.003$) across the studies, and the Higgins I^2 statistic also demonstrated that heterogeneity was high ($I^2 = 80%$). The significant difference between the 95% confidence region and 95% prediction region was large, indicating a high possibility of heterogeneity across the studies (Fig. S2). However, the funnel plot with Egger’s test ($p = 0.277$) and Deeks funnel plot ($p = 0.79$) revealed that the likelihood of publication bias was low (Figs. S3 and S4). Trim and fill analysis estimated that no study was missing (Fig. S5).

Discussion

The current review using RQS found that the overall scientific quality of radiomics studies in osteosarcoma is insufficient, with an average RQS rating of 20.4% and 44.4% for the best

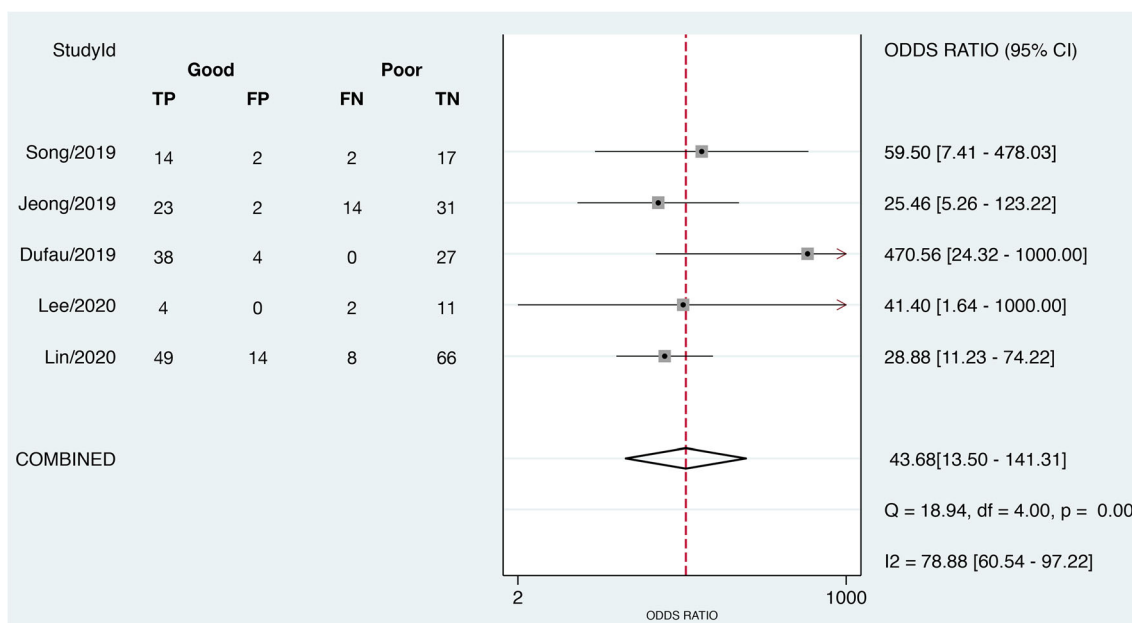


Fig. 4 Forest plot of the effect size calculated as diagnostic odds ratio for studies investigating the diagnostic accuracy of radiomics in neoadjuvant chemotherapy response prediction in osteosarcoma patients. The numbers are pooled estimates with 95% CIs in parentheses; horizontal

lines indicate 95% CI. TP number of good responders correctly diagnosed, FN number of good responders diagnosed as poor, FP number of poor responders diagnosed as good, TN number of poor responders correctly diagnosed

performing study. Although the meta-analysis showed that radiomics had an excellent diagnostic performance (AUC 0.91, 95% CI 0.89–0.94) in predicting patients' response to neoadjuvant chemotherapy, radiomics is far from a clinical applicable tool due to its poor methodological quality.

The mean RQS rating was acceptable (20.4% vs 10.8 to 36.1%) comparing with previous reviews, and the same for the best performing study (44.4% vs 41.2 to 55.6%; Tables S7) [20, 21, 35–42], however, lower than a study using a modified RQS checklist, which included patient selection related criteria from QUADAS [43], and higher than a recent review including studies without a systematic approach [44]. In our review, the main reasons for low RQS rating were lack of validation, absence of prospective study design, and unavailable open data. Further insufficiency in scientific quality of radiomics studies were in feature reproducibility and in analysis of clinical utility. Although the guidelines for machine learning model reporting have not strongly emphasized on publicly available code [45, 46], the open data and code would be preferable for assessing the reproducibility of findings [47].

Despite the promising results of meta-analysis, the repeatability and clinical adoption of those models were uncertain. Only five studies were included and most of them were lack of independent validation. Moreover, neither did studies provide publicly available imaging data with segmentation, nor the code employed for data preparation, feature extraction, and model construction. Both Cochran's Q test and Higgins I^2 statistic showed high heterogeneity, but subgroup analysis was not performed due to limited sample size. The likelihood of publication

bias was low, while negative results were not identified in our review. On the other hand, prognostic studies concerning survival of the patients and metastasis risk were not pooled, due to varied outcomes. Further analysis may be possible, if future studies report the results by similar measurements.

Among the reviewed studies, radiomics analysis was employed mainly in treatment response and prognosis prediction and only one study fell into the diagnostic field that differentiate benign and malignant pulmonary nodules. Some of them accomplished with conventional imaging data, indicating that radiomics may provide novel quantitative imaging markers without new acquisition equipment or tracers. Our study demonstrated that radiomics may be useful in aiding radiologists for answering clinical questions tightly related to practice. To be able to translate these excellent results into clinical radiology, well-designed and properly-conducted studies are indispensable. Therefore, disadvantages in study design, validation, and open science detected by RQS should be avoided. RQS should be used not only as a tool assessing the scientific and reporting quality of published researches but also as a routine self-checklist before manuscript submitting, and even as a guideline for radiomics study design.

During the application of RQS, varying inter-rater agreement was observed [20]. To avoid that, one later study developed a data extraction instrument and introduced a training phase to reach a shared understanding of each parameter before the formal assessment [21]. As a result, agreement for the sum RQS rating (ICC = 0.96) and most items was improved. Other studies discussed topics with initial disagreements and

tailored RQS to the specific research question during the data extract phase to reach a more reproducible assessment [38, 39]; yet, the agreement was not reported. Our study repeated those processes and demonstrated that those efforts allowed reaching a moderate inter-rater agreement in RQS (ICC = 0.95) and shared understanding on most items. Therefore, a similar procedure is deemed to be essential.

However, modifications of RQS in response to practical needs are necessary. Two previous reviews attempt to integrate RQS with six key domains to facilitate the use in radiomics approaches [38, 39], to approach a more precise assessment and appropriate method amelioration. One of them compared RQS with Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist [48], and pointed out that room for improvement was shown in stating study objective in abstract and introduction, blind assessment of outcome, sample size, and missing data categories, to accelerate a more standardized reporting of radiomics researches [39]. Another guideline emphasized on reliable assessment of model validity and consistent interpretation of model outputs, and provided a more clearly defined checklist for assessment of model establishment than RQS, to enable consistent reporting and correct application of model specifications and results [45]. Both RQS and TRIPOD emphasized on validation of the imaging biomarkers [42], but they mainly concerned the dataset used during this process. A recent statement [49] further proposed a structured pipeline for validating based on a three-step technical validation and clinical validation, and pointed out the need of regular updating of validated imaging biomarkers. This process may provide a more practicable and more standardized roadmap for translating radiomics models to clinical applicable tools [9]. Other guidelines concerning artificial intelligence method also provide valuable references for RQS improvement [50, 51].

Some inherent limitations exist in this review. First, radiomics studies investigating osteosarcoma is limited. Hence, only twelve studies were included and five were meta-analyzed. However, osteosarcoma is a rare disease with incidence of several millionth; our review is sufficient to represent the status of this highly specialized field. Second, only one study included in the meta-analysis was validated with an internal dataset. Our results may actually represent a higher performance of radiomics models. Third, the meta-analysis showed high overall heterogeneity, while the subgroup analysis was not performed, since the sample size of the studies was too small to draw any reliable conclusions. Future reviews including more studies and greater sample size may assess the influence of heterogeneity. Finally, the RQS has limitations. Radiomics is still a developing field and so is RQS. It is necessary to improve RQS items in response to actual practical needs. Still, RQS is a timely tool for methodological quality assessment of radiomics research.

In conclusion, radiomics models showed promise for answering clinical questions related to osteosarcoma patients. Especially, for the response to neoadjuvant chemotherapy, the meta-analysis implied moderate performance of radiomics to approach this prediction. However, prospectively designed, well-validated radiomics trials with open data are needed for demonstrating their effectiveness and clinical validity. Moreover, RQS with ongoing improvements may serve as a useful tool to facilitate radiomics towards clinical translation.

Acknowledgments The authors would like to express their gratitude to Prof. Guang Yang and Ms. Chengxiu Zhang for their constructive discussion and suggestions. The authors would like to thank Dr. Guangcheng Zhang for English language editing.

Funding This study has received funding by National Natural Science Foundation of China (81771790) and Medicine and Engineering Combination Project of Shanghai Jiao Tong University (YG2019ZDB09).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Weiwu Yao.

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 authors has significant statistical expertise.

Informed consent Written informed consent was not required for this study because of the nature of our study, which was a systematic review and meta-analysis.

Ethical approval Institutional Review Board approval was not required because of the nature of our study, which was a systematic review and meta-analysis.

Methodology

- retrospective
- diagnostic or prognostic study
- multicenter study

References

1. Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F (2013) World Health Organization classification of tumors: WHO classification of tumours of soft tissue and bone, 4th edn. IARC Press, Lyon
2. Whelan JS, Davis LE (2018) Osteosarcoma, chondrosarcoma, and chordoma. *J Clin Oncol* 36:188–193
3. Casali PG, Bielack S, Abecassis N et al (2018) Bone sarcomas: ESMO-PaedCan-EURACAN clinical practice guidelines for diagnosis, treatment, and follow-up. *Ann Oncol* 29:iv79–iv95
4. National Comprehensive Cancer Network (2019) NCCN clinical practice guidelines in oncology: Bone Cancer, v1.2020. Available

- via https://www.nccn.org/professionals/physician_gls/pdf/bone.pdf. Accessed Apr 2020
5. Link MP, Goorin AM, Miser AW et al (1986) The effect of adjuvant chemotherapy on relapse-free survival in patients with osteosarcoma of the extremity. *N Engl J Med* 314:1600–1606
 6. Rosen G, Murphy ML, Huvos AG, Gutierrez M, Marcove RC (1976) Chemotherapy, en bloc resection, and prosthetic bone replacement in the treatment of osteogenic sarcoma. *Cancer* 37:1–11
 7. Rosen G, Caparros B, Huvos AG et al (1982) Preoperative chemotherapy for osteogenic sarcoma: selection of postoperative adjuvant chemotherapy based on the response of the primary tumor to preoperative chemotherapy. *Cancer* 49:1221–1230
 8. Coffin CM, Lowichik A, Zhou H (2005) Treatment effects in pediatric soft tissue and bone tumors: practical considerations for the pathologist. *Am J Clin Pathol* 123:75–90
 9. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 14:749–762
 10. Castellano G, Bonilha L, Li LM, Cendes F (2004) Texture analysis of medical images. *Clin Radiol* 59:1061–1069
 11. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 48:441–446
 12. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. *Radiology* 278:563–577
 13. Sullivan DC, Obuchowski NA, Kessler LG et al (2015) Metrology standards for quantitative imaging biomarkers. *Radiology* 277:813–825
 14. 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
 15. O'Connor JP, Aboagye EO, Adams JE et al (2017) Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* 14:169–186
 16. Rogers W, Thulasi Seetha S, Refaee TAG et al (2020) Radiomics: from qualitative to quantitative imaging. *Br J Radiol* 93:20190948
 17. McInnes MDF, Moher D, Thombs BD et al (2018) Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 319:388–396
 18. Moher D, Shamseer L, Clarke M et al (2015) Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 4:1
 19. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 155:529–536
 20. 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
 21. Ursprung S, Beer L, Bruining A et al (2020) Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma—a systematic review and meta-analysis. *Eur Radiol* 30(6):3558–3566
 22. Cochrane methods screening and diagnostic tests (2017) Handbook for DTA Reviews. Available via <https://methods.cochrane.org/sdt/handbook-dta-reviews>. Accessed 10 Apr 2020
 23. Bailly C, Leforestier R, Champion 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:e0183841
 24. Cho YJ, Kim WS, Choi YH et al (2019) Computerized texture analysis of pulmonary nodules in pediatric patients with osteosarcoma: differentiation of pulmonary metastases from non-metastatic nodules. *PLoS One* 14:e0211969
 25. 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* 106:983–999
 26. Jeong SY, Kim W, Byun BH et al (2019) Prediction of chemotherapy response of osteosarcoma using baseline 18-F-FDG textural features machine learning approaches with PCA. *Contrast Media Mol Imaging* 2019:3515080
 27. Kayal EB, 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
 28. Lee SK, Jee WH, Jung CK, Im SA, Chung NG, Chung YG (2020) Prediction of poor responders to neoadjuvant chemotherapy in patients with osteosarcoma: additive value of diffusion-weighted MRI including volumetric analysis to standard MRI at 3T. *PLoS One* 15:e0229983
 29. Lin P, Yang PF, Chen S et al (2020) A Delta-radiomics model for preoperative evaluation of neoadjuvant chemotherapy response in high-grade osteosarcoma. *Cancer Imaging* 20:7
 30. Sheen H, Kim W, Byun BH et al (2019) Metastasis risk prediction model in osteosarcoma using metabolic imaging phenotypes: a multivariable radiomics model. *PLoS One* 14:e0225242
 31. Song H, Jiao Y, Wei W et al (2019) Can pretreatment 18-F-FDG PET tumor texture features predict the outcomes of osteosarcoma treated by neoadjuvant chemotherapy? *Eur Radiol* 29:3945–3954
 32. Wu Y, Xu L, Yang P et al (2018) Survival prediction in high-grade osteosarcoma using radiomics of diagnostic computed tomography. *EBioMedicine* 34:27–34
 33. Xu L, Yang P, Yen EA et al (2019) A multi-organ cancer study of the classification performance using 2D and 3D image features in radiomics analysis. *Phys Med Biol* 64:215009
 34. 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
 35. Valdora F, Houssami N, Rossi F, Calabrese M, Tagliafico AS (2018) Rapid review: radiomics and breast cancer. *Breast Cancer Res Treat* 169(2):217–229
 36. Granzier RWY, van Nijnatten TJA, Woodruff HC, Smidt ML, Lobbes MBI (2019) Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: a systematic review. *Eur J Radiol* 121:108736
 37. Wakabayashi T, Ouhmich F, Gonzalez-Cabrera C et al (2019) Radiomics in hepatocellular carcinoma: a quantitative review. *Hepatol Int* 13(5):546–559
 38. Park JE, Kim HS, Kim D et al (2020) A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. *BMC Cancer* 20(1):29
 39. 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
 40. Wang H, Zhou Y, Li L, Hou W, Ma X, Tian R (2020) Current status and quality of radiomics studies in lymphoma: a systematic review. *Eur Radiol*. <https://doi.org/10.1007/s00330-020-06927-1>
 41. Stanzione A, Gambardella M, Cuocolo R, Ponsiglione A, Romeo V, Imbriaco M (2020) Prostate MRI radiomics: a systematic review and radiomic quality score assessment. *Eur J Radiol* 129:109095
 42. Fornaçon-Wood I, Faivre-Finn C, O'Connor JPB, Price GJ (2020) Radiomics as a personalized medicine tool in lung cancer: separating the hope from the hype. *Lung Cancer* 146:197–208
 43. Castillo Tovar JM, Arif M, Niessen WJ, Schoots IG, Veenland JF (2020) Automated classification of significant prostate cancer on MRI: a systematic review on the performance of machine learning applications. *Cancers (Basel)* 12(6):E1606

44. Chetan MR, Gleeson FV (2020) Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives. *Eur Radiol*. <https://doi.org/10.1007/s00330-020-07141-9>
45. Luo W, Phung D, Tran T et al (2016) Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 18(12):e323
46. Jethanandani A, Lin TA, Volpe S et al (2018) Exploring applications of radiomics in magnetic resonance imaging of head and neck cancer: a systematic review. *Front Oncol* 8:131
47. Nagendran M, Chen Y, Lovejoy CA et al (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368:m689
48. 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. *BMJ* 350:g7594
49. European Society of Radiology (ESR) (2020) ESR statement on the validation of imaging biomarkers. *Insights Imaging* 11(1):76
50. CONSORT-AI and SPIRIT-AI Steering Group (2019) Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 25(10):1467–1468
51. Mongan J, Moy L, Kahn CE Jr (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiology Artificial Intelligence* 2(2):e200029

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