European Journal of Nuclear Medicine and Molecular Imaging (2021) 49:345–360
https://doi.org/10.1007/s00259-021-05509-7

REVIEW ARTICLE

Current status and quality of radiomic studies for predicting immunotherapy response and outcome in patients with non-small cell lung cancer: a systematic review and meta-analysis

Qiuying Chen[1,2] · Lu Zhang[1,2] · Xiaokai Mo[1] · Jingjing You[1,2] · Luyan Chen[1,2] · Jin Fang[1] · Fei Wang[1] · Zhe Jin[1,2] · Bin Zhang[1,2] · Shuixing Zhang[1,2]

Received: 15 June 2021 / Accepted: 27 July 2021 / Published online: 17 August 2021

Abstract

Purpose Prediction of immunotherapy response and outcome in patients with non-small cell lung cancer (NSCLC) is challenging due to intratumoral heterogeneity and lack of robust biomarkers. The aim of this study was to systematically evaluate the methodological quality of radiomic studies for predicting immunotherapy response or outcome in patients with NSCLC.
Methods We systematically searched for eligible studies in the PubMed and Web of Science datasets up to April 1, 2021. The methodological quality of included studies was evaluated using the phase classification criteria for image mining studies and the radiomics quality scoring (RQS) tool. A meta-analysis of studies regarding the prediction of immunotherapy response and outcome in patients with NSCLC was performed.
Results Fifteen studies were identified with sample sizes ranging from 30 to 228. Seven studies were classified as phase II, and the remaining as discovery science ($n=2$), phase 0 ($n=4$), phase I ($n=1$), and phase III ($n=1$). The mean RQS score of all studies was 29.6%, varying from 0 to 68.1%. The pooled diagnostic odds ratio for predicting immunotherapy response in NSCLC using radiomics was 14.99 (95% confidence interval [CI] 8.66–25.95). In addition, radiomics could divide patients into high- and low-risk group with significantly different overall survival (pooled hazard ratio [HR]: 1.96, 95%CI 1.61–2.40, $p<0.001$) and progression-free survival (pooled HR: 2.39, 95%CI 1.69–3.38, $p<0.001$).
Conclusions Radiomics has potential to noninvasively predict immunotherapy response and outcome in patients with NSCLC. However, it has not yet been implemented as a clinical decision-making tool. Further external validation and evaluation within clinical pathway can facilitate personalized treatment for patients with NSCLC.

Keywords NSCLC · Radiomics · Immunotherapy · Radiomics quality scoring · Systematic review

Introduction

Lung cancer is the most common cancer and remains the leading cause of cancer-related death, despite continuous progresses in the diagnosis and therapy [1]. Non-small cell lung cancer (NSCLC) accounts for 80–90% of primary lung cancers [1], and around 50% of patients are diagnosed at an advanced stage (stage III or IV), with 5-year survival rate of only 18% [2].

In recent years, immune checkpoint therapy has been a research hotspot in the field of cancer since the increasing understanding of the mechanisms of immune evasion by cancer cells [2]. Susceptibilities in immune checkpoint pathways allow tumor cells to escape immune surveillance, causing tumor propagation. Monoclonal antibodies targeting these pathways reinvigorate the host immunity against

Qiuying Chen and Lu Zhang have contributed equally to this work

This article is part of the Topical Collection on Oncology - Chest.

✉ Bin Zhang
xld_Jane_Eyre@126.com

✉ Shuixing Zhang
shui7515@126.com

[1] Department of Radiology, The First Affiliated Hospital, Jinan University, No. 613, Huangpu West Road, Tianhe District, Guangzhou 510627, Guangdong, People's Republic of China

[2] Graduate College, Jinan University, Guangzhou, Guangdong, People's Republic of China

tumor cells by rescuing pre-existing tumor-specific cytotoxic T cells in the tumor sites and have revolutionized the treatment of NSCLC due to their favorable toxicity profiles and their ability to produce durable clinical responses [3–5]. Nowadays, immune checkpoint inhibitors (ICIs) targeting the programmed cell death ligand 1 (PD-1)/programmed cell death ligand L1 (PD-L1) axis are the standard of care for treatment of patients with advanced NSCLC without targetable genetic alterations [6–8].

As immunotherapy is costly and may lead to immune-related toxicity, it is of great importance to accurately identify the patients who would benefit from immunotherapy. The percentage of tumor cells expressing PD-L1 is the routinely used biomarker to select candidates for this additional therapeutic option [9]. Patients with positive PD-L1 status generally have higher objective response rates [10, 11]. However, the reliability of PD-L1 expression as a biomarker of treatment response is controversial [12–14]. Although PD-L1 expression is positively related with the response to immunotherapy, there are cases of nonresponsive PD-L1–positive tumors and responsive PD-L1–negative tumors [15]. Additionally, identifying PD-L1 expression status via immunohistochemical analysis is time-consuming and cannot reflect dynamic PD-L1 expression. Consequently, the potential of other biomarkers has been investigated.

The emergence of new technologies and the requirements of precision medicine prompt a new promising field, that is, radiomics [16, 17]. Radiomics refers to the comprehensive quantification of tumor phenotypes on radiographic images in a high-throughput manner. The primary goal of radiomics analysis is to develop clinically relevant models that can capture intratumoral heterogeneity using bioinformatics tools. Radiomics is particularly attractive since it represents a non-invasive, repeatable, and cost-effective method of extracting molecular information from medical images. ^{18}F-FDG PET/CT and CT are widely used for baseline staging and response evaluation in NSCLC. The medical images can be analyzed quantitatively with radiomic approach to identify more tumor characterizations beyond human eyes. Unlike traditional biopsy-based assays that represent only a local region of the tumor, images can reflect the entire tumor burden, and thus not subject to sampling bias. This is obvious in NSCLC treated by immunotherapy, where different lesions can have distinct microenvironments, potentially resulting in heterogeneous response patterns [18]. The radiomic features contain information that reflects underlying tumor pathophysiology and allow evaluation of tumor heterogeneity [19, 20].

Recently, a growing body of studies have examined the potential clinical utility of radiomic features derived from CT or ^{18}F-FDG PET/CT images of NSCLC and correlated these features with immunotherapy response or outcome. The purpose of this study was to analyze the current status of radiomic studies for predicting immunotherapy response or outcome in patients with NSCLC via a systematic review and to evaluate the quality of radiomic studies according to the phase classification criteria for image mining studies and the radiomics quality scoring (RQS) tool. In addition, quantitative analysis was also conducted to assess the performance of radiomics in predicting immunotherapy response and outcome.

Materials and methods

This study was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement [21]. The PRISMA checklist is provided in Supplementary Table 1. The study protocol has been registered in International prospective register of systematic reviews (CRD42021246068).

Literature search strategy

A comprehensive literature search for potentially relevant articles was conducted in PubMed and Web of Science databases from the inception to April 1, 2021. The keywords and Medical Subject Headings (MeSH) terms were used as follows: "non-small cell lung cancer," "lung cancer," NSCLC, adenocarcinoma, squamocellular, radiomic, radiomics, textural, texture, histogram, "magnetic resonance imaging", "magnetic resonance", MRI, MR, "computed tomography," CT, "positron emission tomography," PET, PD-1, PD-L1, immunotherapy, "immune checkpoint inhibitor," "immune checkpoint blockade," nivolumab, and pembrolizumab.

Study selection

After the removal of the duplicates, two reviewers (FW and JF) independently performed an initial screening of the identified titles and abstracts; disagreements were solved by consensus or a third reviewer (BZ). We included all eligible studies which evaluated quantitative radiomic features extracted from CT or ^{18}F-FDG PET/CT scans against immunotherapy response or outcome in patients with NSCLC. Full text was available and articles were written in English. The criteria for excluding studies were as follows: (a) studies focused purely on methodological aspects of radiomics; (b) studies in phantom or animal models; and (c) case reports or small case series (≤ 10 patients), reviews, poster presentations, letters, and meeting abstracts. Subsequently, the reviewers retrieved the full text of the selected titles/abstracts and performed an independent second-step selection. Additionally, additional research studies of possible interest were identified from the reference list of relevant articles and reviewed for eligibility.

Data extraction

Two reviewers (QYC and LZ) extracted the information from each included study: publication year, sample size, study population, study design, imaging modality, research question, treatment, software, segmentation, clinical characteristics, imaging features, validation, endpoints, reference standard, and classifiers. Clinical endpoints of interest were overall survival (OS) and progression-free survival (PFS), as well as the power of models to predict immunotherapy response. OS was defined as the time from cancer immunotherapy until death from any cause. PFS was defined as the time from cancer immunotherapy to progression of disease or death from any cause. The evaluation of immunotherapy response was according to the response evaluation criteria in solid tumors 1.1 (RECIST 1.1) and its modified version. The comparison of RECIST 1.1 and iRECIST is shown in Supplementary Table 2.

Quality assessment

Figure 1 shows the workflow of radiomics in NSCLC treated by immunotherapy. The methodological quality of the included studies was independently assessed by the two reviewers (QYC and LZ) using the phase classification criteria for image mining studies [22] and the RQS, which is a radiomics-specific quality assessment tool [23]. The parameters for phase categorization were sample size (< 100 or > 100), study design (retrospective or prospective), type of validation approach (internal or independent), and the development stage (pre- or post-marketing) (Supplementary Table 3). The phase classification criteria assign image mining studies to the discovery science and phases 0–IV. The 16-component RQS tool evaluates the validity and bias of the radiomic studies (Supplementary Table 4). Each study was assigned a number of points per RQS component and summed to give a total score (range − 8 to + 36). A score of − 8 to 0 points correspond to 0% and 36 points correspond to 100%. Mean scores of the two evaluations are presented as a percentage. Agreement between the reviewers was assessed by means of a weighted kappa statistic.

Meta-analysis

Two meta-analyses were performed within the included studies: (1) a meta-analysis of studies investigating the use of radiomics to compare immunotherapy outcome (e.g., PFS and OS) between high- and low-risk group in the validation datasets, which was measured by pooled hazard ratio (HR) and 95% confidence interval (CI), and (2) a meta-analysis of studies investigating the use of radiomics for predicting immunotherapy response in the validation datasets of the optimal radiomic model, which was measured by pooled sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), and diagnostic odds ratio (OR) and corresponding 95%CIs.

Data of all included studies were independently extracted by two reviewers (XKM and ZJ); discrepancy was solved by consensus or a third reviewer (BZ). Studies reported HR and 95%CI directly or their estimation following the methodology as described by Parmar et al. [24] were included in

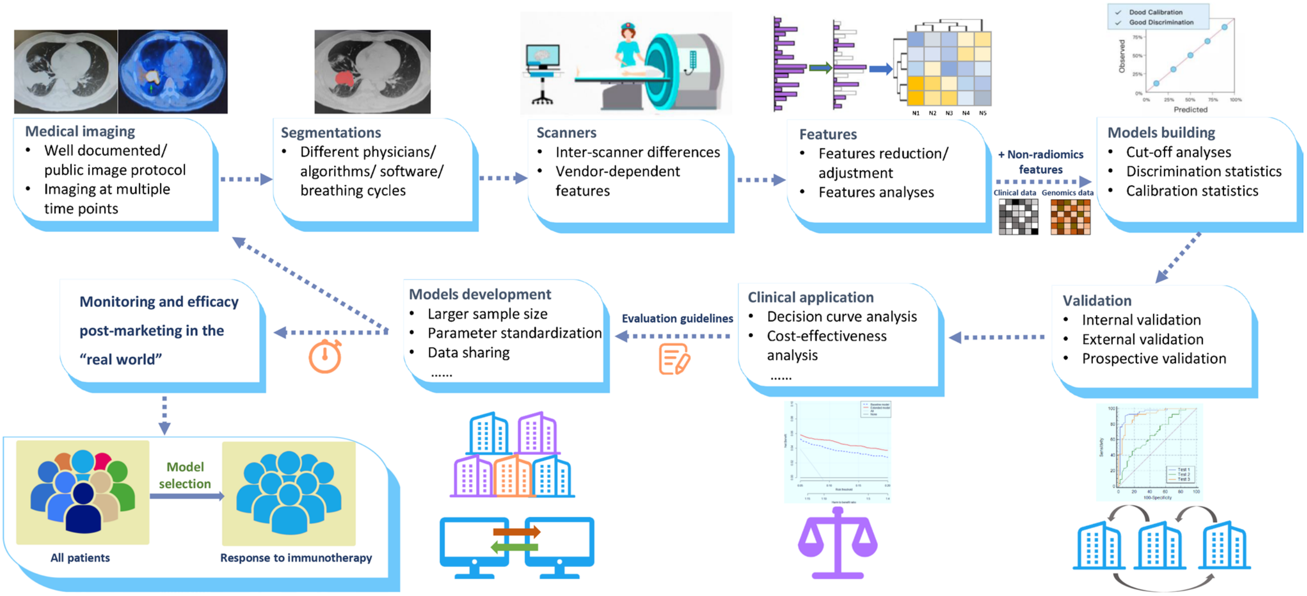


Fig. 1 Workflow of radiomics in NSCLC treated with immunotherapy

the first meta-analysis. Only studies, from which a 2 × 2 contingency table could be directly extracted or reconstructed, were included in the second meta-analysis. If multiple models were reported in a study, only the one with the highest area under the curve (AUC) or Youden's index was extracted. If multiple validation datasets of the optimal model were reported, we extracted the data from each validation dataset. In case of multiple publications deriving from one study, only the article with better methodological quality was included for analysis.

Statistical analysis

Random effects meta-analysis was performed using the Mantel–Haenszel model. Forest plots were used for visualization of the results. We used I^2 metric to assess the heterogeneity across studies; an I^2 value of 0–25% represents insignificant heterogeneity, > 25–50% low heterogeneity, > 50–75% moderate heterogeneity, and > 75% high heterogeneity [25]. Two-sided $p < 0.05$ were considered as statistically significant. All analyses were performed using STATA version 12.0 (Stata Corporation, USA) and Meta-DiSc version 1.4 (https://meta-disc.software.informer.com/1.4/).

Results

Study selection

Figure 2 shows the PRISMA flowchart of the included studies of this systematic review and meta-analysis. The search strategy yielded 42 studies from PubMed and 79 from Web of Science. After exclusion of 38 duplicates, 83 titles/abstracts were screened and 21 eligible studies were retrieved as full text. Finally, 15 peer-reviewed articles published from 2019 to 2021 were included in this systematic review [19, 26–39]. Ten articles were eventually included in the meta-analysis, and reasons for exclusion were as follows: one study [35] only identified patients at risk of hyperprogression and four studies [30, 31, 33, 36] could not extract or reconstruct data.

Study characteristics

Tables 1 and 2 demonstrate the characteristics of the included studies. All studies focused on advanced NSCLC except for three studies [33, 34, 39] that focused on all stages. All patients received anti-PD-1/PD-L1 therapy with at least one ICI agent. The most common imaging modality was contrast-enhanced CT (12 out of 15), followed by ^{18}F-FDG PET/CT (3 out of 15). There were three, eight, and four radiomic studies predicting immunotherapy response, immunotherapy outcome, and both, respectively.

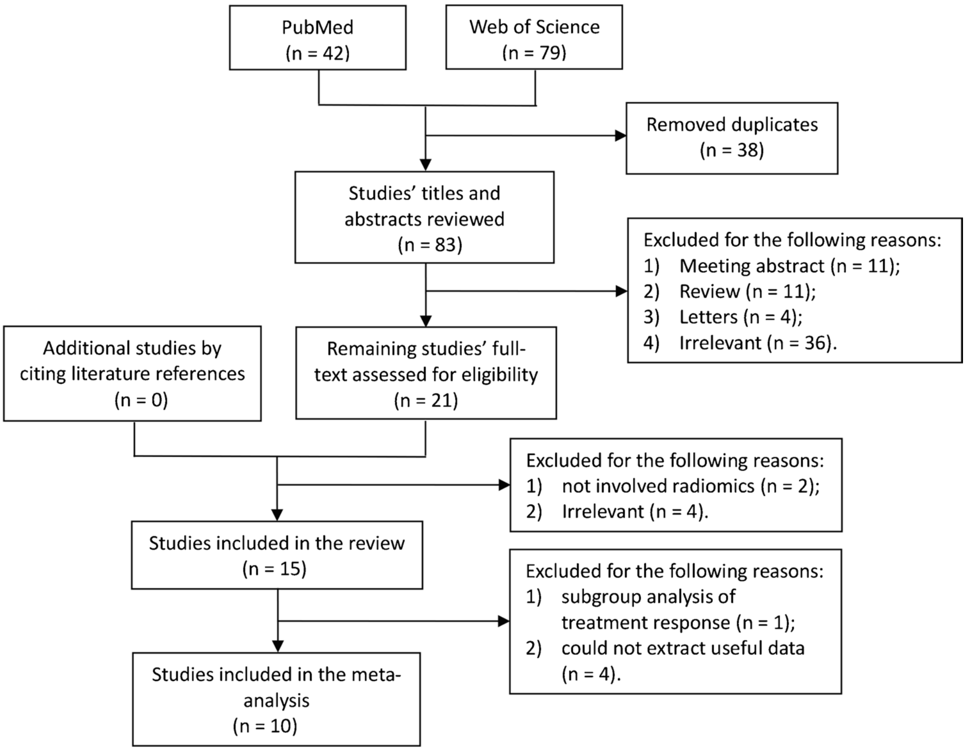

PubMed
(n = 42)
Web of Science
(n = 79)
Removed duplicates
(n = 38)
Studies' titles and abstracts reviewed
(n = 83)
Excluded for the following reasons:
1) Meeting abstract (n = 11);
2) Review (n = 11);
3) Letters (n = 4);
4) Irrelevant (n = 36).
Additional studies by citing literature references
(n = 0)
Remaining studies' full-text assessed for eligibility
(n = 21)
Excluded for the following reasons:
1) not involved radiomics (n = 2);
2) Irrelevant (n = 4).
Studies included in the review
(n = 15)
Excluded for the following reasons:
1) subgroup analysis of treatment response (n = 1);
2) could not extract useful data (n = 4).
Studies included in the meta-analysis
(n = 10)

Fig. 2 PRISMA flowchart of included studies

Table 1 Radiomic studies included in the systematic review

Study ID (Refs.)	Sample size	Population	Study design	Imaging modality	Research question	Phase
Khorrami [27]	139	Advanced NSCLC	Retrospective	CE-CT × 2 baseline, follow-up	Evaluate the performance of DelRADx from within and outside thoracic lesions, in predicting response to multiple ICIs in patients with advanced NSCLC	II
Trebeschi [33]	123	NSCLC	Retrospective	CE-CT × 1 baseline	Explore whether artificial intelligence algorithms can automatically quantify radiographic characteristics that are related to and may therefore act as noninvasive radiomic biomarkers for immunotherapy response	II
Mu [19]	194	Advanced NSCLC	Retrospective + prospective	^{18}F-FDG PET/CT × 1 follow-up	Test the hypothesis that radiomic features from baseline pretreatment ^{18}F-FDG PET/CT scans can predict clinical outcomes of NSCLC patients treated with immunotherapy	III
Tunali [34]	228	NSCLC	Prospective	CE-CT × 1 baseline	Identify clinical and CT-based predictors of rapid disease progression phenotypes in NSCLC patients treated with ICI	Discovery science
Nardone [30]	59	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Evaluate the potential use of radiomics analysis in the identification of patients with NSCLC who may benefit from nivolumab treatment	I
Polverari [31]	57	Advanced NSCLC	Retrospective	^{18}F-FDG PET/CT × 1 baseline	Evaluate the application of radiomics analysis of the primary lesion to identify features predictive of response to immunotherapy	Discovery science
Valentinuzzi [36]	30	Stage IV NSCLC	Retrospective	^{18}F-FDG PET/CT × 3 baseline, 1- and 4-month follow-up	Investigate whether immunotherapy ^{18}F-FDG PET/CT radiomics signature (iRADIOMICS) predicts response of metastatic NSCLC patients to pembrolizumab better than the current clinical standards	0
Vaidya [35]	109	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Evaluate the ability of the image-based radiomics markers extracted from baseline CTs of advanced NSCLC treated with ICI to identify patients at risk of hyperprogressions	II
Liu [29]	46	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Develop a CT-based radiomic model to predict clinical outcomes of advanced NSCLC patients treated with nivolumab	0
Ladwa [28]	47	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Assess CT texture analysis of patients likely to benefit from nivolumab	0
Yang [37]	200	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Develop a unified deep learning model to integrate multimodal serial information from CT with laboratory and baseline clinical information to identify immunotherapy responders and nonresponders	II
Ravanelli [32]	104	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Assess CT histogram analysis as prognostic and predictive factor in platinum-refractory NSCLC treated with nivolumab	II
Dercle [38]	92	Advanced NSCLC	Retrospective	CE-CT × 2 baseline, 8-month follow-up	Define radiomics signatures predicting the sensitivity of NSCLC to nivolumab	0
He [26]	123	Advanced NSCLC	Retrospective	CE-CT × 1 baseline	Investigate the correlation between deep learning radiomic biomarker and TMB, including its predictive value for ICI treatment response in patients with advanced NSCLC	II

Table 1 (continued)

Study ID (Refs.)	Sample size	Population	Study design	Imaging modality	Research question	Phase
Liu [39]	197	NSCLC	Retrospective	CE-CT×2 baseline, follow-up	Identify imaging biomarkers to assess predictive capacity of radiomics nomogram regarding treatment response status in patients with advanced NSCLC undergoing anti-PD-1 immunotherapy	II

Ref reference, *NSCLC* non-small cell lung cancer, *CT* computed tomography, *CE-CT* contrast-enhanced CT, *PET* positron emission tomography, *DelRADx* delta radiomics analysis, *ICI* immune checkpoint inhibitor, *TMB* tumor mutational burden

Quality analysis

The sample size of the included studies ranged from 30 to 228, of which nine (60%) studies enrolled more than 100 patients. Thirteen (86.7%) studies were retrospective, only one study [34] was prospective, and one study [19] was a mix of retrospective training and prospective validation. Notably, two retrospective studies [37, 38] collected data from multicenter clinical trials but retrospectively analyzed. Thirteen (86.7%) studies performed validation analysis, but only two [27, 30] were externally validated. According to the phase classification criteria for image mining studies, seven (46.7%) studies were classified as phase II, and the remaining as discovery science ($n = 2$), phase 0 ($n = 4$), phase I ($n = 1$), and phase III ($n = 1$) (Fig. 3a).

Table 3 shows the single and total RQS scores of all included studies evaluated by two reviewers and the mean RQS score of two evaluations. The mean score of 15 studies was 29.6% (range 0–68.1%), and three studies were assigned a quality score of < 10% (Fig. 3b, c). Most studies reported well-documented image acquisition protocols. Four (26.7%) studies acquired images from the same machine, whereas eight (53.3%) studies obtained images from different scanners. Five (33.3%) studies detected inter-scanner differences and vendor-dependent features to ensure generalizability of the derived predictive models. All studies except for one [19] acquired images from baseline scans, four [27, 36, 38, 39] of which conducted imaging at additional time points during follow-up. Nine (60%) studies used manual segmentation, three (20%) used semiautomatic segmentation, one (6.7%) used automatic segmentation, and two (13.3%) used two segmentation methods. Feature dimension reduction or adjustment was performed in 11 (73.3%) studies. Five (33.3%) studies [19, 27, 34, 37, 39] added clinical features to the radiomic models and three of which suggested that integration of the clinical data and radiomic features could improve the predictive performance of the models. The clinical characteristics included gender, smoking, histology, Eastern Cooperative Oncology Group (ECOG) scale of performance status, metastasis, previous lines of systemic therapies, and blood tests. The correlation between tumor biology and radiomic features were detected and discussed in six (40%) studies. Most studies performed cutoff analysis to stratify patients into low- and high-risk groups. For model assessment, discrimination statistics were usually provided, whereas calibration statistics were less mentioned. Validation of radiomics signatures was performed in 13 (86.7%) studies and two (13.3%) employed external datasets from other institutes. However, only one study [19] prospectively validated radiomic biomarker. Regarding the clinical utility, most studies compared their models with gold standard. Only three studies [19, 26, 39] evaluated whether their models were ready for clinical practice by decision curve

Table 2 The PRISMA literature list

Study ID (Ref)	Treatment	Software	Segmentation	Clinical features in the model	Imaging features	Validation	Endpoints	Reference standard	Classifier
Khorrami [27]	Anti-PD-(L)1 therapy	3D Slicer, MATLAB, R	Manual	Gender, smoking status	Texture and shape features	Split sample, external	Response to immunotherapy, survival	RECIST1.1	LDA
Trebeschi [33]	Anti-PD-1 therapy	NA	Manual	NA	Original images and different image transformations	Split sample	Response to immunotherapy, survival	RECIST	RF
Mu [19]	Anti-PD-(L)1 therapy	MATLAB, R	Semiautomatic	Histology, ECOG scale of performance status, distant metastasis	PET, CT, and KLD features	Split sample	Durable benefit, survival	RECIST1.1	LR
Tunali [34]	Anti-PD-(L)1 therapy	MATLAB, Lung Tumor Analysis software program platform, Stata, R	Automatic + semiautomatic	Hepatic metastasis, bone metastasis, previous systemic therapies, neutrophils to lymphocytes ratio	Radiomic features from tumor and tumor border regions	NA	Ability to predict progression phenotypes	A stringent set of criteria adapted from prior studies	LR
Nardone [30]	Nivolumab	LifeX, X-Tile	Manual	NA	GTV and TA parameters	Split sample, external	Survival	NA	Texture score
Polverari [31]	Anti-PD-(L)1 therapy	LifeX	Manual + semi-automatic	NA	Semi-quantitative PET parameters and radiomic features	NA	Response to immunotherapy	RECIST1.1	NA
Valentinuzzi [36]	Pembrolizumab	3D Slicer, R	Semiautomatic	NA	Radiomic features and texture-based heterogeneity features	Cross-validation	Survival	TPS and iRECIST	LR
Vaidya [35]	Anti-PD-(L)1 therapy	3D Slicer, MATLAB, R	Manual	NA	Texture and QVT features and the angles of each three consecutive points of the vasculature	Split sample	Ability to identify hyper-progressions, survival	RECIST 1.1	RF, LDA, DLDA, QDA, SVM
Liu [29]	Nivolumab	Python, R	Manual	NA	LoG, wavelet, shape, histogram, and texture features	Cross-validation	Survival	RECIST 1.1	SVM, LR, NB

Table 2 (continued)

Study ID (Ref)	Treatment	Software	Segmentation	Clinical features in the model	Imaging features	Validation	Endpoints	Reference standard	Classifier
Ladwa [28]	Nivolumab	NA	Automatic	NA	Texture features	Cross-validation	Survival	RECIST 1.1	General model for combining pairs of texture parameters
Yang [37]	Anti-PD-(L)1 therapy	3D Slicer, PyRadiomics	Manual	Blood tests, other clinical information	Radiomic features	Cross-validation	Response to immunotherapy, survival	RECIST 1.1	DNN
Ravanelli [32]	Nivolumab	TexRAD, R, MedCalc Software	Manual	NA	Histogram features	Cross-validation	Survival	iRECIST	Cox proportional hazards
Dercle [38]	Nivolumab	MATLAB	Semiautomatic	NA	Volume, GLCM IMC1, DWT1, Sigmoid slope	Split sample	Survival	RECIST 1.1	RF
He [26]	Anti-PD-(L)1 therapy	3D Slicer, R, Python	Manual	NA	Radiomic features	Split sample	Survival	RECIST 1.1	DL
Liu [39]	Anti-PD-1 therapy	ITK-SNAP, Python, R	Manual	Distant metastasis	Delta-radiomic features	Split sample	Response to immunotherapy	iRECIST	LR

NA, not applicable; *Ref*, reference; *ECOG*, Eastern Cooperative Oncology Group; *CT*, computed tomography; *PET*, positron emission tomography; *KLD*, Kullback–Leibler divergence; *GTV*, gross tumor volume; *TA*, texture analysis; *QVT*, quantitative vessel tortuosity; *GLCM IMC1,* gray-level co-occurrence matrix; *DWT1*, discrete wavelet transform; *RECIST*, response evaluation criteria in solid tumors; *TPS*, tumor proportion score; *iRECIST*, modified RECIST1.1 for immune-based therapeutics; *RF*, random forest; *LDA*, linear discriminant analysis; *DLDA*, diagonal linear discriminant analysis; *QDA*, quadratic discriminant analysis; *SVM*, support vector machine; *LR*, logistic regression; *NB*, Naïve Bayes; *DNN*, deep neural networks; *DL*, deep learning

Fig. 3 **a** Histogram of the phase of all studies according to the phase classification criteria for image mining studies; **b** bar chart of the mean score of each study according to the radiomics quality scoring tool; and **c** pie chart of the mean score of studies according to the radiomics quality scoring tool

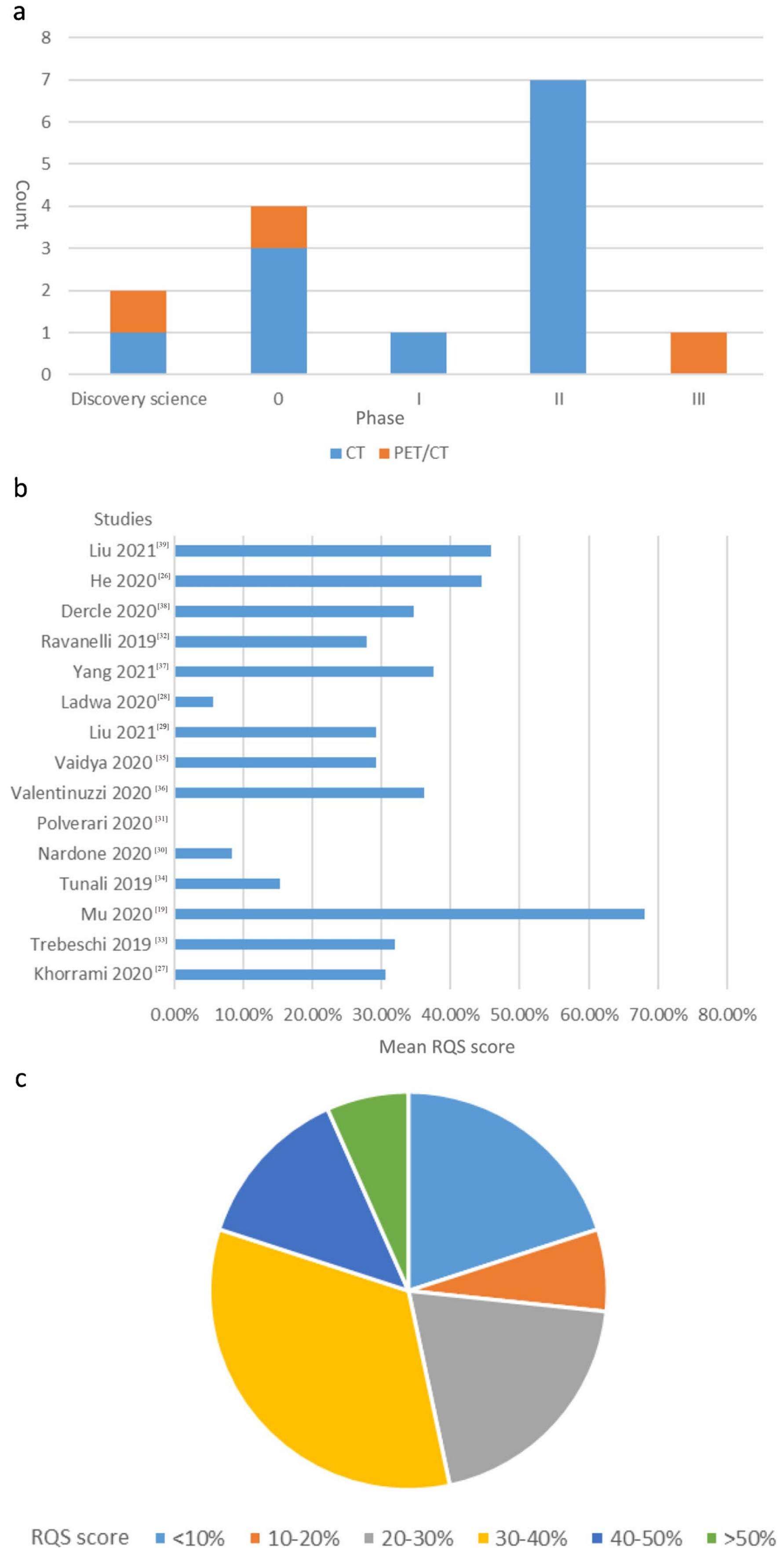


Table 3 Summary of the radiomics quality score of studies

Study ID	Image protocol quality	Multiple segmentations	Phantom study on all scanners	Imaging at multiple time points	Feature reduction or adjustment for multiple testing	Multivariable analysis with non-radiomic features	Detect and discuss biological correlates	Cutoff analyses	Discrimination statistics	Calibration statistics	Prospective study registered in a trial database	Validation	Comparison to gold standard	Potential clinical utility	Cost-effectiveness analysis	Open science and data	Total score	Mean score (%)
Khorrami [27]	1	0	1	1	−3	1	1	1	2	0	0	3	2	0	0	1	11	30.56
	1	0	1	1	−3	1	1	1	2	0	0	3	2	0	0	1	11	
Trebeschi [33]	1	0	0	0	3	0	1	0	2	0	0	2	2	0	0	0	11	31.94
	1	0	0	0	3	1	1	0	2	0	0	2	2	0	0	0	12	
Mu [19]	1	1	1	0	3	1	0	1	2	1	7	2	2	2	0	1	25	68.06
	1	1	1	0	3	1	0	1	2	1	7	2	2	2	0	0	24	
Tunali [34]	1	1	1	0	3	1	0	1	1	0	0	−5	0	0	0	1	5	15.28
	1	1	1	0	3	1	0	1	1	1	0	−5	0	0	0	1	6	
Nardone [30]	1	1	0	0	−3	0	0	1	0	0	0	3	0	0	0	0	3	8.33
	1	1	0	0	−3	0	0	1	0	0	0	3	0	0	0	0	3	
Polverari [31]	1	1	0	0	−3	0	1	0	0	0	0	−5	2	0	0	0	−3	0.00
	1	1	0	0	−3	0	1	0	0	0	0	−5	2	0	0	0	−3	
Valentinuzzi [36]	1	1	0	1	3	0	0	1	2	0	0	2	2	0	0	0	13	36.11
	1	1	0	1	3	0	0	1	2	0	0	2	2	0	0	0	13	
Vaidya [35]	0	0	0	0	3	1	0	0	2	0	0	2	2	0	0	1	11	29.17
	0	0	0	0	3	1	0	0	2	0	0	2	2	0	0	0	10	
Liu [29]	1	1	0	0	3	0	0	1	1	0	0	2	2	0	0	0	11	29.17
	1	1	0	0	3	0	0	0	1	0	0	2	2	0	0	0	10	
Ladwa [28]	0	0	0	0	−3	0	0	1	0	0	0	2	2	0	0	0	2	5.56
	0	0	0	0	−3	0	0	1	0	0	0	2	2	0	0	0	2	
Yang [37]	1	1	0	0	3	1	1	1	2	0	0	2	2	0	0	0	14	37.50
	1	1	0	0	3	1	0	1	2	0	0	2	2	0	0	0	13	

Table 3 (continued)

Study ID	Image protocol quality	Multiple segmentations	Phantom study on all scanners	Imaging at multiple time points	Feature reduction or adjustment for multiple testing	Multivariable analysis with non-radiomic features	Detect and discuss biological correlates	Cutoff analyses	Discrimination statistics	Calibration statistics	Prospective study registered in a trial database	Validation	Comparison to gold standard	Potential clinical utility	Cost-effectiveness analysis	Open science and data	Total score	Mean score (%)
Ravanelli [32]	1	1	0	0	3	0	0	1	0	0	0	2	2	0	0	0	10	27.78
	1	1	0	0	3	0	0	1	0	0	0	2	2	0	0	0	10	
Dercle [38]	1	1	0	1	3	0	0	1	1	0	0	2	2	0	0	0	12	34.72
	1	1	0	1	3	0	0	1	2	0	0	2	2	0	0	0	13	
He [26]	1	1	1	0	3	1	1	1	1	0	0	2	2	2	0	0	16	44.44
	1	1	1	0	3	1	1	1	1	0	0	2	2	2	0	0	16	
Liu [39]	1	1	1	1	3	1	1	0	1	1	0	2	2	2	0	0	17	45.83
	1	1	1	1	3	1	0	0	1	1	0	2	2	2	0	0	16	

For each study, the first row represents the RQS score according to one reviewer, and the second row represents the RQS score according to another reviewer. Each study was assigned a number of points per RQS component and summed to give a total score (range −8 to +36). A score of −8 to 0 points correspond to 0% and 36 points correspond to 100%. Mean scores of the two evaluations are presented as a percentage analysis, but none performed cost-effectiveness analysis. In terms of open science and data, one study [27] obtained radiomic features on a set of representative regions of interest (ROIs), and the calculated features as well as representative ROIs were open access.

Table 4 shows the inter-observer agreement of RQS domains between two reviewers. The inter-reviewer reliability of five domains was strong to very strong (weighted kappa coefficient ranges from 0.60 to 1.00) but moderate in the "open science and data" (weighted kappa coefficient = 0.595). The RQS scores on the remaining domains were completely consistent between the two reviewers.

The value of radiomics in predicting immunotherapy response/outcome

The patients could be stratified into low- and high-risk groups by radiomic models.

The first meta-analysis of comparing the immunotherapy outcome between the two groups showed that the pooled HR was 1.96 (95%CI 1.61–2.40, $p < 0.001$) for OS (five studies, $n = 562$; Fig. 4a) and 2.39 (95%CI 1.69–3.38, $p < 0.001$) for PFS (five studies, $n = 520$; Fig. 4b). The I^2 statistic implied low heterogeneity among the studies ($I^2 = 42.8\%$ and 47.7% for OS and PFS, respectively). The second meta-analysis involving four radiomic studies ($n = 392$) for predicting the response to immunotherapy achieved a pooled diagnostic OR of 14.99 (95%CI 8.66–25.95) (Fig. 5). The sensitivity ranged from 63 to 89% and specificity ranged from 53 to 89%. The pooled sensitivity and specificity were 76% (95%CI 68–83%) and 84% (95%CI 79–89%), respectively (Supplementary Fig. 1a, b). The pooled PLR and NLR were 3.63 (95%CI 2.18–6.04) and 0.29 (95%CI 0.20–0.44), respectively (Supplementary Fig. 1c, d). The I^2 statistic results indicated low heterogeneity in the diagnostic OR ($I^2 = 0\%$) and NLR ($I^2 = 27.0\%$), but moderate heterogeneity in the sensitivity ($I^2 = 55.0\%$), specificity ($I^2 = 71.1\%$), and PLR ($I^2 = 57.8\%$).

Discussion

This present systematic review and meta-analysis explored whether radiomics could predict the immunotherapy response/outcome in patients with NSCLC and evaluated the quality of included studies using the phase classification criteria for image mining studies and the RQS tool. In addition, our meta-analysis, for the first time, combined and interpreted distinct independent investigatory data quantitatively and might provide key clues for its clinical application and further research. Despite promising results, these radiomic studies were far from providing definitive conclusions for clinical implementation and widespread use due to immature phases and relatively poor methodological quality.

The translation of image mining research in the clinical arena is limited by the huge variability of the methods used for image analysis together with the impasse to reproduce the results when tested in a different cohort

Table 4 The inter-observer agreement of the radiomics quality score tool

RQS domains	Weighted kappa*	95%CI	*p* value
Image protocol quality	NA	NA	NA
Multiple segmentations	NA	NA	NA
Phantom study on all scanners	NA	NA	NA
Imaging at multiple time points	NA	NA	NA
Feature reduction or adjustment for multiple testing	NA	NA	NA
Multivariable analysis with non-radiomic features	0.867 (0.127)	0.618–1.116	0.001
Detect and discuss biological correlates	0.706 (0.185)	0.343–1.069	0.004
Ct-off analyses	0.842 (0.151)	0.546–1.138	0.001
Discrimination statistics	0.898 (0.098)	0.706–1.090	< 0.001
Calibration statistics	0.762 (0.223)	0.325–1.199	0.002
Prospective study registered in a trial database	NA	NA	NA
Validation	NA	NA	NA
Comparison to gold standard	NA	NA	NA
Potential clinical utility	NA	NA	NA
Cost-effectiveness analysis	NA	NA	NA
Open science and data	0.595 (0.244)	0.117–1.073	0.012

NA, not applicable

*The value of the weighted kappa ranges from −1.0 to 1.0, and values can be roughly interpreted as poor (< 0.20), fair (0.21–0.40), moderate (0.41–0.60), strong (0.61–0.80), and very strong (0.81–1.00). If all ratings are the same for the two reviewers, the weighted kappa value cannot be calculated

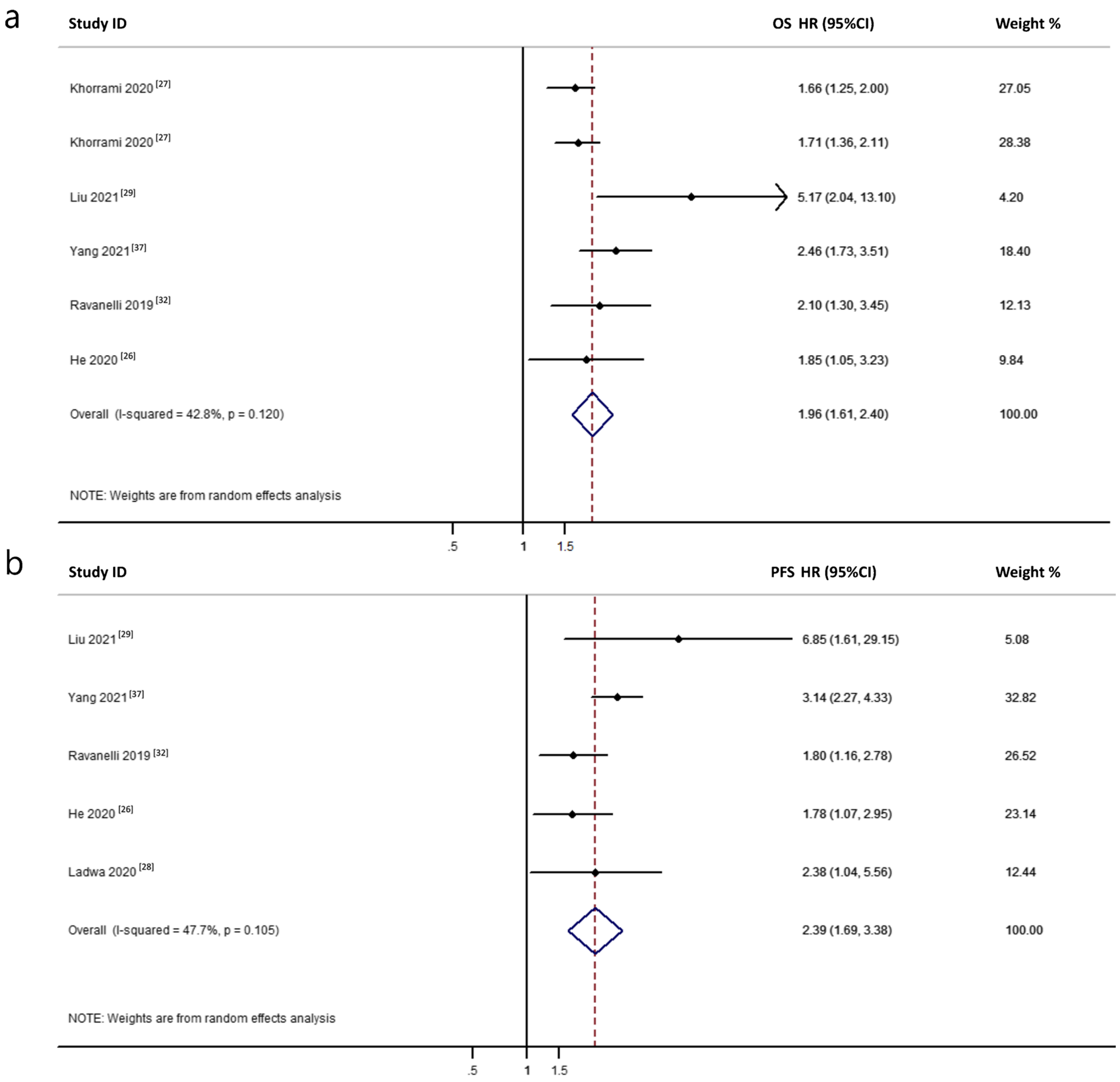
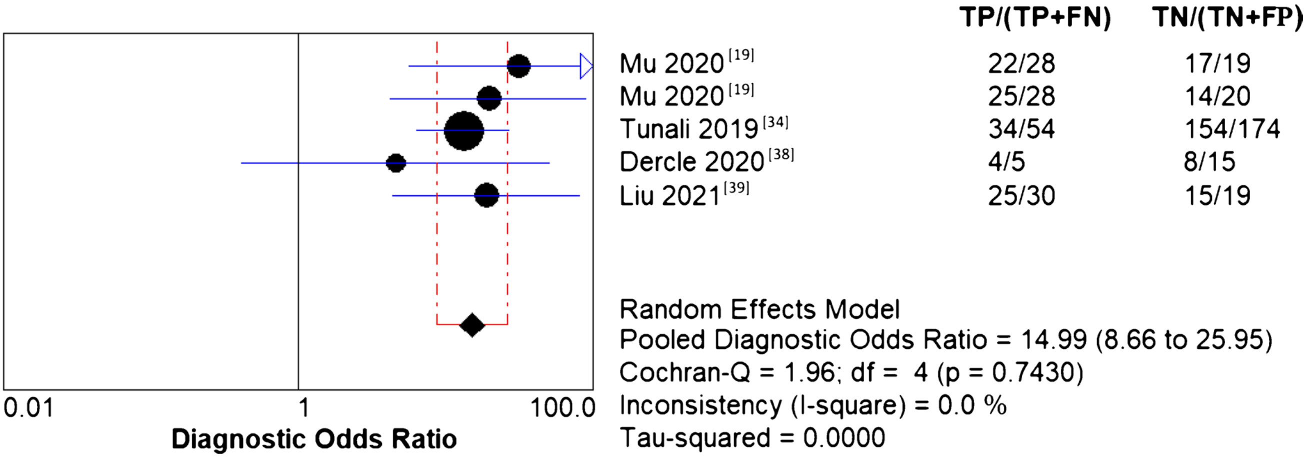


Fig. 4 Forest plots of the predictive performance of radiomics in **a** overall survival and **b** progression-free survival of NSCLC patients treated with immunotherapy. *Note*: Hazard ratio for each study is presented as a black dot, with the horizontal line indicating the 95% confidence interval. Pooled result for all studies is presented as a black diamond

of patients [22]. Researchers should assess whether the model is predictive for the target patient population or just for a specific subset of samples. Validation techniques are useful tools to assess model performance; valid models should exhibit statistical consistency between the training and validation datasets. Therefore, internal validation (e.g., cross-validation and bootstrapping) and external validation (e.g., temporal and geographic) are beneficial to be performed. Typically, the internal validation, used for a preliminary evaluation or for the fine-tuning of the model under development, overestimates the performance [40]. An externally validated model is more reliable than an internally validated model because data obtained from other institutions are considered more independent, which reinforces the validation. External validation is crucial to verify the generalizability of the models [40]; and the random patient selection is an essential prerequisite, as well as the balance in patient characteristics. In particular, the geographic validation, which accounts for technical variability aspects (scanners, acquisition parameters, and



	TP/(TP+FN)	TN/(TN+FP)
Mu 2020[19]	22/28	17/19
Mu 2020[19]	25/28	14/20
Tunali 2019[34]	34/54	154/174
Dercle 2020[38]	4/5	8/15
Liu 2021[39]	25/30	15/19

Random Effects Model
Pooled Diagnostic Odds Ratio = 14.99 (8.66 to 25.95)
Cochran-Q = 1.96; df = 4 (p = 0.7430)
Inconsistency (I-square) = 0.0 %
Tau-squared = 0.0000

Fig. 5 Forest plot of the effect size calculated as diagnostic odds ratio for studies investigating the diagnostic accuracy of radiomics in immunotherapy response prediction in NSCLC patients. Abbreviations: 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. *Note* Diagnostic odds ratio for each study is presented as a black dot, with the horizontal line indicating the 95% confidence interval. Pooled result for all studies is presented as a black diamond

protocols) [40], is expected to be more representative of the clinical setting. However, two studies [31, 34] included in this systematic review were classified as discovery science due to lack of validation analysis. For example, one prospective study [34] with 228 patients only demonstrated the potential utility of radiomics to predict rapid disease progression phenotypes, with the highest AUC of 0.87; these results need to be replicated in the independent validation cohorts. Additionally, only two studies [27, 30] performed external validation from another center. Consequently, the results are promising but still not mature enough for clinical application and widely used as noninvasive image mining tools.

In the era of evidence-based medicine, rigorous research with strict rules is the only way forward to achieve clinical acceptance. The lack of a rigorous procedure largely leads to low RQS scores of the radiomic studies. Two studies [30, 31] only explored the relationship of radiomic features with immunotherapy response in NSCLC, without establishing a simple model to facilitate clinical application. There are only several studies [19, 26, 27, 34, 39] that have analyzed feature robustness considering differences across machines or temporal variability. The RQS scores were low in the domains of prospective study, potential clinical utility, cost-effectiveness analysis, and open science. Only Mu et al. [19] obtained a mean RQS score above 50%. Recently, several guidelines have been proposed to encourage description of radiomic workflows in detail and to provide suggestions for the construction of prediction models [41–43]. These guidelines may help improve the quality of radiomic works and accelerate their clinical applications. However, as radiomics is a work-in-progress field and is developing constantly, the RQS tool may need modification to be a widely accepted tool for radiological research methodologies [44, 45].

Controversy remains regarding the best time point of imaging for decision-making and clinical management. There were 10 (66.7%) studies that obtained radiomic features from images at baseline, and only four studies [27, 36, 38, 39] explored the predictive value of radiomic features deriving from pre- and post-treatment scans. Valentinuzzi et al. [36] found that some features showed high predictive value at baseline but significantly decreased at months 1 and 4 after treatment. Some studies [13, 46, 47] suggested that tumor characteristics, such as histopathology, microenvironment, and immune contexture, may affect the response to immunotherapy. The changes in radiomic features might be associated with treatment response in solid tumors. Longitudinal features deriving from images at multiple time points can provide complementary information and improve prediction performance. Thus, analyzing features from images at different time points (pre-, during, and post-treatment) is useful to illuminate temporal variabilities of radiomic features and to increase the potential of radiomics in NSCLC treatment decision.

There are some limitations in this study. First, the heterogeneity of the studies included in this meta-analysis should be mentioned. The included studies differed in terms of the methodology of the used image reconstruction, feature extraction, and the algorithms used. Second, due to the limited studies included, subgroup analysis was not performed to investigate the influence of various conditions on radiomic findings. Further reviews including more studies with increased sample size are needed to address the issue. Finally, although RQS is a useful tool for the quality assessment of radiomic studies, it has limitations. Radiomics is a young field and RQS is inevitably immature. It is necessary to improve RQS items in response to actual practical needs.

Conclusions

In summary, the radiomic approach shows potential for the prediction of immunotherapy response and outcome in patients with NSCLC. However, the immature phases and unsatisfactory quality of the studies imply that the proposed models are not currently available for clinical implementation. Before radiomics can be successfully introduced into NSCLC clinical settings, further prospective studies with strict adherence to existing guidelines and multicenter validation need to be performed. Additionally, some technical barriers should be faced when considering implementing image mining tools into the everyday practice. Persistent efforts are required to make this tool to be widely used in clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-021-05509-7.

Author contribution QYC and LZ contributed to the conception and design of the study, the analysis and interpretation of data, and the work draft. XKM, FW, JF, and ZJ participated in the data extraction and analysis. JJY and LYC designed figures and tables. BZ and SXZ offered guidance in study design and revised the article critically for important intellectual content. All authors have read and approved the final version of the manuscript.

Funding This work received funding from the National Natural Science Foundation of China (81871323 and 81801665); National Natural Science Foundation of Guangdong Province (2018B030311024); and Scientific Research Cultivation and Innovation Foundation of Jinan University (21620447). The funders had no role in study design, data collection and analysis, preparation of the manuscript, or decision to publish.

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

Code availability Not applicable.

Declarations

Conflict interest The authors declare no conflict interest.

Ethics approval This review was approved by the institutional review board of the First Affiliated Hospital of Jinan University.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7–30.
2. Remon J, Ahn MJ, Girard N, et al. Advanced-stage non-small cell lung cancer: advances in thoracic oncology 2018. J Thorac Oncol. 2019;14(7):1134–55.
3. Creelan BC. Update on immune checkpoint inhibitors in lung cancer. Cancer Control. 2014;21(1):80–9.
4. Luke JJ, Ott PA. PD-1 pathway inhibitors: the next generation of immunotherapy for advanced melanoma. Oncotarget. 2015;6(6):3479–92.
5. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012;12(4):252–64.
6. Mok TSK, Wu YL, Kudaba I, et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. Lancet. 2019;393(10183):1819–30.
7. Murakami S. Durvalumab for the treatment of non-small cell lung cancer. Expert Rev Anticancer Ther. 2019;19(12):1009–16.
8. Socinski MA, Jotte RM, Cappuzzo F, et al. Atezolizumab for first-line treatment of metastatic nonsquamous NSCLC. N Engl J Med. 2018;378(24):2288–301.
9. Akinleye A, Rasool Z. Immune checkpoint inhibitors of PD-L1 as cancer therapeutics. J Hematol Oncol. 2019;12(1):92.
10. Topalian SL, Hodi FS, Brahmer JR, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med. 2012;366(26):2443–54.
11. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet. 2017;389(10066):255–65.
12. Teixido C, Vilarino N, Reyes R, Reguart N. PD-L1 expression testing in non-small cell lung cancer. Ther Adv Med Oncol. 2018;10:1758835918763493.
13. Galon J, Mlecnik B, Bindea G, et al. Towards the introduction of the "Immunoscore" in the classification of malignant tumours. J Pathol. 2014;232(2):199–209.
14. Haragan A, Field JK, Davies MPA, Escriu C, Gruver A, Gosney JR. Heterogeneity of PD-L1 expression in non-small cell lung cancer: implications for specimen sampling in predicting treatment response. Lung Cancer. 2019;134:79–84.
15. Patel SP, Kurzrock R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. Mol Cancer Ther. 2015;14(4):847–56.
16. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–6.
17. Patyk M, Silicki J, Mazur R, Krecichwost R, Sokolowska-Dabek D, Zaleska-Dorobisz U. Radiomics—the value of the numbers in present and future radiology. Pol J Radiol. 2018;83:e171–4.
18. Whiteside TL. The tumor microenvironment and its role in promoting tumor growth. Oncogene. 2008;27(45):5904–12.
19. Mu W, Tunali I, Gray JE, Qi J, Schabath MB, Gillies RJ. Radiomics of (18)F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy. Eur J Nucl Med Mol Imaging. 2020;47(5):1168–82.
20. Scrivener M, de Jong EEC, van Timmeren JE, Pieters T, Ghaye B, Geets X. Radiomics applied to lung cancer: a review. Transl Cancer Res. 2016;5(4):398–409.
21. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
22. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on

artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019;46(13):2656–72.

23. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749–62.
24. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998;17(24):2815–34.
25. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.
26. He BX, Dong D, She YL, et al. Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker. J Immunother Cancer. 2020;8(2):e000550.
27. Khorrami M, Prasanna P, Gupta A, et al. Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non-small cell lung cancer. Cancer Immunol Res. 2020;8(1):108–19.
28. Ladwa R, Roberts KE, O'Leary C, Maggacis N, O'Byrne KJ, Miles K. Computed tomography texture analysis of response to second-line nivolumab in metastatic non-small cell lung cancer. Lung Cancer Manag. 2020;9(3):LMT38.
29. Liu C, Gong J, Yu H, Liu Q, Wang SP, Wang JL. A CT-based radiomics approach to predict nivolumab response in advanced non-small-cell lung cancer. Front Oncol. 2021;11:544339.
30. Nardone V, Tini P, Pastina P, et al. Radiomics predicts survival of patients with advanced non-small cell lung cancer undergoing PD-1 blockade using nivolumab. Oncol Lett. 2020;19(2):1559–66.
31. Polverari G, Ceci F, Bertaglia V, et al. (18)F-FDG Pet parameters and radiomics features analysis in advanced NSCLC treated with immunotherapy as predictors of therapy response and survival. Cancers (Basel). 2020;12(5):1163.
32. Ravanelli M, Agazzi GM, Milanese G, et al. Prognostic and predictive value of histogram analysis in patients with non-small cell lung cancer refractory to platinum treated by nivolumab: a multicentre retrospective study. Eur J Radiol. 2019;118:251–6.
33. Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol. 2019;30(6):998–1004.
34. Tunali I, Gray JE, Qi J, et al. Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: an early report. Lung Cancer. 2019;129:75–9.
35. Vaidya P, Bera K, Patil PD, et al. Novel, non-invasive imaging approach to identify patients with advanced non-small cell lung cancer at risk of hyperprogressive disease with immune checkpoint blockade. J Immunother Cancer. 2020;8(2):e001343.
36. Valentinuzzi D, Vrankar M, Boc N, et al. [18F]FDG PET immunotherapy radiomics signature (iRADIOMICS) predicts response of non-small-cell lung cancer patients treated with pembrolizumab. Radiol Oncol. 2020;54(3):285–94.
37. Yang Y, Yang JC, Shen L, et al. A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer. Am J Transl Res. 2021;13(2):743–56.
38. Dercle L, Fronheiser M, Lu L, et al. Identification of non-small cell lung cancer sensitive to systemic cancer therapies using radiomics. Clin Cancer Res. 2020;26(9):2151–62.
39. Liu Y, Wu M, Zhang Y, et al. Imaging biomarkers to predict and evaluate the effectiveness of immunotherapy in advanced non-small-cell lung cancer. Front Oncol. 2021;11:657615.
40. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018;286(3):800–9.
41. Liu Z, Wang S, Dong D, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. Theranostics. 2019;9(5):1303–22.
42. Kalpathy-Cramer J, Freymann JB, Kirby JS, Kinahan PE, Prior FW. Quantitative imaging network: data sharing and competitive algorithm validation leveraging the cancer imaging archive. Transl Oncol. 2014;7(1):147–52.
43. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
44. Sanduleanu S, Woodruff HC, de Jong EEC, et al. Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score. Radiother Oncol. 2018;127(3):349–60.
45. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30(1):523–36.
46. Yi M, Jiao D, Xu H, et al. Biomarkers for predicting efficacy of PD-1/PD-L1 inhibitors. Mol Cancer. 2018;17(1):129.
47. Zou W, Wolchok JD, Chen L. PD-L1 (B7–H1) and PD-1 pathway blockade for cancer therapy: mechanisms, response biomarkers, and combinations. Sci Transl Med. 2016;8(328):328rv4.

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