Systematic review of radiomic biomarkers for predicting immune checkpoint inhibitor treatment outcomes

Methods 188 (2021) 61–72

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Systematic review of radiomic biomarkers for predicting immune checkpoint inhibitor treatment outcomes

Chong Zhang [1,*], Louise de A. F. Fonseca [1,*], Zhenwei Shi, Cheng Zhu, Andre Dekker, Inigo Bermejo [2], Leonard Wee [2]

Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, the Netherlands

ARTICLE INFO

Keywords:
Solid cancers
Immunotherapy
Treatment response
Immune response
Radiomics
Computed tomography
Systematic review

ABSTRACT

Background: Systemic therapy agents targeting immune checkpoint inhibitors have been approved for use since 2011. This type of therapy aims to trigger a patient's immune response to attack tumor cells, rather than acting against the tumor directly. Radiomics is an automated method of medical image analysis that is now being actively investigated for predictive markers of treatment response in immunotherapy.

Objective: To conduct an early systematic review determining the current status of radiomic features as potential predictive markers of immunotherapy response. Provide a detailed critical appraisal of methodological quality of models, as this informs the degree of confidence about current reports of model performance. In addition, to offer some recommendations for future studies that could establish robust evidence for radiomic features as immunotherapy response markers.

Method: A PubMed citation search was conducted for publications up to and including April 2020, followed by full-text screening. A total of seven articles meeting the eligibility criteria were examined in detail for study characteristics, model information and methodological quality. The review was conducted in the Cochrane style but has not been prospectively registered. Results are reported following Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA) guidelines.

Results: A total of seven studies were examined in detail, comprising non-small cell lung cancer, metastatic melanoma and a diverse assortment of solid tumors. Methodological robustness of reviewed studies varied greatly. Principal shortcomings were lack of prospective registration, and deficiencies in feature selection and dimensionality reduction, model calibration, clinical utility and external validation. A few studies with overall moderate to good methodological quality were identified. These results suggest that current state-of-the-art performance of radiomics in regards to discrimination (area under the curve or concordance index) is in the vicinity of 0.7, but the very small number of studies to date prevents any conclusive remarks to be made. We recommended future improvements in regards to prospective study registration, clinical utility, methodological procedure and data sharing.

Conclusions: Radiomics has a potentially significant role for predicting immunotherapy response. Additional multi-institutional studies with robust methodological underpinning and repeated external validations are required to establish the (added) value of radiomics within the pantheon of clinical tools for decision-making in immunotherapy.

1. Introduction

Immunotherapy has emerged as an important tool in the oncologists' arsenal of cancer treatment options for advanced-stage disease [1] and recalcitrant tumors [2,3]. A subset of modern drugs for immunotherapy, known as immune checkpoint inhibitors, aim to trigger an antitumor immune response from a patient's T cells, instead of acting directly against the tumor itself [4]. Ipilimumab, a cytotoxic T lymphocyte

* Corresponding authors.
E-mail addresses: chong.zhang@maastro.nl (C. Zhang), louise.ferreira@maastro.nl (L. de A. F. Fonseca).
[1] Chong Zhang and Louise de A. F. Fonseca contributed equally and are jointly first authors.
[2] Inigo Bermejo and Leonard Wee contributed equally to the concept and supervision of this review, and are jointly senior authors.

https://doi.org/10.1016/j.ymeth.2020.11.005
Received 20 July 2020; Received in revised form 25 November 2020; Accepted 26 November 2020
Available online 1 December 2020

antigen 4 (CTLA-4), was the first of this new group of drugs to receive Food and Drug Administration (FDA) approval in 2011 for the treatment of metastatic melanoma [5]. In 2014, two other immune checkpoint agents, pembrolizumab and nivolumab, were also approved for metastatic melanoma; these target the programmed cell death protein 1 (PD-1) [5,6] which allowed tumor cells to evade attack by T cells. The 5-year survival rate was shown to increase significantly from 5 to 19% to 26–55% in patients receiving immunotherapy, depending on the metastasis location [7]. In 2016, atezolizumab, a programmed cell death ligand 1 (PD-L1) inhibitor, was approved for use in metastatic urothelial carcinoma, and extended to metastatic NSCLC, showing significant improvement in overall survival [8]. El-Khoueiry *et al.* [9] reported a 20% response rate with immunotherapy for advanced hepatocellular carcinoma patients, in contrast with a 2% response rate observed by Llovet *et al.* [10] for patients treated without immunotherapy.

The current range of approved uses of PD-1/PD-L1 inhibitors now includes metastatic non-small cell lung cancer (NSCLC), metastatic renal cell carcinoma, recurrent or metastatic squamous cell carcinoma (SCC) of head and neck, hepatocellular carcinoma (HCC) and locally advanced or metastatic urothelial carcinoma. Treatment response to immunotherapy from single-agent anti-PD-L1/PD-1 ranges between 10 and 40% for the overall population depending on the patient indication, but seems to be inferior to drugs targeting CTLA-4 [11]. There are also a number of side effects of immunotherapy that need to be taken into consideration, such as diarrhea, colitis, hypophysitis, immune hepatitis and polyarthritis. The cost of immunotherapy treatment is currently high, and when we combine the different response to treatment between patients and possible side effects, it makes a compelling case to use these agents on patients that have the highest probability of a good response to treatment.

A patient's immune profile is currently obtained through biomarker analyses of tissue samples (e.g. biopsy and core biopsy) for inflammatory cytokines, genetic mutation variants, levels of PD-L1, levels of CTLA-4 and the presence of tumor-infiltrating lymphocytes [12]. The expression of these biomarkers is believed to indicate a favorable response to treatment. However, their current predictive power is subject to a number of uncertainties related to spatial heterogeneity in the tumor and changes in the inflammatory microenvironment over time [13,14]. Specifically, small tissue samples might be susceptible to sampling errors and may not capture the full extent of a tumor's heterogeneity. Repeated sampling over time in the same area is generally not feasible.

Radiomics is an emerging new method for non-invasive quantitative analysis of medical images, that encompasses a whole tumor *in situ* with its microenvironment, to define important phenotypes that respond well to immunotherapy. Delta radiomics refers to the application of radiomics techniques in longitudinal studies (comparison of repeated images over time in the same patient) [15]. Radiomics uses computer algorithms to extract a large number of intensity distribution, spatial heterogeneity (i.e. texture) and morphological object metrics (known as "features") directly from a selected region of interest in a given medical image [16]. Radiomics is being actively studied within and outside of the immunotherapy sphere for prognostication and treatment outcome prediction [16–19].

Radiomic features are hypothesized to have close correlation at the molecular level with differences in the tumor immune microenvironment, as shown in a number of studies for lung [20], skin [21], liver [22], brain [23] and breast [24] cancers. It is expected that radiomics could be correlated to gene expressions, the aforementioned immune response biomarkers and pathological tumor grade [20].

To date, the most commonly used radiological imaging modalities for radiomics investigation are computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI) [25]. A collection of individual image features that enables an adequately prognostic characterization of a tumor's imaging phenotype is known as a radiomic signature. Eligible features for signatures might be selected based on its reproducibility, independence from other features or association to the outcome of interest [16]. Several potential pitfalls of radiomics signature development have been identified [26–28]. There have been several systematic reviews about radiomics feature stability [28], methodological quality of radiomics models [29] and cancer-specific performance [30,31]. Significant efforts have been made to standardize definitions of radiomics features, such as the Image Biomarker Standardization Initiative (IBSI), that supports reproducible research and wider generalizability of results [32]. To the authors' knowledge, this is the first systematic review that focuses on radiomic biomarkers for immunotherapy across multiple cancers [18,33,34]. The common element among the cancers and interventions studies is the targeting of one or more specific checkpoint inhibitors expressed by these tumors.

This review aims to summarize the current status of radiomics features as potential predictors of response following immunotherapy treatment. A key motivation within this review is a critical assessment of methodological quality of the models that incorporate radiomic features, since this helps to establish what level of confidence may be derived from the currently reported model performance statistics. By way of synthesis, we will explore what further steps could be taken to establish a high level of evidence that supports the use of radiomics as immunotherapy response markers.

2. Method and materials

2.1. Eligibility criteria

In this study, we considered only the response to systemic treatments (alone or in combination with other therapies) that are based on immune checkpoints inhibitors. Other types of immunotherapy (such as monoclonal antibodies, vaccines, immune system modulators or T-cell transfer therapy) are not within the scope of this review. For the present, we only considered those radiomics studies using features that had been extracted according to pre-defined mathematical formulae, including but not limited to specific features listed in the IBSI. Deep learning neural network features were presently neglected (see exclusion criteria). Only studies on volumetric radiological scans specifically CT, PET and MRI were required. Articles eligible for review were required to describe studies on human subjects, be published as full text in peer-reviewed journals in the English language, published between 1st January 2012 and 30th April 2020.

2.2. Exclusion criteria

Deep learning may be viewed as a subset of radiomics that utilizes multilayered networks of artificial neurons to derive spontaneous features that associate with the desired outcome based on its training data. The body of published research and methodological development in regard to pre-defined radiomics features has arguably more maturity at the time of writing, compared to deep learning radiomics, with is an active and rapidly developing topic of the future. Studies on imaging modalities other than CT, PET or MRI were excluded.

2.3. Search strategy

The search for articles was executed in two phases. An early scoping search was attempted in October 2019 to see if there would be enough material to attempt a systematic review. This scoping review took into consideration the prior knowledge of immuno-oncology experts and provided some of the key search terms which could be used to locate articles in an electronic database.

For the primary search in this review, we sought for eligible publications within the PubMed electronic database after it had been merged with EMBASE. The foundation search strategy was a sensitive search for diagnostic and prognostic studies, using a combination of the broad Haynes [35] and Ingui [36] filters, with the additional modification

proposed by Geersing [37]. To this, we added the MeSH term "cancer", and the following text words anywhere in the title or abstract: "immunotherapy*", "immune checkpoint", "radiomic*", "textur*", "imag*", "computed tomography", "CT", "magnetic resonance", "MR", "positron emission tomography" and "PET". The text words were first combined with the 'OR' operator, then integrated with the MeSH term and the diagnostic studies filter using the 'AND' operator. Neither specific cancer types nor specific outcome types were included as search terms.

Finally, any articles referred to us by experts, known from the prior scoping search and/or found in the references section of the full-text articles we evaluated, were all counted as "prior knowledge".

2.4. Data management

A document repository was created on a shared network folder. Full texts were obtained through electronic subscription services held by Maastricht University. Microsoft Office document templates were used for tracking the article collection and data extraction items.

2.5. Selection process

We followed the methodological conventions of Cochrane systematic reviews and reported our findings according to recommendations of the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) checklist [38]. Three authors worked independently throughout the PubMed records screening process and then selections were combined. After screening by title and abstract alone, full texts were downloaded for the subset of potentially eligible articles for screening in detail. Lastly, articles unanimously deemed eligible were included for detailed review. Disagreements were resolved jointly through re-appraisal, and in case of deadlock, a fourth reviewer was available.

2.6. Data extraction

First, descriptive details of the radiomics studies of immunotherapy response were collected and summarized. These were: primary cancer type, imaging details, cohort description, sample size, immunotherapy target, primary clinical endpoint, software used for radiomics feature extraction, and whether multivariable models with combinations of radiomic and non-radiomic features had been evaluated.

Key reported findings of the included studies were extracted and summarized. These were: clinical outcome(s) predicted by the model(s); number of events and sample sizes used to develop the model; number of radiomic and non-radiomic features initially considered versus retained in the final model; the type of statistical or machine learning models used; reported performance metrics, and whether model calibration was assessed.

2.7. Objectives and prioritizations

2.7.1. Primary objective

The primary objective of interest is to determine the current performance status of radiomic feature for non-invasively predicting response to immunotherapy treatments. The response may be a direct clinically-measurable outcome, overall survival, progression or adverse side effects, or it could be a biological surrogate for the patient's immune system being activated, such as overexpression of known biomarkers. The predictive performance will be assessed primarily through an appropriate discrimination metric for the model in question, such as area under the receiver-operator curve (AUC) for binary predictions, or concordance index (c-index) for time-to-event predictions.

2.7.2. Secondary objectives

To understand what level of confidence may be assigned to the reported predictive performance of a model of immunotherapy response, it is essential to delve deeper into the methodological robustness of the reviewed papers. Of key concern is whether the reported model performance statistics might be at risk of bias, such that the given model performance might be less likely to be reproducible in new and future datasets.

A further secondary objective that comes closely in hand with appraisal of methodological robustness is to explore whether there are certain aspects of the included immunotherapy radiomics studies that might be improved in future. Such critical consideration is needed in order to establish high-quality evidence that supports use of radiomic features as dependable markers of response to immunotherapy.

2.7.3. Methodological quality of radiomics studies

There have been a number of tools proposed to appraise methodological quality of prognostic and diagnostic studies in general, such as QUADAS [39]. Recently, several quality criteria were composed into a Radiomics Quality Score (RQS) by Sanduleanu *et al.* [29] in a systematic review of radiomics models not specifically related to immunotherapy. As has been found in another recent systematic review by Fornacon-Wood *et al.* [31], methodological quality in radiomics modelling is a complex question that is not readily reducible to a single meaningful number. We have thus based a methodological appraisal on the points raised by the RQS, but refrained from assigning a quality score. In its place, we included a brief note of what, in our view, might have compromised some part of the methodological robustness in the study. Each of the three reviewers worked independently on extracting the methodological information, which was afterwards cross-checked by another reviewer. The methodological aspects we sought to extract from the studies were:

i. If the authors of the study had prospectively registered the intended methodology in a study database prior to commencement;
ii. If there was information about the imaging protocol, such as whether it was public or sufficiently detailed (e.g., contrast, slice thickness, reconstruction kernel) that might support reproducibility;
iii. Whether image pre-processing steps (e.g., digital filters, isotropic resampling) prior to radiomics extraction had been described;
iv. Whether features had been assessed, prior to the primary study subjects, for its repeatability and reproducibility (e.g., using phantoms, inter-observer tests, or re-using features tested from other studies);
v. If the authors had applied some form of appropriate feature selection method to minimize the risk of overfitting;
vi. Whether the correlation of radiomic features to biological and/or non-radiomic features had been evaluated;
vii. If the authors had provided clear justification for defining risk groups (e.g. cut-off and operating point) rather than fine-tuning for optimal groups, since the latter might produce overly optimistic results of discrimination;
viii. Whether there was some form of model validation reported, and how likely was the validation likely to be a measure of performance in new data;
ix. Whether the radiomics model performance had been compared with alternative predictors, preferably current 'gold standards' if they exist (such as TNM staging);
x. Whether the clinical utility of the model had been evaluated, through some form of cost-benefit or decision-curve analysis.

3. Results

3.1. Search results

A PRISMA flowchart (see Fig. 1) illustrates the screening process for full-text articles to review. In addition to 399 records located through

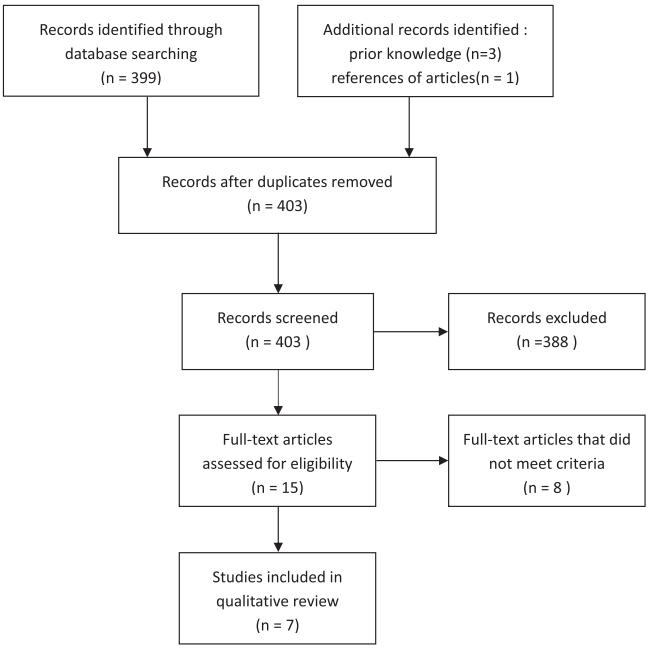

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram illustrating the numbers of records screened and excluded, to arrive at seven articles included in this review.

PubMed, three publications were included based on prior knowledge and one because it was cited by an article. We evaluated 15 full-text articles for eligibility, out of which eight articles were eventually excluded (three articles did not involve any patients treated by immunotherapy, one was for a monoclonal antibody – EGFR, and four articles did not involve radiomic features). The final number of relevant publications to review in detail was seven [12,40–45].

3.2. Overall characteristics of included studies

Characteristics of the articles reviewed in depth are given in Table 1. The majority of studies (4 out of 7) addressed primary non-small cell lung cancer (NSCLC) [12,40–42], including one study that examined both NSCLC and metastatic melanoma [12]. Only one study studied exclusively metastatic melanoma [45]. The remaining two studies [43,44] examined a diverse collection of advanced cancers.

All studies in this review utilized volumetric CT imaging, and almost exclusively contrast-enhanced CT. One study [43] did not explicitly mention the use of a contrast agent. The majority of studies used CT scanners from two or more vendors, which could be important to account for inter-scanner heterogeneity. One study used only a single scanner [40] while another gave little information about the scanners used [43].

Three studies included prospectively enrolled patients from other clinical trials [41,44,45]; one re-used prospective trial data as their training cohort [44] and the other two split them between training and validation [41,45]. The remainder of the studies used retrospective and/or single-institution cohorts.

Patients were mostly in advanced stages (e.g. metastatic, deemed uncurable, or stage IIIA or higher) [12,42,44,45].

The total sample size (training and validation, if any) ranged from 32 [43] to 491 [44]. The latter study [44] also had the biggest training set size of 135 patients, and included genome data from open access repositories for their external validation. Trebeschi *et al.* [12] used genomic data, but only to observe association, not for external validation.

Various immunotherapy checkpoints were targeted by the intervention, including CTLA-4 in two studies [41,45], either PD-1 or PD-L1 in six studies, and a combination of both PD-1/PD-L1 with CTLA-4 in two studies [41,45]. The study by Colen *et al.* [43] included a number of different immune checkpoint targets.

Overall survival was included as a clinical endpoint in three studies [40,44,45], and either short-term progression or hyper-progression in only one study [41]. One single study looked at adverse events, i.e. immunotherapy-induced pneumonitis [43].

Even among a small number of reviewed studies, four distinct radiomic feature extraction software packages are named and used. One of the studies used in-house code generated in Matlab [41], while two other studies [12,43] neglected to give any information about which software was used.

In all except two of the studies [12,43], the predictive performance of radiomics features was combined with non-radiomics features in multivariable models.

3.3. Details and results of radiomics-based models in included studies

The results of radiomics-based models as described in the included studies have been summarized in Table 2, along with modelling details. It is worth noting that reported performances across studies should not be compared directly when referring to different outcomes or cancer types.

The four studies with NSCLC patients reported performance indices in a wide range from 0.61 up to 0.87. The single survival analysis reported a c-index of 0.72 in the validation cohort. For the remaining three studies, combining radiomic features with other predictors either improved the radiomics-model only slightly (AUC increased from 0.66 to 0.67) or moderately (0.72 to 0.80), but none of these added values was demonstrated in external validation datasets. Neither of these results specifically demonstrated the added value of radiomics beyond biomarkers only. One study combining both NSCLC and metastatic melanoma reported overall AUC of 0.76 for both cancers in a validation set.

A single survival analysis on exclusively metastatic melanoma returned a c-index of 0.72 in validation (reproducible from training set, also 0.72) for overall survival, but gave no result for treatment response.

The two studies on various advanced cancers returned conflicting findings; one reported an AUC of 1.00 with perfect sensitivity and specificity [43] – but the finding needs to be called into question due to methodological deficiencies, that we shall discuss in the next section. The other reported AUCs in the range of 0.67–0.76 in external validation datasets, that appeared to be more plausible. The latter study by Sun *et al.* [44] estimated a hazard ratio of 0.58 (95% confidence interval: 0.39–0.87) for the radiomic signature on overall survival.

None of the studies examined model calibration, such as calibration-in-the-large (offset) or calibration slope.

All survival models were based on Cox proportional hazards analysis. The majority of binary outcome models were logistic regression or regularized regression, though one study used random forests [12] and another used anomaly detection [43].

The number of radiomic features considered as candidate variables ranged from 8 up to over 5000. In all the studies, the number of considered features outstripped the patient cohort sizes and numbers of events, on occasion by orders of magnitude. This recapitulates the need to have careful feature selection, dimensionality reduction and model validation to be able to minimize the risk of overfitting. Event rates ranged from as low as 6% of the sample size, up to a maximum of 39%, but in two studies [40,45] the numbers of events were unreported. Only the study by Tunali *et al.* [41] reported using synthetic minority oversampling (SMOTE) as a means to address the highly unbalanced rate of events.

3.4. Methodological quality of the included studies

An overview of the methodological quality of the included studies,

Table 1
Summary of general study characteristics.

Reference	Primary cancer	Imaging protocol details of relevance for radiomics	Cohort description	Sample size	Immuno-therapy target(s)	Primary clinical endpoint(s)	Radiomics extraction software	Combined with non-radiomics predictors
Tang *et al.*, 2018	**NSCLC**	Pre-treatment **contrast-enhanced CT** scans. Only single scanner. No further scan acquisition details given in manuscript or suppl. materials.	Retrospective institutional cohort of non-metastatic NSCLC treated by definitive surgery without induction therapy. (Training) Treated December 2000 through February 2012. (Validation) Treated January 2006 to December 2009.	114 (training) 176 (validation)	PD-L1	Overall survival	**IBEX**, version not specified	Yes; clinical risk factors and use of adjuvant therapy
Tunali *et al.*, 2019	**NSCLC**	**Contrast-enhanced CT** within 30 days prior to initiation of immunotherapy. Details in suppl. materials. Scanners: GE/ Siemens0.80x0.80 mm; 3 mm slices7 convolution kernels, majority were B41f.	Prospectively enrolled in industry-sponsored clinical trials of either anti-PD-1 or anti-PD-L1 as single agent, or in combination with CTLA4 as second agent. Treated June 2011-June 2016.	228	Either PD-1/ PD-L1, or in combination with CTLA4	Progression or hyper-progression at 2 months	**In-house**, based on Matlab v2016b and C++ (detailed in suppl. materials)	Yes: clinical and conventional biomarkers
Yoon *et al.*, 2020	**NSCLC (adenocarcinoma only)**	**Contrast-enhanced CT** of chest. Multiple CT scanners In-plane spatial resolution not reported; 1–1.25 mm slices (axial) and 2.5–3 mm slices (coronal) Soft tissue-optimized kernels.	Retrospective institutional cohort of advanced stage (>IIIA) lung adenocarcinoma with primary lesion distinguishable in CT. Diagnosed between January 2016 to August 2018.	153	PD-L1	PD-L1 expression	**AVIEW Research**, version unknown (Coreline Soft Inc., Seoul, South Korea)	Yes; clinical factors and EGFR status
Trebeschi *et al.*, 2019	**NSCLC or melanoma**	(Radiomics biomarker) **Contrast-enhanced CT** before and 12 weeks after start of treatment. Scanners: Toshiba/Siemens. 0.75x0.75 mm; 1 mm slices. (Genome association) **Contrast-enhanced CT** acquired within 60 days of diagnosis, no further details found in article or suppl. materials. (Anti-association) Same settings as for radiomics biomarker cohort.	(Radiomics) Retrospective institutional cohort of (stage unknown) NSCLC and metastatic melanoma receiving anti-PD-1 immunotherapy between2014 and 2016. (Genomic association only) Single institutional cohort from another country, surgically-treated NSCLC patients between 2006 and 2009. (Anti-association with chemotherapy outcome) Stage IV NSCLC treated with cytotoxic chemotherapy treated between 2015 and 2016.	203 (training: 133; testing: 70) 262 (genomics) 39 (chemotherapy)	PD-1	Radiologically-assessed (RECIST) immunotherapy response	Not reported in either article or suppl. materials.	No
Schraag *et al.*, 2019	**Melanoma**	**Contrast-enhanced whole-body CT** prior to immunotherapy and within 2 months of clinical data registration. CT image quality visually pre-screened before use. No details of	Prospectively registered patients between June 2006 to June 2016 with metastatic melanoma treated with anti-PD-1, anti-CTLA4 or a combination of both, with available baseline demographics and	69 (training) 34 (validation)	PD-1 and/or CTLA 4	Overall survival and radiologically-assessed (RECIST) immunotherapy response	**Mint Lesion™ v3.0** (Mint Medical, Dossenheim, Germany)	Yes, with conventional biomarkers (LDH/S100B) and tumor burden

(*continued on next page*)

Table 1 (*continued*)

Reference	Primary cancer	Imaging protocol details of relevance for radiomics	Cohort description	Sample size	Immuno-therapy target(s)	Primary clinical endpoint(s)	Radiomics extraction software	Combined with non-radiomics predictors
		scanner(s) and acquisition protocol (s) in article.	biomarkers (LDH and S100B).					
Colen *et al.*, 2018	**Multiple types of advanced cancers**	**Unspecified chest CT** according to institution protocol. Scanners: GE/ Philips/Siemens 0.78x0.78 mm, 2–2.5 mm slices Unspecified kernel.	Retrospective institutional cohort of patients treated on early phase immunotherapy clinical trials with at least one immunotherapeutic agent (including, among others immune checkpoint inhibitors). Treated between January 2010 and July 2015.	Case-control design (2 positive cases; 30 controls)	Multiple targets including immune checkpoint (s)	Immunotherapy-induced pneumonitis	Not reported in either article or suppl. materials	No
Sun *et al.*, 2018	**Multiple types of advanced solid tumors**	(Radiomics development) **Contrast enhanced CT** from 4 different scanners (predominantly GE) and various tube potentials (mainly 120kVp). Axial resolution 0.67–0.82 mm; slice thickness 1.25–2 mm. Additional details of image acquisition settings (also for validation cohorts) were provided in suppl. materials.	(Training) Re-use of prospective clinical trial of patients with various types of uncurable or metastatic solid tumors. Enrolled between May 2012 and March 2016. (Validation) Open access data of The Cancer Genome Atlas (TCGA) comprising lung, liver, bladder and head-and-neck tumors. Archive accessed until June 30, 2017. (Immune phenotype) Retrospective institutional cohort with most extreme tumor immune phenotypes. Treated between Aug 2005 and November 2015. (Treatment outcome) Re-use of data from five distinct Phase I clinical trials of either anti-PD-1 or anti-PD-L1 monotherapy against advanced solid tumors. Patients enrolled between December 2011 and January 2014.	135 (training) 119 (validation) 100 (immune phenotype) 137 (treatment outcome)	PD1/PD-L1	(1) Presence of tumor-infiltrating CD8 + cells, (2) association with immune phenotype, and (3) overall survival after immunotherapy	**LIFEx, v3.44**	Yes; with region of interest location as variable.

Abbreviations used in the table – NSCLC: non-small-cell lung carcinoma; HCC: hepatocellular carcinoma; CT: computed tomography; MRI: magnetic resonance imaging; PD-L1: programmed death-ligand 1; PD-1: programmed cell death protein; ; CTLA-4: cytotoxic T-lymphocyte-associated protein 4; EGFR: epidermal growth factor receptor; LDH: lactate dehydrogenase; RECIST: response evaluation criteria in solid tumors; ADC: apparent diffusion coefficient; DCE: dynamic contrast enhanced; suppl.: supplementary; approx.: approximately

with brief explanatory comments, has been provided in Table 3. None of the studies had been prospectively registered in a study database, and none of the studies provided an evaluation of potential clinical utility by way of a cost-benefits or decision curve analysis.

Methodological quality across the included studies was found to be highly heterogeneous with respect to risk of biased estimates for discriminative performance. The studies that included NSCLC subjects, viewed as a whole, were qualitatively of higher methodological quality compared to the rest, in part due to clarity of imaging protocols, inclusion of appropriate pre-processing, and tests of feature repeatability or reproducibility apart from the main study. Among the non-NSCLC work, the report of Sun *et al.* [44] were subjectively rated good or moderate on more of the methodological aspects.

Sufficient detail of imaging protocol to support reproducibility and validation were rated moderate or good in five studies, and poor in two studies [40,45] that only used a single CT scanner in the institution and/or did not provide details of image acquisition settings. The better studies on this aspect utilized multiple vendors' scanners during model development and provided information for others to attempt to reproduce their scans.

Image pre-processing steps prior to radiomics extraction, such as isotropic voxel resampling and intensity normalization, are important

Table 2
Summary of radiomics-based prediction model characteristics described in included studies.

Reference	Predicted outcome(s)	Number of events (Number of samples)	Number of features: considered (radiomics)/in final model (radiomics)	Type of model	Reported performance	Model calibration
Tang *et al.*, 2018	OS ‡, clustering †	NR (114)	(12)/(4) *Intensity histogram features - mean, standard deviation, uniformity; Textural feature - GLCM homogeneity.*	k-nearest neighbors Cox proportional hazards	Clustering: p = 0.002 OS: C-index: 0.72	No
Tunali *et al.*, 2019	TTP < 2 months † HPD †	TTP: 54 (228) HPD: 15 (172)	625 (600)/8 (4) *Texture features in TTP model - Radial gradient border SD-2D, 3D Laws E5L5E5, border 3D Laws E5E5L5 and border quartile coefficient of dispersion.* *Texture features in HPD model – NGTDM strength.*	Logistic regression	TTP: AUC: 0.717 (radiomics only) AUC: 0.804 (combined model) Sen: 67.91%; Spe: 74.44% (combined model) HPD: AUC: 0.865 (combined model) Sen: 91.34%; Spe: 66.14% (combined model)	No
Yoon *et al.*, 2020	PD-L1 expression (positive/negative) †	53 (153)	63 (58)/8 (4) *Texture only - GLCM_ASM, GLRLM_RV, GLRLM_RE and GLRLM_SRHGE*	Logistic regression	AUC: 0.661 (95% CI 0.580–0.735) (radiomics) AUC: 0.667 (95% CI 0.575–0.760) (combined) Sen: 52.8%, Spe: 76.0% (radiomics only) ¶	No
Trebeschi *et al.*, 2019	Treatment response†	NSCLC: 135 (266) Melanoma:77 (274)	5865/68 *Radiomic features appearing in the final model were not individually listed.*	Random forest	AUC: 0.76 (p < 0.001) for both cancers.	No
Schraag *et al.*, 2019	OS ‡ Treatment response†	OS: NR (69) TR: 27 (42)	8 (5)/OS: 3(2)/TR: 0 # *Intensity histogram - kurtosis; classical image measure – tumor burden.*	Cox proportional hazards, Logistic regression	OS: C-index: 0.720 (training) 0.716 (validation) Response: - #	No
Colen *et al.*, 2018	Immunotherapy-induced pneumonitis†	2 (32)	1860/2 *Intensity histogram features only – skewness & angular variance of sum of squares*	Anomaly detection algorithm	AUC, Sen, Spe: 1.00 (LOOCV)	No
Sun *et al.*, 2018	CD8 cells (high vs low)† OS ‡	Training: 135 Validation*: 119/100/53(137)	84 (78)/8 (5) *Intensity histogram – tumour minimum; textural GLRLM features - SRLGE, SRHGE, LGRE and LRLGE.*	Linear elastic net	CD8*: AUC: 0.74 (95% CI 0.66–0.82)/ AUC: 0.67 (95% CI 0.57–0.77)/ AUC: 0.76 (95% CI 0.66–0.86) Sen: 36%, Spe: 90% ¶ OS: HR 0.58, 95% CI 0.39–0.87; p = 0.0081	No

Abbreviations - AUC: Area under the curve; CI: confidence interval; HR: Hazard ratio; HPD: Hyperprogressive disease; LOOCV: Leave-one-out cross-validation; NR: Not reported; OR: Odds ratio; OS: overall survival; Sen: sensitivity; Spe: specificity; TR: treatment response; TTP: time to progression; GLCM : grey-level co-occurrence matrix; GLRLM : grey-level run-length matrix; NGTDM : neighbouring grey-tone difference matrix. Symbols - *multiple validation datasets; #: no radiomic feature predicted TR, no performance reported; ¶ based on optimal cutoff point; †: Binary Classification ‡: Time-to-event

for improving feature standardization, particularly textural radiomic features. This aspect has been rated as good for four studies that documented appropriate pre-processing, but three studies[42,43,45] were devoid of such information.

Generally, *a priori* feature selection on the basis of repeatability (e.g., test–retest) and reproducibility (e.g., inter-observer study) were more commonly performed in the NSCLC studies. Three studies used some degree of feature pre-selection [40–42], but four did not report any test of feature stability.

Overall, methods reported by all seven studies for feature selection or dimensionality reduction, which is one of the key steps to reduce risk of overfitting, were rated as suboptimal for a variety of reasons. Most studies failed to assess the internal validity of the feature selection by adding cross-validation or bootstrap resampling. Three studies used exclusively univariate association with an outcome to select features for the model [41,42,45] without adjustment for multiple-testing; this leads to a high risk of biased model results. In one of these, risk of overfitting was partly ameliorated by regularization with repeated cross-validation [42] and another study used backwards stepwise regression but then did not use either resampling and cross-validation [41] to check for

Table 3
Assessment of methodological quality of included studies.

Reference	Study pre registered	Imaging protocol	Image pre-processing	Feature reproducibility	Dimensionality reduction	Correlations with non-radiomic biomarkers	Justification of risk groupings	Independent validation	Compared radiomics to alternatives	Clinical utility evaluated
Tang *et al*, 2018	No	**Poor**: only single scanner, details neither in article nor suppl.	**Good**: details provided in suppl.	**Good**: multi-user delineations and selected model with only reproducible features.	**Moderate**: re-used 12 features from another study then hierarchical clustering to select model; however, no correction for multiple testing.	**Good**: clusters correlated with PD-L1 and CD3 + from pathology analysis.	**Poor:** risk groups defined only by clustering.	**Moderate**: validation set in same center but split by time.	**Good**: tested against immune-pathology features	Not evaluated
Tunali *et al,* 2019	No	**Good**: two scanners, main details provided in suppl.	**Good**: details provided in suppl.	**Good**: Test-retest (RIDER set) used to select repeatable features.	**Poor**: Univariate selection to choose features for multivariable modelling without multiple-testing correction; pairwise and volume-correlation elimination; but then backwards stepwise selection without either resampling or cross-validation.	**Poor**: Only combined clinical and radiomic models but did not test their correlation.	**Moderate**: median cut-offs were used for clinical model, but then optimal cutoffs for the rest.	**Poor**: lacking an independent test or validation cohort.	**Good**: tested against clinical variables.	Not evaluated
Yoon et al, 2020	No	**Moderate**: scanners from 3 vendors, details in article but did not specify pixel spacing.	**Poor**: nothing stated for image pre-processing steps.	**Good**: Interobserver reproducibility test to initially select features	**Moderate**: Univariate selection of reproducible features using the outcome but no correction for multiple testing, final lasso regularization step was appropriate using repeated cross-validation.	**Poor**: no correlation testing against non-radiomics features.	**Poor**: optimized cut-off used to select operating point for classification.	**Moderate**: Internal validation by bootstrapped optimism correction.	**Good**: tested against clinical variables.	Not evaluated
Trebeschi *et al*, 2019	No	**Good**: Two scanners, main details provided in suppl.	**Good**: Details provided in the suppl.	**Poor**: no check of either repeatability or reproducibility of features.	**Moderate:** Random forest with wrapper-type feature elimination, but no detailed information provided.	**Good**: tested against genetic profile for significant associations	**Poor**: No details about cut-off used for survival risk groups.	**Good**: two unique cohorts for validation.	**No**	Not evaluated
Schraag *et al,* 2019	No	**Poor**: image quality visually screened but no further details of scanner(s) or acquisition.	**Poor**: nothing stated for image pre-processing steps.	**Poor**: no check of either repeatability or reproducibility of features.	**Poor:** exclusively univariate Cox regression against the outcome was used to select individual features for modelling.	**Poor**: no correlation analyses with non-radiomic biomarkers.	**Poor**: optimal threshold tuning used to define risk groups.	**Poor**: validation set exists: not stated if the cohort was split in a non-random manner.	**Good**: tested against clinical factors and tumor burden.	Not evaluated
Colen *et al*, 2018	No	**Moderate**: no details about scanners, some settings given in detail.	**Poor**: no details on digital filters or resampling, intensity discretization reported.	**Poor**: no check of either repeatability or reproducibility of features.	**Poor**: MRMR method used to select two radiomics features for two events, no justification for effect of patient sampling or gross imbalance in number of events.	**Poor**: no correlation analyses with non-radiomic biomarkers.	**Poor**: no risk group analysis.	**Poor**: internal cross-validation only without accounting for unbalanced sample.	**No**	Not evaluated
Sun *et al*, 2018	No	**Good**: multiple independent scanners,	**Good**: isotropic voxel resampling described: no	**Poor**: no prior test of repeatability or	**Moderate**: appropriate use of elastic-net regularized	**Good**: assessed association between	**Good**: median used as group cut-offs for	**Good**: three cohorts (one external) used	**No**	Not evaluated

(*continued on next page*)

Table 3 (*continued*)

Reference	Study pre registered	Imaging protocol	Image pre-processing	Feature reproducibility	Dimensionality reduction	Correlations with non-radiomic biomarkers	Justification of risk groupings	Independent validation	Compared radiomics to alternatives	Clinical utility evaluated
		summary of imaging settings, only missing contrast infusion detail.	enhancement filters used.	reproducibility of features.	regression, but no details of cross-validation during feature selection.	radiomic signature and genomic signature, and tumor-infiltrating lymphocytes	CD8 and radiomic score	for testing and validation.		

robustness. Tang *et al.* [40] only started out with a relatively small set of twelve pre-selected radiomic features taken from a previous study. Two studies used wrapper-type feature elimination with either random forests [12] or elastic-net regularization [44], but lacked details about resampling and cross-validation during the process. In one study [43], the number of radiomic features retained equaled the number of events in the small dataset, with no procedure used to address imbalance in outcomes; this almost certainly ensured that model overfitting would occur.

Associative or correlation studies of image-derived features against known biological markers are very helpful to provide a sound rationale and explainable basis for radiomics. However, only three studies [12,40,44] assessed the correlation between radiomic and non-radiomic features, such as pre-existing genetic or immune-response biomarker. Tang *et al.* [40] reported significant correlations between radiomics-based clusters with PD-L1 and CD3+ expressions. Sun *et al.* [44] reported that their radiomic signature was correlated with levels of tumor-infiltrating lymphocytes and genetic features. Trebeschi *et al.* [12] reported that top-ranking genes with significant association to the radiomic signature were involved in mitosis and cell-cycle progression.

Justifications of risk groups were generally poorly executed among the reviewed papers. Reporting of model performance based on optimally-tuned risk group cutoffs present a high risk of overly-optimistic results and irreproducible performance. Sun *et al.* [44] and Tunali *et al.* [41] were the only two that used an unbiased cut-off for dichotomizing risk scores, such as the median. However, Tunali *et al.* [41] only did this for the clinical model, but used tuning-optimized cut-offs for the radiomics models. Sun *et al.* [44] also used optimal cutoff tuning to select an operating point for sensitivity and specificity, instead of a clinically-relevant justification. Other studies only used clustering to define risk groups [40], or exclusively used optimally-tuned cutoffs [42,45], or did not report [12,43].

Model validation on non-randomly sampled cohorts, preferably on multiple datasets derived by fully independent institutions and unrelated investigators, remains the benchmark test of robust model performance. Such validation was only performed by Trebeschi *et al.* [12] and Sun *et al.* [44]. Tang *et al.* [40] and Yoon *et al.* [42] were rated as moderate, because the former used a non-random (temporally) split cohort from the same institution for validation, and Yoon *et al.* [42] used a repeated bootstrapping method test for over-optimism. Independent validation was absent from Tunali *et al.* [41], and insufficient detail about the validation cohort was given by Schraag *et al.* [45]. The perfect results reported by Colen *et al.* [43] had been derived by internal cross-validation on 32 patients (of which there were only 2 pneumonitis events) without accounting for the low rate event rate and extremely unbalanced sample.

In regards to attempting to estimate the added value of radiomic features for predicting immunotherapy treatment response, four studies [40–42,45] compared radiomic predictors against alternative predictors such as clinical risk factors.

4. Discussion

The objective of this systematic review was to determine the current status of radiomics models (including clinical models that incorporate radiomic features) for predicting response to systemic treatment with an immune checkpoint-targeting agent. Given the potential toxicity and the foreseeable cost of immunotherapy, the availability of robust prediction models would allow personalizing immunotherapy towards patients who are most likely to derive net positive benefit from it.

This review was conducted in the Cochrane style, as closely as pragmatically possible, and we have reported these findings according to PRISMA guidelines (a completed PRISMA checklist is provided in our Supplementary Materials).

We selected seven full-length articles to review after a careful screening and selection process, resulting in a total pool of 1648 subjects from a variety of advanced, late stage or incurable solid tumors. A pooled statistical metanalysis was not possible given the very low number of studies as well as heterogeneity in population characteristics, modelling methodology and clinical endpoints used. The risks of irreproducibility and various pitfalls concerning radiomics models have been highlighted by previous articles. Thus, a valuable and informative part of our review is a deeply detailed analysis of the methodological quality of the included studies, specifically in the direction of potential for over-optimism in the reported results and the risk that published discriminative metrics might not be reproducible in future studies. Following a qualitative synthesis of our overall findings, we propose a number of recommendations that could help future investigations of immunotherapy radiomics obtain more robust estimates of prediction performance.

4.1. Qualitative synthesis

The included studies can be grossly divided into NSCLC, metastatic melanoma and a diverse group of advanced solid tumors. Methodological robustness was overall higher in studies with a major proportion of NSCLC patients compared to the rest, most noticeably in the reporting of imaging protocol, appropriate use (and documentation of) digital image pre-processing, use of external datasets (such as the RIDER Test-Retest [46], and inter-observer delineation studies) to partially pre-select radiomic features, and comparing radiomics predictors against other (clinical or biomarker) predictors to estimate the added value of radiomics.

Tang *et al.* [40] has reported a c-index of 0.72 for the prediction of overall survival in NSCLC with reasonably good methodology, but this estimate is significantly weakened by being a single-institution study employing only a single CT scanner, with clinical risk groups defined purely by optimal cluster separation. The random forest model by Trebeschi *et al.* [12] reported an AUC of 0.76 for NSCLC and metastatic melanoma patients, from a reasonably robust study with good external validation, but leaving 68 radiomic features in the final model presents

some questions about model robustness. Yoon *et al.* [42] and Tunali *et al.* [41] were comparable in methodological quality, and we detected different methodological flaws in each. These studies reported conflicting results about the *added value* of radiomics. The higher discrimination and greater added value of radiomics appeared to be for clinical progression endpoints, but given that no independent validation was conducted, reported performance estimates are likely to be over-optimistic.

A discriminative metric (either AUC or c-index) around the 0.7 value, therefore, seems to be the likely current status of performance of radiomics models for predicting overall survival or treatment response, with the caveat that the number of studies is too small to be conclusive.

The performance of radiomics models for immunotherapy in melanoma is not presently possible to judge. We located only two radiomic studies including metastatic melanoma, and the analysis in one of these was combined with NSCLC patients. We detected several methodological flaws in the one melanoma-only study [45], which then reported AUC in a validation set randomly sampled from the same cohort as the training set. This has been repeatedly pointed out by statisticians as the weakest method of estimating of out-of-sample predictive performance, since the validation and training sets are almost always guaranteed to have the same distribution of factors.

Among two studies comprising multiple solid tumors, the results by Sun *et al.* [44] appear to be dependable due to reasonably robust methodology. In multiple independent validation sets, their radiomics model had AUCs in the range 0.67–0.76, and they estimated a hazard ratio for overall survival of 0.58 for their radiomics signature. The perfect discrimination result reported in the only other multiple-tumors study is difficult to take seriously, due to the use of a case-control study design and multiple major flaws in modelling methodology.

As above, we propose that a discriminative index in the vicinity of 0.7 might also be valid for the current status of radiomics-immunotherapy studies on multiple tumors, with the same caveat of there being too few studies presently to be conclusive.

4.2. Limitations of the present review

The present study is an early systematic review in an emerging "hot topic" and clinically impactful question; could radiomics play a role in immunotherapy response prediction? Systemic therapies targeting one or more immune checkpoint inhibitors were approved for use in humans not so long ago, and we clearly need more time for high-quality evidence to emerge. Applying strict criteria to answer our selected question, we only reviewed a total of seven eligible peer-reviewed full-text articles. This is clearly not enough to make any conclusive statements, but we have attempted to summarize a robust state-of-present-art of radiomics for discriminating immunotherapy response based around a detailed analysis of methodological quality. Due to the low number of studies, methodological heterogeneity and variation of clinical endpoints, we have not attempted a statistical meta-analysis.

Additionally, though this review was conducted by a dedicated team, we did not have the resources to exhaustively sift the grey literature (e.g. unpublished reports, conference abstracts, non-peer reviewed publications) to locate everything on this topic. We only searched a single electronic database, though large, that does not include all possible journals of interest, and is known to have a time lag from publication to indexation. We elected to skip hand-searching of printed indices or manual search of journal tables of content, as these required resources we did not presently have. We also acknowledge that our automated search needs further refinement in future, but we presently made use of clinical experts and prior knowledge to supplement our search. At the end, we successfully found a few focused studies on this question, presenting robust modelling results, with moderate and/or good methodology overall.

Activity in deep learning analysis of medical images for prognosis or prediction, i.e. "deep radiomics", is likely to grow exponentially in the next few years. We recognize this will be a future hot topic for review, and also acknowledge that deep learning has potential to push boundaries much further in immunotherapy response prediction. As mentioned in the exclusion criteria, we left out the deep learning studies for the present review, because the radiomics community needs more time to mature with rapidly evolving deep learning techniques, while a better understanding of methodological robustness already exists for modelling with hand-crafted features. We foresee that methodological pitfalls in deep radiomics will be not any less troublesome with regards to risk of biased model performance and loss of external validity. Deep learning may likely exacerbate the severity of the aforementioned methodological flaws, and potentially render model performance interpretation even more difficult than it already is.

Lastly, we have not performed a careful analysis of selective reporting within studies, nor of publication bias across all studies. We can only note that all reviewed articles, irrespective of methodological quality, reported "positive" findings for their models, i.e. all discriminative indices (AUC or c-index) were reported to be greater than 0.5. There were no articles reporting "negative" results of modelling among the eligible studies to review.

4.3. Current challenges and recommendations for future studies

The majority of studies we reviewed made pragmatic use of retrospective patient data, and most consisted of institutional cohorts. These were generally limited in terms of the number of patients, unbalanced in terms of outcomes, and model validation cohorts were likely to be homogenous with the model development cohort. The latter implies that little useful knowledge about future performance in unseen data would be derived by randomly sampling a hold-out subset for validation.

The majority of the reviewed studies lacked adequate (non-randomly assigned) external validation and therefore their reported model performance would be difficult to widely generalize. Furthermore, all of the studies had focused exclusively on model discrimination performance, but had neglected to report on model calibration, calibration-in-the-large or other estimates of "goodness" of model fit.

In retrospective studies, investigators have no choice other than to use the radiological images "as is"; it is not possible to re-image the patient nor is it simple to assure for standardization of image acquisition protocols. A further notable methodological deficiency in the present studies pertains to feature selection, dimensionality reduction and safeguards against overfitting. The radiomics community, on the whole, concedes that parsimonious models, using as few standardized robust radiomic features as needed, is preferable to a complex but potentially over-fitted model. However, only the minority of reviewed studies appeared to keep this in mind.

Our recommendations for future studies into the potential of radiomics for predicting immunotherapy treatment response are as follows.

First, the model development protocol should be prospectively registered and/or made openly accessible on a study registry. This could be used to support (a) constructive engagement with other radiomics researchers towards improving modelling methodology, and (b) transparency of study outcomes by way of reporting all results (both positive and negative).

Second, studies should explicitly state where a radiomics-empowered model might fit in the immunotherapy workflow and what would be its anticipated impact on the clinical decision. Since an obvious role for radiomics might be to identify patients that would maximally benefit from immunotherapy, a decision-analysis curve or an estimate of cost-effectiveness would be especially informative.

Third, investigators and reviewers should be cognizant of good practices regarding methodology, and particularly to watch for well-known pitfalls when modelling with very large numbers of candidate features relative to either event rate or sample size. Mature knowledge is readily available in the epidemiological and biostatistical community, and guidance already exists for diagnostic/prognostic modelling in

general [39], and for radiomics in particular [29].

Lastly, data sharing and data re-usability should be strongly encouraged among immunotherapy-radiomics investigators, since this is expected to enhance standardization, methods harmonization and, perhaps most importantly, robust external validation. Imaging data should be either openly or privately shared through major repositories such as The Cancer Imaging Archive [47], and compliance with Findable-Accessible-Interoperable-Reusable (FAIR) data principles [48] is relatively straightforward.

5. Conclusion

New developments in immunotherapy with checkpoint inhibitor agents have shown impressive results in the long-running battle against advanced cancer and metastatic disease. However, this comes with a heavy financial cost and a risk of treatment-related adverse events. Radiomics has the potential to assume a highly significant role in clinical decision-making by identifying non-invasive image-based biomarkers for either clinical response or immune response to such systemic therapy, or otherwise determining indicators for high risk of treatment-induced injury. This systematic review has examined the small volume of literature available to date, and reports that the present state-of-the-art discriminative performance of radiomics for immunotherapy response is around 0.7 (for either AUC or c-index). The extremely low number of high-quality studies to date prevents any conclusive statements to be made. We have made some recommendations, pertaining to improving the overall methodological quality and to advocate for data sharing, that we hope will accelerate developments on this important question and push beyond our present level.

Funding

The production of this review was not connected to any specific funding the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2020.11.005.

References

[1] S. Kruger, et al., Advances in cancer immunotherapy 2019 – latest trends, J. Exp. Clin. Cancer Res. 38 (1) (2019) 268.
[2] D.W.S. Mak, S. Li, A. Minchom, Challenging the recalcitrant disease-developing molecularly driven treatments for small cell lung cancer, Eur. J. Cancer 119 (2019) 132–150.
[3] V.P. Balachandran, G.L. Beatty, S.K. Dougan, Broadening the impact of immunotherapy to pancreatic cancer: challenges and opportunities, Gastroenterology 156 (7) (2019) 2056–2072.
[4] F. Teng, et al., Progress and challenges of predictive biomarkers of anti PD-1/PD-L1 immunotherapy: A systematic review, Cancer Lett. 414 (2018) 166–173.
[5] J. Eno, Immunotherapy through the years, J. Adv. Pract. Oncol. 8 (7) (2017) 747–753.
[6] M. Abbott, Y. Ustoyev, Cancer and the immune system: the history and background of immunotherapy, Semin. Oncol. Nurs. 35 (5) (2019), 150923.
[7] A. Sandru, et al., Survival rates of patients with metastatic malignant melanoma, J. Med. Life 7 (4) (2014) 572–576.
[8] L. Fehrenbacher, et al., Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial, Lancet 387 (10030) (2016) 1837–1846.
[9] A.B. El-Khoueiry, et al., Nivolumab in patients with advanced hepatocellular carcinoma (CheckMate 040): an open-label, non-comparative, phase 1/2 dose escalation and expansion trial, Lancet 389 (10088) (2017) 2492–2502.
[10] J.M. Llovet, et al., Sorafenib in advanced hepatocellular carcinoma, N. Engl. J. Med. 359 (4) (2008) 378–390.
[11] D.S. Chen, I. Mellman, Elements of cancer immunity and the cancer-immune set point, Nature 541 (7637) (2017) 321–330.
[12] S. Trebeschi, et al., Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers, Ann. Oncol. 30 (6) (2019) 998–1004.
[13] D. Saeed-Vafa, et al., Combining radiomics and mathematical modeling to elucidate mechanisms of resistance to immune checkpoint blockade in non-small cell lung cancer, bioRxiv (2017), 190561.
[14] S. Koyama, et al., Adaptive resistance to therapeutic PD-1 blockade is associated with upregulation of alternative immune checkpoints, Nat. Commun. 7 (2016) 10501.
[15] X. Fave, et al., Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer, Sci. Rep. 7 (1) (2017) 588.
[16] P. Lambin, et al., Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (4) (2012) 441–446.
[17] H.J. Aerts, The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review, JAMA Oncol 2 (12) (2016) 1636–1642.
[18] A. Guerrisi, et al., Novel cancer therapies for advanced cutaneous melanoma: The added value of radiomics in the decision making process-A systematic review, Cancer Med. 9 (5) (2020) 1603–1612.
[19] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images are more than pictures they are data, Radiology 278 (2) (2016) 563–577.
[20] A.M. Rutman, M.D. Kuo, Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging, Eur. J. Radiol. 70 (2) (2009) 232–241.
[21] A. Esteva, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[22] M. Wu, et al., Predicting the grade of hepatocellular carcinoma based on non-contrast-enhanced MRI radiomics signature, Eur. Radiol. 29 (6) (2019) 2802–2811.
[23] Y. Li, et al., MRI features can predict EGFR expression in lower grade gliomas: A voxel-based radiomic analysis, Eur. Radiol. 28 (1) (2018) 356–362.
[24] N.M. Braman, et al., Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI, Breast Cancer Res. 19 (1) (2017) 57.
[25] G.L. Banna, et al., The promise of digital biopsy for the prediction of tumor molecular features and clinical outcomes associated with immunotherapy, Front Med (Lausanne) 6 (2019) 172.
[26] M. Hatt, et al., Multicentric validation of radiomics findings: challenges and opportunities, EBioMedicine 47 (2019) 20–21.
[27] Mattea L. Welch, C.M., Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, David A. Jaffray, Vulnerabilities of radiomic signature development: The need for safeguards. Radiotherapy & Oncology, 2018. 130: p. 2-9.
[28] A. Traverso, et al., Repeatability and Reproducibility of Radiomic Features: A Systematic Review, Int. J. Radiat. Oncol. Biol. Phys. 102 (4) (2018) 1143–1158.
[29] S. Sanduleanu, et al., Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score, Radiother. Oncol. 127 (3) (2018) 349–360.
[30] A. Stanzione, et al., Prostate MRI radiomics: A systematic review and radiomic quality score assessment, Eur. J. Radiol. 129 (2020), 109095.
[31] I. Fornacon-Wood, et al., Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype, Lung Cancer 146 (2020) 197–208.
[32] A. Zwanenburg, et al., The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology 295 (2) (2020) 328–338.
[33] M. Sinigaglia, et al., Imaging-guided precision medicine in glioblastoma patients treated with immune checkpoint modulators: research trend and future directions in the field of imaging biomarkers and artificial intelligence, EJNMMI Res 9 (1) (2019) 78.
[34] G. Lee, et al., Measurement variability in treatment response determination for non-small cell lung cancer: improvements using radiomics, J. Thorac. Imaging 34 (2) (2019) 103–115.
[35] R.B. Haynes, et al., Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey, BMJ 330 (7501) (2005) 1179.
[36] B.J. Ingui, M.A. Rogers, Searching for clinical prediction rules in MEDLINE, J. Am. Med. Inform. Assoc. 8 (4) (2001) 391–397.
[37] G.-J. Geersing, et al., Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews, PLoS ONE 7 (2) (2012).
[38] D. Moher, et al., PRISMA statement, Epidemiology 22 (1) (2011) 128.
[39] P. Whiting, et al., The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews, BMC Med. Res. Method. 3 (1) (2003) 25.
[40] C. Tang, et al., Development of an Immune-Pathology Informed Radiomics Model for Non-Small Cell Lung Cancer, Sci. Rep. 8 (1) (2018) 1922.
[41] I. Tunali, et al., Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report, Lung Cancer 129 (2019) 75–79.
[42] J. Yoon, et al., Utility of CT radiomics for prediction of PD-L1 expression in advanced lung adenocarcinomas, Thorac Cancer 11 (4) (2020) 993–1004.

[43] R.R. Colen, et al., Radiomics to predict immunotherapy-induced pneumonitis: proof of concept, Invest. New Drugs 36 (4) (2018) 601–607.
[44] R. Sun, et al., A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study, Lancet Oncol. 19 (9) (2018) 1180–1191.
[45] A. Schraag, et al., Baseline clinical and imaging predictors of treatment response and overall survival of patients with metastatic melanoma undergoing immunotherapy, Eur. J. Radiol. 121 (2019), 108688.
[46] J.E. Park, 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. 30 (1) (2020) 523–536.
[47] S. Rizzo, et al., Radiomics: the facts and the challenges of image analysis, Eur. Radiol. Exp. 2 (1) (2018) 36.