nature reviews clinical oncology

https://doi.org/10.1038/s41571-022-00707-0

Review article

Check for updates

Criteria for the translation of radiomics into clinically useful tests

Erich P. Huang[1]✉, James P. B. O'Connor[2], Lisa M. McShane[1], Maryellen L. Giger[3], Philippe Lambin[4], Paul E. Kinahan[5], Eliot L. Siegel[6] & Lalitha K. Shankar[1]

Abstract

Computer-extracted tumour characteristics have been incorporated into medical imaging computer-aided diagnosis (CAD) algorithms for decades. With the advent of radiomics, an extension of CAD involving high-throughput computer-extracted quantitative characterization of healthy or pathological structures and processes as captured by medical imaging, interest in such computer-extracted measurements has increased substantially. However, despite the thousands of radiomic studies, the number of settings in which radiomics has been successfully translated into a clinically useful tool or has obtained FDA clearance is comparatively small. This relative dearth might be attributable to factors such as the varying imaging and radiomic feature extraction protocols used from study to study, the numerous potential pitfalls in the analysis of radiomic data, and the lack of studies showing that acting upon a radiomic-based tool leads to a favourable benefit–risk balance for the patient. Several guidelines on specific aspects of radiomic data acquisition and analysis are already available, although a similar roadmap for the overall process of translating radiomics into tools that can be used in clinical care is needed. Herein, we provide 16 criteria for the effective execution of this process in the hopes that they will guide the development of more clinically useful radiomic tests in the future.

Sections

[1]Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, MD, USA. [2]Division of Radiotherapy and Imaging, Institute of Cancer Research, London, UK. [3]Department of Radiology, University of Chicago, Chicago, IL, USA. [4]Department of Precision Medicine, Maastricht University, Maastricht, Netherlands. [5]Department of Radiology, University of Washington, Seattle, WA, USA. [6]Department of Diagnostic Radiology, University of Maryland, Baltimore, MD, USA. ✉e-mail: erich.huang@nih.gov

Key points

- Despite tens of thousands of radiomic studies, the number of settings in which radiomics is used to guide clinical decision-making is limited, in part owing to a lack of standardization of the radiomic measurement extraction processes and the lack of evidence demonstrating adequate clinical validity and utility.
- Processes to acquire and process source images and extract radiomic measurements should be established and harmonized.
- A radiomic model should be tested on external data not used for its development or, if no such dataset is available, tested using proper internal validation techniques.
- Model outputs should be shown to guide disease management decisions in a way that leads to a favourable risk–benefit balance for patients.
- Clinical performance should be assessed periodically in its intended clinical setting (task and population) after model lockdown.
- A list of 16 criteria for the optimal development of a radiomic test has been compiled herein and should hopefully guide the implementation of future radiomic analyses.

Introduction

For decades, computer-aided diagnosis (CAD) algorithms have made use of computer-extracted tumour characteristics for improved disease detection and diagnosis, treatment planning, and follow-up[1], with some particularly notable developments in breast and lung cancer screening[2,3]. More recently, radiomics, involving high-throughput computer-extracted quantitative characterization of healthy or pathological structures and processes as captured by in vivo medical imaging, has emerged as an extension of CAD[4]. Similar to other 'omics' technologies, the extraction of such large quantities of information from images obtained during standard clinical workflows enables extensive tumour characterization and facilitates assessments of both within-tumour and between-tumour heterogeneity and longitudinal changes[1]. Interest in both CAD and radiomics (two terms that are occasionally used interchangeably) has increased substantially within the past two decades; a PubMed search for "(computer-aided diagnosis) OR CAD OR radiomic OR radiomics AND (cancer OR tumours OR tumours)" yields over 44,000 publications since 1967, over 85% of which are from 2005 onwards (Fig. 1).

Similar to CAD, radiomics can assist with clinical decision-making. Radiomic features, namely measurements extracted from medical images (currently usually CT, MRI or digital radiography), are combined with data on clinical characteristics and from other omics analyses to detect disease, predict the likelihood of death, disease progression and/or recurrence by a specific time point, evaluate response to therapy or identify an appropriate course of treatment. The ultimate goal of radiomic analyses should be the development of a test, defined by the FDA–NIH Biomarker Working Group as a system comprising materials for measurement, procedures for measurement, and methods or criteria for interpretation[5], that can be used to guide medical decision-making as in disease diagnosis and management.

Despite a dramatic increase in research output over the past two decades (Fig. 1), the vast majority of radiomic studies have not yet led to clinically useful tests. Across all medical indications, 343 artificial intelligence and machine learning-based tests currently have FDA clearance, only a small proportion of which are based on radiomics[6]. This lack of clinical translation might be attributable to several factors. The vast majority of radiomic studies assess correlations between certain radiomic features and a biological or clinical end point of interest; therefore, the added value of the radiomic test (such as improved clinical performance or reduced invasiveness) is often neglected as is clinical utility, namely that acting upon the information provided leads to a favourable benefit–risk balance for the patient. Additionally, as established in the statistical and machine learning literature, analyses of high-throughput data, such as those obtained using radiomics, are fraught with potential issues, including insufficient data for development and validation and improper application of statistical methodology for the specific purpose of the test. Furthermore, different studies have used widely varying protocols for image acquisition and feature extraction. Several studies have shown the effects of differences in data acquisition, image reconstruction and image post-processing on downstream analyses; different software platforms or even different versions of the same software can produce widely varying results regarding the strength and direction of the associations between features and outcomes[7].

Existing guidelines on the acquisition and analysis of radiomic data include a radiomic quality score to evaluate the completeness and appropriateness of such an analysis[8], computational procedures for commonly used types of features[9], and protocols for image acquisition, feature extraction and statistical analysis[10,11]. However, radiomics would also benefit from a roadmap for the entire process of translating radiomic data into clinically useful tools for guiding clinical care, encompassing not only recommendations for image acquisition and processing, feature extraction, and statistical analysis but also aspects such as test lockdown and demonstrating clinical utility. Such a roadmap has yet to be published for radiomics, although similar criteria and guidelines have been compiled for other omics technologies[12].

Herein, we present a 16-point list of criteria for the translation of radiomics into clinically useful tests. These criteria (Box 1) were developed by radiologists, physicists and statisticians with extensive experience with radiomics and other omics technologies, and are based on analogous recommendations developed for other omics technologies[12]. These criteria are also adapted to accommodate issues that are unique to radiomics, such as vendor-driven changes in imaging technology and software and the dynamic nature of certain models, and are intended to help researchers to navigate the translation process and catalyse an increase in the number of clinically useful radiomic tests.

Clinical application

Prior to any formal development and validation, the intended clinical use of the radiomic test and the target population should be established (criterion 1). The use of the test in clinical care should be expected to guide disease assessment and management decisions in a way that leads to a favourable benefit–risk tradeoff and offers advantages over other tests designed to serve the target population in the same role (criterion 2). The intended clinical use will have important implications for the subsequent stages of development and validation, including which features to extract from the imaging data, the optimal imaging time points and the design of the clinical trial to directly assess the performance of the test in its intended role.

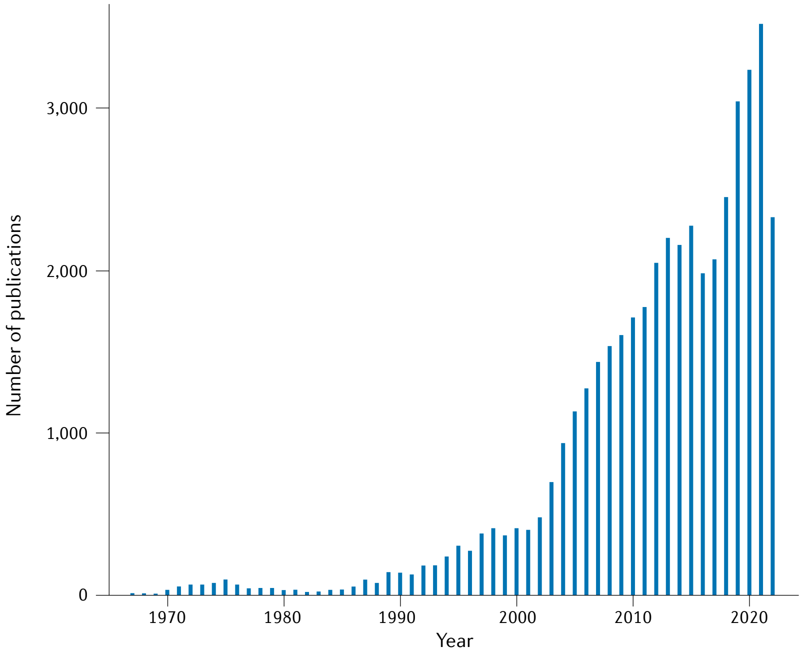

Fig. 1 | Number of publications per year since 1967. A PubMed search of "(computer-aided diagnosis) OR CAD OR radiomic OR radiomics AND (cancer OR tumours OR tumours)" was performed. The number of items published each year is presented as of 20 September 2022.

Criterion 1: intended role and target population

Radiomics is often used for either screening or cancer diagnosis. For example, MRI radiomics is useful for the diagnosis of breast abnormalities[13] and CT radiomics for the detection of lesions in various organs, including lungs, brain and prostate[14]. The use of radiomics in prognostication, namely predicting the clinical outcomes of patients undergoing standard therapy, is an area of increasing research interest[15]; for example, CT-based radiomics might be a useful method of predicting the outcomes of patients with head and neck squamous cell carcinomas or non-small-cell lung cancer receiving standard-of-care therapies[16]. Radiomic tests can also be used for treatment selection, namely as assays designed to indicate benefit, or lack thereof, from a specific class of therapies; for example, a model of oestrogen receptor expression based on tumour size, shape and entropy features on dynamic contrast-enhanced MRI (DCE-MRI) has been developed to inform treatment selection for patients with breast cancer[17]. Radiomic tests might also be used to assess response to treatment and monitor disease status[18–20].

Roles in which radiomic tests could serve have been summarized comprehensively elsewhere[21]. In certain scenarios, the same radiomic test can have more than one role; for example, the aforementioned model of oestrogen receptor expression might also be useful for prognostication[17]. However, 'off label' use of radiomic tests, namely application in a role other than the one for which the test has been shown to be clinically useful, is discouraged. The criteria for clinical performance depend strongly on the intended role (see criterion 14) as is typical in the regulatory clearance and approval processes applied to both new drugs and medical devices. Diagnostic radiomic tests should have an adequate level of accuracy in detecting disease. Prognostic radiomic tests should have an adequate ability to predict death, disease recurrence or progression depending on the intended role of the test. Tests designed for therapy selection should also be sufficient to predict outcomes, such as death or disease progression, in patients receiving the therapy of interest. If the goal is to guide the choice between one treatment and a designated alternative approach, the outcomes of patients receiving each therapy need to be studied. However, if the predictive goal is merely to identify those patients who are most likely to respond to a particular therapy, then the test should have adequate ability to predict either a response or a level of expression of an established predictive biomarker sufficient to indicate a response to the treatment of interest. The translation process outlined in this Review should therefore be applied for each role in which a specific radiomic test is likely to be useful.

Aspects of the target population to specify include those pertaining to disease characteristics (such as primary tumour types and grades, disease stage, molecular subtypes, risk groups and receptor expression status) and treatment history. A radiomic test might also be useful in multiple target populations; the test based on the model described by Aerts et al.[16], for example, might be useful for predicting the outcomes of patients with head and neck cancer or non-small-cell lung cancer receiving standard-of-care therapies. However, researchers are encouraged not to assume, without appropriate evidence, that the utility of a radiomic test extends across target populations because the technical performance of the imaging device and feature extraction software and the clinical performance of the test might not be consistent across different populations.

Criterion 2: patient benefit from use of the test in clinical care

The benefit of using a radiomic test should be clearly specified in the context of available treatments for the target population and access to other tests serving similar roles. A radiomic test might be used to stratify patients to optimize the choice of therapy for each individual, thus sparing patients from receiving ineffective or unnecessary treatments. A predictive test designed to guide treatment selection might differentiate between patients who are likely to derive clinical benefit (such as a longer median progression-free survival (PFS) or

Box 1

The 16 criteria for the translation of radiomics into a clinically useful tool

Clinical use of the test

- Criterion 1: the intended role of the radiomic test in clinical decision-making and the target population is clearly specified.
- Criterion 2: use of the test in clinical care is expected to lead to patient benefit beyond that resulting from a standard-of-care testing procedure.

Imaging and feature extraction

- Criterion 3: protocols for image acquisition, processing and normalization are standardized to optimize image quality and minimize variability owing to factors such as device or operator.
- Criterion 4: list of quantities to extract from the images and computational procedures to do so are specified.
- Criterion 5: the technical validity of feature measurements is adequately robust.
- Criterion 6: procedures to correct for technical artefacts in the feature measurements are in place.

Model development and validation

- Criterion 7: images, outcomes and other relevant data are obtained for patients in the target population through prospective clinical trials or retrospectively via completed studies, databases or image repositories.
- Criterion 8: the radiomic model is developed with guards against overfitting.
- Criterion 9: the ability of the model to predict an end point of interest is shown to be sufficiently robust using proper model validation techniques.
- Criterion 10: imaging, feature extraction, data preprocessing and computational procedures are fully specified and locked down to the greatest extent possible.
- Criterion 11: each possible value of the test output is associated with an unambiguous interpretation with regard to clinical management decisions.
- Criterion 12: the reproducibility of the test output with regard to variables, such as imaging centre or operator, is sufficiently high.
- Criterion 13: measures are in place to monitor drift, permit lockdown relaxation as appropriate and re-evaluate test performance as needed.

Justifying the use of the test in clinical care

- Criterion 14: the test outputs are shown to be clinically valid.
- Criterion 15: prospective or prospective–retrospective studies are conducted to directly evaluate the performance of the radiomic test in the context of its intended use.
- Criterion 16: the patient benefit resulting from acting upon the result of the radiomic test is sufficiently substantial to warrant its use in clinical care.

overall survival duration) from a specific therapy or class of therapies and those that will not. A prognostic test could identify patients with particularly poor outcomes on standard-of-care therapy who might consider a more intensive regimen; however, such a test will probably only be useful if a suitable, alternative treatment is available[22]. Moreover, a radiomic test might help to direct clinical management in a way that treatment-related toxicities, including financial toxicities, are reduced; prognostic tests might also identify patients whose outcomes on standard well-tolerated regimens are so good that they need not consider additional highly aggressive or toxic treatments or may consider treatment de-escalation.

The decision to use a radiomic test over other tests addressing the same clinical problem should be supported by a compelling reason. The radiomic test could have superior clinical performance to a standard test serving in the same role. The radiomic test might be able to identify underlying characteristics that cannot be detected as easily using other means; for example, assessing intratumour and intertumour heterogeneity of oestrogen receptor expression might be much less difficult when using radiomic tests compared with immunohistochemistry assays. Alternatively, the radiomic test might have a similar level of clinical performance but reduced invasiveness (such as biopsy avoidance), a reduced financial burden, greater convenience, or a reduction of one or more associated risks (potential harms, discomforts or exposures inherent to the testing procedure).

Imaging and feature extraction

Standard operating procedures for imaging, including protocols for the administration of contrast or imaging agents, specifications for image acquisition, procedures for image processing, and the timing of the scans should be in place (criterion 3) as should those for feature extraction, including a list of quantities to compute from the imaging data, segmentation algorithms, and computational algorithms and software to compute these quantities (criterion 4). The resulting feature measurements should also have been shown to have adequate technical validity (criterion 5). In most cases, this would entail each feature exhibiting strong repeatability and reproducibility or, if feasible, robust agreement with a standardized reference measurement of the underlying characteristic. A procedure to correct feature measurements for technical artefacts (the effects of factors such as imaging centre, device, operator or device-calibration settings on the distribution of the feature measurements) should also have been developed (criterion 6).

Criterion 3: standard operating procedures for image acquisition and processing

Image acquisition parameters should be specified in order to optimize image quality (for example, by keeping imaging noise to an acceptably low level or ensuring that the spatial, contrast and/or temporal resolution is adequate) and should be standardized to maximize reproducibility across imaging centres, devices and operators. Numerous studies have demonstrated the strong dependency of the resulting feature measurements on the imaging protocol[23,24]. Standard operating procedures for image acquisition could be based on established imaging guidelines such as those provided by the American College of Radiology[25], the Society of Nuclear Medicine and Molecular Imaging[26], the European Association of Nuclear Medicine[27] or the Quantitative Imaging Biomarker Alliance[28].

Image acquisition protocols will depend on the intended use as well as on the imaging modality and features that will be extracted. If the radiomic test is intended for diagnosis, spatial resolution will be an important consideration[29]. In theory, tests involving the analysis of morphological features will depend more on spatial resolution[30], whereas kinetic features, such as those derived from fast DCE-MRI, will depend more on temporal resolution[31]. In practice, perfect standardization is infeasible as is optimization of the protocol with respect to all the features to be extracted. The optimal resolution in DCE-MRI for breast cancer diagnosis often involves a compromise between spatial and temporal parameters to obtain measurements of morphological and kinetic features with adequate technical validity[32]. Furthermore, image acquisition protocols, particularly those applied to standard-of-care imaging approaches, are often determined in an ad hoc manner.

The time points at which patients undergo imaging should also be specific to the intended use. Radiomic tests intended for treatment selection will involve scans obtained prior to intervention. Those intended for response assessment will involve scans obtained not only prior to the intervention but also at specified time points during and after therapy (often termed 'delta-radiomics'[33]). The timing of response assessment can vary substantially; for example, radiomic tests designed to measure early metabolic response could involve imaging at baseline and then at a certain number of days to weeks following the initiation of treatment[34] whereas assessment of the effects of certain classes of therapies, such as anti-vascular agents, might occur in a timescale of hours to days[31].

Standard operating procedures should include processes designed to normalize the intensity values of images obtained from different patients and from the same patient. Normalization techniques include image resampling with filtering[35], normalizing voxel intensity values relative to a histogram or global and local intensities on a reference image[36,37], or harmonizing across different scans obtained from different populations or acquisition sites[38]. For certain features (such as second-order textural features), discretization through methods such as grey-level resampling and histogram binning is also needed[11,39]. Although the grey-level and standardized uptake value discretization methods used vary from centre to centre, these values can be normalized relative to a reference set of measurements[40,41]. Alternatively, standardized image preprocessing methods can be applied[42]. A comprehensive summary of imaging harmonization methods is provided elsewhere[43].

Criterion 4: standard operating procedures for feature extraction

Prior to formal test development, a list of quantities that will be extracted from the imaging data should be established. Traditionally, radiomic features are human-engineered and are extracted through delineation of the tumour from surrounding tissues using manual, semi-automated or fully automated segmentation[44–46] followed by application of pre-specified computational procedures to the voxel data within the region of interest[10]. Human-engineered features include those quantifying size (tumour dimension), shape (3D geometry), morphology (margin characteristics), enhancement texture (the extent of heterogeneity within the texture of the tumour and/or contrast uptake), quantifications of kinetic curves (shape of the curve and quantifications of the physiological process of uptake and washout of the contrast agent) and enhancement-variance kinetics (such as the time course of spatial variance of enhancement within the tumour)[47–50].

Extraction of such features will typically involve conversion and harmonization of the imaging data (criterion 6), post-processing (such as interpolation to cubic voxels, denoising, and correction of intensity and partial volume effects), image segmentation, region-of-interest extraction, and feature computation[9]. Existing guidelines and recommendations can serve as a starting point for the development of a standard operating procedure for feature extraction but will often require adaptation to suit both the target population and the imaging modality[51].

Alternatively, features of interest can be computer learned, namely extracted by direct application of computer algorithms to voxel data without the need for human intervention such as those computed using deep learning networks[52,53]. In this approach, a deep learning network can be applied to the voxel-level data and the last layer of the underlying convolutional neural network is taken as a set of features, similar to those used by Li et al. to predict *IDH1* mutation status in patients with low-grade gliomas[54]. An illustration of the differences between such features and human-engineered ones is provided in Fig. 2. Computer-learned features have been considered in conjunction with operator-dependent features[55] or even as a replacement. Such features are often less transparent in their computation and less interpretable; nonetheless, they might capture information that human-engineered features cannot, often resulting in more reproducible feature extraction and models with improved performance[54]. Fully automated extraction of such features enables the processing and computation of larger volumes of data with reductions in the variability of test output values owing to the elimination of human error during processes such as manual delineation and segmentation[52].

Criterion 5: technical validity of the feature measurements

Adequate technical validity typically entails assessing the repeatability and reproducibility of the feature measurements. Repeatability describes the precision when a specific imaging and feature extraction standard operating procedure are applied multiple times to the same patient at the same centre by the same operators within a short period of time. Reproducibility describes the precision of repeat measurements when factors such as imaging centre and operator are allowed to vary[56–58]. Study designs and the statistical methodology for studies assessing repeatability and reproducibility have been summarized in detail elsewhere[59]. Strong technical validity is important for model development and the establishment of the clinical utility of a radiomic test given that poor feature reproducibility, as mentioned previously, can produce widely varying results regarding the strength and direction of the association between features and outcomes[7] and result in models with insufficient levels of performance[60].

Ideally, repeatability and reproducibility would be assessed using clinical data. In such clinical studies, patients undergo repeat scans with the feature extraction standard operating procedure then applied to each image. Such studies have been conducted[61,62], although they are often difficult in practice as patients can be reluctant to participate owing to a lack of direct benefit, the inconvenience of undergoing multiple scans and, with certain techniques, additional exposure to contrast agents or ionizing radiation. An alternative approach involves different operators extracting features from the same set of images,

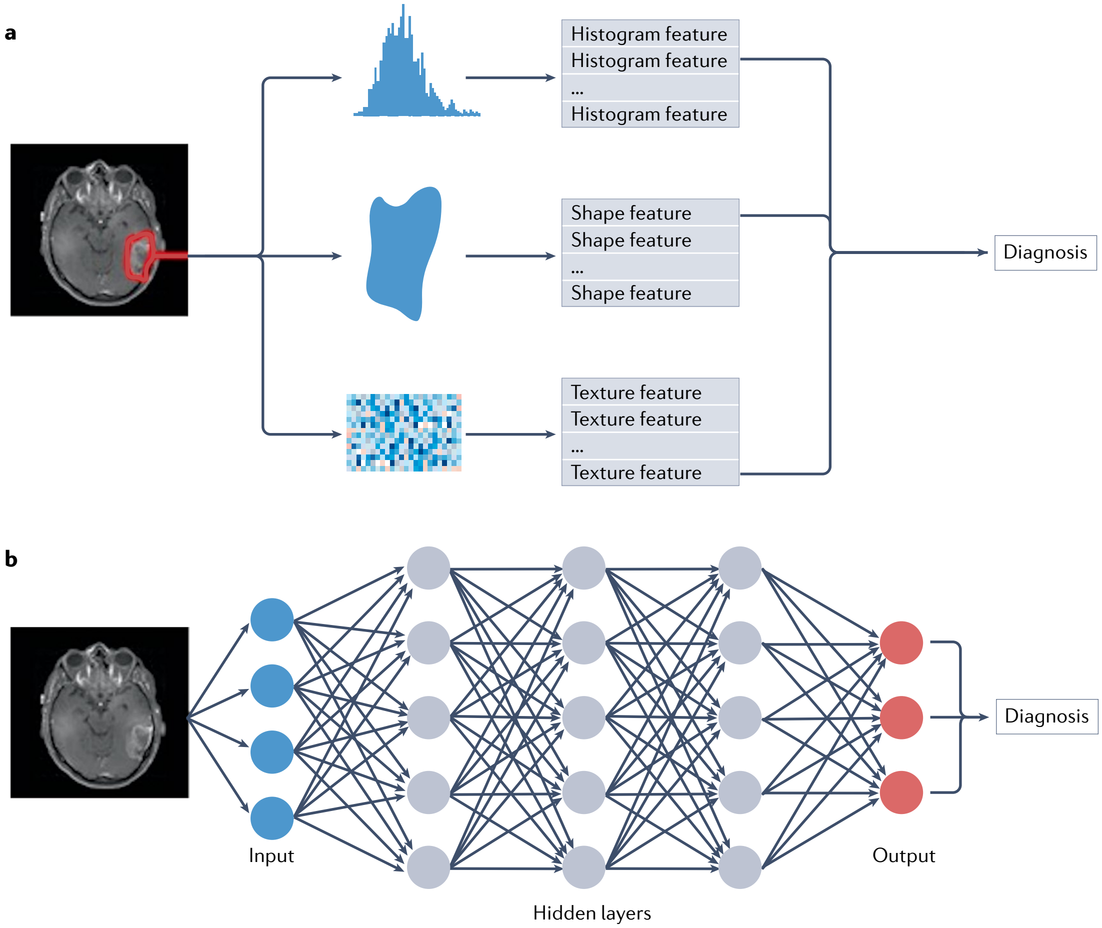

Fig. 2 | Types of radiomic analysis. a, Analyses using human-engineered features. Different types of features (such as histogram, shape or texture) are extracted from the images according to a pre-specified computational procedure. Variable selection techniques are used to identify which of these features are important in diagnosing a medical condition. The values of these selected variables are combined into a model to produce a diagnosis. **b**, Analyses using machine learning and artificial intelligence algorithms. The voxel-level data are fed into a convolutional neural network consisting of multiple hidden layers whose output is used to produce a diagnosis.

possibly at different centres; however, this approach, although also feasible as a retrospective method, only enables the assessment of variability attributed to the feature extraction process[51,56,58].

As an alternative approach, some components of technical validation can be conducted using in vitro or in silico phantoms, simulated digital reference images or synthetic data such as those produced by generative adversarial network systems[40,63]. However, conclusions based on data obtained using phantoms and digital reference images will be overly optimistic regarding their technical validity given that they cannot fully capture the complexity of actual patients. Several authors have provided recommendations on the minimum technical validity requirements of phantoms and digital reference images[59,64].

Technical validity can also be assessed using the level of agreement between feature measurements and certain comparator quantities (for example, with a measurement of the underlying biological characteristic according to an independent in vitro assay), bias (the mean difference between the measurement and the true value of the characteristic being measured), and the linearity of the relationship between the feature measurement and the true value[59]. However, assessing agreement is often not possible for computer-learned features owing to a lack of an appropriate biological correlate. Assessing bias or the relationship between the measurement and the true value of the feature being measured is generally only possible with phantoms and digital reference images.

Repeatability and reproducibility can be used as screening criteria to immediately eliminate features with poor technical validity from further consideration for inclusion in the model. Filtering out features in this manner has been shown to improve the level of power in settings with large numbers of features, of which only a small proportion are associated with the outcome of interest[65]. Such filtering must be done solely on the basis of technical validity and must not use outcome data that will also be used to assess performance of the model under development (criterion 9).

Technical validity criteria are much less well developed for computer-learned features such as those described by Li et al.[54]; their methodology produced 16,384 dimensional descriptors arranged in 128 × 64 × 2 arrays, for which applying the technical validity assessment methods described above is clearly not feasible. Regardless of the type of feature used, researchers are encouraged to assess the technical validity of the output of any radiomic models based on these features (criterion 12).

Criterion 6: feature measurement correction for technical artefacts

Technical artefacts, namely the effects of factors related to variables such as imaging centre, operator and/or device configurations on the distributions of the feature measurements, can potentially confound the results of subsequent radiomic analyses. For example, a feature with no association with survival might seem to predict outcome if patients who undergo imaging in one location have substantially better outcomes than those undergoing imaging in another centre and if the median feature measurement differs between the two sites owing to variations in image acquisition and processing. Thus, procedures designed to correct the variations in feature measurements created by such factors should be established prior to the development and validation of a radiomic model.

In addition to the image normalization methods described previously (criterion 3), the feature measurements themselves can also be standardized following extraction. These measurements can be normalized relative to a reference set of measurements[66] or according to a harmonization model[67,68], similar to the approaches used in other omics settings. Features strongly associated with variation from these technical artefacts might then be removed from consideration before model construction[69,70].

Model development and validation

Patient-level data, including images, outcomes, standard clinical variables, measurements of in vitro biomarkers and other relevant data, should be obtained from the target population; these data can be obtained prospectively or retrospectively from already completed studies, imaging repositories or health-care databases (criterion 7). A radiomic model should be developed using appropriate statistical or machine learning techniques incorporating safeguards designed to avoid overfitting (criterion 8). The performance of a model in predicting an end point of interest must be shown to be adequately robust using proper model validation techniques (criterion 9). By the end of its development, all aspects of the radiomic test, including the feature preprocessing steps, mechanisms of imputing missing data, the underlying computational procedures, any cut points in the feature measurements themselves and/or the model outputs, must be fully specified (criterion 10). Each possible output value of the test is then linked to an unambiguous interpretation with regard to clinical care (criterion 11) and the reproducibility of these outputs should be shown to be sufficiently strong (criterion 12). Processes designed to address drift in the performance of the radiomic test, which refers to changes arising from factors such as the evolution of image acquisition and processing protocols and feature extraction procedures over time, software upgrades and obsolescence, and replacement of devices with newer models, should be established, including monitoring processes and procedures to perform further technical validation and model adjustment as necessary (criterion 13).

Criterion 7: imaging, outcome and other relevant data from the target population

Data on the performance of radiomic analyses can be acquired prospectively, most often as part of a clearly stated secondary objective in a phase II or phase III trial involving the target population, with standard operating procedures for image acquisition and processing at the desired time points and a feature extraction protocol, guided by the points described previously, written into the protocol. Alternatively, data can be acquired retrospectively from imaging data repositories, health-care databases, or datasets from completed clinical trials, subject to inclusion and/or exclusion criteria involving image acquisition and processing protocols, image quality, and the availability of images acquired at the relevant time points. For example, The Cancer Genome Atlas Breast Imaging Research Group identified patients from The Cancer Imaging Archive repository[71,72] for whom gene expression analysis and pretreatment standard-of-care breast MRIs obtained with 1.5 Tesla GE Medical Systems devices were available[17,18,73]. Any clinical data to be obtained should be matched with the images via unique patient ID numbers.

Sample sizes should be determined according to factors such as the number of events (patients with disease versus without, or observed number of deaths), the type of model to be fitted to the data, the expected strength of the relationship between the features and the outcome, the desired standard error of the performance metric, the variance of the model outputs and their concordance with observed event probabilities[74–76]. Logistic and Cox regression models constructed using data from too few patients often have lower performance relative to models constructed using larger sample sizes[60]. Deep learning classifiers can require data from thousands of patients per class owing to their complexity (in preprint[77]); however, dataset sizes can be reduced with the use of transfer learning through feature extraction or fine-tuning methods[78]. Smaller numbers of patients can be used for model fitting if the relationship between the features and the outcome is particularly strong. Notwithstanding, sample sizes are often constrained by the amount of data available from the completed studies, image repositories or databases from which they were acquired or, if the radiomic study is a secondary objective of a clinical trial assessing a therapeutic intervention, by the number of patients required to meet the primary objective, which will often be much smaller than what is needed for the radiomic analysis.

Ideally, prospective studies should involve multiple centres and retrospectively acquired data should be obtained from multiple studies or repositories and then combined. Using multiple imaging centres, as opposed to a single one, not only facilitates more rapid accrual of data and accumulation of a sufficient number of patients for reliable statistical modelling and validation but can also result in the acquisition of data from a broader population. However, this approach comes with the risk of introducing technical artefacts into the data that will need to be corrected prior to model development and validation (criterion 6).

Criterion 8: development of the radiomic model with guards against overfitting

The range of model-fitting techniques proposed in the statistical and machine learning literature has been described in detail elsewhere[79]. The literature suggests that no single model-fitting technique is uniformly superior to any other[80] although, regardless of the approach used, care should always be taken to avoid overfitting, that is, fitting an overly complex model to noise in the data and thus producing a model that is only poorly predictive when applied to completely new data. Overfitting risk is high when using more complex models, such as those based on neural networks[81] or non-parametric regression, as opposed to simpler ones such as those based on logistic or Cox regression. These simpler models have also been shown to often perform as well, if not better, than their more complex counterparts, especially when the number of variables is large and the underlying relationship between the radiomic features and the end point is neither strong nor complex[45,60,82,83].

Inclusion of too many features in the model, which can be viewed as another form of model complexity, is another common cause of overfitting. Models based on high-dimensional data, such as those typically encountered in radiomic settings, are particularly prone to this issue. Eliminating any features with subpar levels of technical validity (such as poor reproducibility) or those associated with batch processing before any formal model development takes place (criteria 5 and 6) might reduce the likelihood of overfitting as will the use of variable selection techniques. Several authors have described common variable selection techniques in greater detail elsewhere[79,84]. Of note, many of these techniques require the selection of a tuning parameter controlling the stringency of the inclusion criteria for variables in the model (such as a *P* value below which univariate associations between individual features and the outcome must lie to be included, the number of variables to be included, or regularization parameters in LASSO regression techniques[85]). The optimal tuning parameter value is typically identified using the data (see criterion 9).

Criterion 9: model validation

Once a model has been developed, with mitigation against possible overfitting, the model should then be shown to be capable of predicting an end point of interest, be it a clinical event or state or a biological characteristic, with a sufficient level of accuracy. Robust model performance does not necessarily imply usefulness in guiding medical decision-making; for example, as mentioned previously, a radiomic test with a high level of diagnostic accuracy or a robust ability to predict treatment response or an end point of interest will not be clinically useful if the improvement in clinical performance is not substantial enough to justify its use over standard-of-care diagnostic workups. The broad principles described in this subsection, as well as those regarding lockdown, clinical validity and clinical utility in subsequent sections, apply to both more traditional human-engineered features and computer-learned features.

The area under the receiver operating characteristic curve[86] (AUC) of the model outputs or their sensitivity and specificity can be used to quantify the ability of the model to discriminate between patients with a specific health condition from those without. A related metric to the AUC is the c-index[87], which quantifies the ability of the model to predict survival (the probability that among two randomly chosen patients, the one with the higher model output has the shorter survival time). Additionally, assessments of model performance should include calibration, namely the concordance between the predicted and expected probabilities of an event of interest[88–90]. Calibration curves, namely plots of the observed frequencies of the event versus predicted probabilities, are also used to examine whether the model predictions are consistently either too high or too low[91]. As emphasized during the discussion of criterion 1, the most appropriate performance metric will depend on the intended use of the radiomic test.

Ideally, model validation should be accomplished by applying the newly developed model, without any alterations to any aspect, to a completely external dataset that was not used in any part of the model development process. External data should be acquired from patients in the target population from whom imaging data were obtained under similar imaging, processing and feature extraction protocols to the data used in model development. Variations in imaging centre, operating personnel, scan acquisition date, and certain methods of imaging and feature extraction (such as device and software version) between the training and validation datasets might be permitted to enable evaluation of the robustness of the model to variability in these factors.

However, adequate external validation is not always performed, primarily owing to the logistical challenges associated with accessing data from an independent cohort. In our experience, the performance of the model is often assessed through internal validation, namely the use of a single dataset for both model development and evaluation. Internal validation involves carefully splitting or subsampling the data to avoid overlap with the data used to develop the model (the training set) and those used to evaluate the performance of the model (the validation set). Internal validation can provide reasonable estimates of the predictive accuracy of the radiomic model, although results obtained in this way might not necessarily be generalizable to completely new data. If model development and internal validation were performed on data that were obtained using obsolete image acquisition and processing protocols or that involved a cohort that was not completely representative of the entire target population (such as patients from a location at which a disproportionate percentage had a poor prognosis), then the results will reflect performance in this setting; performance might be diminished in other settings such as those with updated image acquisition and processing protocols[12].

Internal validation methods include split-sample validation[92], cross-validation[93] or bootstrap validation[94]; these various techniques have been summarized in detail elsewhere[79,95]. Cross-validation is usually preferable to split-sample validation when only small sample sizes are available; the latter produces estimates of model performance that are often pessimistically biased (that is, estimates of model performance that are substantially lower than those obtained from external validation) when sample sizes are of about 200 or fewer individuals[60,96].

Appropriate internal validation requires the maintenance of strict separation of data used to specify any aspect of the model from those used to evaluate its performance. Any violation of this strict separation results in overly optimistic estimates of the performance[97,98]. In this regard, full resubstitution, in which the entire dataset is used for both development and validation of the same model, provides the most egregious example. Partial cross-validation, in which the entire dataset is used to select features based on their significant univariate association with outcome followed by cross-validation of the model using only this restricted feature set, is another variant of this inappropriate approach to validation. In a comprehensive review of internal validation approaches, data from simulation studies are presented indicating that, even in a scenario in which the variables have no relationship with an outcome, inappropriate internal validation techniques can still produce an AUC estimate of 0.7–0.8 (ref.[97]).

The selection of tuning parameters during model development (criterion 8) is yet another stage at which problems in model validation can arise. Often, for each candidate from a list of tuning parameter values, the model is fitted using the training set and then applied to the validation set to obtain a performance metric estimate. The tuning parameter value associated with the optimal performance metric estimate is then identified and this metric estimate is then reported. However, in this approach, some aspects of the model development (the identification of the tuning parameter) took place on data used to estimate the performance metric. Such approaches can lead to biased estimates of the performance metric[98]. Appropriate validation techniques for use when tuning parameter selection is also involved include a three-way split of the data into training, validation and test sets (the training and validation sets are used to identify the tuning parameter and fully specify the model, which is then applied to the test set to obtain a performance metric estimate)[92] or nested cross-validation[98].

Criterion 10: radiomic test lockdown

Once the model has been developed and shown to have reasonable predictive accuracy, all components of the test, as described in the Introduction of this Review, should be locked down. In radiomics, procedures for measurement will include both standard operating procedures for image acquisition and processing (criterion 3) as well as those for feature extraction (criterion 4) and calculation of model output. Outputs are then associated with specific clinical interpretations (criterion 11).

All computational aspects of the model (for example, the mathematical expression, including regression coefficients, weightings, cutoffs and any other parameters) should be locked down to the greatest extent possible. In situations in which concise model descriptions are not feasible, such as for those based on deep learning, the underlying computational algorithm and software platform should be closed to further changes and any crucial components, such as the random number generator seeds used to generate the model or the output, should be fixed. Interpretations of the inputs of the model (for example, the variables included in a logistic or Cox regression model involving human-engineered features) are often of interest to researchers as they can provide insights into the degree of importance of each feature in predicting an outcome. For computationally derived model inputs, such as features obtained using deep learning algorithms, methods to aid interpretability include visualizing the latent space discovered through the learning process, post hoc highlighting of the regions of the input images that the model labelled as important and visualization of features from different filters in the convolutional neural network[99].

The locked-down model could still be affected by any remaining biases inherent to the data on which it was fitted and validated (such as technical artefacts and distributions of radiomic feature values and outcomes that differ substantially from those of the target population). Allowing the model to evolve over time as new data become available will alleviate some of these effects (criterion 13).

Criterion 11: interpretation of test outputs

Models based on techniques such as support vector machines will produce outputs consisting of discrete categories[78], each of which can be linked to a specific clinical interpretation and decision. However, models constructed via most other techniques will produce a quantitative output such as the predicted probability of a specific event of interest. Binning these continuous outputs into a limited number of discrete categories might be desired for the purposes of interpretation and clinical decision-making. For example, a test output value that falls below a prescribed cutoff value might indicate a good prognosis and that additional treatment will not be needed and/or that the likelihood of a response to a treatment is high. Alternatively, a test output value above a prescribed cutoff could indicate a high risk of mortality and that the patient might survive longer on an alternative regimen.

Sometimes, these cutoffs are set arbitrarily to specific quantiles, such as the median, in order to define high-risk versus low-risk groups; however, this approach ignores associations with clinical outcomes. Cutoff optimization and comparisons of the outcomes of patients in each category defined by the cutoffs should be done on separate datasets so as not to violate the principle of separation of data used for model development from those used for validation. When cutoff optimization and outcome comparisons are done using the same data (for example, by applying various cutoffs to a dataset, computing the log-rank test *P* values of the resulting groups and choosing the cutoffs associated with the lowest *P* values), the risk of a type I error is increased[100]. To ensure the test can be applied to one patient at a time, cutoff values should be specified as absolute values rather than as percentiles that would need to be recalculated on the availability of new data.

Analytical approaches that consider the consequences of specific treatment decisions based on the test output have also been proposed as a method for cutoff selection. These methods aim to balance the risks (adverse consequences) of incorrect test results against the benefits (positive consequences) of correct test results. The risk–benefit balance can then be compared to that of the standard-of-care approach for the specific clinical indication or any other competing tests or to the use of no test at all. Such approaches include the decision curve analysis method[101]. This methodology has been applied to a radiomic study involving features obtained from preoperative CT images in conjunction with images from intraoperative frozen sections and clinical data to differentiate invasive lung adenocarcinomas from preinvasive lesions or minimally invasive adenocarcinomas[102].

Criterion 12: test output reproducibility

The reproducibility of the test outputs should be shown to be sufficiently robust to ensure that the radiomic test will produce similar results regardless of where it is performed or by whom. One approach involves having patients undergo repeat scans using an established standard operating procedure without interventions in between. Multiple operators, also possibly at different imaging centres, would then apply feature extraction according to standard operating procedures and the radiomic model to the repeat scans independently of one another. Finally, the model or algorithm underlying the test is applied to the images and feature data. Reproducibility metrics for individual features can also be considered at this stage.

This assessment of reproducibility encompasses variability potentially owing to all aspects of image acquisition and processing, feature extraction, and application of the model; however, this approach is rarely feasible in practice for reasons that include a lack of availability of repeat imaging in many scenarios and the unwillingness of many patients to undergo multiple scans within a short space of time. Alternatively, both the feature extraction process and the model can be applied repeatedly to the same set of images, possibly by different operators at different locations. This approach can be applied to retrospectively acquired data but can only produce an assessment of reproducibility that encompasses variability owing to feature extraction and application of the model (and not factors that influence raw data acquisition). If estimates of the repeatability and reproducibility of individual features are known, error propagation models and simulation approaches can be used to estimate the reproducibility of the test output[60].

Criterion 13: processes to address data and radiomic test drift

The computational procedures underlying most radiomic tests are likely to evolve over time. Imaging hardware and computational software are likely to be upgraded. Furthermore, the model itself could change after fitting to new data[102]. Monitoring for such changes in a way that enables their effects to be assessed should be in place. Certain changes might also require a return to previous steps in model development and validation. Changes not related to drift, such as application of the test in a different patient population or indication or the addition of new features, should necessitate a return to model development and validation and might also require the re-establishment of standard operating procedures for feature extraction with re-assessments of the technical validity of individual features.

Assessments of technical and clinical validity and clinical utility (criteria 14 and 15) should be performed periodically for tests for which the underlying computational procedure is expected to evolve over time. Changes to the standard operating procedures for image acquisition and processing or upgrades to the feature extraction software should be followed by assessments of the level of agreement between feature measurements obtained under the previous and the new versions. Researchers should proceed with the new versions of the standard operating procedures and software platforms based on the degree of agreement; however, empirical guidelines on what constitutes a sufficiently strong level of agreement are not available and are probably dependent on both the feature itself and the context. If this agreement in feature measurements is inadequate, the level of concordance between the test outputs computed using the two versions can be assessed (for example, by demonstrating that the mean squared difference of the outputs from the two versions is lower than some meaningful threshold). High concordance between the test outputs indicates that the two versions produce similar results and that the new one could therefore safely replace the previous one, although poor concordance might also reflect the superior clinical performance of the new version. In some scenarios, the model might need to be refitted; such changes can alter the significance and sometimes even the direction of the association of the features with an outcome of interest[7].

Justifying use in clinical care

Robust performance of the underlying model in predicting an end point of interest does not automatically mean that the test will be clinically useful (meaning that acting upon the results of the radiomic test leads to patient benefit via improved outcomes or quality of life, reduction in toxicity, invasiveness, risk of complications, financial burden, or the avoidance of ineffective or unnecessary treatments). After the radiomic model has been validated and the test has been locked down, its clinical validity, namely the ability of the outputs to provide information regarding the presence or absence of a condition or the risk of an event of interest[103] (for example, sensitivity, specificity, or positive and negative predictive values in detecting disease or the proportion of patients classified as low-risk who remain progression free at 5 years), should be assessed in the context of its intended use and clinical setting (criterion 14). The clinical utility of the test should then be assessed using a prospective study or an appropriately designed prospective–retrospective study, in which the performance of the test in its intended clinical setting is directly assessed (criterion 15) and the risk–benefit balance for the patient when acting upon the results of the radiomic test is shown to be sufficiently favourable to justify use in clinical care (criterion 16). Note that such scenarios often do not reflect the standalone performance of the radiomic model but rather how the test influences the end user (for example, the clinician) when they make clinical decisions with and without the test results as was often used in assessing CAD systems[104,105].

Criterion 14: clinical validity of the test

Clinical validation goes beyond model validation (criterion 9) in that the former involves the evaluation of model performance with greater specificity to the clinical setting and intended use. For example, model validation of a prognostic radiomic test might involve showing that the level of concordance between overall survival and model outputs is above some pre-specified and meaningful threshold. Clinical validation, meanwhile, might involve demonstrating that patients who have been classified in a low-risk category have a very high (>90%) 5-year PFS on a well-tolerated standard therapy regimen whereas those in other risk categories have substantially worse outcomes. This may suggest that patients in the low-risk category may potentially consider foregoing additional highly invasive or toxic treatments. Alternatively, showing clinical validity of a prognostic radiomic test might entail demonstrating that the association between test output and clinical outcome remains statistically significant even after adjusting for standard clinical or pathological variables with known prognostic value. The robustness of such a finding to the effects of potential confounders, such as variations in the operator of the feature extraction or the imaging centre in which the extraction and test were performed, should also be established. Different approaches have been summarized in detail elsewhere[21].

The radiomic test should be fully locked down and the data used to determine clinical validity should be independent from any data used in model development and validation. Such data could come from prospective clinical trials. For example, to estimate the 5-year PFS of patients with low-risk disease according to the radiomic test, such a cohort could receive standard-of-care therapy and comparisons of the outcomes of the different risk groups could be made after 5 years of follow-up monitoring. Alternatively, data might also be acquired retrospectively from completed clinical trials or imaging data repositories such as The Cancer Imaging Archive[71] or the sequestered commons from Medical Imaging and Data Resource Center[106], from which testing data can be drawn based on the clinical question and population of interest. Again, this approach assumes that imaging data for a sufficient number of patients from the target population were acquired using protocols similar to the previously established standard operating procedures (criterion 3).

Criterion 15: direct evaluation of performance of the test in its clinical use

The optimal design, end points and statistical analyses to assess the benefits of using a radiomic test to guide clinical disease management differ widely depending on the intended use of the test[21]. For example, for a radiomic test expected to outperform an in vitro prognostic assay currently in widespread use, patients whose treatment decisions were based on the radiomic test should be shown to have substantially improved outcomes compared to those of patients for whom clinical care was dictated by the in vitro assay.

Prospective studies have numerous desirable qualities, including enabling researchers to have full control over the features to measure, image acquisition and processing, the study design, and sample size. However, such studies are likely to be time consuming and costly, particularly for disease settings with already favourable outcomes that require a large sample size and/or lengthy follow-up duration to observe sufficient events (such as death, disease recurrence or progression) for adequate statistical power. Prospective–retrospective studies can reduce or even eliminate many of the delays and costs associated with image acquisition and follow-up assessments[107]. For prospective–retrospective studies, data from standard-of-care images, clinical outcomes and other data, such as standard clinical variables, are acquired from patients in completed clinical trials that satisfy the appropriate inclusion and/or exclusion criteria regarding the patient population, treatment approach, image acquisition and processing specifications, and

Glossary

Biomarker
A characteristic indicating non-pathological or pathological biological processes and/or an increased likelihood of a response to an exposure or intervention[5].

Clinical utility
The degree to which acting upon the results of the radiomic test leads to a favourable benefit–risk balance for the patient.

Clinical validity
The adequacy of the clinical performance of the radiomic test for its intended purpose.

Deep learning
A class of machine learning based on neural networks.

Model
A computational algorithm applied to extracted image features or voxel-level image data themselves.

Model outputs
The result of a computational algorithm applied to the extracted image features or voxel-level data themselves; a quantity to be used in guiding clinical management.

Model validation
Establishment of the ability of a model to predict an outcome of interest when applied to new data.

Neural network
A type of computational algorithm based on the operation of biological neural systems in animals that feeds the input (in this context, feature measurements or voxel-level data) through a series of nodes that perform mathematical operations on the outputs of preceding nodes to produce an output. In a convolutional neural network, these mathematical operations involve applying convolutional kernels to the outputs of preceding nodes.

Normalization
A process for adjusting the voxel intensity values of an image for differences resulting from variability in image acquisition and processing parameters.

Omics
The study of related sets of biological molecules in a comprehensive fashion with examples including genomics, transcriptomics, proteomics, metabolomics and epigenomics[109]. Radiomics naturally extends this definition to include quantification of radiological imaging features for the purposes of characterization and measurement of structure, function and interaction between biological molecules in a comprehensive and high-throughput manner.

Overfitting
The process of fitting an overly complex model to noise in the data, thus producing a model that is only poorly predictive when applied to completely new data.

Performance metric
A quantity indicating the ability of a model to predict an outcome of interest.

Phantoms
An object that is imaged to measure the technical performance of an imaging device.

Radiomic features
Quantities computed from voxel-level image data.

Radiomic test
A system comprising materials, methods and procedures for image acquisition, processing and feature extraction, and methods or criteria for interpretation of the image data for use in guiding clinical management.

Technical artefacts
The effects of factors, such as imaging centre, device, operator or device-calibration settings, on the distribution of the feature measurements.

Technical validity
The quality of the feature measurements in terms of their accuracy in assaying an underlying characteristic of interest or their variability when the feature extraction process is applied repeatedly to the same patient.

Test lockdown
Full specification of all image acquisition, processing and feature extraction procedures, all aspects of the underlying model, and interpretations of the output.

availability of the necessary images. Both the feature extraction and the test are applied prospectively. Similar to a prospective study, the radiomic test, the statistical analysis plan, sample size, level of power, and the inclusion and exclusion criteria should be fully specified in a protocol before the initiation of a prospective–retrospective study. Criteria for establishing clinical utility through prospective–retrospective studies for other omics approaches have already been published[12,107]. These criteria include the stipulation that two such studies must produce similar results, an approach that can also be adapted for radiomic tests. In silico clinical trials using patient-specific models to develop a simulated cohort might provide an alternative approach[108], although these simulated patients might not entirely reflect the complexities of real-life patients.

Criterion 16: benefit versus risk balance from use of the radiomic test

The benefit–risk balance associated with use of a radiomic test will encompass not only the risks and benefits associated with performing the test but also those associated with the clinical decisions directed by the test results. If the intended use of a test is to choose a therapy that provides superior clinical outcomes compared with other available options, then the improvement in clinical outcome should not only be statistically significant but also large enough in magnitude to justify use of the radiomic test. Alternatively, a favourable benefit–risk balance might emerge when use of the radiomic test leads to non-inferior outcomes while being associated with reduced risks, including those inherent in the standard testing procedure, or if the toxicities of unnecessary or ineffective treatment can be avoided. For example, even if the radiomic test leads to treatment decisions that are similar to those based on standard diagnostic workups, the former might nevertheless have clinical utility if the information it provides enables patients to undergo fewer subsequent scans or biopsies while still leading to similar outcomes.

Finally, a radiomic test does not have clinical utility if it separates patients into groups for which the outcomes are statistically different but the recommended clinical management would be the same. Even if one patient group has an inferior outcome on standard therapy, another treatment might be available that is more effective for that group.

Conclusions

The 16 recommended criteria provided herein aim to guide the translation of radiomic tests into clinically useful tools and are expected to be relevant across a range of imaging modalities and scenarios.

Many of these recommendations share common themes with other published guidelines for radiomics; adherence to these recommendations addresses many components of the radiomic quality score[4], for example.

The statistical considerations regarding model development and validation and the design of studies for the assessment of clinical utility have numerous parallels to those for in vitro test considerations[12,109]. Several components of our recommendations are based on these sources. However, some important and consequential differences specific to radiomics also merit consideration. Radiomic approaches increasingly utilize multiple machine learning and deep learning methods, which introduces new issues regarding standard operating procedures for feature extraction, test lockdown, machine learning interpretability, correlations with biology, regulatory considerations and assessments of analytical validity. These criteria are likely to further evolve in the future as researchers become aware of additional issues and as more radiomic models become locked down, validated and evaluated for clinical utility. We emphasize that these recommendations pertain to the conduct and analysis of radiomic studies and are not intended as reporting guidelines for radiomic and CAD studies in the vein of REMARK for tumour prognostic studies[110] or other reporting guidelines catalogued by the EQUATOR project[111]. However, some of these recommendations are expected to serve as the basis of such radiomic-specific reporting guidelines.

Radiomics is increasingly likely to involve full machine learning-based image analysis such as deep learning-based features or the application of artificial intelligence and machine learning algorithms directly to voxel-level data. Such a transformation, as mentioned before, is expected to eliminate much of the variability created by human error and improve model performance in many scenarios, although it will also benefit from integration with clinical information to better personalize the test result to each patient. For example, this type of test might be used not only to detect cancer but also to do so in the presence of additional comorbidities (for example, examining a renal finding in the presence of diabetes mellitus, chronic inflammatory processes and/or hypertension). The increased availability of different types of data should facilitate these types of improvements.

Published online: 28 November 2022

References

1. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
2. Giger, M. L. Update on the potential of computer-aided diagnosis for breast cancer. *Fut. Oncol.* **6**, 1–4 (2010).
3. Doi, K. Computer-aided diagnosis in medical imaging: historical review, current status, and future potential. *Comput. Med. Imaging Graph.* **31**, 198–211 (2007).
4. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
5. FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and other Tools) Resource* (Food and Drug Administration and National Institutes of Health, 2016).
6. FDA. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Devices* https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. (2022).
7. Fornacon-Wood, I. M. et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur. Radiol.* **30**, 6241–6250 (2020).
8. Radiomics. *Radiomics Quality Score – RQS 2.0* https://www.radiomics.world/rqs2 (2022).
9. Zwanenburg, A. et al. The image biomarker standardization initiative: standardized quantitative radiomics for high throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
10. Kumar, V. et al. Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).
11. Fournier, L. et al. Incorporating radiomics into clinical trials: expert consensus endorsed by the European society of radiology on considerations for data-driven compared to biologically driven quantitative biomarkers. *Eur. Radiol.* **31**, 6001–6012 (2021).
12. McShane, L. M. et al. Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration. *BMC Med.* **11**, 220 (2013).
13. Jiang, Y., Edwards, A. V. & Newstead, G. M. Artificial intelligence applied to breast MRI for improved diagnosis. *Radiology* **298**, 39–46 (2021).
14. Data Science Institute, American College of Radiology. *FDA Cleared AI Algorithms* https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms, (2022).
15. Clark, G. M. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. *Mol. Oncol.* **1**, 406–412 (2008).
16. Aerts, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
17. Li, H. et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *NPJ Breast Cancer* **2**, 16012 (2016).
18. Li, H. et al. MRI radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of gene assays of MammaPrint, Oncotype DX, and PAM50. *Radiology* **281**, 382–391 (2016).
19. Cha, K. H. et al. Bladder cancer treatment response assessment in CT using radiomics with deep learning. *Nat. Sci. Rep.* **7**, 8738 (2017).
20. Drukker, K. et al. Most-enhancing tumor volume by mri radiomics predicts recurrence-free survival "Early On" in neoadjuvant treatment of breast cancer. *Cancer Imaging* **18**, 12 (2018).
21. Huang, E. P., Lin, F. I. & Shankar, L. K. Beyond correlations, sensitivities, and specificities: a roadmap for demonstrating utility of advanced imaging in oncology treatment and clinical trial design. *Acad. Radiol.* **24**, 1036–1049 (2017).
22. Subramanian, J. & Simon, R. What should physicians look for in evaluating prognostic gene-expression signatures? *Nat. Rev. Clin. Oncol.* **7**, 327–334 (2010).
23. Shafiq-Ul-Hassan, M. et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys.* **44**, 1050–1062 (2017).
24. Berenguer, R. et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology* **288**, 407–415 (2018).
25. American College of Radiology. *ACR Appropriateness Criteria* https://www.acr.org/Clinical-Resources/ACR-Appropriateness-Criteria (2022).
26. Society of Nuclear Medicine and Medical Imaging. *Procedure Standards* https://www.snmmi.org/ClinicalPractice/content.aspx?ItemNumber=6414. (2022).
27. European Association of Nuclear Medicine. *Guidelines* https://www.eanm.org/publications/guidelines/ (2022).
28. QIBQ Wiki. *Profiles* http://qibawiki.rsna.org/index.php/Profiles (2022).
29. Fass, L. Imaging and cancer: a review. *Mol. Oncol.* **2**, 115–152 (2008).
30. Zhao, B. et al. Exploring intra- and inter-reader variability in unidimensional, bidimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals. *Eur. J. Radiol.* **82**, 959–968 (2013).
31. O'Connor, J. P. B., Jackson, A., Parker, G. J. M., Roberts, C. & Jayson, G. C. Dynamic contrast-enhanced MRI in clinical trials of anti-vascular therapies. *Nat. Rev. Clin. Oncol.* **9**, 167–177 (2012).
32. Tudorica, L. A. et al. QIN: a feasible high spatiotemporal resolution breast DCE-MRI protocol for clinical settings. *Magn. Reson. Imaging* **30**, 1257–1267 (2012).
33. Nardone, V. et al. Delta radiomics: a systematic review. *Radiol. Med.* **126**, 1571–1583 (2021).
34. Pinker, K., Riedl, C. & Weber, W. A. Evaluating tumor response with FDG-PET: updates on PERCIST, comparison with EORTC criteria and clues to future development. *Eur. J. Nucl. Med. Mol. Imaging* **44**, 55–66 (2017).
35. Mackin, D. et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS ONE* **12**, e0178524 (2017).
36. Madabhushi, A., Udupa, J. K. & Souza, A. Generalized scale: theory, algorithms, and application to image inhomogeneity correction. *Comput. Image Vis. Underst.* **101**, 100–121 (2006).
37. Madabhushi, A. & Udupa, J. K. New methods of MR image intensity standardization via generalized scale. *Med. Phys.* **33**, 3426–3434 (2006).
38. Whitney, H. M. et al. Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. *J. Med. Imaging* **7**, 012707 (2020).
39. Duron, L. et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS ONE* **14**, e0213459 (2019).
40. Larue, R. T. H. M. et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents, and slice thicknesses: a comprehensive phantom study. *Acta Oncol.* **56**, 1544–1553 (2017).
41. Leijenaar, R. T. et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. *Nat. Sci. Rep.* **5**, 11075 (2015).
42. Willemink, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
43. Mali, S. A. et al. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *J. Per. Med.* **11**, 842 (2021).
44. Lin, Y. et al. Deep learning for fully automated tumor segmentation and extraction of magnetic resonance radiomics features in cervical cancer. *Eur. Radiol.* **30**, 1297–1305 (2020).

45. Parmar, C., Grossman, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine learning methods for quantitative radiomic biomarkers. *Nat. Sci. Rep.* **5**, 13087 (2015).
46. Primakov, S. P. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat. Commun.* **13**, 3423 (2022).
47. Gilhuijs, K. G. A., Giger, M. L. & Bick, U. Automated analysis of breast lesions in three dimensions using dynamic magnetic resonance imaging. *Med. Phys.* **25**, 1647–1654 (1998).
48. Chen, W., Giger, M. L., Lan, L. & Bick, U. Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics. *Med. Phys.* **31**, 1076–1082 (2004).
49. Chen, W., Giger, M. L., Bick, U. & Newstead, G. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Med. Phys.* **33**, 2878–2887 (2006).
50. Chen, W., Giger, M. L., Li, H., Bick, U. & Newstead, G. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magn. Reson. Med.* **58**, 562–571 (2007).
51. van Timmeren, J. E. et al. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomography* **2**, 361–365 (2016).
52. Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A. & Benali, H. From hand-crafted to deep learning-based cancer radiomics: challenges and opportunities. *IEEE Signal. Process. Mag.* **36**, 132–160 (2019).
53. Sahiner, B. et al. Deep learning in medical imaging and radiation therapy. *Med. Phys.* **46**, e1–e36 (2019).
54. Li, Z., Wang, Y., Yu, J., Guo, Y. & Cao, W. Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Nat. Sci. Rep.* **7**, 1–11 (2017).
55. Antropova, N., Huynh, B. Q. & Giger, M. L. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med. Phys.* **44**, 5162–5171 (2017).
56. International Organization for Standardization. *Guidance for the Use of Repeatability, Reproducibility, and Trueness Estimates in Measurement Uncertainty Evaluation* https://www.iso.org/obp/ui/#iso:std:iso:21748:ed-2:v1:en (2020).
57. Drukker, K., Pesce, L. & Giger, M. L. Repeatability in computer-aided diagnosis: application to breast cancer diagnosis on sonography. *Med. Phys.* **37**, 2659–2669 (2010).
58. Kessler, L. G. et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat. Methods Med. Res.* **24**, 9–26 (2015).
59. Raunig, D. L. et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat. Methods Med. Res.* **24**, 27–67 (2015).
60. Huang, E. P. et al. Multiparametric quantitative imaging in risk prediction: recommendations for data acquisition, technical performance assessment, and model development and validation. *Acad. Radiol.* https://doi.org/10.1016/j.acra.2022.09.018 (2022).
61. McHugh, D. J. et al. Image contrast, image preprocessing, and T1-mapping affect MRI radiomic feature repeatability in patients with colorectal cancer liver metastases. *Cancers* **13**, 240 (2021).
62. Jha, A. K. et al. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Sci. Rep.* **11**, 2055 (2021).
63. Bissoto, A., Perez, F., Valle, E. & Avila, S. Skin lesion synthesis with generative adversarial networks. *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. OR 2.0 First International Workshop, CARE Fifth International Workshop, CLIP Seventh International Workshop, ISIC Third International Workshop*. Springer Lecture Notes in Computer Science (Springer, 2019).
64. Sullivan, D. C. et al. Metrology standards for quantitative imaging biomarkers. *Radiology* **277**, 813–825 (2015).
65. Hackstadt, A. J. & Hess, A. M. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* **10**, 11 (2009).
66. Luo, J. et al. A comparison of batch effect removal methods for enhancement of prediction performance using MACQ-II microarray gene expression data. *Pharmacogenomics J.* **10**, 278–291 (2010).
67. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127 (2007).
68. Orlhac, F. et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. *J. Nucl. Med.* **59**, 1321–1328 (2018).
69. Parker, H. S. & Leek, J. T. The practical effect of batch on genomic prediction. *Stat. Appl. Genet. Mol. Biol.* **11**, 10 (2012).
70. Robinson, K., Li, H., Lan, L., Schacht, D. & Giger, M. Radiomics robustness assessment and classification evaluation: a two-stage method demonstrated on multivendor FFDM. *Med. Phys.* **46**, 2145–2156 (2019).
71. *The Cancer Imaging Archive* http://cancerimagingarchive.net (2020).
72. Clark, K. et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digital Imaging* **26**, 1045–1057 (2013).
73. Zhu, Y. et al. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. *Nat. Sci. Rep.* **5**, 17787 (2015).
74. Riley, R. D. et al. Minimum sample size for developing a multivariable prediction model: part II — binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2018).
75. Riley, R. D. et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).
76. Riley, R. D. et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat. Med.* **41**, 1280–1295 (2022).
77. Cho, J., Lee, K., Shin, E., Choy, G. & Do, S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? Preprint at https://doi.org/10.48550/arXiv.1511.06348 (2015).
78. Whitney, H., Li, H., Ji, Y., Liu, P. & Giger, M. L. Comparison of breast MRI tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods. *Proc. IEEE* **108**, 163–177 (2020).
79. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* 2nd edn (Springer, 2009).
80. Deist, T. M. et al. Machine learning algorithms for outcome prediction in (chemo) radiotherapy: an empirical comparison of classifiers. *Med. Phys.* **45**, 3449–3459 (2018).
81. Haykin S. *Neural Networks: A Comprehensive Foundation* (Prentice Hall, 1994).
82. Ben-Dor, A. et al. Tissue classification with gene expression profiles. *J. Comput. Biol.* **7**, 559–583 (2000).
83. Dudoit, S., Fridlyand, J. & Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77–87 (2002).
84. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection — a review and recommendations for the practicing statistician. *Biom. J.* **60**, 431–449 (2018).
85. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
86. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
87. Harrell, F. E. Jr., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *J. Am. Med. Assoc.* **247**, 2543–2546 (1982).
88. Hosmer, D. W. & Lemeshow, S. Goodness of fit tests for the multiple logistic regression model. *Commun. Stat. Theory Methods* **9**, 1043–1069 (1980).
89. Lemeshow, S. & Hosmer, D. A review of goodness of fit statistics for use in the development of logistic regression model. *Am. J. Epidemiol.* **115**, 92–106 (1982).
90. van Calster, B. & Steyerberg, E. W. *Wiley StatsRef: Statistics Reference Online* (John Wiley and Sons, Ltd., 2018).
91. Bröcker, J. & Smith, L. A. Increasing the reliability of reliability diagrams. *Weather Forecast.* **22**, 651–661 (2007).
92. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition* (John Wiley and Sons, 2002).
93. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B* **36**, 111–147 (1974).
94. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
95. Molinaro, A. M., Simon, R. & Pfeffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307 (2005).
96. Dobbin, K. K. & Simon, R. M. Optimally splitting cases for training and testing high-dimensional classifiers. *BMC Med. Genomics* **4**, 31 (2011).
97. Sachs, M. C. & McShane, L. M. Issues in developing multivariable molecular signatures for guiding clinical care decisions. *J. Biopharm. Stat.* **26**, 1098–1110 (2016).
98. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006).
99. Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput. Biol. Med.* **140**, 105111 (2022).
100. Hilsenbeck, S. G., Clark, G. M. & McGuire, W. L. Why do so many prognostic factors fail to pan out? *Breast Cancer Res. Treat.* **22**, 197–206 (1992).
101. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574 (2006).
102. Wu, G. et al. Preoperative CT-based radiomics combined with intraoperative frozen section is predictive of invasive adenocarcinoma in pulmonary nodules: a multicenter study. *Eur. Radiol.* **30**, 2680–2691 (2020).
103. Hayes, D. F. Defining clinical utility of tumor biomarker tests: a clinician's viewpoint. *J. Clin. Oncol.* **39**, 238–249 (2021).
104. Saha, A., Hosseinzadeh, M. & Huisman, H. End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Med. Image Anal.* **73**, 102155 (2021).
105. Hosseinzadeh, M. et al. Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *Eur. Radiol.* **32**, 2224–2234 (2022).
106. Baughan, N. et al. *Sequestration of Imaging Studies in MIDRC: A Multi-institutional Data Commons. Medical Imaging 2002; Image Perception, Observer Performance, and Technology Assessment*, vol. 12035 (SPIE, 2022).
107. Simon, R. M., Paik, S. & Hayes, D. F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl Cancer Inst.* **101**, 1446–1452 (2009).
108. Pappalardo, F., Gusso, G., Tshinanu, F. M. & Viceconti, M. In silico clinical trials: concepts and early adoptions. *Brief. Bioinforma.* **20**, 1699–1708 (2019).
109. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, Institute of Medicine. *Evolution of Translational Omics: Lessons Learned and the Path Forward* (The National Academies Press, 2012).

110. Altman, D. G., McShane, L. M., Sauerbrei, W. & Taube, S. E. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *PLoS Med.* **9**, e1001216 (2012).
111. Equator Network. *Enhancing the Quality and Transparency of Health Research* (EQUATOR) https://www.equator-network.org/ (2022).

Acknowledgements

P.L. acknowledges support for the publication of this work from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement CHAIMELEON No. 952172, EuCanImage No. 952103, IMI-OPTIMA No. 101034347 and ERC advanced grant (ERC-ADG-2015 No. 694812 – Hypoximmuno). P.K. acknowledges support for the publication of this work from NCI grant P50 CA228944.

Competing interests

M.G. has acted as a scientific adviser of Quantitative Insights (now Qlarity Imaging), is the contact Principal Investigator for MIDRC (funded by NIBIB COVID-19 Contract 75N92020D00021), receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi and Toshiba, holds stocks in R2/Hologic, is a shareholder in Qview, and is a co-founder of and equity holder in Quantitative Insights (now Qlarity Imaging). P.L. is a co-founder, minority shareholder and member of the advisory board of Oncoradiomics, and is listed as a co-inventor on several licensed patents in radiomics. E.P.H., J.P.B.O.-C., L.M.M., P.E.K., E.L.S. and L.K.S. declare no competing interests.

Additional information

Correspondence should be addressed to Erich P. Huang.

Peer review information *Nature Reviews Clinical Oncology* thanks K. Bera. J.-E. Bibault, J. Tian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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