nature reviews clinical oncology

https://doi.org/10.1038/s41571-025-01067-1

Review article

Check for updates

Radiomics Quality Score 2.0: towards radiomics readiness levels and clinical translation for personalized medicine

Philippe Lambin[1,2] ✉, Henry C. Woodruff[1,2], Shruti Atul Mali[1], Xian Zhong[1,3], Sheng Kuang[1], Elizaveta Lavrova[1,4], Hamza Khan[1,5], Karim Lekadir[6,7], Alex Zwanenburg[8,9,10], Joseph Deasy[11], Maciej Bobowicz[12], Luis Marti-Bonmati[13], Andrew Maidment[14], Michel Dumontier[15], Paul E. Kinahan[16,17], J. Martijn Nobel[2,18], Sina Amirrajab[1] & Zohaib Salahuddin[1]

Abstract

Radiomics is a tool for medical imaging analysis that could have a relevant role in precision oncology by offering precise quantitative support for clinical decision-making. The Radiomics Quality Score (RQS) is a tool developed to assess the rigour of radiomics studies that has now been widely adopted by researchers. Although RQS version 1.0 established a benchmark, an updated framework is required to account for evolving knowledge and ensure optimal evaluation of the quality of radiomics studies through the inclusion of fairness, explainability, rigorous quality control and harmonization. In this Review, we introduce the updated RQS 2.0, which maintains the scientific rigour of its predecessor and addresses these contemporary needs, and therefore could potentially accelerate clinical translation. Moreover, we introduce the radiomics readiness levels, inspired by the technology readiness level framework, which are integrated in RQS 2.0 and reflect nine distinct levels of incremental improvement in radiomics research with the ultimate aim of clinical implementation. We also detail anticipated future directions in radiomics, outlining a strategic vision to advance precision oncology, which is the ultimate aim of RQS 2.0.

Sections

A full list of affiliations appears at the end of the paper. ✉e-mail: philippe.lambin@maastrichtuniversity.nl

Key points

- Radiomics is a quantitative image analysis method that enhances disease diagnosis, tumour characterization and prediction of treatment response, supporting its application in precision oncology.
- The Radiomics Quality Score, originally introduced in 2017 to improve methodological and reporting rigour, has driven improvements in study quality but lacked sufficient coverage of certain aspects such as deep learning-specific challenges, cost effectiveness, prospective design and real-world feasibility.
- Radiomics approaches can be broadly categorized into handcrafted and deep learning-based methods, each with distinct workflows, challenges and requirements for clinical translation.
- Key barriers to clinical translation of radiomics include methodological inconsistencies, limited external and prospective validation, lack of transparency and insufficient consideration of model fairness, robustness, explainability and usability in clinical settings.
- Radiomics Quality Score 2.0 addresses the limitations of its predecessor by incorporating updated criteria for both handcrafted and deep learning-based radiomics and introducing radiomics readiness levels to guide incremental progress towards clinical deployment.
- Future advances in radiomics will be driven particularly by the integration of multiomics data, the adoption of foundation models, data standardization and the use of federated learning and synthetic data to enhance model generalizability and data accessibility.

Introduction

Precision medicine takes into account the individual characteristics of each patient, such as unique genetic makeup, individual phenotype, environmental influences and lifestyle choices, to deliver the appropriate treatment for each patient at the right time[1]. Radiomics is the quantitative analysis of medical images using complex algorithms to uncover patterns not always easily discernible by the naked eye[2,3]. Radiomics has proven to be a powerful tool in precision oncology, advancing areas such as diagnosis, tumour characterization, detection of genetic mutations, stratification of patients into different risk groups and, ultimately, prediction of response to treatment[4,5].

Radiomics approaches can be broadly categorized into handcrafted and deep learning-based radiomics. Handcrafted radiomics involves predefining (handcrafting) features to be extracted from a region of interest (ROI), followed by selecting such features and developing a machine learning model[6,7]. Deep learning-based radiomics use algorithms such as convolutional neural networks (CNNs), transformer-based models or graph neural networks to automatically learn the most important features from the images to achieve the task at hand[8]. The rise of big data and enhanced computational capabilities combined with promising performance outcomes has led to a growing interest in radiomics.

In 2017, we introduced the Radiomics Quality Score (RQS) to assess the methodological and reporting rigour of individual radiomics studies[4]. The RQS takes into account crucial aspects such as protocol standardization, clinical relevance, statistical validity and model performance, which are essential for developing robust and clinically relevant radiomics models. The wide adoption of the RQS by the scientific community as a tool for evaluating the quality of radiomics studies is reflected in the growing number of citations per year of the score (Supplementary Fig. 1a). Moreover, systematic reviews have extensively used the RQS to evaluate the quality of radiomics research in different organs and modalities (Table 1 and Supplementary Fig. 1b).

Since the introduction of RQS version 1.0, the overall quality of radiomics studies has modestly but measurably improved. A meta-analysis of 3,258 RQS evaluations across 130 systematic reviews found a significant positive correlation between RQS and publication year (Pearson $r = 0.3$; $P < 0.01$), indicating that more recent studies tend to have higher scores[9]. Notably, the average RQS nearly doubled after 2017: studies published before 2018 had a mean score of ~5.6 out of 36 versus ~10.1 for those from 2018 onwards. This improvement aligns with the growing adoption of the RQS and suggests that increased awareness of reporting standards has helped to improve methodological rigour in the field.

The RQS criteria related to phantom studies, test–retest analysis, external validation, prospective design, cost-effectiveness evaluation and open science are consistently under-addressed in radiomics studies[9]. However, as most radiomics research remains in the early stages of development, some of the aforementioned criteria (such as cost effectiveness and prospective design) are often not applicable. These limitations highlight the need for an improved version of RQS that reflects the different phases of research and encourages incremental progress towards clinical translation. Moreover, some studies have identified additional limitations of RQS 1.0, such as the need for specific criteria related to deep learning methods and the over-penalization of preliminary retrospective studies[10]. Also, certain requirements from RQS 1.0 that are rarely applied in real-world clinical practice are too strict[11].

Despite the increasing number of studies proving the effectiveness of radiomics, substantial challenges remain that must be addressed to maximize the scope of radiomics research, paving the way for reproducible and trustworthy clinical adoption[12,13]. The low clinical translation rate of radiomics studies can be attributed to various factors, including methodological challenges and limited clinical validation[13]. Further challenges that currently impede clinical translation, for example, relating to harmonization[14], algorithm fairness[15], explainability[16] and performance drift[17], need to be addressed.

We now propose an updated RQS, RQS version 2.0, in which we address the limitations of RQS 1.0 and incorporate guidelines aimed at accelerating clinical translation. RQS 2.0 differentiates between handcrafted radiomics and deep learning-based approaches to address these challenges, is further aimed at improving the quality and reliability of radiomics research and reporting, and increases the potential for clinical applications. Further increasing the stringency of RQS 2.0 criteria could pose practical challenges for laboratories and research teams, whereas adding more stringent requirements could complicate the implementation of the entire radiomics pipeline, potentially discouraging participation and stalling valuable research efforts. Therefore, to strike a balance between rigour and feasibility, we have introduced radiomics readiness levels (RRLs) as a structured approach to alleviate challenges related with adoption of the RQS 2.0 (ref. 18).

In this Review, we present RQS 2.0 as a tool that not only integrates contemporary considerations in the field but also incorporates RRLs to provide a systematic framework for the incremental improvement of

radiomics projects, including the integration of components necessary for their regulatory approval[19,20], leading up to their clinical application. Moreover, we explore promising future directions with potential to accelerate the incorporation of radiomics into clinical practice.

RQS 2.0

The initial version of the RQS established a benchmark for assessing the quality of radiomics studies and their reporting. Since the introduction of RQS 1.0, developments in artificial intelligence (AI) have brought about new challenges in the field of radiomics, such as the need for increased fairness, universality, traceability, usability, robustness and explainability[21]. RQS 2.0 presents substantial advances in the methodological assessment of radiomics research, specially designed to tackle these contemporary challenges. RQS 2.0 distinguishes between handcrafted and deep learning-based radiomics by acknowledging the unique requirements of each approach and their distinct workflows (Fig. 1). The RQS 2.0 is the result of discussions among a diverse group of experts in the field.

We have incorporated RRLs into the RQS 2.0 framework to promote a systematic and incremental approach to assessing research in the field (Table 2 and Box 1). RRLs are modelled after technology readiness levels, a systematic framework that is used to evaluate the readiness and maturity of technologies for deployment[22,23]. Technology readiness levels were initially introduced by the National Aeronautics and Space Administration (NASA) for assessing the maturity of technologies for space exploration, but they have been widely adapted to other fields, including machine learning systems[24]. Hence, RRLs provide a stepwise framework that enables researchers to assess and communicate the readiness of their radiomics research, guiding them from initial exploration to full clinical implementation in addition to accommodating variations in resources and expertise. This approach not only streamlines the evaluation process but also ensures that radiomics research can progress systematically and effectively in response to the growing demands of the field. Several checklists for radiomics studies have been proposed, such as CLEAR[25], ARISE[26] and METRICS[27], although they focus solely on the evaluation of scientific research. By contrast, the RQS 2.0 and RRLs provide a more holistic and comprehensive approach that not only enables the evaluation of scientific quality in radiomics studies but also establishes checkpoints to ensure effective clinical translation of this research.

As in RQS 1.0, in RQS 2.0 weights were assigned to different criteria on the basis of expert opinion gathered through multiple rounds of discussion; accordingly, the criteria considered the most important were given higher weights (Fig. 2). RRLs encompass different stages, starting from foundational exploration through validation and verification processes to eventual clinical deployment (Fig. 2 and Supplementary Fig. 2). The integration of RRLs into the RQS 2.0 framework underscores the necessity of research studies that collectively address all the criteria outlined in the RQS 2.0, thereby facilitating the clinical translation of radiomics tools.

Foundational exploration

RRL1 establishes the foundational elements needed to set the stage for future radiomics research. In the initial phase of radiomics research,

Table 1 | Selection of systematic reviews using the RQS 1.0

Title	Organ or cancer type	Modality	Mean RQS score out of 36 ±STD (%)	Maximum RQS score out of 36 (%)	Publication year	Refs.
Prostate MRI radiomics: a systematic review and radiomics quality score assessment	Prostate cancer	MRI	7.9±5.1 (23%±13%)	18 (50%)	2020	158,159
A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation	Osteosarcoma	MRI, PET	6.9±6.0 (20.4%±16.7%)	16 (44%)	2020	160,161
MRI-based radiomics in nasopharyngeal cancer: systematic review and perspectives using RQS assessment	Head and neck cancer	MRI	7.5±5.4 (21.3%±14.3%)	20 (56%)	2021	162,163
Radiomics in renal cell carcinoma — a systematic review and meta-analysis	RCC	CT, MRI, PET–MRI	4.9±4.6 (13.6%±12.8%)	15 (42%)	2021	164,165
A systematic review of the current status and quality of radiomics for glioma differential diagnosis	Glioma	MRI	8.7±5.6 (24.2%±15.6%)	19 (53%)	2022	166,167
Radiomics models for preoperative prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis	HCC	CT, MRI	13.6±4.1 (37.7%±11.4%)	23 (64%)	2022	168,169
MRI-based radiomics methods for predicting Ki-67 expression in breast cancer: a systematic review and meta-analysis	Breast cancer	MRI	5.9±4.0 (16.4%±11.1%)	12 (33%)	2023	170,171
A systematic review and meta-analysis of CT and MRI radiomics in ovarian cancer: methodological issues and clinical utility	Ovarian cancer	CT, MRI	11±4.8 (30.7%±13.3%)	22 (61%)	2023	172,173
Systematic review, meta-analysis and Radiomics Quality Score assessment of CT radiomics-based models predicting tumour *EGFR* mutation status in patients with non-small-cell lung cancer	NSCLC	CT	15.3±2.5 (42.3%±7%)	24 (67%)	2023	105,174
MRI-based radiomics in bladder cancer: a systematic review and Radiomics Quality Score assessment	Urothelial cancer	MRI	11.7±4.8 (32.5%±13.3%)	19 (53%)	2023	175–177
Current status and quality of radiomic studies for predicting KRAS mutations in colorectal cancer patients: a systematic review and meta-analysis	CRC	MRI, CT, PET	9.6±3.9 (26.5%±10.8%)	17 (47%)	2023	178,179

CRC, colorectal cancer; HCC, hepatocellular carcinoma; NSCLC, non-small-cell lung cancer; RCC, renal cell carcinoma; RQS, Radiomics Quality Score.

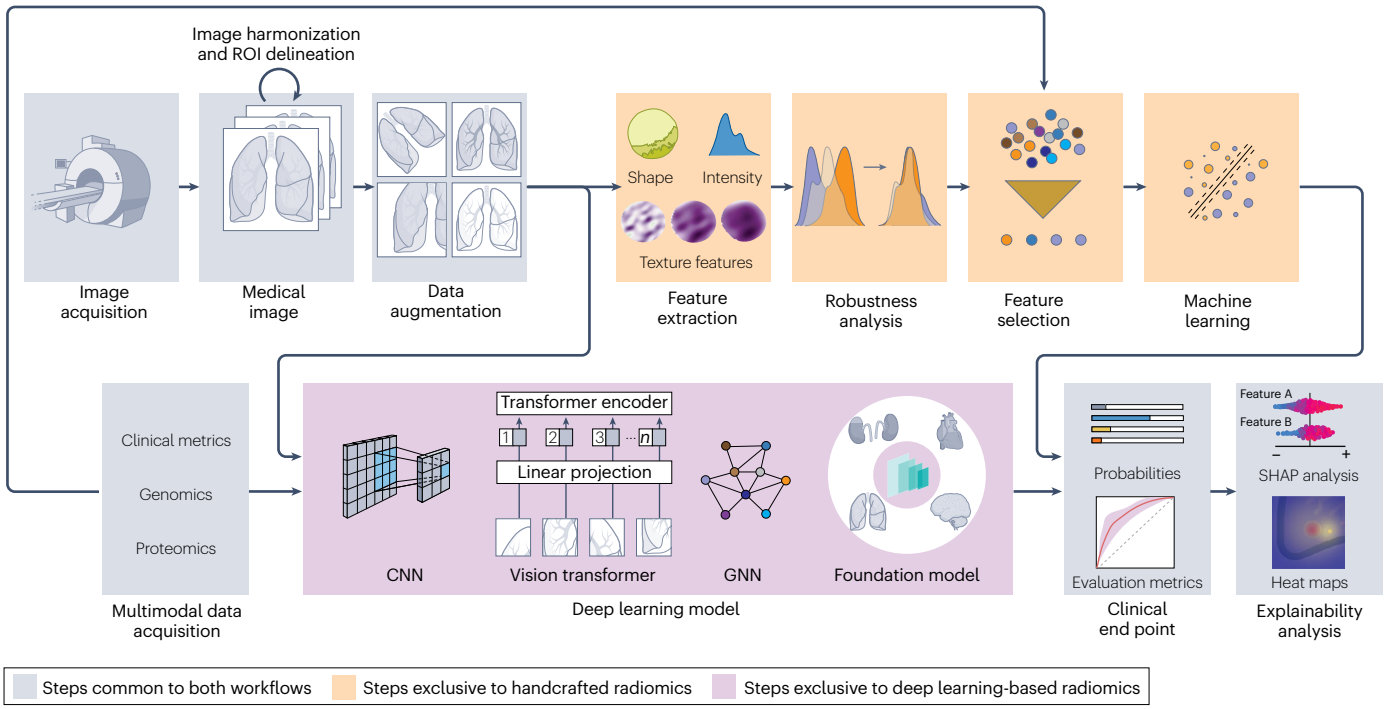

Fig. 1 | Handcrafted and deep learning-based radiomics. General workflows for handcrafted and deep learning-based approaches to radiomics. Some steps are common to both workflows and others are exclusive to each approach. Deep learning models include foundation models, which are generalist, pretrained models developed on large-scale datasets and designed to be adaptable across a variety of downstream tasks. CNN, convolutional neural network; GNN, graph neural network; ROI, region of interest; SHAP, Shapley additive explanations.

recognizing and agreeing upon an unmet clinical need is essential. In 2006, the European Commission defined unmet clinical need as a condition for which no satisfactory diagnostic, preventive or treatment methods exist or for which new medicinal products will offer substantial benefit relative to existing solutions[28]. The agreement of experts from multiple centres and countries on the fact that a certain condition is a unmet clinical need is crucial. This approach broadens the relevance and applicability of the research because it considers diverse healthcare systems and patient demographics. A structured consensus approach, such as the Delphi method, is recommended for defining a unmet clinical need[29] and to ensure methodological rigour of the process, with comprehensive and systematic integration of expert opinions.

Detailed documentation of the radiological imaging hardware used in radiomics studies, including model, manufacturer and specifications, facilitates reproducibility and aids in assessing heterogeneity. The imaging protocols should be thoroughly reported and validated to enhance their robustness and applicability[30]. We propose five levels of image protocol quality in terms of transparent reporting of medical image acquisition for future-proof radiomics: level 0 indicates protocols without formal approval; level 1 includes protocols approved with a reference number within the host institution; and level 2 refers to protocols approved with formal quality assurance, which can be conducted either by an external certifying body or internally within the institution, provided that a formal and documented protocol is in place. These protocols are suitable for prospective trials; level 3 protocols are internationally established and published in guidelines and white papers, and level 4, the highest level, denotes future-proof protocols that follow level 3 standards, adhere to FAIR (findable, accessible, interoperable and reusable) principles for scientific data management[31] and retain raw imaging data, ensuring robustness and applicability for future research. Moreover, the establishment of clear patient inclusion and exclusion criteria ensures that the study population accurately represents the targeted clinical scenario, thereby enhancing the relevance and applicability of the research findings[32]. In medical AI, the focus on diversity and distribution is essential to ensure that data and findings accurately represent the wider population[33]. Identifying and addressing potential biases relevant to the clinical issue at hand is crucial, necessitating a comprehensive evaluation of various factors, such as demographic, socioeconomic, geographical and clinical features[21,34].

Data preparation

RRL2 emphasizes the key steps for formulating the inputs and concepts for use in routine clinical settings. Handcrafted radiomics features can be sensitive to scanner differences, acquisition parameters, image reconstruction settings and segmentation[35–37]. Phantom studies are experiments that use standardized physical models designed to simulate human tissue, allowing researchers to assess the reproducibility of radiomics features in a controlled environment. These studies can facilitate assessment of the reproducibility of radiomics features within a controlled environment, particularly in terms of changes in image acquisition and reconstruction parameters[38]. Test–retest studies help to evaluate the consistency of features over time by acquiring images under identical conditions at two or more timepoints[39,40]. The precise and reproducible delineation of the ROI for handcrafted radiomics

Table 2 | Calculation of RQS 2.0 and alignment with RRLs

RRL[a]	Criterion[a]	Points	Applies to HCR and/or DLR?
RRL1	1. Unmet clinical need	Agreed upon and defined by more than one centre (+1) Defined using an established consensus method (+2)	Both
RRL1	2. Hardware description	+1	Both
RRL1	3. Image protocol quality	Level 1[a] and Level 2[a] (+1) Level 3[a] and Level 4[a] (+2)	Both
RRL1	4. Inclusion and exclusion criteria	+1	Both
RRL1	5. Diversity and distribution	+1	Both
Cumulative maximum total score (RRL1): HCR 7 points, DLR 7 points			
RRL2	6. Feature robustness	+1	HCR
RRL2	7. Preprocessing of images	+1	Both
RRL2	8. Harmonization	+1	Both
RRL2	9. Compliance with international standards in radiomics	+1	HCR
RRL2	10. Automatic segmentation	+1	Both
Cumulative maximum total score (RRL1–RRL2): HCR 7+5=12 points, DLR 7+3=10 points			
RRL3	11. Feature reduction	+1	HCR
RRL3	12. Feature robustness for feature selection	+1	HCR
RRL3	13. Combination of HCR and DLR	+1	Both
RRL3	14. Multivariable analysis	+2	Both
Cumulative maximum total score (RRL1–RRL3): HCR 12+5=17 points, DLR 10+3=13 points			
RRL4	15. Single-centre validation	+1	Both
RRL4	16. Cut-off analyses	+1	Both
RRL4	17. Discrimination statistics	Discrimination statistic and its significance reported (+1) Resampling method technique also applied (+1)	Both
RRL4	18. Calibration statistics	+1	Both
RRL4	19. Failure mode analysis	+1	Both
RRL4	20. Open science and data	Data (including scans) are open source (+1) ROI segmentations are open source (+1) Code is open source (+1)	Both
Cumulative maximum total score (RRL1–RRL4): HCR 17+9=26 points, DLR 13+9=22 points			
RRL5	21. Multicentre validation	Data from one external institution (+1) Data from ≥2 external institutions (+2) Validation carried out on a third-party platform using completely unseen data (+3)	Both
RRL5	22. Comparison with current clinical standard	+2	Both
RRL5	23. Comparison with previous work	+1	Both
RRL5	24. Potential clinical utility	+2	Both
Cumulative maximum total score (RRL1–RRL5): HCR 26+8=34 points, DLR 22+8=30 points			
RRL6	25. Explainability	+1	Both
RRL6	26. Evaluation of explainability	+1	Both
RRL6	27. Biological correlates	+1	Both
RRL6	28. Evaluation of fairness and plan for mitigation of bias	Fairness evaluation (+1) Appropriate bias correction methods applied, if necessary (+1)	Both
Cumulative maximum total score (RRL1–RRL6): HCR 34+5=39 points, DLR 30+5=35 points			
RRL7	29. Usability for clinicians	+1	Both
RRL7	30. Sample size calculation	+1	Both
RRL7	31. Clinical trial preregistration	+1	Both
RRL7	32. Prospective validation	+3	Both
RRL7	33. Real-world clinical assessment	+1	Both
Cumulative maximum total score (RRL1–RRL7): HCR 39+7=46 points, DLR 35+7=42 points			
RRL8	34. Software traceability	+1	Both
RRL8	35. Software safeguards	+1	Both
RRL8	36. Cost-effectiveness analysis	+2	Both
RRL8	37. Performance drift	+1	Both
RRL8	38. Continual learning	+1	Both
Cumulative maximum total score (RRL1–RRL8): HCR 46+6=52 points, DLR 42+6=48 points			
RRL9	39. Define the level of automation in clinical practice	+1	Both
RRL9	40. Quality management system	+1	Both
RRL9	41. Regulatory requirements	+1	Both
RRL9	42. Product on the market	+1	Both
Cumulative maximum total score (RRL1–RRL9): HCR 52+4=56 points, DLR 48+4=52 points			

[a]Defined in Box 1. DLR, deep learning-based radiomics; HCR, handcrafted radiomics; ROI, region of interest; RQS, Radiomics Quality Score; RRL, radiomics readiness level.

feature extraction is a key step. Indeed, interobserver and intraobserver differences in segmentation can introduce substantial variation in radiomics features that affects the reliability of these features. Therefore, identifying radiomics features that are sufficiently stable despite segmentation differences is important[37,41,42]. Image preprocessing is an important step towards reducing variability between different images and improving the robustness of the analyses[43,44]. Exploring different

Box 1 | Key criteria of the Radiomics Quality Score (RQS) 2.0 and integration within the radiomics readiness levels (RRLs) framework

RRL1: foundational exploration

1. Unmet clinical need: unmet clinical need is defined.
2. Hardware description: detailed description of the imaging hardware used, including model, manufacturer and technical specifications.
3. Image protocol quality: five levels of image protocol quality are defined for transparent reporting of medical image acquisition for future-proof radiomics:
   - Level 0 indicates that the protocol has not been formally approved with a reference number.
   - Level 1 indicates that the protocol has been approved with a reference number in the archive of the institution.
   - Level 2 indicates that the protocol has been approved with formal quality assurance (recommended minimum level for prospective trials).
   - Level 3 indicates that the protocol is established internationally and has been published in guideline documents and peer-reviewed studies.
   - Level 4 indicates that the protocol is future proof: it follows transparent reporting of medical image acquisition level 3, FAIR (findable, accessible, interoperable and reusable) principles[31] and retains raw data.
4. Inclusion and exclusion criteria: detailed criteria for patient selection in studies, outlining the rationale behind inclusion and exclusion.
5. Diversity and distribution: before the start of the project, identify potential biases including demographic attributes (such as sex, gender, age or ethnicity), socioeconomic and geographic backgrounds, and medical profiles (such as comorbidities or disabilities).

RRL2: data preparation

6. Feature robustness: the robustness of the radiomics features can be assessed using any of the methods below or other approaches to ensure their robustness and repeatability:
   - Collecting images at various timepoints to analyse feature robustness against temporal variabilities, such as in test–retest studies.
   - Implementing segmentation variations by different physicians, algorithms, software or introducing perturbations such as noise and at different breathing cycles (that is, variations in anatomical position owing to inhalation and exhalation) to assess robustness against segmentation variabilities.
   - Conducting phantom studies on all scanners to identify interscanner differences and vendor-dependent features, evaluating feature robustness against these variabilities.
7. Preprocessing of images: apply preprocessing steps to standardize images, providing clear reasoning for each step.
8. Harmonization: explore image-level harmonization and/or feature-level harmonization techniques to mitigate variability across multicentre acquisitions.
9. Compliance with international standards in radiomics: adhere to the use of implementations that adhere to international standards, such as the Image Biomarker Standardization Initiative[46,47], for standardized radiomic feature extraction and analysis.
10. Automatic segmentation: an automated segmentation algorithm is used for defining the region of interest (ROI).

RRL3: prototype model development

11. Feature reduction: feature reduction decreases the risk of overfitting. Overfitting is highly likely if the number of features exceeds the number of samples. Check for correlation with other features such as volume.
12. Feature robustness for feature selection: integrate an evaluation of feature robustness into the feature selection process, using data from prior research or experimental findings. This could include leveraging insights from previously published test–retest, phantom and/or segmentation studies or methodologies outlined in point 6.
13. Combination of handcrafted radiomics and deep learning-based radiomics: compare and explore the synergistic combination of handcrafted radiomics and deep learning-based radiomics. Evaluate each type of model as well as the consensus of both types.
14. Multivariable analysis: multivariable analysis incorporating non-radiomics features, such as clinical factors (for example, tumour–node–metastasis (TNM) staging and age), genomics and proteomics, is expected to yield a more holistic model. This enables effective correlation and inference between radiomics and non-radiomics features.

RRL4: internal validation

15. Single-centre validation: the validation is performed without retraining and without adaptation of the cut-off value on the data from the same institute, providing crucial information about credible clinical performance.
16. Cut-off analyses: identify optimal thresholds for analysis in various study types. Utilize methods such as Youden's index[66] to determine the optimal operating point, particularly in classification tasks. In survival analysis, apply suitable cut-offs for effective risk stratification.
17. Discrimination statistics: report discrimination statistics (for example, receiver operator characteristic curve, sensitivity, specificity) and their statistical significance (such as *P* values and confidence intervals). One can also apply a resampling method (for example, bootstrapping or cross-validation)[180].
18. Calibration statistics: report calibration statistics (for example, calibration-in-the-large/slope or calibration plots)[67].
19. Failure mode analysis: document the limitations of the model limitations with examples of edge-case scenarios illustrating failure cases.
20. Open science and data: make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study.

RRL5: capability testing

21. Multicentre validation: validation conducted with data from multiple institutes, ensuring no overlap with training data from those institutes.

(continued from previous page)

22. Comparison with current clinical (gold) standard: assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM staging for survival prediction). This comparison shows the added value of radiomics.
23. Comparison to previous work: performance comparison with previously published handcrafted radiomics signatures and/or deep learning algorithms for the use case if prior work is available using an identical dataset.
24. Potential clinical utility: report on the current and potential application of the models in a clinical setting (for example, decision curve analysis).

RRL6: assessment of trustworthiness

25. Explainability: apply explainability tools such as Shapley additive explanations (SHAP)[75] for handcrafted radiomics and gradient-weighted class activation mapping (Grad-CAM)[76] for deep learning-based radiomics to understand the underlying decision-making processes of the models, enhancing clarity and trust in the results.
26. Explainability evaluation: conduct qualitative and quantitative evaluation of interpretability methods to ensure the reliability and validity of explanations. For example, evaluating explanation consistency to adversarial perturbations, correlating key features identified through SHAP analysis with established clinical knowledge.
27. Biological correlates: detect and discuss biological correlates. The demonstration of phenotypic differences (possibly associated with underlying gene and/or protein expression patterns) deepens understanding of radiomics and biology.
28. Fairness evaluation and mitigation: evaluation of model performance concerning previously identified biases (see point 5).

RRL7: prospective validation

29. Usability for clinicians: evaluate the usability of the radiomics tool usability, focusing on clinician interface, workflow integration and ease of use.
30. Sample size calculation: perform sample size calculation before the start of the prospective validation to ensure statistical validity and robustness of the study.
31. Clinical trial preregistration: the prospective clinical trial, including its statistical plan, is registered in a clinical trial database (such as ClinicalTrials.gov), with any post hoc changes to the protocol being tracked.
32. Prospective validation: prospective validation, which might include in silico studies, is carried out to ensure the clinical validity and usefulness of the radiomics biomarker.
33. Real-world clinical assessment: human-in-the-loop assessments are conducted to evaluate the practical application and effect of the radiomics model in real-world clinical settings[85].

RRL8: applicability and sustainability

34. Software traceability: implement and document a robust software traceability process. This should detail the development, changes and version control of the software used in the radiomics workflow.
35. Software safeguards: implement appropriate software checks to prevent out-of-scope use or the utilization of unreliable input.
36. Cost-effectiveness analysis: report on the cost effectiveness of the clinical application (for example, quality-adjusted life years generated).
37. Performance drift: define a strategy to evaluate the model performance periodically owing to data shifts.
38. Continuous learning: define a strategy for continuous learning to learn from errors and improve over time.

RRL9: clinical deployment

39. Define the level of automation in clinical practice of your artificial intelligence (AI) solution in clinical practice:
    - Level 0 (no automation): a clinician performs the clinical task without using the radiomics model.
    - Level 1 (clinical assistance): the clinician uses the prediction of the radiomics model for a part of the clinical task.
    - Level 2 (partial automation): the clinician considers the radiomics model's prediction for the clinical task before making the final recommendation.
    - Level 3 (conditional automation): the radiomics model provides predictions for the clinical task under supervision and the clinician can intervene at any time.
    - Level 4 (high automation): the radiomics model provides the predictions and the intervention of the clinician is required for special (out-of-distribution) cases.
    - Level 5 (full automation): the radiomics model provides predictions for the clinical task without human intervention.
40. Quality management system: implement and maintain a quality management system (such as ISO 9001) to ensure consistent quality and compliance in the radiomics workflow.
41. Regulatory requirements: evaluate the alignment of the AI solution with the requirements of the chosen regulatory pathway (for example, 510(k) or Premarket Approval for Food and Drug Administration (FDA), conformity assessment for European Medicines Agency and European Union AI Act.
42. Product on the market: successfully introduce the radiomics product to the market, ensuring regulatory approval and clinical adoption.

preprocessing methods that reduce noise without compromising resolution is essential to promote standardization[45].

The reproducibility of radiomics is further challenged by the absence of standardized definitions for radiomics features across different software implementations. Adhering to an international standard, such as the Image Biomarker Standardization Initiative, is crucial to ensure the use of common nomenclature and definitions[46,47]. Such definitions ensure consistency, standardization and comparability across radiomics studies, enabling broader applicability and validation. Variability in imaging devices and techniques limits the generalizability of radiomics models. Harmonization provides a means to mitigate the effects of this variability, enabling robust model development[14,48]. Image harmonization methods can be classified into two categories depending on whether they refer to the image domain or the feature

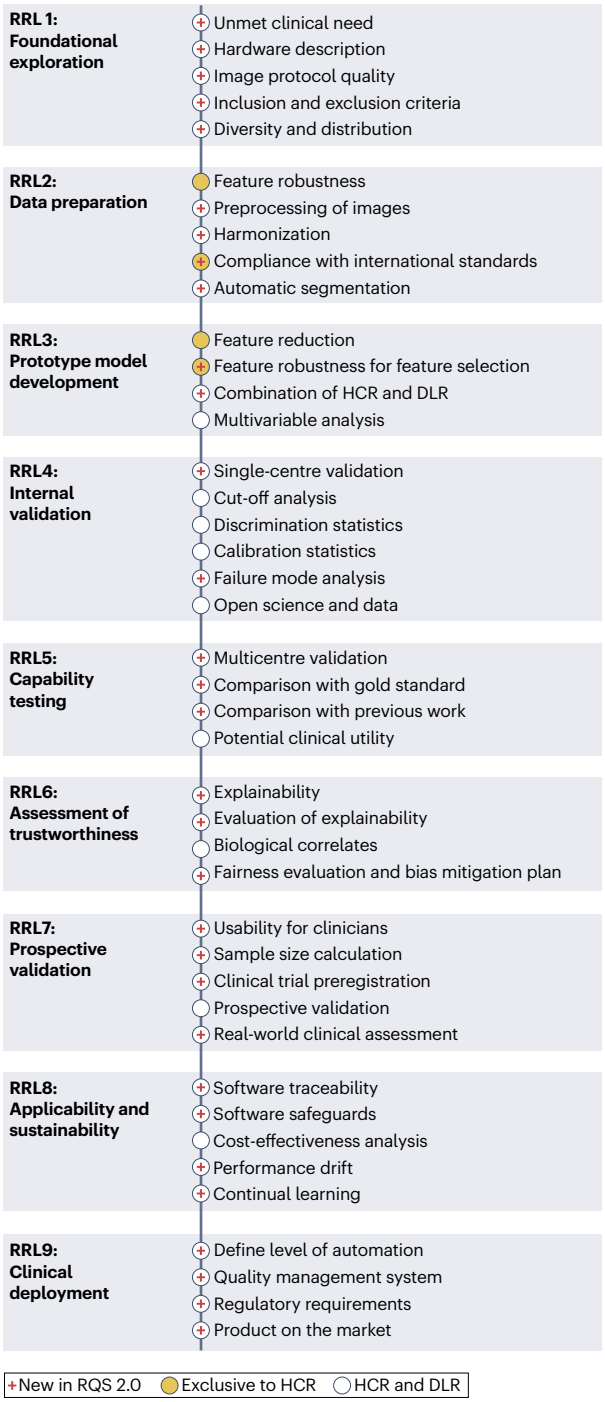

Fig. 2 | Incorporation of RRLs in RQS 2.0. Radiomics readiness levels (RRLs) are structured in analogy to the technology readiness levels[22,23] and encompass nine stages from initial exploration to full clinical deployment, including methodological refinement, model validation, interpretability and regulatory considerations. This flowchart presents the incorporation of RRLs in the Radiomics Quality Score (RQS) 2.0 and shows the necessary steps that RQS 2.0 rewards or penalizes to encourage best scientific practice. The radiomics workflow applies to handcrafted and deep learning-based radiomics approaches. DLR, deep learning-based radiomics; HCR, handcrafted radiomics.

domain[14]. Image domain methods modify images using deep learning techniques, such as generative adversarial networks (GANs) and style transfer to normalize data across scanners, acquisition settings or modalities. For example, CycleGANs have been used for unsupervised kernel conversion in CT and MRI harmonization across sites[49,50]. By contrast, feature domain harmonization adjusts extracted radiomic features using statistical methods such as ComBat, which removes scanner-related effects in addition to preserving biological variability[51].

AI-based segmentation algorithms can be used to define an ROI in handcrafted radiomics studies in addition to maintaining consistency and eliminating interobserver variability[52]. These algorithms, particularly those based on CNNs, have demonstrated high agreement with expert delineations[53–55]. Automated segmentation algorithms have also been proved to be beneficial for deep learning-based radiomics, in which cropping around the ROI focuses analysis by the CNN algorithm on relevant areas of medical images. This approach not only improves efficiency but also potentially increases accuracy by excluding irrelevant background information[56].

Development of prototype model

RRL3 focuses on methodological refinements for the development of proof-of-concept radiomics tools. In handcrafted radiomics, reducing the number of features is essential to minimize the risk of overfitting (which occurs when an algorithm fits the training data too closely and underperform on other datasets)[57,58]. In addition, radiomics features must be examined for correlations with certain features, such as volume and image noise, to reduce redundancy and prevent confounding effects[59,60]. The assessment of feature robustness and reproducibility, conducted at RRL2, should be incorporated into the feature selection phase. For example, assessing the reproducibility of features through repeated segmentation by various annotators enables the identification and exclusion of features that fall below a predefined threshold of intraclass correlation coefficients[61]. Handcrafted and deep learning-based radiomics can offer complementary insights into the clinical problem and investigating their synergistic combination through an ensemble or fusion methodology might be useful[62,63]. The incorporation of a diverse array of data, including clinical factors (such as tumour–node–metastasis (TNM) stage and age), genomics and proteomics data, to radiomics could result in a more holistic model. Such integrative approaches can enable an effective correlation between radiomics and non-radiomics features, enhancing the overall performance of the model[64,65].

Internal validation

RRL4 emphasizes in-depth model evaluation and validation, with a particular focus on thorough single-centre processes. Cut-off analysis has a crucial role in identifying the ideal threshold for various study types. This process involves the use of statistical methods, such as Youden's index, which are instrumental in establishing an appropriate operating point for classification tasks[66]. Moreover, each output value of a radiomics test should be mapped to a clear clinical action to ensure unambiguous interpretation. For example, risk stratification based on cut-off values can guide treatment intensification, de-escalation or continuation of standard care. Validation must be conducted without retraining the model or adjusting the cut-off value. Discriminative metrics, such as the area under the receiving operating characteristic curve, sensitivity and specificity, must be reported along with their statistical significance. Validation should also use resampling techniques, such as bootstrapping or cross-validation, to ensure the robustness and reliability of these statistical measures.

Furthermore, calibration statistics are essential to assess how well the probabilities of an event predicted by the model align with its actual occurrence[67]. The limitations of the model must be outlined, providing examples of edge-case scenarios to highlight potential failure points, thereby offering a comprehensive evaluation of the performance of the model. To drive advances in research and ensure transparency, code and data must be made publicly accessible. Indeed, model transparency facilitates knowledge sharing and fosters the reproducibility of studies[68].

Capability testing

RRL5 encompasses multicentre independent validation and comparative analysis. For external validation, using datasets from multiple institutions in addition to ensuring that the data from each institution has not been included in the training phase of the model is crucial to demonstrate the generalizability of radiomics models. Moreover, independent validation on external datasets is important for unbiased evaluation of the performance of the model. Such validation can be carried out by either uploading the model's predictions, generated without access to ground truth labels, to a third-party platform for scoring or, preferably, by uploading the model itself in a containerized format to the platform, which then executes the model on a hidden dataset and evaluates its performance automatically[69]. Furthermore, comparing the performance of a model against existing handcrafted and deep learning-based methodologies using an identical dataset is imperative to engage in incremental research. To evaluate the added value of radiomics, the performance of the model must also be compared with the current clinical standard, for example, TNM staging for prognosis. The potential clinical utility of the radiomics model should be described by outlining how its outputs could support specific clinical decisions in real-world practice. For example, a decision curve analysis can be used to weigh the risks and benefits of using a predictive radiomics model for guiding clinical decisions[70,71].

Assessment of trustworthiness

RRL6 incorporates the essential elements of explainability and fairness into radiomics studies. Research on the biological meaning of radiomics studies is needed to understand the relationship between radiomics signatures and their biological basis and thus foster clinical translation[72,73]. This research can provide insight on the association between radiomics and the complex area of genomics and proteomics[74]. Radiomics models involve intricate decision-making processes and, therefore, explainability approaches need to be adopted to provide insight into how decisions are made[16]. The application of explainability methods, such as Shapley additive explanations (SHAP) for handcrafted radiomics and gradient-weighted class activation mapping (Grad-CAM) for deep learning models, can help to unravel the reasoning process of these algorithms[75,76]. However, the explanations of AI models generated by certain explainability methods can be misleading because they might not accurately reflect the true behaviour of the model[77,78]. Therefore, these explanations must be assessed by correlating them with prior clinical knowledge and, ideally, validating them quantitatively using established imaging biomarkers. For example, researchers used clinically relevant metrics, such as the cardiothoracic ratio, to evaluate the output of a deep learning counterfactual explanations model developed for the classification of chest radiography[77,79]. Clinicians can perform analyses of explanations by evaluating multiple aspects, such as understandability and decision justification[79]. Quantitative analysis with metrics such as demographic parity and equalized odds is essential for assessing the fairness of radiomics models regarding biases identified at RRL1. This analysis helps to identify and correct biases to ensure model fairness[15].

Prospective validation

RRL7 marks a meaningful phase in the development and validation of radiomics as a dependable biomarker in clinical settings through prospective validation. The calculation of sample size for statistical power to ensure that the study is adequate to detect the anticipated effects is a crucial step before starting the validation process[80]. The prospective clinical validation study, including a statistical plan, should be preregistered in a clinical trial database such as ClinicalTrials.gov[81]. Any changes to the clinical trial protocol should all be tracked to ensure transparency and accountability. Prospective validation is conducted to confirm the clinical relevance and utility of the radiomics biomarker after training it on prospectively collected data. In silico trials (ISTs) can be used for this purpose, leveraging digital data for their speed and cost efficiency[82]. In addition to not considered the gold standard for prospective validation in oncology, ISTs offer tight control and the ability to simulate various scenarios with a known ground truth to address research queries and provide important preliminary insights[83]. For example, an IST might be used to assess how the recommendations from a radiomics tool would influence the predictions of a radiologist[84]. Human-in-the-loop assessments, which are workflows that incorporate human input in predictions from AI models, are integral for evaluating how radiomics models function in real-world clinical settings, focusing on their utility, safety and integration into routine workflows[85]. Evaluations for prospective validation range from early phase, small-cohort clinical studies to large-scale randomized controlled trials that assess effectiveness and safety across diverse scenarios.

Applicability and sustainability

RRL8 aims to evaluate the applicability and sustainability of radiomics tools in clinical settings, ensuring their effectiveness and long-term viability. Implementing a detailed traceability process throughout the lifecycle of the radiomics tool is essential to guarantee transparency and accountability. Such a process encompasses the development, modification and version control (that is, reporting of changes in the source code) of the radiomics software. The usability of the radiomics tool should be evaluated, with an emphasis on the user interface, workflow integration and ease of use. This assessment can be conducted using questionnaires, such as the System Usability Scale, or through user trials to evaluate effectiveness, efficiency and satisfaction[86,87]. Evaluating the radiomics tool in an actual clinical environment, including human-in-the-loop evaluations, is crucial; this approach will enable assessment of the extent to which the tool provides tangible benefits for patient care and supports the clinical workflow[85]. Cost-effectiveness analyses are necessary to determine the economic value of the tool in relation to clinical outcomes. Quantitative measures, such as quality-adjusted life years and incremental cost-effectiveness ratio, can be used to assess the potential financial viability of the radiomics tool if it was to be implemented in the clinic[88].

Over time, radiomics models can be affected by variations in image acquisition owing to hardware and software updates. Such data shifts can negatively influence the performance of the model, potentially leading to detrimental outcomes in clinical decision-making processes[89]. Automated recalibration approaches are crucial to maintain the performance of AI tools despite changes in data acquisition and reconstruction methods[17]. Therefore, researchers must define

strategies to manage performance drift to safeguard the consistency and reliability of radiomics tools[90]. A study on mammography-based breast cancer screening and histopathology data showed that performance drift can be mitigated using linear piecewise cumulative distribution matching[17]. Continual learning is an approach that aims at improving AI-based models by integrating new data in addition to maintaining previously acquired knowledge[91]. Continual learning is particularly advantageous for radiomics models, enabling them to evolve by learning from past mistakes and thereby enhancing their performance, although the introduction of new data poses potential challenges because it might lead to the incorporation of errors, biases and a potential decline in overall performance[92]. Continual learning can also help to address the problem of performance drift caused by evolving imaging protocols and data quality. A continual learning approach has been developed that utilizes dynamic memory to store a small, diverse subset of past training data and capture new styles in the continuous data stream using a style-based metric[93]. This approach helped to maintain model performance on previously learned domains in addition to adapting to evolving imaging scenarios[93]. In addition, the introduction of the new data increases the risk of cybersecurity threats such as adversarial attacks with the goal to manipulate the training data so that the model gives a substantially wrong output[94]. Adversarial attacks can contain visually imperceptible perturbations and pass manual or simple data checks[95]. Therefore, developing a strategy for incremental learning that includes thorough quality assurance measures is crucial to effectively use new data and enhance the performance of radiomics models.

Clinical deployment

RRL9, the final milestone in the clinical deployment of a radiomics solution, is focused on ensuring quality and obtaining regulatory compliance to successfully introduce the product to the market. A clear definition of the level of automation for the AI solution in clinical practice is needed. We propose six levels from level 0 or 'no automation', in which clinicians perform tasks without using a radiomics model, to level 5 or 'full automation', which refers to a model operating entirely without human intervention. Level 0 serves as a theoretical baseline, as it does not involve a model and is therefore not applicable to radiomics deployment in practice. Intermediate levels include level 1 or 'clinical assistance', in which the clinician uses the model predictions to perform part of the clinical task; level 2 or 'partial automation', in which the clinician considers predictions before making a final decision; level 3 or 'conditional automation', in which the model provides predictions under clinician supervision, permitting intervention at any time; and level 4 or 'high automation', in which the model provides predictions and clinician intervention is only required in exceptional cases.

Implementing and maintaining a Quality Management System, such as ISO 9001, IEC 62304 or an alternative depending on the available resources, is crucial for ensuring consistent quality and complying with regulatory standards throughout the radiomics workflow[96,97]. The radiomics solution must be evaluated for its alignment with the requirements of the chosen regulatory pathway, such as 510(k) clearance or premarket approval for the Food and Drug Administration (FDA) and conformity assessments for the European Medicines Agency (EMA) and European Union AI Act[98–100]. This evaluation ensures that the radiomics solution adheres to all essential safety, efficacy and regulatory standards.

Regulatory approval faces substantial challenges owing to inconsistencies in standards across major agencies, such as the FDA, EMA and others, whereby each regulator applies different frameworks for classifying software as a medical device and defining clinical evidence requirements[101]. Navigating these divergent regulatory pathways, with aspects such a sharp contrast between the 510(k) clearance process and the marking requirements for the Conformité Européenne under the EMA Medical Device Regulation and the emerging European Union AI Act, is a major challenge that creates substantial compliance burdens[99,102]. The rapidly evolving regulatory landscape, in which agencies continue to refine their approaches to AI and/or machine learning-based medical software, adds further complexity, particularly in relation to continual learning and appropriate validation protocols[92]. Moreover, international differences in regulations relating to data privacy and sharing requirements, such as the stricter controls imposed by General Data Protection Regulation in Europe relative to those from the Health Insurance Portability and Accountability Act in the USA, further complicate the development and validation of radiomics models across multiple jurisdictions[103,104].

The market launch of the radiomics product marks the culmination of RRL9. A radiomics solution that has reached this RRL complies with regulatory standards and is acknowledged and can be utilized by healthcare professionals in clinical settings. The postmarket surveillance process involves human oversight and continuous monitoring of radiomics tools to assess their performance in real-world clinical scenarios and to identify any issues that might not have been evident during the initial testing of RRLs.

Interpretation of RQS 2.0 and RRLs

We have developed a web-based scoring tool to assess radiomics studies using the RQS 2.0 and aligned RRLs. This tool allows users to select a target RRL and calculate the percentage of RQS compliance up to that level. The tool also generates visual summaries showing cumulative scores across RRL levels and the proportion of criteria fulfilled at each stage (Fig. 3). These visualizations provide valuable insights not only into the total compliance up to a selected RRL level but also regarding which specific criteria are met or missed at each stage. Given that each RRL reflects a key milestone in the path to clinical translation, this layered view helps research teams to identify gaps and prioritize improvements either within a particular RRL or incrementally across levels. This approach encourages a structured progression of radiomics research, supporting the development of studies that advance stepwise towards clinical application.

We evaluated the RQS 2.0 of a particular study focused on predicting *EGFR* mutations and Ki-67 proliferation index in non-small-cell lung cancer, which had a reported RQS 1.0 score of 24 of 36 (ref. 105) (Fig. 3 and Supplementary Table 1). The RQS 2.0 tool was used to assess compliance up to RRL9. The study achieved high scores across the early readiness levels (RRL1–RRL5) but did not meet any criteria in RRL7–RRL9, which pertain to prospective validity, applicability and sustainability, and clinical deployment, respectively. At RRL6, the study only met one out of five points related to ethical considerations, indicating a need for further work in areas such as explainability and fairness assessment. At RRL5, the study had a compliance of 65% (22 out of 34 points). Although this score indicates robust performance in the early to middle stages of development, the radiomics model remains far from clinical integration, with an RRL9 compliance of only 41% (23 out of 56 points). This example highlights the value of interpreting RQS scores in the context of RRLs, which provides a more accurate picture of the clinical readiness of a study and helps to prevent overestimating its applicability based solely on the total score.

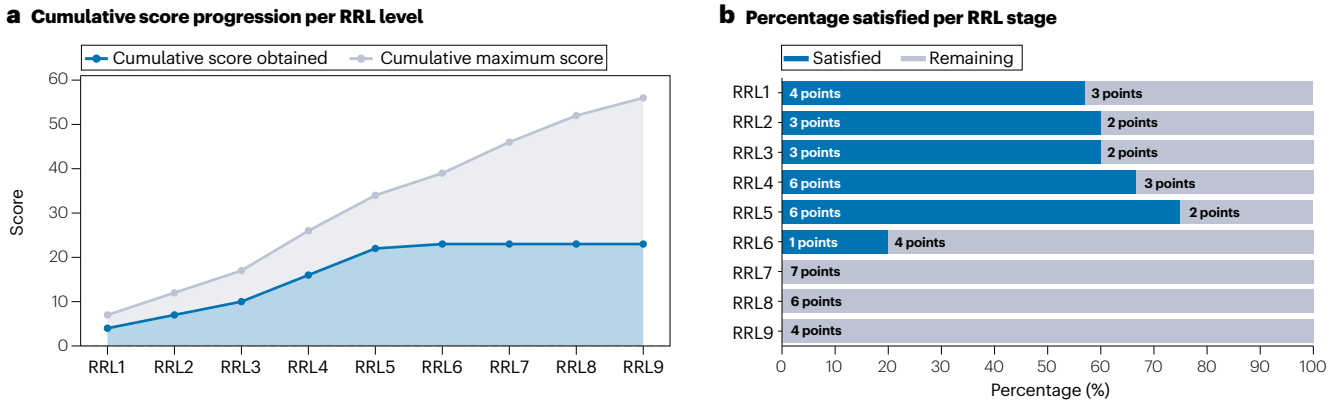

Fig. 3 | Assessment of RQS 2.0 for a research study. a, Cumulative Radiomics Quality Score (RSQ) progression by radiomics readiness level (RRL) level. **b**, The percentage of criteria satisfied per RRL level. The RQS 2.0 was calculated in alignment with the RRLs for a particular study[105].

The way forward for radiomics

The future of radiomics is set for substantial advances with the incorporation of foundation models and multiomics, which will continue to enhance diagnostic accuracy and further enable the delivery of personalized patient care. This evolution will be facilitated by a thorough clinical evaluation framework, federated learning and synthetic data generation to address data-sharing challenges, and the standardization of data via common data models (CDMs).

Multiomics integration

The integration of radiomics with diverse omics data, including genomics, proteomics, transcriptomics, epigenomics, microbiomics, metabolomics and pathomics, facilitates the creation of comprehensive models that offer a holistic view of the biological landscape of tumours. For example, a radiopathomics model has been developed for the prediction of pathological complete response in patients with locally advanced rectal cancer using data from pretreatment MRI and haematoxylin and eosin-stained biopsy-derived tissue[106]. Novel deep learning architectures that integrate multiomics information outperformed traditional models in the prediction of the outcomes of patients with non-small-cell lung cancer after radiotherapy[107]. Many studies have confirmed the relationship between radiomics features and cellular or molecular features, indicating the potential application of a non-invasive radiomics approach for visualizing molecular functions[108,109]. Such an approach can help to establish novel diagnostic, prognostic and predictive biomarkers that can be assessed non-invasively. Multiomics integration provides clinicians with unprecedented insights into the intricate mechanisms of biological processes at both the mesoscopic and microscopic scales. For example, intratumour heterogeneity can be assessed through dynamic contrast enhanced MRI radiomic features and validated using genomic, transcriptomic, metabolomic and digital pathological data[110]; a radiogenomic signature has been developed by mapping radiomics features to genomic subclones correlated with intratumoural heterogeneity[111].

In the era of precision oncology, the investigation of multiomics data assisted by AI tools, such as machine learning and deep learning, has the potential to revolutionize cancer subtyping, risk stratification, prognostication, prediction and clinical decision-making. However, translating multiomics research into clinical practice necessitates consistent efforts to standardize the collection and analysis of omics data, build computational infrastructure for data storage and sharing, develop advanced methods for data fusion and interpretability, and ultimately, conduct large-scale prospective clinical trials to bridge the gap between research findings and clinical benefit[112].

Foundation models based on large-scale data

Foundation models are a category of AI tools typically trained through self-supervised learning on extensive datasets[113]. The use of large-scale data ensures that foundation models can capture a wide range of variations and complexities present in medical data. Therefore, foundation models can be fine-tuned for various downstream tasks and serve as the bedrock upon which advanced radiomic tools will be built. A notable advantage of foundation models lies in their ability to make zero-shot or few-shot predictions (that is, on the basis of variables that they have not been exposed to during training) for a new task without requiring additional training on an annotated dataset[114], diminishing the need for extensive data annotation efforts. Moreover, foundation models have demonstrated proficiency in managing multimodal datasets, including processing both image and text[115], which can be obtained from medical images and electronic health records (EHRs).

Currently, large language models are the primary focus of research on foundation models but other domains are also receiving attention. Visual language foundation models (VLMs), for example, specialize in acquiring visual representations from extensive image datasets. These learned representations are then applied to subsequent tasks in computer vision or vision–language processing. Examples of foundation models include the segment anything model[116] and contrastive language–image pretraining[117]. Variants of the segment anything model[54] and VLMs[118] have also been introduced into medical image analysis. In addition, VLMs based on contrastive language–image pretraining have been proposed for zero-shot diagnosis of diseases across various medical specialties[114,115] and some multimodal medical foundation models have also been developed for this purpose[119,120]. For example, a foundation model trained on CT images was found to be more stable to input variations and outperformed conventional pretraining and supervised strategies[121]. Foundation models have shown great potential in transforming vast amounts of medical information into a valuable resource for AI models. This transformation enhances the effectiveness of AI models and facilitates knowledge sharing.

Comprehensive clinical evaluation system for radiomics

Over the past decades, radiomics has demonstrated notable potential efficacy in various clinical tasks as predicted by preclinical in silico studies but limited high-quality evidence from clinical studies is available on whether radiomics improves clinician performance or patient outcomes[122]. Therefore, establishing a comprehensive evaluation system to assess the role of radiomics in real-world clinical scenarios is imperative.

The clinical evaluation framework for radiomics models should include multiple stages. The first stage is in silico evaluation, which is predicated around assessing model performance[123] and diagnostic accuracy[124] in controlled, non-clinical settings, focusing on evaluating the capabilities of an AI model before its introduction into real-world clinical environments. The second stage is early clinical evaluation, which involves the initial assessment of AI systems as interventions in actual clinical settings, albeit on a small scale. This stage primarily focuses on analyses of clinical feasibility, safety considerations and human factors influencing the integration of AI into clinical workflows[125,126]. The final stage is comparative prospective clinical evaluation, which entails comprehensive summative evaluation through large-scale randomized controlled trials. The primary metrics for assessment should be effectiveness and safety to provide robust insight on the effect of AI-based interventions in diverse clinical scenarios. RRL7 highlights the need for comprehensive clinical evaluation of radiomics tools in a prospective manner. The usability assessment and in silico prospective validation criteria established in RRL7 correspond to the first stage in our proposed clinical evaluation framework, whereas the real-world clinical assessment criteria (RRL7) address the second and third stages.

By systematically progressing through these evaluation stages, researchers can thoroughly understand the performance of radiomics tools in real-world clinical settings. This approach not only addresses the technical aspects of AI but also considers the practical implications and human factors crucial for successful integration into healthcare practices.

Generation of synthetic data

The robustness and generalizability of AI models, including deep neural networks (DNNs), are closely related to the quantity and quality of training data; however, technical, legal and ethical considerations often prevent clinical centres from easily sharing their data[112]. Generative models, a class of AI models that typically use DNNs such as GANs and diffusion models can address issues related to data scarcity[127]. Conditional generative models are adopted to synthesize a substantial number of medical images with corresponding labels to aid segmentation network generalization and adaptation in multivendor, multidisease scenarios, making them more representative of real-world clinical settings[128,129]. GANs can be conditioned by an under-represented data category to generate synthetic images that can help in addressing problems associated with class imbalance and dataset bias[130]. Two studies have revealed that augmenting real-world data with synthetic samples enhances model robustness across different medical tasks and promotes fairness by boosting diagnostic accuracy in under-represented groups, particularly in out-of-distribution cases[131,132]. Synthetic images can also be used to pretrain models and enhance their performance. For example, GAN-generated nodule images have been used to pretrain CNNs, resulting in improved accuracy in the classification of lung nodules from CT images[133]. Generative models can also be used to harmonize multicentre medical imaging data by mapping diverse datasets to a common reference domain, which improves cross-site generalization and performance of downstream tasks, such as prediction of brain age and disease classification[48,134].

Synthetic images generated from the training dataset can serve as a substitute for real data, and the legal considerations for sharing such synthetic data are less stringent[135,136]. However, generative models can memorize and reproduce training data during the generation process, which could compromise privacy by making models vulnerable to data leakage through adversarial attacks[137]. Therefore, further research is necessary to ensure that the generation of synthetic images effectively preserves privacy[138].

Multimodal images can be crucial for clinical tasks and, in some scenarios, one of the modalities might be missing[139]. For instance, DNNs can be used to synthesize pseudo-CT from T1-weighted MRI to enable accurate PET–magnetic resonance attenuation correction[140], generate CT from MRI to support organ segmentation[141] and impute absent MRI sequences such as T1-weighted and fluid attenuated inversion recovery to preserve brain-tumour segmentation accuracy[142]. Moreover, generative models are prone to hallucinations and can introduce artefacts or remove features in the synthetic images[143]. Therefore, integrating both qualitative and quantitative performance evaluations within synthetic image generation is essential to ensure the reliability of such images. A systematic review of generative AI models developed to synthesize various types of medical data revealed a major gap in the generation of data for purposes beyond augmentation, such as validation and evaluation of medical AI models[144]. The systematic review also emphasized that the absence of standardized evaluation methodologies for medical images hinders clinical applications, highlighting the need for comprehensive evaluation approaches and collaborative benchmarking[144].

Federated learning

Data-driven radiomics studies require extensive datasets from multiple centres to mitigate overfitting and improve model generalizability. Federated learning has been proposed as an effective method to address challenges derived from patient data sharing and foster collaboration among centres in addition to upholding governance and data privacy[145,146]. This approach involves training machine learning models within multiple nodes, each on their local dataset[147]. Thus, external sharing of data from patients is avoided, and clients only share derivative data, such as model updates (for example, coefficients and weight parameters), to develop the final model.

Federated learning has been used for several radiology tasks performed on CT and MRI images for the detection of coronavirus disease 2019 and brain tumour segmentation, showing great promise in achieving performance equivalent to that in a centralized setting[148,149]. However, the theoretical formulation and practical implementation of these approaches pose numerous challenges. The first challenge is data heterogeneity, in which different data distributions among centres deteriorate the accuracy of the federated learning model[150]. Data harmonization before processing might help to alleviate this effect. Further technical studies should be carried out to find the optimum technique for updating the central model with heterogeneous data. The second challenge is bias, whereby the machine learning model is inclined towards a particular node owing to the data distribution or size. The third challenge is the lack of access to standardized data, which refers to data inconsistencies such as non-uniform DICOM tags, annotation schemas or feature nomenclature that prevent straightforward integration of data across sites. Many institutions lack the infrastructure to process images following a standardized imaging

pipeline. Moreover, a universal method to organize and manage other data, such as EHRs, does not exist. In addition, data privacy and security issues remain a concern associated with federated learning[151]. Healthcare institutions also need cloud-based or on-premises computational facilities and robust network connections, which are still limited in some centres, for model training and data transfer.

CDM

The lack of standardization makes it difficult to conduct federated learning among different centres. Data from distinct centres originate from diverse sources or varied technologies. Hence, metadata might be missing, incorrect or non-harmonized, and key information could be difficult to identify for federated learning. The CDM structure, which applies the same data structure to run an identical analysis code for each data holder, has been developed for data standardization[152]. The Observational Medical Outcomes Partnership (OMOP) is a CDM developed by the Observational Health Data Sciences and Informatics consortium as a repository of all vocabularies used in the community, for their standardization and mapping of their use in research[153]. OMOP has the potential to address many challenges associated with the representation and use of data from EHRs.

OMOP advances the field of data integration and standardization; however, it cannot fully cover all aspects of medical data management without task-tailored extensions. Radiology (R)-CDM is an extension of OMOP developed for the standardization of data from digital imaging and communications in medicine (DICOM), a widely adopted standard for storing and transmitting medical imaging information[154]. R-CDM has achieved standardization in the extract-transform-load process, a data processing pipeline in which raw data are extracted from source systems, transformed into a standardized format and then loaded into a structured database. In addition, R-CDM was designed to be linked with OMOP to achieve a seamless link between data from EHRs and medical imaging, which might help in large-scale radiomics research by enabling the collection and completeness of medical imaging data from multiple institutions. Although R-CDM effectively standardizes DICOM metadata and enables linkage between imaging and clinical data, it does not explicitly capture detailed radiomic features. Some other CDMs such as the DICOM structured reporting and the annotation and image markup scheme have the potential to support structured representation of radiomic features[155,156]. Genomic CDM is another extension of OMOP for effective integration of genomic data with standardized clinical data, permitting data sharing across institutes[157].

Conclusion

Radiomics holds promise for revolutionizing clinical decision-making in oncology but must overcome substantial hurdles for effective clinical adoption. RQS is a widely adopted benchmark for assessing the quality of radiomics studies. We have now updated this score and present RQS 2.0, which includes considerations and guidelines to address contemporary challenges in the field, aiming to enhance the quality of radiomics studies and facilitate their clinical translation. Moreover, we propose nine RRLs seamlessly integrated into RQS 2.0 to promote step-by-step improvement of radiomics tools. Researchers can now develop tools that align with specific RRLs, with research quality assessed accordingly up to the targeted level. Researchers can also use a web-based scoring tool to quantitatively assess radiomics studies aligned with specific RRLs, facilitating empirical validation and providing insights into their methodological rigour and translational readiness for clinical application. By making incremental improvements to reach RRL9, researchers can enhance the quality of their work for clinical deployment.

Published online: 03 September 2025

References

1. Larry Jameson, J. & Longo, D. L. Precision medicine — personalized, problematic, and promising. *Obstet. Gynecol. Surv.* **70**, 612 (2015).
2. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
3. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
4. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).
5. Ding, H. et al. Radiomics in oncology: a 10-year bibliometric analysis. *Front. Oncol.* **11**, 689802 (2021).
6. Hosny, A., Aerts, H. J. & Mak, R. H. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *Lancet Digit. Health* **1**, e106–e107 (2019).
7. Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A. & Benali, H. From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. *IEEE Signal. Process. Mag.* **36**, 132–160 (2019).
8. Jiang, X., Hu, Z., Wang, S. & Zhang, Y. Deep learning for medical image-based cancer diagnosis. *Cancers* **15**, 3608 (2023).
9. Barry, N. et al. Evaluating the impact of the radiomics quality score: a systematic review and meta-analysis. *Eur. Radiol.* **35**, 1701–1713 (2025).
10. Spadarella, G. et al. Systematic review of the radiomics quality score applications: an EuSoMII radiomics auditing group initiative. *Eur. Radiol.* **33**, 1884–1894 (2023).
11. Park, C. J. et al. Quality of radiomics research on brain metastasis: a roadmap to promote clinical translation. *Korean J. Radiol.* **23**, 77–88 (2022).
12. Li, H., El Naqa, I. & Rong, Y. Current status of radiomics for cancer management: challenges versus opportunities for clinical practice. *J. Appl. Clin. Med. Phys.* **21**, 7–10 (2020).
13. Huang, E. P. et al. Criteria for the translation of radiomics into clinically useful tests. *Nat. Rev. Clin. Oncol.* **20**, 69–82 (2023).
14. Mali, S. A. et al. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *J. Pers. Med.* **11**, 842 (2021).
15. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
16. 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 (2021).
17. Roschewitz, M. et al. Automatic correction of performance drift under acquisition shift in medical image classification. *Nat. Commun.* **14**, 6608 (2023).
18. Martínez-Plumed, F., Gómez, E. & Hernández-Orallo, J. Futures of artificial intelligence through technology readiness levels. *Telemat. Inform.* **58**, 101525 (2021).
19. Hillis, J. M. et al. The lucent yet opaque challenge of regulating artificial intelligence in radiology. *NPJ Digit. Med.* **7**, 69 (2024).
20. Veale, M. & Borgesius, F. Z. Demystifying the draft EU artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach. *Comput. Law Rev. Int.* **22**, 97–112 (2021).
21. Lekadir, K. et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025).
22. Mankins, J. C. *Technology Readiness Levels: A White Paper* (Office of Space Access and Technology, NASA, 6 April 1995).
23. Kimmel, W. M. et al. *Technology Readiness Assessment Best Practices Guide. NASA Special Publication SP-20205003605* (NASA, 30 June 2020).
24. Lavin, A. et al. Technology readiness levels for machine learning systems. *Nat. Commun.* **13**, 6039 (2022).
25. Kocak, B. et al. CheckList for evaluation of radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging* **14**, 75 (2023).
26. Kocak, B. et al. Assessment of RadiomIcS rEsearch (ARISE): a brief guide for authors, reviewers, and readers from the scientific editorial board of European radiology. *Eur. Radiol.* **33**, 7556–7560 (2023).
27. Kocak, B. et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging* **15**, 8 (2024).
28. European Commission. Commission Regulation (EC) No 507/2006 of 29 March 2006 on the conditional marketing authorisation for medicinal products for human use falling within the scope of Regulation (EC) No 726/2004 of the European Parliament and of the Council. *Off. J. Eur. Union* **50**, 6–9 (2006).
29. Ab Latif, R., Mohamed, R., Dahlan, A. & Mat Nor, M. Z. Using Delphi technique: making sense of consensus in concept mapping structure and multiple choice questions (MCQ). *Educ. Med. J.* **8**, 89–98 (2016).
30. Cobo, M., Menéndez Fernández-Miranda, P., Bastarrika, G. & Lloret Iglesias, L. Enhancing radiomics and deep learning systems through the standardization of medical imaging workflows. *Sci. Data* **10**, 732 (2023).

31. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
32. Sollini, M., Antunovic, L., Chiti, A. & Kirienko, M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *Eur. J. Nucl. Med. Mol. Imaging* **46**, 2656–2672 (2019).
33. Zou, J. & Schiebinger, L. Ensuring that biomedical AI benefits diverse populations. *eBioMedicine* **67**, 103358 (2021).
34. Galadima, H. et al. Machine learning as a tool for early detection: a focus on late-stage colorectal cancer across socioeconomic spectrums. *Cancers* **16**, 540 (2024).
35. Hatt, M. et al. Joint EANM/SNMMI guideline on radiomics in nuclear medicine: jointly supported by the EANM physics committee and the SNMMI physics, instrumentation and data sciences council. *Eur. J. Nucl. Med. Mol. Imaging* **50**, 352–375 (2023).
36. Chen, Y. et al. Robustness of CT radiomics features: consistency within and between single-energy CT and dual-energy CT. *Eur. Radiol.* **32**, 5480–5490 (2022).
37. Haarburger, C. et al. Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. *Sci. Rep.* **10**, 12688 (2020).
38. Alsyed, E., Smith, R., Bartley, L., Marshall, C. & Spezi, E. A heterogeneous phantom study for investigating the stability of PET images radiomic features with varying reconstruction settings. *Front. Nucl. Med.* **3**, 1078536 (2023).
39. Schwier, M. et al. Repeatability of multiparametric prostate MRI radiomics features. *Sci. Rep.* **9**, 9441 (2019).
40. Zhang, J. et al. Comparing effectiveness of image perturbation and test retest imaging in improving radiomic model reliability. *Sci. Rep.* **13**, 18263 (2023).
41. Tixier, F., Um, H., Young, R. J. & Veeraraghavan, H. Reliability of tumor segmentation in glioblastoma: impact on the robustness of MRI-radiomic features. *Med. Phys.* **46**, 3582–3591 (2019).
42. Mahmood, U. et al. Quality control of radiomic features using 3D-printed CT phantoms. *J. Med. Imaging* **8**, 033505 (2021).
43. Demircioğlu, A. The effect of preprocessing filters on predictive performance in radiomics. *Eur. Radiol. Exp.* **6**, 40 (2022).
44. Heidari, M. et al. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int. J. Med. Inform.* **144**, 104284 (2020).
45. Salvi, M., Acharya, U. R., Molinari, F. & Meiburger, K. M. The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. *Comput. Biol. Med.* **128**, 104129 (2021).
46. Zwanenburg, A. et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
47. Whybra, P. et al. The image biomarker standardization initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology* **310**, e231319 (2024).
48. Bashyam, V. M. et al. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *J. Magn. Reson. Imaging* **55**, 908–916 (2022).
49. Yang, S., Kim, E. Y. & Ye, J. C. Continuous conversion of CT kernel using switchable CycleGAN with AdaIN. *IEEE Trans. Med. Imaging* **40**, 3015–3029 (2021).
50. Roca, V. et al. IGUANe: A 3D generalizable CycleGAN for multicenter harmonization of brain MR images. *Med. Image Anal.* **99**, 103388 (2025).
51. Orlhac, F., Frouin, F., Nioche, C., Ayache, N. & Buvat, I. Validation of A method to compensate multicenter effects affecting CT radiomics. *Radiology* **291**, 53–59 (2019).
52. Jin, J. et al. The accuracy and radiomics feature effects of multiple U-net-based automatic segmentation models for transvaginal ultrasound images of cervical cancer. *J. Digit. Imaging* **35**, 983–992 (2022).
53. Primakov, S. P. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat. Commun.* **13**, 3423 (2022).
54. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
55. Andrearczyk, V. et al. Automatic head and neck tumor segmentation and outcome prediction relying on FDG-PET/CT images: findings from the second edition of the HECKTOR challenge. *Med. Image Anal.* **90**, 102972 (2023).
56. Salahuddin, Z. et al. From head and neck tumour and lymph node segmentation to survival prediction on PET/CT: an end-to-end framework featuring uncertainty, fairness, and multi-region multi-modal radiomics. *Cancers* **15**, 1932 (2023).
57. Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. *Learning from Data: A Short Course* (AMLBook.com, 2012).
58. Hua, J., Xiong, Z., Lowey, J., Suh, E. & Dougherty, E. R. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* **21**, 1509–1515 (2005).
59. Roy, S. et al. Optimal co-clinical radiomics: sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging. *eBioMedicine* **59**, 102963 (2020).
60. Volpe, S. et al. Impact of image filtering and assessment of volume-confounding effects on CT radiomic features and derived survival models in non-small cell lung cancer. *Transl. Lung Cancer Res.* **11**, 2452–2463 (2022).
61. Arthur, A. et al. A CT-based radiomics classification model for the prediction of histological type and tumour grade in retroperitoneal sarcoma (RADSARC-R): a retrospective multicohort analysis. *Lancet Oncol.* **24**, 1277–1286 (2023).
62. Refaee, T. et al. Diagnosis of idiopathic pulmonary fibrosis in high-resolution computed tomography scans using a combination of handcrafted radiomics and deep learning. *Front. Med.* **9**, 915243 (2022).
63. Beuque, M. P. L. et al. Combining deep learning and handcrafted radiomics for classification of suspicious lesions on contrast-enhanced mammograms. *Radiology* **307**, e221843 (2023).
64. Hatamikia, S. et al. Ovarian cancer beyond imaging: integration of AI and multiomics biomarkers. *Eur. Radiol. Exp.* **7**, 50 (2023).
65. Kang, W. et al. Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis. *J. Transl. Med.* **21**, 598 (2023).
66. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
67. Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
68. Park, J. E., Park, S. Y., Kim, H. J. & Kim, H. S. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J. Radiol.* **20**, 1124–1137 (2019).
69. Armato, S. G. 3rd, Drukker, K. & Hadjiiski, L. AI in medical imaging grand challenges: translation from competition to research benefit and patient care. *Br. J. Radiol.* **96**, 20221152 (2023).
70. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574 (2006).
71. Wang, H. et al. Radiomics-clinical model based on 99mTc-MDP SPECT/CT for distinguishing between bone metastasis and benign bone disease in tumor patients. *J. Cancer Res. Clin. Oncol.* **149**, 13353–13361 (2023).
72. Grossmann, P. et al. Defining the biological basis of radiomic phenotypes in lung cancer. *eLife* **6**, e23421 (2017).
73. Zhang, G. et al. Biological and clinical significance of radiomics features obtained from magnetic resonance imaging preceding pre-carbon ion radiotherapy in prostate cancer based on radiometabolomics. *Front. Endocrinol.* **14**, 1272806 (2023).
74. Tomaszewski, M. R. & Gillies, R. J. The biological meaning of radiomic features. *Radiology* **299**, E256 (2021).
75. Wang, Y. et al. The radiomic-clinical model using the SHAP method for assessing the treatment response of whole-brain radiotherapy: a multicentric study. *Eur. Radiol.* **32**, 8737–8747 (2022).
76. Zhang, Y. et al. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J. Neurosci. Methods* **353**, 109098 (2021).
77. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. & Kim, B. Sanity checks for saliency maps. In *Proc. Advances in Neural Information Processing Systems 31* (eds Bengio, S. et al.) (NeurIPS, Canada, 2018).
78. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
79. Singla, S., Eslami, M., Pollack, B., Wallace, S. & Batmanghelich, K. Explaining the black-box smoothly-a counterfactual approach. *Med. Image Anal.* **84**, 102721 (2023).
80. Baeza-Delgado, C. et al. A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data. *Eur. Radiol. Exp.* **6**, 22 (2022).
81. Trofimova, A. V. & Bluemke, D. A. Prospective clinical trial registration: a prerequisite for publishing your results. *Radiology* **302**, 1–2 (2022).
82. Badano, A. In silico imaging clinical trials: cheaper, faster, better, safer, and more scalable. *Trials* **22**, 64 (2021).
83. Abadi, E. et al. Virtual clinical trials in medical imaging: a review. *J. Med. Imaging* **7**, 042805 (2020).
84. Boverhof, B.-J. et al. Radiology AI deployment and assessment rubric (RADAR) to bring value-based AI into radiological practice. *Insights Imaging* **15**, 34 (2024).
85. Bodén, A. C. S. et al. The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice. *Histopathology* **79**, 210–218 (2021).
86. Lewis, J. R. The system usability scale: past, present, and future. *Int. J. Hum. Comput. Interact.* **34**, 577–590 (2018).
87. Nielsen, J. *Usability Engineering* (Morgan Kaufmann, 1994).
88. Di Pilla, A. et al. A cost-effectiveness analysis of an integrated clinical-radiogenomic screening program for the identification of BRCA 1/2 carriers (e-PROBE study). *Sci. Rep.* **14**, 928 (2024).
89. Jenkins, D. A. et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn. Progn. Res.* **5**, 1 (2021).
90. Feng, J. et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit. Med.* **5**, 66 (2022).
91. Lee, C. S. & Lee, A. Y. Clinical applications of continual learning machine learning. *Lancet Digit. Health* **2**, e279–e281 (2020).
92. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit. Health* **3**, e337–e338 (2021).
93. Perkonigg, M. et al. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nat. Commun.* **12**, 1–12 (2021).
94. Sorin, V. et al. Adversarial attacks in radiology - a systematic review. *Eur. J. Radiol.* **167**, 111085 (2023).
95. Dong, J., Chen, J., Xie, X., Lai, J. & Chen, H. Adversarial attack and defense for medical image analysis: methods and applications. *ACM Comput. Surv.* **57**, 79 (2024).
96. Sutherland, E. *QMS Manual ISO9001* (Eric Sutherland T/A Trog Associates, 2007).
97. Rust, P., Flood, D. & McCaffery, F. Creation of an IEC 62304 compliant software development plan. *J. Softw. Evol. Process.* **28**, 1005–1010 (2016).

98. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
99. Muehlematter, U. J., Daniore, P. & Vokinger, K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit. Health* **3**, e195–e203 (2021).
100. Vemula, A. *EU AI Act Explained: A Guide to the Regulation of Artificial Intelligence in Europe* (Anand Vemula, 2024).
101. Larson, D. B. et al. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. *J. Am. Coll. Radiol.* **18**, 413–424 (2021).
102. Aboy, M., Minssen, T. & Vayena, E. Navigating the EU AI Act: implications for regulated digital medical products. *NPJ Digit. Med.* **7**, 1–6 (2024).
103. Cohen, I. G., Evgeniou, T., Gerke, S. & Minssen, T. The European artificial intelligence strategy: implications and challenges for digital health. *Lancet Digit. Health* **2**, e376–e379 (2020).
104. Pesapane, F., Volonté, C., Codari, M. & Sardanelli, F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* **9**, 745–753 (2018).
105. Dong, Y. et al. Development and validation of novel radiomics-based nomograms for the prediction of EGFR mutations and Ki-67 proliferation index in non-small cell lung cancer. *Quant. Imaging Med. Surg.* **12**, 2658–2671 (2022).
106. Feng, L. et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *Lancet Digit. Health* **4**, e8–e17 (2022).
107. Cui, S., Ten Haken, R. K. & El Naqa, I. Integrating multiomics information in deep learning architectures for joint actuarial outcome prediction in non-small cell lung cancer patients after radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **110**, 893–904 (2021).
108. Fan, M. et al. Radiogenomic analysis of cellular tumor-stroma heterogeneity as a prognostic predictor in breast cancer. *J. Transl. Med.* **21**, 851 (2023).
109. 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).
110. Su, G.-H. et al. Radiogenomic-based multiomic analysis reveals imaging intratumor heterogeneity phenotypes and therapeutic targets. *Sci. Adv.* **9**, eadf0837 (2023).
111. Fan, M., Xia, P., Clarke, R., Wang, Y. & Li, L. Radiogenomic signatures reveal multiscale intratumour heterogeneity associated with biological functions and survival in breast cancer. *Nat. Commun.* **11**, 4861 (2020).
112. Wei, L. et al. Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. *Br. J. Radiol.* **96**, 20230211 (2023).
113. Jung, K.-H. Uncover this tech term: foundation model. *Korean J. Radiol.* **24**, 1038–1041 (2023).
114. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
115. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
116. Kirillov, A. et al. Segment anything. *In Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003 (IEEE, France, 2023).
117. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) 139, 8748–8763 (PMLR, 2021).
118. Zhou, J. et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 50043 (2024).
119. Zhang, K. et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. Preprint at https://doi.org/10.48550/arXiv.2305.17100 (2023).
120. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Preprint at https://doi.org/10.48550/arXiv.2308.02463 (2023).
121. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
122. Freeman, K. et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* **374**, n1872 (2021).
123. Collins, G. S. et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open.* **11**, e048008 (2021).
124. Sounderajah, V. et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. *Nat. Med.* **26**, 807–808 (2020).
125. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
126. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit. Health* **2**, e537–e548 (2020).
127. Kazerouni, A. et al. Diffusion models in medical imaging: a comprehensive survey. *Med. Image Anal.* **88**, 102846 (2023).
128. Amirrajab, S. et al. Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. *Comput. Med. Imaging Graph.* **101**, 102123 (2022).
129. Al Khalil, Y. et al. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Med. Image Anal.* **84**, 102688 (2023).
130. Li, X., Cui, Z., Wu, Y., Gu, L. & Harada, T. Estimating and improving fairness with adversarial learning. Preprint at https://doi.org/10.48550/arXiv.2103.04243 (2021).
131. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).
132. Ktena, I. et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30**, 1166–1173 (2024).
133. Onishi, Y. et al. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. *Int. J. Comput. Assist. Radiol. Surg.* **15**, 173–178 (2020).
134. Liu, S. & Yap, P.-T. Learning multi-site harmonization of magnetic resonance images without traveling human phantoms. *Commun. Eng.* **3**, 1–10 (2024).
135. Osuala, R. et al. Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging. *Med. Image Anal.* **84**, 102704 (2023).
136. Rajotte, J.-F. et al. Synthetic data as an enabler for machine learning applications in medicine. *iScience* **25**, 105331 (2022).
137. Carlini, N. et al. Extracting training data from diffusion models. In *Proc. 32nd USENIX Security Symposium* 5253–5270 (USENIX Security, USA, 2023).
138. Sun, C., van Soest, J. & Dumontier, M. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *J. Biomed. Inform.* **143**, 104404 (2023).
139. Lyu, Q. & Wang, G. Conversion between CT and MRI images using diffusion and score-matching models. Preprint at https://doi.org/10.48550/arXiv.2209.12104 (2022).
140. Liu, F., Jang, H., Kijowski, R., Bradshaw, T. & McMillan, A. B. Deep learning MR imaging-based attenuation correction for PET/MR imaging. *Radiology* **286**, 676–684 (2018).
141. Huo, Y. et al. SynSeg-net: synthetic segmentation without target modality ground truth. *IEEE Trans. Med. Imaging* **38**, 1016–1025 (2018).
142. Conte, G. M. et al. Generative adversarial networks to synthesize missing T1 and FLAIR MRI sequences for use in a multisequence brain tumor segmentation model. *Radiology* **299**, 313–323 (2021).
143. Wolterink, J. M., Kamnitsas, K., Ledig, C. & Išgum, I. in *Handbook of Medical Image Computing and Computer Assisted Intervention* (eds Zhou, S. K., Rueckert, D. & Fichtinger, G.) 547–574 (Academic Press, 2020).
144. Ibrahim, M. et al. Generative AI for synthetic data across multiple medical modalities: a systematic review of recent developments and challenges. *Comput. Biol. Med.* **189**, 109834 (2025).
145. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
146. Liu, J. et al. From distributed machine learning to federated learning: a survey. *Knowl. Inf. Syst.* **64**, 885–917 (2022).
147. Darzidehkalani, E., Ghasemi-Rad, M. & van Ooijen, P. M. A. Federated learning in medical imaging: part I: toward multicentral health care ecosystems. *J. Am. Coll. Radiol.* **19**, 969–974 (2022).
148. Dayan, I. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**, 1735–1743 (2021).
149. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
150. Zhao, Y. et al. Federated learning with non-IID data. Preprint at https://doi.org/10.48550/arXiv.1806.00582 (2018).
151. Zhang, Y. et al. The secret revealer: generative model-inversion attacks against deep neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 250–258 (IEEE, USA, 2020).
152. Park, R. W. Sharing clinical big data while protecting confidentiality and security: observational health data sciences and informatics. *Healthc. Inform. Res.* **23**, 1–3 (2017).
153. Callahan, T. J. et al. Ontologizing health systems data at scale: making translational discovery a reality. *NPJ Digit. Med.* **6**, 89 (2023).
154. Park, C. et al. Development and validation of the radiology common data model (R-CDM) for the international standardization of medical imaging data. *Yonsei Med. J.* **63**, S74–S83 (2022).
155. Fedorov, A. et al. DICOM re-encoding of volumetrically annotated lung imaging database consortium (LIDC) nodules. *Med. Phys.* **47**, 5953–5965 (2020).
156. Levy, M. A. et al. Informatics methods to enable sharing of quantitative imaging research data. *Magn. Reson. Imaging* **30**, 1249–1256 (2012).
157. Shin, S. J. et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. *J. Med. Internet Res.* **21**, e13249 (2019).
158. Stanzione, A. et al. Prostate MRI radiomics: a systematic review and radiomic quality score assessment. *Eur. J. Radiol.* **129**, 109095 (2020).
159. Sun, Y. et al. Automatic stratification of prostate tumour aggressiveness using multiparametric MRI: a horizontal comparison of texture features. *Acta Oncol.* **58**, 1118–1126 (2019).
160. Zhong, J. et al. A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. *Eur. Radiol.* **31**, 1526–1535 (2021).
161. Lin, P. et al. A Delta-radiomics model for preoperative evaluation of Neoadjuvant chemotherapy response in high-grade osteosarcoma. *Cancer Imaging* **20**, 7 (2020).

162. Spadarella, G. et al. MRI based radiomics in nasopharyngeal cancer: systematic review and perspectives using radiomic quality score (RQS) assessment. *Eur. J. Radiol.* **140**, 109744 (2021).
163. Zhang, L.-L. et al. Pretreatment MRI radiomics analysis allows for reliable prediction of local recurrence in non-metastatic T4 nasopharyngeal carcinoma. *eBioMedicine* **42**, 270–280 (2019).
164. Mühlbauer, J. et al. Radiomics in renal cell carcinoma-a systematic review and meta-analysis. *Cancers* **13**, 1348 (2021).
165. Li, Z.-C. et al. Differentiation of clear cell and non-clear cell renal cell carcinomas by all-relevant radiomics features from multiphase CT: a VHL mutation perspective. *Eur. Radiol.* **29**, 3996–4007 (2019).
166. Brancato, V., Cerrone, M., Lavitrano, M., Salvatore, M. & Cavaliere, C. A systematic review of the current status and quality of radiomics for glioma differential diagnosis. *Cancers* **14**, 2731 (2022).
167. Chen, Y. et al. Primary central nervous system lymphoma and glioblastoma differentiation based on conventional magnetic resonance imaging by high-throughput SIFT features. *Int. J. Neurosci.* **128**, 608–618 (2018).
168. Zhong, X. et al. Radiomics models for preoperative prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis. *Abdom. Radiol.* **47**, 2071–2088 (2022).
169. He, M. et al. Radiomic feature-based predictive model for microvascular invasion in patients with hepatocellular carcinoma. *Front. Oncol.* **10**, 574228 (2020).
170. Tabnak, P., HajiEsmailPoor, Z., Baradaran, B., Pashazadeh, F. & Aghebati Maleki, L. MRI-based radiomics methods for predicting Ki-67 expression in breast cancer: a systematic review and meta-analysis. *Acad. Radiol.* **31**, 763–787 (2023).
171. Liu, W. et al. Preoperative prediction of Ki-67 status in breast cancer with multiparametric MRI using transfer learning. *Acad. Radiol.* **28**, e44–e53 (2021).
172. Huang, M.-L. et al. A systematic review and meta-analysis of CT and MRI radiomics in ovarian cancer: methodological issues and clinical utility. *Insights Imaging* **14**, 117 (2023).
173. Song, X.-L., Ren, J.-L., Yao, T.-Y., Zhao, D. & Niu, J. Radiomics based on multisequence magnetic resonance imaging for the preoperative prediction of peritoneal metastasis in ovarian cancer. *Eur. Radiol.* **31**, 8438–8446 (2021).
174. Felfli, M. et al. Systematic review, meta-analysis and radiomics quality score assessment of CT radiomics-based models predicting tumor EGFR mutation status in patients with non-small-cell lung cancer. *Int. J. Mol. Sci.* **24**, 11433 (2023).
175. Boca, B. et al. MRI-based radiomics in bladder cancer: a systematic review and radiomics quality score assessment. *Diagnostics* **13**, 2300 (2023).
176. Li, L. et al. An MRI-based radiomics nomogram in predicting histologic grade of non-muscle-invasive bladder cancer. *Front. Oncol.* **13**, 1025972 (2023).
177. Zheng, J. et al. Development of a noninvasive tool to preoperatively evaluate the muscular invasiveness of bladder cancer using a radiomics approach. *Cancer* **125**, 4388–4398 (2019).
178. Jia, L.-L., Zhao, J.-X., Zhao, L.-P., Tian, J.-H. & Huang, G. Current status and quality of radiomic studies for predicting KRAS mutations in colorectal cancer patients: a systematic review and meta-analysis. *Eur. J. Radiol.* **158**, 110640 (2023).
179. Xue, T. et al. Preoperative prediction of KRAS mutation status in colorectal cancer using a CT-based radiomics nomogram. *Br. J. Radiol.* **95**, 20211014 (2022).
180. Cawley, G. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).

Acknowledgements

Some of the authors acknowledge financial support from the European Union's Horizon research and innovation programme under the following grant agreements: AIDAVA (HORIZON-HLTH-2021-TOOL-06) (grant no. 101057062 to P.L. and S.A.), CHAIMELEON (grant no. 952172 to P.L., H.C.K., S.A.M. and L.M-B.), EUCAIM (DIGITAL-2022-CLOUD-AI-02) (grant no. 101100633 to P.L., L.M-B. and S.A.), EuCanImage (grant no. 952103 to P.L., H.C.W., S.A.M., H.K., K.L. and Z.S.), GLIOMATCH (grant no. 101136670 to P.L. and Z.S.), IMI-OPTIMA (grant no. 101034347), ImmunoSABR (grant no. 733008) and REALM (HORIZON-HLTH-2022-TOOL-11) (grant no. 101095435 to P.L.), and RADIOVAL (HORIZON-HLTH-2021-DISEASE-04-04) (grant no. 101057699 to P.L., K.L. and S.A.). The research of X.Z. is partially supported by the Guangzhou basic and applied basic research foundation (grant no. SL2023A04J02221). The research of H.C.W. and S.K. is partially supported by the Dutch Cancer Society (KWF Kankerbestrijding) (project no. 2021-PoC/14449). The research of P.E.K. is supported by NIH (grant nos. R01CA258298 and U24CA264044).

Author contributions

P.L., H.C.W., S.A.M., X.Z., S.K., E.L., H.K., S.A. and Z.S. researched data for the article. All authors contributed substantially to discussion of the content, wrote, reviewed and/or edited the manuscript before submission.

Competing interests

P.L. has received grants and sponsored research agreements from Convert Pharmaceuticals SA, LivingMed Biotech srl and Radiomics SA; has received presenter fees and/or reimbursement of travel costs or consultancy fees (all in cash or in kind) from AstraZeneca, BHV srl and Roche; holds or has held minority shares in Bactam srl, Convert Pharmaceuticals SA, Comunicare SA, LivingMed Biotech srl and Radiomics SA; is a co-inventor on two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/050728) licensed to Radiomics SA, one issued patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, one granted patent on LSRT (PCT/P126537PC00, US patent no. 12,102,842) licensed to Varian, one issued patent on prodrugs (WO2019EP64112) without royalties, one non-issued, non-licensed patent on deep learning radiomics (N2024889) and three non-patented inventions (software) licensed to Health Innovation Ventures, ptTheragnostic/DNAmito and Radiomics SA. P.L. confirms that none of these disclosures are related to the current manuscript and none of the above entities were involved in the preparation of this Review. H.C.W. owns minority shares in Radiomics SA, and confirms that this entity was not involved in the preparation of this manuscript. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41571-025-01067-1.

Peer review information *Nature Reviews Clinical Oncology* thanks L. Dercle, W. Hsu and E. Huang for their contribution to the peer review of this work.

[1]The D-Lab, Department of Precision Medicine, GROW Research Institute for Oncology and Reproduction, Maastricht University, Maastricht, Netherlands. [2]Department of Radiology and Nuclear Medicine, GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, Netherlands. [3]Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [4]GIGA Cyclotrone Research Center, University of Liège, Liège, Belgium. [5]University MS Center, Biomedical Research Institute (BIOMED) & Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium. [6]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain. [7]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [8]OncoRay — National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany. [9]National Center for Tumour Diseases, NCT/UCC Dresden, Dresden, Germany. [10]German Cancer Research Center, Heidelberg, Germany. [11]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]2nd Division of Radiology, Medical University of Gdansk, Gdansk, Poland. [13]La Fe Health Research Institute, Biomedical Imaging Research Group and Imaging La Fe node, Distributed Network for Biomedical Imaging Unique Scientific and Technical Infra-structures, Valencia, Spain. [14]Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA. [15]Department of Advanced Computing Sciences, Institute of Data Science, Maastricht University, Maastricht, Netherlands. [16]Department of Bioengineering, University of Washington, Seattle, WA, USA. [17]Department of Radiology, University of Washington, Seattle, WA, USA. [18]Department of Radiology and Nuclear Medicine, Care and Public Health Research Institute, Maastricht University Medical Center+, Maastricht, Netherlands.