Review

Why Radiomics Rarely Reaches the Clinic: Reproducibility, Validation and Evidence Gap – A Critical Narrative Review

Jacopo Pozzi *, Jacopo D'Argenzio, Serena Carriero, Maurizio Cè, Dario D'Arrigo, Carolina Lanza, Pierpaolo Biondetti, Salvatore Alessio Angileri, Matilde Pavan, Rossella Catona, Gianpaolo Carrafiello

Posted Date: 3 June 2026

doi: 10.20944/preprints202606.0240.v1

Keywords: radiomics; reproducibility; clinical translation; external validation; methodological quality; meta-research; data leakage; reporting guidelines; clinical prediction models; evidence-based radiology

 doi:10.20944/preprints202606.0240.v1

Review

Why Radiomics Rarely Reaches the Clinic: Reproducibility, Validation and Evidence Gap – A Critical Narrative Review

Jacopo Pozzi [1,*], Jacopo D'Argenzio [1], Serena Carriero [2], Maurizio Cè [2], Dario D'Arrigo [1], Carolina Lanza [2], Pierpaolo Biondetti [2], Salvatore Alessio Angileri [2], Matilde Pavan [1], Rossella Catona [1] and Gianpaolo Carrafiello [2,3]

[1] Postgraduate School in Radiodiagnostics, Università degli Studi di Milano, 20122 Milan, Italy
[2] Department of Diagnostic and Interventional Radiology, Fondazione IRCCS Ca' Granda—Ospedale Maggiore Policlinico, Via Francesco Sforza 35, 20122 Milan, Italy
[3] Department of Oncology and Hemato-Oncology, Università degli Studi di Milano, 20122 Milan, Italy
* Correspondence: jacopo.pozzi@unimi.it

Abstract

Radiomics has produced tens of thousands of publications yet almost no tools in routine clinical use, and the reasons are increasingly understood to be problems of reproducibility and clinical translation rather than of algorithms. This critical narrative review argues that the field systematically generates *paper-grade evidence*—findings sufficient to publish—far faster than *decision-grade evidence*—findings sufficient to change clinical practice. Drawing on meta-scientific research, we describe seven fragility mechanisms (publication bias, analytical flexibility, underpowering, HARKing [hypothesizing after the results are known], citation distortion, cognitive bias, and misaligned incentives) and show why radiomics is structurally exposed to all of them simultaneously: high-dimensional feature spaces, acquisition-dependent measurement instability, segmentation variability, retrospective single-centre data, small samples, and leakage-prone validation. We then summarise empirical evidence on the radiomics literature, which remains pervaded by suboptimal methodological quality, near-absent negative results, limited external validation, sparse calibration and clinical-utility assessment, low data and code sharing, and a measurable retraction signal. We interpret these patterns as the output of a self-reinforcing system rather than isolated errors, and argue that better algorithms alone cannot resolve them. Finally, we argue that closing this gap requires not better models but evidentiary discipline: the consistent, enforceable application of standards the field already has, and the calibration of published claims to the strength of the underlying evidence.

Keywords: radiomics; reproducibility; clinical translation; external validation; methodological quality; meta-research; data leakage; reporting guidelines; clinical prediction models; evidence-based radiology

1. Introduction

A PubMed query for the term "radiomics" (search conducted in May 2026) returned approximately 19,000 indexed records, of which roughly two-thirds were published in 2023 or later [1]. The premise is compelling: extracting hundreds to thousands of quantitative features from routine medical images to capture tumour biology non-invasively [2–4]. Yet this literature has produced virtually no tools in routine clinical use. Deep-learning detection, triage, and quantification systems have entered routine practice, with FDA clearances of AI radiological devices surpassing 1000 by 2025 [5]. However, no handcrafted radiomic signature has matched this deployment [6–8]. The question is not whether radiomics works in principle, but why a field generating thousands of papers per year has failed to produce decision-grade evidence—predictions robust enough to change clinical behaviour.

 doi:10.20944/preprints202606.0240.v1

This disconnect is not unique to radiomics; similar patterns appear across high-dimensional biomedical disciplines [9–13]. Meta-scientific research has identified structural mechanisms (publication bias, analytical flexibility, underpowered designs, and weak validation norms) that cause literatures to become populated with individually plausible but collectively unreliable findings [14–17]. Chalmers and Glasziou [18] famously estimated that a large share — on the order of 85% — of biomedical research investment may be avoidable waste; although the precise figure is debated, it conveys the scale of the concern. Smaldino and McElreath [19] formalised the key dynamic: when career success depends on publication productivity and publication depends on significant results, methods maximising publishability spread through the scientific population even if they degrade reliability. This could lead to a literature in which *paper-grade evidence* can accumulate much faster than *decision-grade evidence*. We use *paper-grade evidence* to denote findings sufficient to support publication of a radiomic association or model, but insufficient to support clinical decision-making. Such evidence is typically retrospective, internally evaluated, discrimination-centred, and vulnerable to analytical flexibility. By contrast, *decision-grade evidence* denotes evidence sufficient to justify a change in clinical behaviour: the model or signature must be externally validated, calibrated, analytically locked, benchmarked against available clinical alternatives, and shown to provide decision-relevant incremental value. This is not, however, a uniform characterisation. Programmes with standardised multicentric protocols, IBSI-compliant extraction, and independent external validation produce evidence of a substantially higher order than the typical retrospective single-centre study. Nor do the mechanisms we describe implicate individual researchers: they are properties of the research ecosystem, and they can produce fragile evidence even in the absence of misconduct [19,20].

2. Materials and Methods: Scope, Evidence Selection, and Narrative Synthesis

This review synthesizes meta-scientific evidence on the conditions under which radiomics research is generated, validated, and translated into clinical practice. We assembled sources through structured searches of PubMed/MEDLINE and Google Scholar, supplemented by citation tracing of methodological papers, meta-research studies, consensus documents, and reporting guidelines. We prioritized five categories of evidence: (i) radiomics-specific meta-research studies; (ii) systematic reviews and methodological audits assessing study quality, validation, reporting, reproducibility, or clinical translation; (iii) empirical studies quantifying specific failure modes such as leakage, acquisition dependence, feature instability, and insufficient sample size; (iv) consensus statements and reporting or appraisal frameworks relevant to radiomics, imaging AI, and clinical prediction modelling; and (v) foundational meta-scientific literature on publication bias, analytical flexibility, underpowered research, citation distortion, cognitive bias, and scientific incentives. Source selection was purposive rather than exhaustive, and sources were included when they contributed directly to one of three analytic functions: defining a mechanism of evidentiary fragility, documenting an empirical signature of that mechanism in radiomics or imaging AI, or supporting a proposed governance response.

3. Meta-Scientific Mechanisms of Evidentiary Fragility

Several mechanisms known from meta-scientific research can make biomedical literatures fragile, describing conditions under which plausible, statistically significant, or technically sophisticated findings may accumulate faster than robust, transportable, clinically useful evidence. They include: publication bias [21–24], analytical flexibility and the garden of forking paths [25,26], underpowering and the winner's curse [27–29], HARKing [30–32], citation distortion [33], and cognitive biases including motivated reasoning and confirmation bias [34,35].

These mechanisms are interrelated and can act as a self-reinforcing system (Figure 1). Publication bias rewards positive findings [21,24], which increases the payoff from analytical flexibility [25,26]. High-dimensional pipelines may amplify this: the larger the space of defensible analytic choices, the more paths can lead to a publishable result. Underpowering ensures that any significant finding overestimates the true effect [27,28]. Cognitive mechanisms — motivated

 doi:10.20944/preprints202606.0240.v1

reasoning [34], confirmation bias [35], and, in AI-assisted settings, automation bias [36] — guide researchers through the garden of forking paths toward configurations that yield significance, without deliberate deception. The resulting paper reports inflated performance. Citation distortion [33] can propagate these inflated claims: high-performing models are cited without reference to validation status, progressively hardening provisional findings into accepted facts. Finally, incentive structures determine which research behaviours are rewarded. When career advancement, publication success, and institutional prestige depend more on producing positive and novel outputs than on generating negative results, independent replications, or externally validated tools, methods that maximise publishability can spread even if they do not maximise reliability [19,20].

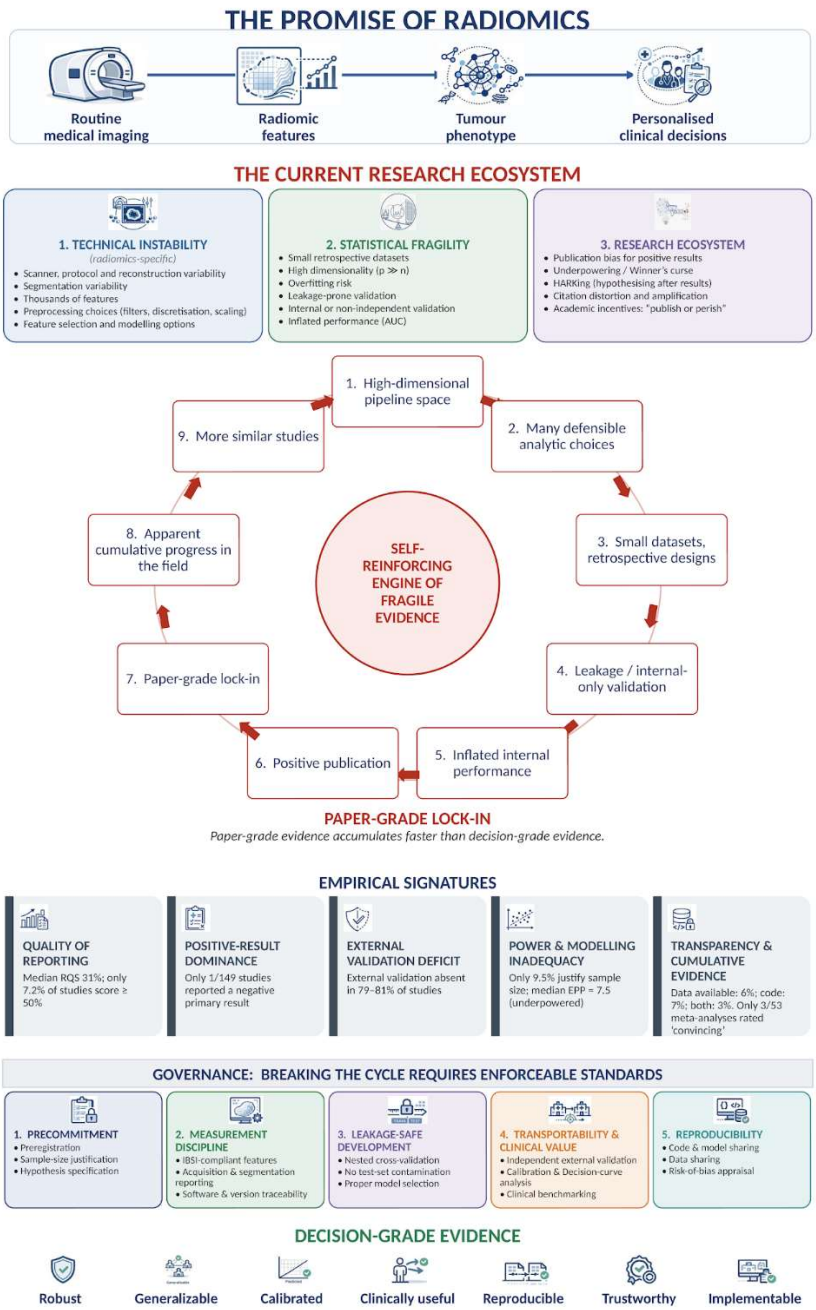

Figure 1. From the promise of radiomics to decision-grade evidence. The idealised radiomics pipeline (top) turns routine images into quantitative features expected to capture tumour phenotype and guide personalised care. In practice, interacting families of mechanisms — technical instability, statistical fragility, and ecosystem-

 doi:10.20944/preprints202606.0240.v1

level incentives (top centre) — leave measurable signatures in the published literature (lower centre)**:** uneven methodological quality, positive-result dominance, scarce external validation, underpowered designs, and limited transparency. These sustain a self-reinforcing cycle (centre): a high-dimensional pipeline with many defensible analytic choices, applied to small retrospective datasets under weak validation, inflates apparent performance and favours positive publication, locking in paper-grade findings that spawn further similar studies. As a result, paper-grade evidence accumulates faster than decision-grade evidence. Breaking the cycle requires not better algorithms but enforceable governance (bottom): safeguards that deliver robust, generalisable, calibrated, clinically useful, reproducible, trustworthy, and implementable decision-grade evidence.

4. Why Radiomics Is Structurally Exposed

The general mechanisms described above find an extreme case in radiomics, a field in which every structural vulnerability is simultaneously present. A high-dimensional feature space, physically unstable measurements, predominantly small retrospective monocentric samples, and leakage-prone validation converge to maximise analytical flexibility and structurally weaken evidentiary discipline.

4.1. Combinatorial Analytical Flexibility

A standard radiomic pipeline extracts hundreds to thousands of features from segmented image regions [37–39]. Each subsequent step — feature selection, normalisation, classifier choice, and hyperparameter tuning — is individually defensible; in combination, however, these choices generate a vast combinatorial space of plausible workflows. In radiomics, this multiplicity can produce a "vibration of effects," whereby materially different results emerge from the same data depending on how the analysis is specified. Such flexibility makes unstable or spurious associations increasingly easy to obtain unless the pipeline is rigorously constrained and analytical multiplicity is explicitly controlled [40]. Methodological variants further expand this space: delta-radiomics, for example, multiplies the configuration count by the number of time-points and delta-metric choices [41].

4.2. Acquisition-Dependent Measurement Instability

Radiomic features are computational derivatives of images that are themselves products of complex acquisition and reconstruction chains, and their values inherently depend on the technical conditions under which images are generated, including scanner platform, acquisition protocol, and voxel geometry [42–45]. Controlled CT experiments have shown that tube current, noise index, and iterative reconstruction levels substantially alter feature reproducibility, especially for texture features that depend on spatial intensity distributions [46]. This technical dependence persists in patient-level data: in a same-patient study of liver metastases reconstructed across multiple dose levels, section thicknesses, kernels, and reconstruction settings, only 11% of tested radiomic features remained reproducible across technical variations, with reconstructed section thickness producing the largest single-parameter effect [47]. Zhu et al. [48] tested 93 features across five CT systems: scan–rescan (test–retest) repeatability was excellent (97.1% repeatable), and intra-system reproducibility across dose levels was high (mean ICC = 0.945), but inter-system reproducibility was near zero (mean ICC = 0.157; 0% of features with ICC > 0.90). Zhang et al. [49] showed that on photon-counting detector CT, 0% of features were robust to slice-thickness changes. A systematic review of 481 studies confirmed that acquisition introduces more feature variability than segmentation, particularly for MRI [50]. Repeatability concerns extend beyond CT and MRI: a test–retest analysis of deep-learning-based PSMA-PET segmentation—cited here as a comparator—documented non-trivial inter-scan variability even when the segmentation algorithm was deterministic [51]. The Image Biomarker Standardisation Initiative (IBSI) Phase 1 [39] and Phase 2 [52] have standardised computational definitions but cannot eliminate the physical dependence on acquisition conditions. Statistical

 doi:10.20944/preprints202606.0240.v1

harmonisation methods such as ComBat can reduce inter-site variability [53,54], but their corrective capacity remains incomplete and context-dependent [54,55].

4.3. Segmentation Variability

Radiomic feature extraction often relies on manually delineated regions of interest; inter-reader variability can propagate directly into feature values and their stability, although standardisation recommendations now exist [56–58]. A meta-analysis of CT-based machine-learning studies in renal tumours further found that pooled diagnostic performance differed according to phase-selection and manual segmentation strategy, with contour-focused single-phase approaches showing the highest pooled AUC [59].

4.4. Retrospective, Monocentric Data Dominance

In the 2023 NEVER study [60], 95% of 149 sampled publications were retrospective, 75% were single-centre, and 91% used private data; in the 2024 self-reporting meta-research [61], 94% were retrospective and 68% single-centre. Such designs are especially exposed to all preceding vulnerabilities: when development and evaluation data originate from the same institutional environment, they often share scanners, protocols, and local clinical workflows, creating a weak test of generalisability and increasing the risk that models capture centre-dependent regularities rather than transportable biological signal [62,63]. Overlapping or reused patient cohorts have also been documented across radiomics publications, further complicating the independence and cumulative interpretation of the evidence base [64,65].

4.5. Small Samples Relative to Analytic Complexity

Zhong et al. [66] examined 116 radiomics studies from seven leading journals published in 2023: only 9.5% justified their sample size, the median events per predictor parameter (EPP) was 7.5, and, under the Riley et al. criteria [29], only 10.3% had a sufficient training sample size, with a median deficit of 268 patients. Horvat, Papanikolaou, and Koh [67] further noted that fewer than 20 published radiomics studies had incorporated clinical trial data, and that none had prospectively implemented radiomics as a clinical decision-support tool.

4.6. Leakage-Prone Validation

Leakage is not a single error but a family of train–test contamination mechanisms. In radiomics, recurrent forms include: (1) feature selection performed on the full dataset before cross-validation; (2) oversampling procedures such as SMOTE applied before data splitting; (3) harmonisation fitted on the entire dataset rather than within the training loop; (4) preprocessing statistics, including normalisation parameters, estimated using all available observations; and (5) hyperparameter optimisation evaluated through non-nested cross-validation [68–72]. Kapoor and Narayanan [69] documented leakage in at least 294 papers across 17 scientific fields, showing how easily it can generate overoptimistic claims. Within imaging pipelines, Marzi et al. [71] demonstrated that harmonisation before data splitting itself creates leakage and inflates apparent performance, while Gidwani et al. [72] showed that inconsistent partitioning across normalisation, feature selection, hyperparameter selection, and model assessment can markedly idealise radiomic models. The magnitude of the bias is substantial: oversampling leakage inflated AUC by up to +0.343 and produced AUCs as high as 0.90 on random data [70], whereas feature-selection leakage increased AUC-ROC by up to +0.15 across ten public radiomics datasets [68]. Even transparent external validation can expose the fragility of internally optimised signatures: a PET-radiomics model for recurrence-site prediction after head-and-neck re-irradiation declined from a reported balanced accuracy of 84.5% to 70% in an independent cohort, recovering only partially to 78% after cut-off recalibration [73].

 doi:10.20944/preprints202606.0240.v1

4.7. Weak Linkage to Clinical Decision-Making

Calibration assessment and clinical utility evaluation remain consistently underreported in RQS-based appraisals [74,75]. In HPV-status prediction for oropharyngeal cancer, only 5% of studies reported calibration statistics [76]; in MRI-radiomics studies for MGMT promoter methylation prediction in glioma, only 8% did so [64]. By contrast, discrimination-focused reporting was far more common: 89% of HPV studies reported discrimination statistics [76], and all MGMT studies reported AUC or accuracy [64]. However, even a model achieving a high internal AUC offers limited clinically actionable information if its calibration is unreported, its net benefit has not been assessed [77,78], and its performance has not been benchmarked against simpler or clinically available predictors [37].

5. Empirical Signatures of Evidentiary Fragility

Across the literature, each fragility mechanism described above leaves a quantifiable trace: Table 1 collects these signatures alongside the mechanism each one reflects.

5.1. Quality, Publication Bias, and Validation

Two large 2024–2025 evidence syntheses converge on the finding that radiomics study quality has improved over time, but remains persistently low overall [74,75]. These concerns had already been documented by Park et al. in 2020 [79] and were subsequently reinforced by Spadarella et al. [80] in 2023. Using the Radiomics Quality Score (RQS), Kocak et al. [74] found a median RQS of 31% across 1574 unique publications, with a strong positive temporal trend (Kendall's tau = 0.908, $p < 0.001$); and Barry et al. [75] reported a mean RQS of 26.1% ± 17.8% across 3258 RQS assessments (with only 7.2% reaching ≥50% of the maximum score). Complementary METRICS (METhodological RadiomICs Score [38], with an explanation-and-elaboration companion [81]) audits show that methodological quality remains, at best, moderate across several subfields: prostate MRI, 52% [82]; cardiac CT/MRI, 54.5% [83]; and glioma radiomics, 57% [84]. The cardiovascular METRICS audit also illustrates why reported performance should not be equated with evidentiary robustness: despite a pooled AUC of 0.81, Cavallo et al. reported high heterogeneity, statistically significant funnel-plot asymmetry, and only moderate average methodological quality, with 9 papers eligible for the meta-analysis [83]. Direct evidence of positive-result bias comes from the NEVER meta-research study: only 1 of 149 radiomics articles published in Q1 clinical radiology journals reported negative results (0.7%; rounded to 1% by the authors) [60]. Consistently, a glioma radiomics synthesis found that 26 of 27 studies (96%) reported positive effects, which the authors interpreted as evidence of high non-statistical publication bias [84]. Across two independent radiomics meta-research samples, external validation was absent in 81% (121/149) [60] and 79% (93/117) [61] of studies. In the broader radiologic deep-learning literature (invoked here as a comparator to handcrafted radiomics), externally validated algorithms also frequently lose performance: Yu et al. [85] found at least some external performance decrease in 70 of 86 algorithms (81%). Zhong et al. [86] re-performed 53 meta-analyses: only 3 of 53 (5.7%) reached a convincing level of evidence, whereas 43 of 53 (81%) were rated as weak. A recent deep-learning diagnostic synthesis likewise paired QUADAS-2 with METRICS in its quality appraisal; subgroup analyses showed significant performance differences according to study quality, indicating that methodological rigor can materially influence pooled estimates [87].

5.2. Reporting Transparency

Transparency remains exceptional rather than routine in radiomics research. Among 117 radiomics papers, only seven (6%) reported the use of any checklist or quality-scoring instrument [61], and adoption of the CLEAR (CheckList for EvaluAtion of Radiomics research) guideline across eligible literature has reached only 2% [88]. In 257 radiomics studies published in leading radiology journals, 16 (6%) shared data or used publicly available datasets, 20 (7%) shared code, and only seven (3%) did both [89]; in the broader radiology and nuclear medicine AI literature, just one of 161 private datasets was made available [90]. Reproducibility is further constrained by software dependence:

 doi:10.20944/preprints202606.0240.v1

nominally equivalent radiomic features can differ materially across extraction packages because of discrepancies in preprocessing and implementation [91], a broader problem in quantitative imaging, as illustrated for instance by reported cross-platform differences in cardiac CT measurements [92]. The IBSI framework [39,52] provides the technical basis for comparability through consensus definitions, reference values, and benchmarking; however, reporting and transparency practices have not kept pace [88].

5.3. Benchmarks and Translation

The translational deficit is also reflected in how radiomics claims are benchmarked. In a meta-research analysis of radiomics studies published in leading radiology journals, 44% made no comparison with non-radiomic approaches [60]. In treatment-response radiomics for non-small-cell lung cancer, comparison with the current gold standard was absent in all but two studies [93]. Even when direct comparisons are performed, incremental value is often limited or absent: in prostate radiotherapy, adding MRI radiomics increased the C-index only from 0.69 to 0.70 relative to a clinical-only model [94]; in advanced melanoma, CT radiomics failed to improve on a simpler clinical model for predicting benefit from checkpoint inhibitors [95]; and in locally advanced rectal cancer, MRI-based radiomic models showed no definite added value over clinical models for predicting pathological complete response after neoadjuvant chemoradiotherapy [96]. The problem, therefore, is not simply that radiomics has been slow to reach clinical practice, but that much of the literature still stops short of establishing why practice should change. This is the translational gap described by Kocak, Pinto dos Santos, and Dietzel [6], and it persists despite measurable improvements in formal study quality: Barry et al. showed that the dimensions most consequential for clinical adoption remain among the least developed [75]. RQS 2.0 and its radiomics readiness levels were introduced precisely to make that gap visible, shifting appraisal from isolated methodological adequacy to staged translational maturity [97,98].

5.4. Retractions as an Ecosystem Signal

Retractions provide a complementary ecosystem-level indicator of evidentiary fragility in radiomics. Demircioğlu identified 93 retracted radiomics publications across six databases, corresponding to an estimated mean retraction rate of 6.7 per 10,000 publications—a rate that, in absolute terms, is comparable to or below several biomedical baselines and is therefore best read as a weak, indirect signal; among the 20 cases examined in detail, 11 retraction notices (55%) did not clearly distinguish misconduct from error or assign responsibility, and no major radiological or oncological journal appeared to have retracted a radiomics publication [99]. Against a background of persistently low code- and data-sharing rates, which hinder independent scrutiny of published work [89], this pattern suggests that formal retractions may underestimate the broader burden of problematic radiomics studies [99].

6. From Recurrent Patterns to Systemic Interpretation

Sections 3–5 described fragility mechanisms, their structural amplification in radiomics, and their empirical signatures. We interpret these not as independent issues, but as the coupled components of a self-reinforcing system. Where Ioannidis [15] distinguishes useful from merely publishable research, Huang et al. [7] specify the criteria for clinical translation, and RQS 2.0 [97] grades study-level readiness, the paper-grade/decision-grade distinction locates the gap in the interaction between fragility mechanisms and the publication ecosystem rather than in any single study.

6.1. Paper-Grade Evidence as the Product of a Self-Reinforcing System

Radiomics is structurally exposed because the vulnerabilities described above do not merely coexist: high dimensionality and analytical flexibility, acquisition- and segmentation-dependent measurement, retrospective single-centre designs, small samples, leakage-prone workflows, and

 doi:10.20944/preprints202606.0240.v1

limited sharing of code and data tend to co-occur and may jointly lower the threshold for exploratory findings to become publishable before they become clinically reliable [52,66,68,70,75,89,100–107].

Radiomic pipelines create many defensible analytical routes: preprocessing, discretisation, harmonisation, feature selection, classifier choice, hyperparameter tuning, and validation design can each be varied in ways that may appear individually reasonable. In such a setting, cognitive mechanisms such as motivated reasoning and confirmation bias may influence which routes are explored, retained, or interpreted as most compelling, without requiring deliberate misconduct [34,35]. When positive or apparently high-performing results are more likely to enter the visible literature, this flexibility can favour the accumulation of attractive findings over the accumulation of robust evidence, as illustrated in radiomics by the scarcity of negative studies [60].

Once published, these findings can acquire cumulative force, as provisional radiomic claims are transmitted through a literature already skewed toward positive findings rather than through a balanced evidentiary record. Broader meta-research shows that positive or statistically significant findings are preferentially cited, while citation networks can further transform qualified claims into apparent authority [33,108,109]. In a radiomics literature already enriched for positive results [60], this may create a secondary amplification layer: internally promising models can become part of the cumulative narrative before their external validity, calibration, and clinical utility have been adequately established. The same process is reinforced by institutional and publication incentives. Systems that reward novel, positive, technically sophisticated, and publishable outputs more strongly than correction, replication, external validation, or negative evidence can select for research behaviours that maximise publication success even when they do not maximise reliability [19,20].

The result is not a literature of false claims, but a literature in which paper-grade evidence can become visible, citable, and cumulatively influential before the harder tests required for decision-grade evidence have been satisfied [19,33–35,60]. Paper-grade evidence—typically retrospective, internally validated, and discrimination-centred—may justify publication, but not clinical action. Decision-grade evidence requires independent external validation [7,67], adequate calibration [77], comparison with clinical alternatives, and incremental value demonstrated through decision-analytic evaluation [78,110]. Claim–evidence alignment should discipline the language of radiomics translation: exploratory studies should remain explicitly exploratory; internally validated models should be framed as candidates for external testing; generalisability should not be claimed without external validation; and clinical usefulness should not be claimed without calibration, benchmarking against clinically available alternatives, incremental value, and decision-analytic assessment of clinical utility—otherwise limited evidence is converted into overstatement, as described in the literature on biomedical spin, diagnostic-test overinterpretation, and unsupported AI performance claims [111–113].

7. Standards and Safeguards Across the Radiomics Pipeline

Because the fragility described above is a property of the ecosystem rather than of any algorithm, the response is methodological discipline applied sequentially across the radiomics pipeline—from measurement and model development to validation and open science (Table 2).

7.1. Measurement, Reporting, and Statistical Planning

Stable measurement is the first requirement for credible radiomics. IBSI's original feature-standardisation effort established consensus definitions and reference values for 169 radiomic features [39], while the subsequent filter-standardisation initiative provided reference implementations and verification resources for commonly used convolutional filters [52]. Measurement stability also depends on upstream image acquisition. Although not radiomics-specific, professional-society efforts to harmonise CT protocols, such as the SIRM position papers [114,115], address an acquisition layer that is directly relevant to radiomic feature robustness, given the well-documented sensitivity of CT radiomics to scanner, reconstruction, slice-thickness, and discretisation choices [43–45]. Reporting standards are equally essential. The CLEAR guideline requires authors to

 doi:10.20944/preprints202606.0240.v1

report the software, version, feature definitions, segmentation and preprocessing details, and extraction parameters needed to assess reproducibility [37]; CLEAR-E3 provides worked explanations and examples for applying these requirements consistently [116]. Statistical planning must then match the effective complexity of the model-development process. Riley et al. [29] provide practical methods for sample-size calculation in prediction-model development, and radiomics analyses should accordingly align candidate model complexity with the information content of the available data.

7.2. Leakage-Free Model Development

Data-dependent operations must be confined to the training process and re-estimated independently within each resampling fold or validation split. Radiomics-specific studies have quantified the inflation caused by feature-selection leakage [68] and by oversampling before cross-validation [70], while broader machine-learning methodology has documented leakage as a pervasive source of irreproducible scientific claims [69]. Leakage can arise at multiple stages of the pipeline, including preprocessing, feature selection, resampling, harmonisation, and model or hyperparameter selection. The radiomics literature already provides direct empirical evidence for some of these mechanisms [68]; separate work in medical imaging has shown that harmonisation performed before data splitting can also induce leakage and inflate downstream performance estimates [71]. Nested cross-validation helps separate model tuning from model evaluation [117]. PROBAST+AI [118] and TRIPOD+AI [119] should then be used, respectively, to assess risk of bias and to ensure complete reporting of the resulting prediction-model study.

7.3. Validation Hierarchy, Calibration, and Clinical Utility

Independent external validation is the minimum requirement for claims of generalisability beyond the source data, and prospective clinical evaluation becomes increasingly important as tools approach implementation [120,121]. For systems entering early live clinical evaluation, DECIDE-AI provides stage-specific reporting guidance tailored to AI-based decision-support studies [122]. Discrimination alone is insufficient for clinical credibility. For studies making clinical applicability claims, calibration should be assessed and reported explicitly [77,123], and decision-analytic utility should be evaluated through net-benefit approaches such as decision-curve analysis [78,124]. These model-centred assessments should be complemented by clinically meaningful and patient-centred outcome measures, which remain less frequently foregrounded in the evaluation literature on AI in radiology [125].

7.4. Open Science and Preregistration

Reproducibility requires preservation and, where feasible, sharing of the analytic materials needed to interrogate a radiomics claim: code, extraction configurations, model specifications, and reusable data resources [89,126,127]. The current literature remains far from this standard. In two independent radiomics meta-research samples, 91% and 89% of studies relied on private data [60,61], while a dedicated audit of sharing practices identified limited availability of models, code, and datasets as a major obstacle to clinical translation [89]. Preregistration addresses a complementary vulnerability: the blurring of confirmatory and exploratory analysis. By fixing hypotheses and analysis plans before outcomes are inspected, it helps distinguish prediction from postdiction and limits the scope for selective analytical adaptation [128,129].

7.5. Testing Robustness to Analytic Flexibility

Reporting standards document the analytic pipeline that was chosen, but they cannot establish whether a result reflects a stable signal or only one path through the space of defensible choices. Multiverse and specification-curve analyses make that dependence measurable: the model is re-estimated across the full set of reasonable specifications (discretisation, harmonisation, feature

 doi:10.20944/preprints202606.0240.v1

selection, and classifier), and the distribution of results is reported in place of a single pipeline [130], while the specification curve orders these fits to show how far the signature depends on any individual choice [131]. This converts the vibration of effects from an unmeasured threat into a reported quantity: a result stable across this space is credible, while one that emerges only under a narrow configuration might be an artefact of analytic flexibility.

7.6. The Shifting Technical Frontier: Foundation Models and Generative AI

Self-supervised imaging foundation models are the most credible technical answer to handcrafted radiomics' small samples and weak transportability [132], yet they relocate rather than remove the problem, concentrating capability and raising validation and governance challenges that, by their developers' own account, strain current medical-AI evaluation [133]. Generative methods are similarly double-edged: synthetic-data augmentation and harmonisation can enlarge scarce datasets and attenuate scanner effects [134], but in controlled phantom data image-level generative harmonisation improved appearance while reducing radiomic feature stability [135], so such steps must be validated on the downstream radiomic endpoint, not assumed beneficial.

8. Implications for the Field: From Available Standards to Routine Adoption

Although a substantial methodological infrastructure now exists [136–138] and formal quality has improved, its uptake remains marginal (Section 5.2). This reflects not the absence of standards but an incentive asymmetry: methodological safeguards impose immediate costs on individual investigators (time, expertise, larger samples, and reduced analytical flexibility), while their benefits accrue collectively, through more reliable cumulative knowledge. The result is a collective-action problem, mapping onto the distinction between the rewards of getting work published and those of getting it right [139]; as long as publication and advancement track productivity and positive findings, the privately rational choice diverges from the collectively useful one. Compliance will not become routine until methodological rigour is made consequential at the point of publication and evaluation [128,139].

Building on this incentive logic [139], reform should follow four design principles: interventions should be structural rather than exhortatory, composable rather than monolithic, enforceable rather than voluntary, and incentive-compatible rather than virtue-demanding. These principles are proposals, not validated remedies: their effect on radiomics translation, as distinct from reporting quality, has not been tested, and each can fail in predictable ways. Entry requirements may favour groups already equipped to meet them, and mandates may produce formal rather than substantive compliance—a pattern already suggested by the divergence between self-reported and expert-confirmed CLEAR adherence [88]. They are accordingly best assessed against the outcome indicators set out below.

The principles translate into distinct responsibilities. Authors should pre-specify analytic plans, report against CLEAR, share code and data where feasible, and calibrate conclusions to the evidentiary tier actually reached [116,126–129]. Reviewers should apply PROBAST+AI and METRICS and assess whether a study's claims match that tier [38,118]. At the editorial level, screening submissions against minimum methodological requirements before peer review, and weighing whether a study's claims are commensurate with the evidence it attains, would serve the same end [97]. Registered Reports are the most promising structural lever for confirmatory work: in psychology they reduced positive first-hypothesis results from 96% of standard reports to 44%, illustrating how decoupling publication from result direction reshapes the visible evidence base [140,141]. Funders should require data-sharing and, where appropriate, preregistration as conditions of award [126–128], and institutions should reward rigour and transparency over venue prestige, in line with DORA [142]. Progress should be tracked against concrete indicators: a rising share of published negative results, increasing median quality scores, external validation becoming routine, and Registered Reports becoming a normal pathway for confirmatory studies.

 doi:10.20944/preprints202606.0240.v1

9. Limitations and Boundary Conditions

Several limitations bound the claims of this review. By design it is a critical narrative synthesis with purposive, question-driven source selection rather than a systematic, bias-controlled survey; it is therefore itself exposed to the selection effects it describes, and its conclusions are interpretive rather than quantitative. The supporting evidence is also of uneven provenance: some is direct, such as the controlled leakage-inflation experiments; some is indirect, such as the retraction signal; and some is imported by analogy from deep-learning and other high-dimensional fields. The evidence base is also concentrated, as much of the appraisal infrastructure behind the quality critique derive from a small number of investigators and groups, so that common provenance may carry common assumptions, and literature's apparent convergence rests on a comparatively narrow foundation. In addition, most of the empirical signatures are snapshots of 2020–2025 period, whereas most of the standards against which they are read (CLEAR, METRICS, TRIPOD+AI, RQS 2.0) postdate much of the literature they appraise: part of the observed under-adoption is therefore temporal, and this review necessarily describes the field with a delay, understating improvements already under way.

Moreover, robust work at the frontier of the field is not in dispute. An overview that re-performed 53 radiomics meta-analyses found three associations supported by convincing and seven by highly suggestive evidence [86]; an umbrella review of artificial intelligence in cancer imaging graded roughly one-third of pooled estimates as moderate-certainty [143]; and methodological quality is rising over time [74,75]. Decision-grade radiomics therefore exists — produced by standardized, externally validated programmes that sit well above the typical retrospective single-centre study.

Finally, nor is the entire translation gap attributable to evidentiary fragility. Imaging-biomarker translation is intrinsically slow and capital-intensive, and its regulatory frameworks are still consolidating: the European Union Artificial Intelligence Act phases in obligations for high-risk medical AI through 2026–2027 [144], and the United States Food and Drug Administration finalised its predetermined-change-control guidance only in late 2024 [145]. Limited data-sharing is often driven by privacy and governance constraints rather than reluctance, and for some clinical endpoints the incremental value of radiomics over established predictors is genuinely marginal, itself a substantive finding and not a failure of effort.

10. Conclusions

Radiomics does not suffer primarily from a shortage of algorithms or features; it suffers from an evidentiary architecture in which paper-grade findings accumulate faster than the decision-grade evidence required to change clinical practice. The mechanisms responsible—analytical flexibility, measurement instability, underpowered and single-centre designs, leakage-prone validation, selective publication, and incentive misalignment—are mutually reinforcing, which is why purely technical fixes have not closed the translational gap. The constructive implication is that the field already possesses most of the standards it needs; what is missing are the composable, enforceable, and incentive-compatible mechanisms that make their sequential application routine and consequential, together with the editorial norm that claims be matched to the evidentiary tier actually achieved. Realigning publication with reliability is the path from paper-grade to decision-grade radiomics.

 doi:10.20944/preprints202606.0240.v1

Table 1. Meta-scientific mechanisms of evidentiary fragility and their empirically observed signatures in contemporary radiomics research.

Mechanism of evidentiary fragility	Expected ecosystem-level signature	Representative measured evidence in radiomics
1. Persistently low and uneven methodological maturity	Formal standards, reporting frameworks, and quality tools expand, yet the average methodological quality of the published literature remains low and varies across radiomics subdomains.	Large-scale RQS evidence remains consistently poor: mean RQS 26.1% across 3258 assessments, with only 7.2% reaching ≥50% of the maximum score [75]; median RQS 31% across 1574 publications [74]; delta-radiomics median RQS 25%, with 51.2% of studies scoring <25% [41]; endometrial MRI radiomics mean RQS 13.77 [146]
2. Positive-result selection and publication bias	Null, negative, and non-superior findings are selectively underrepresented, while positive claims dominate the published record and are often insufficiently benchmarked against simpler alternatives.	Positive findings dominate the literature: NEVER found only 1/149 negative studies (0.7%) and no non-radiomic comparator in 44% [60]; glioma radiomics reported positive findings in 26/27 studies (96%) [84]; cardiovascular radiomics showed funnel-plot asymmetry by Egger's test ($z = -2.39$, $p = 0.017$) [83].
3. Leakage-prone analytical flexibility	Researcher degrees of freedom and workflow errors that compromise separation between model development and evaluation inflate apparent performance.	Empirical leakage studies show substantial optimism: feature selection outside cross-validation inflated AUC by up to 0.15 [68]; oversampling before cross-validation biased AUC by up to 0.34, produced AUCs up to 0.90 with random outcomes, and was likely present in 5/34 radiomics papers from 2023 [70]
4. Underpowered model development and winner's curse	Sample sizes remain too small relative to model complexity, favouring unstable estimates, optimistic effect sizes, and poor transportability.	In 116 binary-outcome radiomics prediction studies, only 11/116 (9.5%) justified sample size and only 6 included an a priori calculation; median training size was 150, median EPP 7.5, median Riley-based shortfall 268 patients, and only 12/116 (10.3%) met all adequacy criteria [66].
5. External-validation deficit and fragile generalisability	Models are rarely tested on genuinely independent populations; when transported beyond development data, apparent performance commonly attenuates.	External validation was absent in 121/149 NEVER studies (81%) [60]; across 1574 radiomics publications, only 14% reported independent external validation and 32% lacked any separate validation set [74]. In radiologic deep learning, external validation reduced median AUC by −0.046, with decline in 70/86 algorithms (81%) [85].
6. Acquisition-driven measurement instability	Apparent radiomic signal fails to transport across scanners, platforms, or acquisition settings, indicating vulnerability to non-biological measurement variation.	Acquisition effects markedly impair reproducibility: across five CT systems, 97.1% of features were repeatable on test–retest (scan–rescan), but inter-system reproducibility was poor (mean ICC/CCC 0.157 ± 0.174), with no feature reaching ICC/CCC >0.90 across systems

 doi:10.20944/preprints202606.0240.v1

Mechanism of evidentiary fragility	Expected ecosystem-level signature	Representative measured evidence in radiomics
		[48]; on photon-counting detector CT, no features were robust to high-pitch acquisition or slice-thickness changes [49]; and acquisition effects exceeded segmentation effects in a 481-study reliability review [50].
7. Reporting opacity, self-assessment inflation, and weak open-science practice	Methodological errors remain difficult to detect, formal self-audit is rare, and independent verification is constrained by incomplete reporting and limited sharing of code and data.	Transparency remains limited: only 7/117 studies (6%) included a self-reported checklist/quality score [61]; CLEAR adoption was 2%, with self-reported adherence exceeding expert-confirmed adherence by 21 percentage points [88]; among 257 studies, 6% shared data/open datasets, 7% shared code, and 3% shared both [89]; 0/195 empirical radiology articles shared analysis scripts [147].
8. Quality-sensitive heterogeneity in evidence synthesis	Methodological quality is not merely a descriptive deficit: it becomes a measurable source of variation in pooled performance estimates.	Study quality measurably affects pooled evidence: in endometrial MRI radiomics, higher RQS was associated with lower QUADAS-2 risk, more recent publication year, and higher reported performance [146]; and in CT hematoma-expansion deep learning, subgroup analyses showed significant performance differences by segmentation technique and study quality [87].
9. Weak cumulative evidence and stalled clinical translation	Local methodological fragilities accumulate into low-certainty evidence at synthesis level and limited progression toward clinically embedded use.	Meta-evidence remains low-certainty: among 53 re-performed radiomics meta-analyses, only 3/53 associations (5.7%) were convincing and 43/53 (81%) were weak [86]. Fewer than 20 oncologic radiomics studies used clinical-trial data, and no published model had been prospectively implemented as routine clinical decision support [67]; this is consistent with recent analyses describing a widening publication–translation gap [6].

Table 2. Methodological vulnerabilities, minimum evidentiary requirements, and available standards across the radiomics translational pipeline.

Translational stage	Principal vulnerability / failure mode	Minimum operational requirement for a non-exploratory clinical claim	Primary standard(s), guideline(s), or framework(s)	Empirical evidence that the requirement remains insufficiently met
1. Study conception	Analytical flexibility, post hoc	Pre-specify the clinical question, population, endpoint, candidate predictors, analysis plan,	Preregistration [128] Registered Reports [140],	In broader meta-research, positive findings were reported in 44% of Registered Reports versus 96% of

 doi:10.20944/preprints202606.0240.v1

Translational stage	Principal vulnerability / failure mode	Minimum operational requirement for a non-exploratory clinical claim	Primary standard(s), guideline(s), or framework(s)	Empirical evidence that the requirement remains insufficiently met
	hypothesis shaping, selective reporting	validation strategy, and primary performance metrics	TOP guidelines [127]	standard publication models [141]
2. Imaging measurement, feature extraction, and software traceability	Acquisition-, preprocessing-, filter-, and software-dependent feature instability	Use IBSI-compliant definitions; report acquisition, reconstruction, preprocessing, interpolation, discretisation, filters, extraction software, and software version	IBSI Phase 1, 2 [39,52]; documented open implementation: PyRadiomics [148]	Across photon-counting and dual-energy CT systems, mean inter-system ICC was 0.157, and no feature reached ICC >0.90 at matched dose [48]. Feature values differed across radiomics software implementations [91]. Vendor-dependent quantitative CT differences were also reported outside radiomics [92].
3. Sample-size adequacy and study positioning	Underpowered model development; unstable estimates; inflated apparent performance	Provide a formal sample-size justification using prediction-model criteria	Riley et al. [29]	Sample-size justification was absent in 90.5% of studies; only 10.3% met strict Riley-based criteria; median shortfall: 268 patients [66].
4. Model development and leakage-safe internal validation	Data leakage; optimistic bias; non-nested feature selection; misuse of resampling	Nest preprocessing, feature selection, resampling, hyperparameter tuning, and model selection within training folds; use leakage-safe internal validation.	Radiomic-signature safeguards [149]; nested cross-validation principles [117]; PROBAST+AI [118].	AUC inflation reached +0.34 when oversampling preceded cross-validation [70] and +0.15 when feature selection was applied before cross-validation [68].
5. Reporting transparency of radiomics and AI methods	Incomplete or non-verifiable pipeline reporting	Report item-by-item against the framework appropriate to the study scope: CLEAR, CLEAR-E3, TRIPOD+AI, and CLAIM where applicable	CLEAR [37]; CLEAR-E3 [116]; TRIPOD+AI [119]; CLAIM 2024 [150]	Only 7/117 radiomics papers (6%) included a self-reported reporting checklist or quality-scoring document [61]. CLEAR adoption reached 2%; self-reported versus expert-confirmed adherence was 91% vs 66% (mean gap, 21 percentage points) [88].

 doi:10.20944/preprints202606.0240.v1

Translational stage	Principal vulnerability / failure mode	Minimum operational requirement for a non-exploratory clinical claim	Primary standard(s), guideline(s), or framework(s)	Empirical evidence that the requirement remains insufficiently met
6. Methodological appraisal	Conflation of reporting quality, methodological quality, and translational maturity	Use structured tools for methodological appraisal and translational readiness, rather than relying on discrimination metrics or narrative claims alone	METRICS [38]; RQS 2.0 and Radiomics Readiness Levels [97]	Median RQS was 31% of the maximum across 1574 publications [74]. In a 2025 diagnostic-accuracy synthesis, study quality assessed with METRICS emerged as a significant source of between-study differences in subgroup analyses [87].
7. Open science and computational reproducibility	Unavailable code; inaccessible datasets; non-reproducible computational workflows	Share code and data where feasible; otherwise state access restrictions and provide sufficient computational detail for independent re-analysis	FAIR principles [126]; TOP guidelines [127]	In 257 radiomics papers published in leading journals, only 6% shared data and 7% shared code [89]. Private data were used in 91% of papers in the NEVER study [60] and 89% in a separate meta-research sample [61]. In broader radiology and nuclear medicine AI, only 1/161 private-data studies shared the dataset [90].
8. External validation and transportability	Internal-only evidence; poor transportability across institutions, scanners, and protocols	Perform external validation on independent data; distinguish internal, internal–external, and external validation	Validation hierarchy and clinical prediction-model guidance [120,121]	External validation was absent in 81% [60] and 79% [61] of radiomic studies; only 14% of 1574 publications included external validation [74].
9. Calibration, clinical benchmarking, and incremental value	Discrimination-only evaluation; absent calibration; untested incremental clinical value	Report calibration alongside discrimination; compare against clinical, non-radiomic, or standard-of-care baselines; quantify added value	Calibration principles [77,123]; CLEAR comparison requirements [37]	Calibration was reported in 1/19 HPV-prediction studies [76] and 2/26 MGMT-prediction studies [64]. 44% of radiomics studies included no non-radiomic comparator [60]. In prostate radiotherapy, adding MRI radiomics increased the C-index from 0.69 to 0.70 over a clinical model [94].

 doi:10.20944/preprints202606.0240.v1

Translational stage	Principal vulnerability / failure mode	Minimum operational requirement for a non-exploratory clinical claim	Primary standard(s), guideline(s), or framework(s)	Empirical evidence that the requirement remains insufficiently met
10. Early clinical evaluation and trial-level evidence	Retrospective performance claims substituted for early clinical evaluation; incomplete protocol and trial reporting	Evaluate live clinical performance, safety, workflow effects, and human-factor consequences before trial-level claims; use AI-specific reporting extensions for interventional protocols and trial reports	DECIDE-AI [122]; SPIRIT-AI [151]; CONSORT-AI [152]	In oncology AI, median SPIRIT-AI concordance was 78.2% across 12 RCT protocols [153]; median combined CONSORT 2010/CONSORT-AI concordance was 82% across 57 RCT reports [154]
11. Deployment governance, regulatory readiness, and real-world translation	Clinical-readiness claims without deployability, lifecycle governance, or regulatory fitness	Address trustworthiness, robustness, fairness, explainability, traceability, post-deployment monitoring, and applicable regulatory requirements before claiming clinical readiness	FUTURE-AI [155]; GMLP / IMDRF [156]; applicable medical-device regulation, including EU MDR where relevant [157]	The research–clinical translation gap has been described as widening [6]. Routine oncologic implementation remains limited [7,65]. Progress in cost-effectiveness analysis is minimal or insignificant across radiomics studies [75].

Author Contributions: Conceptualization, J.P. and G.C.; methodology, J.P., M.C..; investigation, J.P., J.D., D.D., M.P. and R.C.; resources, S.C., M.C., C.L., P.B., S.A.A. and G.C.; data curation, J.P., J.D., D.D., M.P. and R.C.; writing—original draft preparation, J.P.; writing—review and editing, J.P., J.D., S.C., M.C., D.D., C.L., P.B., S.A.A., M.P., R.C. and G.C.; visualization, J.P.; supervision, S.C., M.C., C.L., P.B., S.A.A. and G.C.; project administration, J.P. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AUC, area under the receiver operating characteristic curve; CCC, concordance correlation coefficient; CLEAR, CheckList for EvaluAtion of Radiomics research; CT, computed tomography; EPP, events per predictor parameter; ICC, intraclass correlation coefficient; IQR, interquartile range; METRICS, METhodological RadiomICs Score; Q1/Q3, first/third quartile; RQS, Radiomics Quality Score.

 doi:10.20944/preprints202606.0240.v1

References

1. Kocak B, Baessler B, Cuocolo R, Mercaldo N, Pinto Dos Santos D. Trends and statistics of artificial intelligence and radiomics research in Radiology, Nuclear Medicine, and Medical Imaging: bibliometric analysis. Eur Radiol. 2023 Nov;33(11):7542–55. https://doi.org/10.1007/s00330-023-09772-0
2. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017 Dec;14(12):749–62. https://doi.org/10.1038/nrclinonc.2017.141
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563–77. https://doi.org/10.1148/radiol.2015151169
4. Ferrari R, Trinci M, Casinelli A, Treballi F, Leone E, Caruso D, et al. Radiomics in radiology: What the radiologist needs to know about technical aspects and clinical impact. Radiol Med. 2024 Dec;129(12):1751–65. https://doi.org/10.1007/s11547-024-01904-w
5. U.S. Food and Drug Administration, Center for Devices and Radiological Health. Artificial Intelligence-Enabled Medical Devices [Internet]. Silver Spring (MD): FDA; [cited 2026 May]. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
6. Kocak B, Pinto Dos Santos D, Dietzel M. The widening gap between radiomics research and clinical translation: rethinking current practices and shared responsibilities. European Journal of Radiology Artificial Intelligence. 2025 Jan;1:100004. https://doi.org/10.1016/j.ejrai.2025.100004
7. Huang EP, O'Connor JPB, McShane LM, Giger ML, Lambin P, Kinahan PE, et al. Criteria for the translation of radiomics into clinically useful tests. Nat Rev Clin Oncol. 2023 Feb;20(2):69–82. https://doi.org/10.1038/s41571-022-00707-0
8. Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Annals of Oncology. 2017 Jun;28(6):1191–206. https://doi.org/10.1093/annonc/mdx034
9. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001 Nov;29(3):306–9. https://doi.org/10.1038/ng749
10. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digit Med. 2022 Apr 12;5(1):48. https://doi.org/10.1038/s41746-022-00592-y
11. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. eLife. 2021 Dec 10;10:e71601. https://doi.org/10.7554/eLife.71601
12. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021 Mar 15;3(3):199–217. https://doi.org/10.1038/s42256-021-00307-0
13. Carp J. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. Front Neurosci. 2012;6. https://doi.org/10.3389/fnins.2012.00149
14. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005 Aug 30;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124
15. Ioannidis JPA. Why Most Clinical Research Is Not Useful. PLoS Med. 2016 Jun 21;13(6):e1002049. https://doi.org/10.1371/journal.pmed.1002049
16. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie Du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017 Jan 10;1(1):0021. https://doi.org/10.1038/s41562-016-0021
17. Van Calster B, Steyerberg EW, Wynants L, Van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023 Feb 24;21(1):70. https://doi.org/10.1186/s12916-023-02779-w
18. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet. 2009 Jul;374(9683):86–9. https://doi.org/10.1016/S0140-6736(09)60329-9
19. Smaldino PE, McElreath R. The natural selection of bad science. R Soc open sci. 2016 Sep;3(9):160384. https://doi.org/10.1098/rsos.160384

 doi:10.20944/preprints202606.0240.v1

20. Edwards MA, Roy S. Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. Environmental Engineering Science. 2017 Jan;34(1):51–61. https://doi.org/10.1089/ees.2016.0223
21. Dickersin K. The Existence of Publication Bias and Risk Factors for Its Occurrence. JAMA. 1990 Mar 9;263(10):1385. https://doi.org/10.1001/jama.1990.03440100097014
22. Song F, Parekh S, Hooper L, Loke Y, Ryder J, Sutton A, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess. 2010 Feb;14(8). https://doi.org/10.3310/hta14080
23. Dwan K, Gamble C, Williamson PR, Kirkham JJ, the Reporting Bias Group. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review. Boutron I, editor. PLoS ONE. 2013 Jul 5;8(7):e66844. https://doi.org/10.1371/journal.pone.0066844
24. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. N Engl J Med. 2008 Jan 17;358(3):252–60. https://doi.org/10.1056/NEJMsa065779
25. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychol Sci. 2011 Nov;22(11):1359–66. https://doi.org/10.1177/0956797611417632
26. Gelman A, Loken E. The Statistical Crisis in Science. Am Sci. 2014;102(6):460. https://doi.org/10.1511/2014.111.460
27. Ioannidis JPA. Why Most Discovered True Associations Are Inflated. Epidemiology. 2008 Sep;19(5):640–8. https://doi.org/10.1097/EDE.0b013e31818131e7
28. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013 May;14(5):365–76. https://doi.org/10.1038/nrn3475
29. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar 18;m441. https://doi.org/10.1136/bmj.m441
30. Kerr NL. HARKing: Hypothesizing After the Results are Known. Pers Soc Psychol Rev. 1998 Aug;2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4
31. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. Psychol Sci. 2012 May;23(5):524–32. https://doi.org/10.1177/0956797611430953
32. Banks GC, Rogelberg SG, Woznyj HM, Landis RS, Rupp DE. Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly. J Bus Psychol. 2016 Sep;31(3):323–38. https://doi.org/10.1007/s10869-016-9456-7
33. Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. BMJ. 2009 Jul 23;339(jul20 3):b2680–b2680. https://doi.org/10.1136/bmj.b2680
34. Kunda Z. The case for motivated reasoning. Psychological Bulletin. 1990;108(3):480–98. https://doi.org/10.1037/0033-2909.108.3.480
35. Nickerson RS. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology. 1998 Jun;2(2):175–220. https://doi.org/10.1037/1089-2680.2.2.175
36. Kim SH, Schramm S, Riedel EO, Schmitzer L, Rosenkranz E, Kertels O, et al. Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. Radiol med. 2025 Feb 12;130(4):555–66. https://doi.org/10.1007/s11547-025-01964-6
37. Kocak B, Baessler B, Bakas S, Cuocolo R, Fedorov A, Maier-Hein L, 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. 2023 May 4;14(1):75. https://doi.org/10.1186/s13244-023-01415-8
38. Kocak B, Akinci D'Antonoli T, Mercaldo N, Alberich-Bayarri A, Baessler B, Ambrosini I, et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insights Imaging. 2024 Jan 17;15(1):8. https://doi.org/10.1186/s13244-023-01572-w
39. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020 May;295(2):328–38. https://doi.org/10.1148/radiol.2020191145

 doi:10.20944/preprints202606.0240.v1

40. Buvat I, Orlhac F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. Journal of Nuclear Medicine. 2019 Nov 1;60(11):1543–4. https://doi.org/10.2967/jnumed.119.235325
41. Nardone V, Reginelli A, Rubini D, Gagliardi F, Del Tufo S, Belfiore MP, et al. Delta radiomics: an updated systematic review. Radiol med. 2024 Jul 17;129(8):1197–214. https://doi.org/10.1007/s11547-024-01853-4
42. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International Journal of Radiation Oncology*Biology*Physics. 2018 Nov;102(4):1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053
43. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018 Aug;288(2):407–15. https://doi.org/10.1148/radiol.2018172361
44. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features: Investigative Radiology. 2015 Nov;50(11):757–65. https://doi.org/10.1097/RLI.0000000000000180
45. Larue RTHM, Van Timmeren JE, De Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncologica. 2017 Nov 2;56(11):1544–53. https://doi.org/10.1080/0284186X.2017.1351624
46. Midya A, Chakraborty J, Gönen M, M.d RKGD, Simpson AL. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. JMI. 2018 Feb 15;5(1):011020. https://doi.org/10.1117/1.JMI.5.1.011020
47. Meyer M, Ronald J, Vernuccio F, Nelson RC, Ramirez-Giraldo JC, Solomon J, et al. Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings. Radiology. 2019 Dec;293(3):583–91. https://doi.org/10.1148/radiol.2019190928
48. Zhu L, Dong H, Sun J, Wang L, Xing Y, Hu Y, et al. Robustness of radiomics among photon-counting detector CT and dual-energy CT systems: a texture phantom study. Eur Radiol. 2024 Jul 24;35(2):871–84. https://doi.org/10.1007/s00330-024-10976-1
49. Zhang H, Lu T, Wang L, Xing Y, Hu Y, Xu Z, et al. Robustness of radiomics within photon-counting detector CT: impact of acquisition and reconstruction factors. Eur Radiol. 2025 Jan 31;35(8):4661–73. https://doi.org/10.1007/s00330-025-11374-x
50. Xue C, Yuan J, Lo GG, Chang ATY, Poon DMC, Wong OL, et al. Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review. Quant Imaging Med Surg. 2021 Oct;11(10):4431–60. https://doi.org/10.21037/qims-21-86
51. Kendrick J, Francis RJ, Hassan GM, Ong JSL, Jeraj R, Barry N, et al. Deep learning-based PSMA PET segmentation repeatability: A post-hoc analysis of a single-center, prospective, test–retest trial. Radiol med. 2025 Oct 30;131(2):320–31. https://doi.org/10.1007/s11547-025-02137-1
52. Whybra P, Zwanenburg A, Andrearczyk V, Schaer R, Apte AP, Ayotte A, et al. The Image Biomarker Standardization Initiative: Standardized Convolutional Filters for Reproducible Radiomics and Enhanced Clinical Insights. Radiology. 2024 Feb 1;310(2):e231319. https://doi.org/10.1148/radiol.231319
53. Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol. 2021 Apr;31(4):2272–80. https://doi.org/10.1007/s00330-020-07284-9
54. Da-ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci Rep. 2020 Jun 24;10(1):10248. https://doi.org/10.1038/s41598-020-66110-w
55. Demircioğlu A. Reproducibility and interpretability in radiomics: a critical assessment. dir. 2024 Oct 21. https://doi.org/10.4274/dir.2024.242719
56. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. Woloschak GE, editor. PLoS ONE. 2014 Jul 15;9(7):e102107. https://doi.org/10.1371/journal.pone.0102107

 doi:10.20944/preprints202606.0240.v1

57. Saha A, Harowicz MR, Mazurowski MA. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. Medical Physics. 2018 Jul;45(7):3076–85. https://doi.org/10.1002/mp.12925
58. deSouza NM, Van Der Lugt A, Deroose CM, Alberich-Bayarri A, Bidaut L, Fournier L, et al. Standardised lesion segmentation for imaging biomarker quantitation: a consensus recommendation from ESR and EORTC. Insights Imaging. 2022 Oct 4;13(1):159. https://doi.org/10.1186/s13244-022-01287-4
59. Song H, Wang X, Wu R, Liu W. The influence of manual segmentation strategies and different phases selection on machine learning-based computed tomography in renal tumors: a systematic review and meta-analysis. Radiol med. 2024 May 13;129(7):1025–37. https://doi.org/10.1007/s11547-024-01825-8
60. Kocak B, Bulut E, Bayrak ON, Okumus AA, Altun O, Borekci Arvas Z, et al. NEgatiVE results in Radiomics research (NEVER): A meta-research study of publication bias in leading radiology journals. European Journal of Radiology. 2023 Jun;163:110830. https://doi.org/10.1016/j.ejrad.2023.110830
61. Kocak B, Akinci D'Antonoli T, Ates Kus E, Keles A, Kala A, Kose F, et al. Self-reported checklists and quality scoring tools in radiomics: a meta-research. Eur Radiol. 2024 Jan 5;34(8):5028–40. https://doi.org/10.1007/s00330-023-10487-5
62. Halligan S, Menu Y, Mallett S. Why did European Radiology reject my radiomic biomarker paper? How to correctly evaluate imaging biomarkers in a clinical setting. Eur Radiol. 2021 Dec;31(12):9361–8. https://doi.org/10.1007/s00330-021-07971-1
63. Bleker J, Yakar D, van Noort B, Rouw D, de Jong IJ, Dierckx RAJO, et al. Single-center versus multi-center biparametric MRI radiomics approach for clinically significant peripheral zone prostate cancer. Insights Imaging. 2021 Oct 21;12(1):150. https://doi.org/10.1186/s13244-021-01099-y
64. Doniselli FM, Pascuzzo R, Mazzi F, Padelli F, Moscatelli M, Akinci D'Antonoli T, et al. Quality assessment of the MRI-radiomics studies for MGMT promoter methylation prediction in glioma: a systematic review and meta-analysis. Eur Radiol. 2024 Feb 3;34(9):5802–15. https://doi.org/10.1007/s00330-024-10594-x
65. Malcolm JA, Tacey M, Gibbs P, Lee B, Ko HS. Current state of radiomic research in pancreatic cancer: focusing on study design and reproducibility of findings. Eur Radiol. 2023 Oct;33(10):6659–69. https://doi.org/10.1007/s00330-023-09653-6
66. Zhong J, Liu X, Lu J, Yang J, Zhang G, Mao S, et al. Overlooked and underpowered: a meta-research addressing sample size in radiomics prediction models for binary outcomes. Eur Radiol. 2025 Jan 9;35(3):1146–56. https://doi.org/10.1007/s00330-024-11331-0
67. Horvat N, Papanikolaou N, Koh DM. Radiomics Beyond the Hype: A Critical Evaluation Toward Oncologic Clinical Use. Radiology: Artificial Intelligence. 2024 Jul 1;6(4):e230437. https://doi.org/10.1148/ryai.230437
68. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights Imaging. 2021 Dec;12(1):172. https://doi.org/10.1186/s13244-021-01115-1
69. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023 Sep;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804
70. Demircioğlu A. Applying oversampling before cross-validation will lead to high bias in radiomics. Sci Rep. 2024 May 21;14(1):11563. https://doi.org/10.1038/s41598-024-62585-z
71. Marzi C, Giannelli M, Barucci A, Tessa C, Mascalchi M, Diciotti S. Efficacy of MRI data harmonization in the age of machine learning: a multicenter study across 36 datasets. Sci Data. 2024 Jan 23;11(1):115. https://doi.org/10.1038/s41597-023-02421-7
72. Gidwani M, Chang K, Patel JB, Hoebel KV, Ahmed SR, Singh P, et al. Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models. Radiology. 2023 Apr;307(1):e220715. https://doi.org/10.1148/radiol.220715
73. Beddok A, Grogg K, Nioche C, Rozenblum L, Orlhac F, Calugaru V, et al. Predicting tumor recurrence site after reirradiation in head and neck cancer: a retrospective external validation of a published [18F]-FDG PET radiomic signature. Radiol med. 2025 Aug 20;130(11):1854–63. https://doi.org/10.1007/s11547-025-02072-1

 doi:10.20944/preprints202606.0240.v1

74. Kocak B, Keles A, Kose F, Sendur A. Quality of radiomics research: comprehensive analysis of 1574 unique publications from 89 reviews. Eur Radiol. 2024 Sep 6;35(4):1980–92. https://doi.org/10.1007/s00330-024-11057-z
75. Barry N, Kendrick J, Molin K, Li S, Rowshanfarzad P, Hassan GM, et al. Evaluating the impact of the Radiomics Quality Score: a systematic review and meta-analysis. Eur Radiol. 2025 Jan 10;35(3):1701–13. https://doi.org/10.1007/s00330-024-11341-y
76. Spadarella G, Ugga L, Calareso G, Villa R, D'Aniello S, Cuocolo R. The impact of radiomics for human papillomavirus status prediction in oropharyngeal cancer: systematic review and radiomics quality score assessment. Neuroradiology. 2022 Aug;64(8):1639–47. https://doi.org/10.1007/s00234-022-02959-0
77. On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec;17(1):230. https://doi.org/10.1186/s12916-019-1466-7
78. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Making. 2006 Nov;26(6):565–74. https://doi.org/10.1177/0272989X06295361
79. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020 Jan;30(1):523–36. https://doi.org/10.1007/s00330-019-06360-z
80. Spadarella G, Stanzione A, Akinci D'Antonoli T, Andreychenko A, Fanni SC, Ugga L, et al. Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. Eur Radiol. 2022 Oct 25;33(3):1884–94. https://doi.org/10.1007/s00330-022-09187-3
81. Kocak B, Ammirabile A, Ambrosini I, Akinci D'Antonoli T, Borgheresi A, Cavallo AU, et al. Explanation and Elaboration with Examples for METRICS (METRICS-E3): an initiative from the EuSoMII Radiomics Auditing Group. Insights Imaging. 2025 Aug 13;16(1):175. https://doi.org/10.1186/s13244-025-02061-y
82. Cavallo AU, Stanzione A, Ponsiglione A, Trotta R, Fanni SC, Ghezzo S, et al. Prostate cancer MRI methodological radiomics score: a EuSoMII radiomics auditing group initiative. Eur Radiol. 2024 Dec 30;35(3):1157–65. https://doi.org/10.1007/s00330-024-11299-x
83. Cavallo AU, Ponsiglione A, Pereira B, Di Donna C, Koltsakis E, Vernuccio F, et al. CT and MRI radiomics in cardiovascular risk prediction: a systematic review and meta-analysis by the EuSoMII Radiomics Auditing Group. Eur Radiol. 2025 Dec 24;36(5):4049–60. https://doi.org/10.1007/s00330-025-12236-2
84. Kocak B, Mese I, Ates Kus E. Radiomics for differentiating radiation-induced brain injury from recurrence in gliomas: systematic review, meta-analysis, and methodological quality evaluation using METRICS and RQS. Eur Radiol. 2025 Feb 12;35(8):4490–505. https://doi.org/10.1007/s00330-025-11401-x
85. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiology: Artificial Intelligence. 2022 May 1;4(3):e210064. https://doi.org/10.1148/ryai.210064
86. Zhong J, Lu J, Zhang G, Mao S, Chen H, Yin Q, et al. An overview of meta-analyses on radiomics: more evidence is needed to support clinical translation. Insights Imaging. 2023 Jun 19;14(1):111. https://doi.org/10.1186/s13244-023-01437-2
87. Ahmadzadeh AM, Ashoobi MA, Broomand Lomer N, Elyassirad D, Gheiji B, Vatanparast M, et al. Application of Deep Learning for Predicting Hematoma Expansion in Intracerebral Hemorrhage Using Computed Tomography Scans: A Systematic Review and Meta-Analysis of Diagnostic Accuracy. Radiol med. 2025 Sep 11;130(12):1973–85. https://doi.org/10.1007/s11547-025-02089-6
88. Kocak B, Ponsiglione A, Stanzione A, Ugga L, Klontzas ME, Cannella R, et al. CLEAR guideline for radiomics: Early insights into current reporting practices endorsed by EuSoMII. European Journal of Radiology. 2024 Dec;181:111788. https://doi.org/10.1016/j.ejrad.2024.111788
89. Akinci D'Antonoli T, Cuocolo R, Baessler B, Pinto Dos Santos D. Towards reproducible radiomics research: introduction of a database for radiomics studies. Eur Radiol. 2023 Aug 12;34(1):436–43. https://doi.org/10.1007/s00330-023-10095-3
90. Kocak B, Yardimci AH, Yuzkan S, Keles A, Altun O, Bulut E, et al. Transparency in Artificial Intelligence Research: a Systematic Review of Availability Items Related to Open Science in Radiology and Nuclear Medicine. Academic Radiology. 2023 Oct;30(10):2254–66. https://doi.org/10.1016/j.acra.2022.11.030

 doi:10.20944/preprints202606.0240.v1

91. Foy JJ, Robinson KR, Li H, Giger ML, Al-Hallaq H, Armato SG. Variation in algorithm implementation across radiomics software. J Med Imag. 2018 Dec 4;5(4):044505. https://doi.org/10.1117/1.JMI.5.4.044505
92. Challa AB, Radike M, Rizvi A, Weber NM, Wamil M, Poigai Arunachalam S, et al. Interobserver and intraobserver variability among different vendors for mitral valve assessment: implications for transcatheter mitral valve repair. Radiol med. 2025 Jan 15;130(3):296–301. https://doi.org/10.1007/s11547-025-01950-y
93. Chetan MR, Gleeson FV. Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives. Eur Radiol. 2021 Feb;31(2):1049–58. https://doi.org/10.1007/s00330-020-07141-9
94. Zhong J, Davey A, Frood R, McWilliam A, Shortall J, Reardon M, et al. Combining MRI radiomics, hypoxia gene signature score and clinical variables for prediction of biochemical recurrence-free survival after radiotherapy in prostate cancer. Radiol med. 2025 Jul 2;130(8):1139–48. https://doi.org/10.1007/s11547-025-02037-4
95. Ter Maat LS, van Duin IAJ, Elias SG, Leiner T, Verhoeff JJC, Arntz ERAN, et al. CT radiomics compared to a clinical model for predicting checkpoint inhibitor treatment outcomes in patients with advanced melanoma. Eur J Cancer. 2023 May;185:167–77. https://doi.org/10.1016/j.ejca.2023.02.017
96. Peng W, Wan L, Wang S, Zou S, Zhao X, Zhang H. A multiple-time-scale comparative study for the added value of magnetic resonance imaging-based radiomics in predicting pathological complete response after neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Front Oncol. 2023;13:1234619. https://doi.org/10.3389/fonc.2023.1234619
97. Lambin P, Woodruff HC, Mali SA, Zhong X, Kuang S, Lavrova E, et al. Radiomics Quality Score 2.0: towards radiomics readiness levels and clinical translation for personalized medicine. Nat Rev Clin Oncol. 2025 Nov;22(11):831–46. https://doi.org/10.1038/s41571-025-01067-1
98. McGale J, Beddok A, Schwartz LH, Dercle L. Radiomics Quality Score 2.0: what changed from version 1.0 and why it matters. Nat Rev Clin Oncol. 2026 Jan;23(1):84–5. https://doi.org/10.1038/s41571-025-01098-8
99. Demircioğlu A. Retractions of publications in radiomics: An underestimated problem? Eur Radiol. 2025 Dec 20;36(5):3778–87. https://doi.org/10.1007/s00330-025-12231-7
100. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magnetic Resonance Imaging. 2012 Nov;30(9):1234–48. https://doi.org/10.1016/j.mri.2012.06.010
101. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014 Jun 3;5:4006. https://doi.org/10.1038/ncomms5006
102. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging. 2020 Aug 12;11(1):91. https://doi.org/10.1186/s13244-020-00887-2
103. Lee J, Steinmann A, Ding Y, Lee H, Owens C, Wang J, et al. Radiomics feature robustness as measured using an MRI phantom. Sci Rep. 2021 Feb 17;11(1):3973. https://doi.org/10.1038/s41598-021-83593-3
104. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncologica. 2018 Aug 3;57(8):1070–4. https://doi.org/10.1080/0284186X.2018.1445283
105. Granzier RWY, Verbakel NMH, Ibrahim A, van Timmeren JE, van Nijnatten TJA, Leijenaar RTH, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. Sci Rep. 2020 Aug 25;10(1):14163. https://doi.org/10.1038/s41598-020-70940-z
106. Poirot MG, Caan MWA, Ruhe HG, Bjørnerud A, Groote I, Reneman L, et al. Robustness of radiomics to variations in segmentation methods in multimodal brain MRI. Sci Rep. 2022 Oct 6;12:16712. https://doi.org/10.1038/s41598-022-20703-9
107. Park JE, Kim HS, Kim D, Park SY, Kim JY, Cho SJ, et al. A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. BMC Cancer. 2020 Dec;20(1):29. https://doi.org/10.1186/s12885-019-6504-5

 doi:10.20944/preprints202606.0240.v1

108. Jannot AS, Agoritsas T, Gayet-Ageron A, Perneger TV. Citation bias favoring statistically significant studies was present in medical research. J Clin Epidemiol. 2013 Mar;66(3):296–301. https://doi.org/10.1016/j.jclinepi.2012.09.015
109. Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017 Aug;88:92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002
110. Vickers AJ, Woo S. Decision curve analysis in the evaluation of radiology research. Eur Radiol. 2022 Mar 29;32(9):5787–9. https://doi.org/10.1007/s00330-022-08685-8
111. Chiu K, Grundy Q, Bero L. "Spin" in published biomedical literature: a methodological systematic review. PLoS Biology. 2017;15(9):e2002173. https://doi.org/10.1371/journal.pbio.2002173
112. McGrath TA, McInnes MDF, van Es N, Leeflang MMG, Korevaar DA, Bossuyt PMM. Overinterpretation of research findings: evidence of "spin" in systematic reviews of diagnostic accuracy studies. Clinical Chemistry. 2017;63(8):1353–1362. https://doi.org/10.1373/clinchem.2017.271544
113. Oh YK. Position: State-of-the-Art Claims Require State-of-the-Art Evidence. arXiv. 2026. arXiv:2605.17273v2 [cs.LG]. https://doi.org/10.48550/arXiv.2605.17273
114. Di Cesare E, Esposito A, Lo Casto A, Mazzei MA, Polonara G, Sverzellati N, et al. CT acquisition protocols by pathology, SIRM position paper part 1: head and neck, brain and spine, chest, cardiovascular. Radiol med. 2025 Jul 29;130(10):1594–601. https://doi.org/10.1007/s11547-025-02018-7
115. Di Cesare E, Ascenti G, Cappabianca S, Granata C, Reginelli A, Trinci M, et al. CT acquisition protocols by pathology, SIRM position paper part 2 (Abdominal and Oncologic Imaging, Urology, Paediatric). Radiol med. 2025 Oct 28;131(2):292–301. https://doi.org/10.1007/s11547-025-02123-7
116. Kocak B, Borgheresi A, Ponsiglione A, Andreychenko AE, Cavallo AU, Stanzione A, et al. Explanation and Elaboration with Examples for CLEAR (CLEAR-E3): an EuSoMII Radiomics Auditing Group Initiative. Eur Radiol Exp. 2024 May 14;8(1):72. https://doi.org/10.1186/s41747-024-00471-z
117. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006 Dec;7(1):91. https://doi.org/10.1186/1471-2105-7-91
118. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025 Mar 24;388:e082505. https://doi.org/10.1136/bmj-2024-082505
119. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378. https://doi.org/10.1136/bmj-2023-078378
120. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. Cham: Springer International Publishing; 2019 [cited 2026 May 17]. (Statistics for Biology and Health). Available from: http://link.springer.com/10.1007/978-3-030-16399-0 https://doi.org/10.1007/978-3-030-16399-0
121. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology. 2016 Jan;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005
122. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022 May;28(5):924–33. https://doi.org/10.1038/s41591-022-01772-9
123. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020 Apr 1;27(4):621–33. https://doi.org/10.1093/jamia/ocz228
124. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016 Jan 25;i6. https://doi.org/10.1136/bmj.i6
125. Park SH, Han K, Lee JG. Conceptual review of outcome metrics and measures used in clinical evaluation of artificial intelligence in radiology. Radiol med. 2024 Sep 3;129(11):1644–55. https://doi.org/10.1007/s11547-024-01886-9

 doi:10.20944/preprints202606.0240.v1

126. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018. https://doi.org/10.1038/sdata.2016.18
127. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015 Jun 26;348(6242):1422–5. https://doi.org/10.1126/science.aab2374
128. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proc Natl Acad Sci USA. 2018 Mar 13;115(11):2600–6. https://doi.org/10.1073/pnas.1708274114
129. Wagenmakers EJ, Wetzels R, Borsboom D, Van Der Maas HLJ, Kievit RA. An Agenda for Purely Confirmatory Research. Perspect Psychol Sci. 2012 Nov;7(6):632–8. https://doi.org/10.1177/1745691612463078
130. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. Perspect Psychol Sci. 2016 Sep;11(5):702–12. https://doi.org/10.1177/1745691616658637
131. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nat Hum Behav. 2020 Nov;4(11):1208–14. https://doi.org/10.1038/s41562-020-0912-z
132. Pai S, Bontempi D, Hadzic I, Prudente V, Sokač M, Chaunzwa TL, Bernatz S, Hosny A, Mak RH, Birkbak NJ, Aerts HJWL. Foundation model for cancer imaging biomarkers. Nat Mach Intell. 2024 Mar;6(3):354–367. https://doi.org/10.1038/s42256-024-00807-9
133. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. Nature. 2023 Apr;616(7956):259–265. https://doi.org/10.1038/s41586-023-05881-4
134. Koetzier LR, Wu J, Mastrodicasa D, Lutz A, Chung M, Koszek WA, Pratap J, Chaudhari AS, Rajpurkar P, Lungren MP, Willemink MJ. Generating synthetic data for medical imaging. Radiology. 2024 Sep;312(3):e232471. https://doi.org/10.1148/radiol.232471
135. Mali SA, Mohammadian Rad N, Woodruff HC, Depeursinge A, Andrearczyk V, Lambin P. Harmonizing CT scanner acquisition variability in an anthropomorphic phantom: a comparative study of image-level and feature-level harmonization using GAN, ComBat, and their combination. PLoS One. 2025;20(5):e0322365. https://doi.org/10.1371/journal.pone.0322365
136. Floca R, Bohn J, Haux C, Wiestler B, Zöllner FG, Reinke A, et al. Radiomics workflow definition & challenges - German priority program 2177 consensus statement on clinically applied radiomics. Insights Imaging. 2024 Jun 3;15(1):124. https://doi.org/10.1186/s13244-024-01704-w
137. Santinha J, Pinto Dos Santos D, Laqua F, Visser JJ, Groot Lipman KBW, Dietzel M, et al. ESR Essentials: radiomics—practice recommendations by the European Society of Medical Imaging Informatics. Eur Radiol. 2024 Oct 25;35(3):1122–32. https://doi.org/10.1007/s00330-024-11093-9
138. Avanzo M, Soda P, Bertolini M, Bettinelli A, Rancati T, Stancanello J, et al. Robust radiomics: a review of guidelines for radiomics in medical imaging. Front Radiol. 2026 Jan 12;5:1701110. https://doi.org/10.3389/fradi.2025.1701110
139. Nosek BA, Spies JR, Motyl M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. Perspect Psychol Sci. 2012 Nov;7(6):615–31. https://doi.org/10.1177/1745691612459058
140. Chambers CD, Tzavella L. The past, present and future of Registered Reports. Nat Hum Behav. 2021 Nov 15;6(1):29–42. https://doi.org/10.1038/s41562-021-01193-7
141. Scheel AM, Schijen MRMJ, Lakens D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. Advances in Methods and Practices in Psychological Science. 2021 Apr;4(2):25152459211007467. https://doi.org/10.1177/25152459211007467
142. Cagan R. The San Francisco Declaration on Research Assessment. Dis Model Mech. 2013 Jul;6(4):869–70. https://doi.org/10.1242/dmm.012955
143. Xu HL, Gong TT, Song XJ, et al. Artificial Intelligence Performance in Image-Based Cancer Identification: Umbrella Review of Systematic Reviews. J Med Internet Res. 2025;27:e53567. https://doi.org/10.2196/53567
144. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024 Jul 12;OJ L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

 doi:10.20944/preprints202606.0240.v1

145. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and FDA staff. Silver Spring (MD): FDA; 2024 Dec 3. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence
146. Huang ML, Ren J, Jin ZY, Liu XY, Li Y, He YL, et al. Application of magnetic resonance imaging radiomics in endometrial cancer: a systematic review and meta-analysis. Radiol med. 2024 Feb 13;129(3):439–56. https://doi.org/10.1007/s11547-024-01765-3
147. Wright BD, Vo N, Nolan J, Johnson AL, Braaten T, Tritz D, et al. An analysis of key indicators of reproducibility in radiology. Insights Imaging. 2020 May 11;11(1):65. https://doi.org/10.1186/s13244-020-00870-x
148. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 2017 Nov 1;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339
149. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: The need for safeguards. Radiotherapy and Oncology. 2019 Jan;130:2–9. https://doi.org/10.1016/j.radonc.2018.10.027
150. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiology: Artificial Intelligence. 2024 Jul 1;6(4):e240300. https://doi.org/10.1148/ryai.240300
151. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020 Sep;26(9):1351–63. https://doi.org/10.1038/s41591-020-1037-7
152. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ. 2020 Sep 9;m3164. https://doi.org/10.1136/bmj.m3164
153. Chen D, He E, Pace K, Chekay M, Raman S. Concordance with SPIRIT-AI guidelines in reporting of randomized controlled trial protocols investigating artificial intelligence in oncology: a systematic review. Oncologist. 2025 May 8;30(5):oyaf112. https://doi.org/10.1093/oncolo/oyaf112
154. Chen D, Arnold K, Sukhdeo R, Farag Alla J, Raman S. Concordance with CONSORT-AI guidelines in reporting of randomised controlled trials investigating artificial intelligence in oncology: a systematic review. BMJ Oncol. 2025;4(1):e000733. https://doi.org/10.1136/bmjonc-2025-000733
155. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025 Feb 5;388:e081554. https://doi.org/10.1136/bmj-2024-081554
156. International Medical Device Regulators Forum, Artificial Intelligence/Machine Learning-enabled Medical Devices Working Group. Good Machine Learning Practice for Medical Device Development: Guiding Principles. IMDRF/AIML WG/N88 FINAL:2025. 2025 Jan 27. Available from: https://www.imdrf.org/sites/default/files/2025-01/IMDRF_AIML%20WG_GMLP_N88%20Final_0.pdf
157. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). OJ L [Internet]. 2017 Apr 5. Available from: http://data.europa.eu/eli/reg/2017/745/oj