Medical AI Has Changed. Has the Basis of Clinical Trust Kept Pace?

From narrow computer vision to foundation models, agentic workflows, and the unresolved question of how imaging observations acquire meaning

Vladimir A. Novikov, MD

Radiologist · Head of Radiology Department · Co-founder, ResetRay

Perspective · Article I of II · First published July 2026

Correspondence: dr.vladimir.novikov@gmail.com · resetray.com

Abstract

Artificial intelligence in radiology no longer fits the image of a single algorithm looking for a single abnormality. Today's systems detect and segment anatomy, prioritize examinations, reconstruct images, produce quantitative measurements, compare current and prior studies, assist with reporting, and increasingly participate in multimodal and agentic workflows. In several well-defined tasks, they already provide genuine clinical value.

Daily practice, however, is less controlled than a benchmark. Reliability must survive different institutions, populations, scanners, acquisition protocols, software versions, and regulatory environments. Medical AI does not learn disease as an abstract objective entity; it learns from the representations of disease made available by human observers, institutions, devices, measurement conventions, reference standards, and datasets. The same lesion may be contoured, measured, classified, or clinically weighted differently by different radiologists, and models trained in different environments can produce similarly divergent representations.

This perspective brings together recent developments in radiology AI, evidence of clinical benefit, the public strategies of GE HealthCare, Philips, and Siemens Healthineers, and the scientific, technical, and human factors that determine whether an output deserves clinical trust. Its central argument is that technical interoperability is not the same as semantic equivalence, and that trust belongs not to the model alone but to the full lifecycle through which data become observations and observations enter clinical work. The article ends with a question rather than a product claim: what if part of the limitation lies not only in the model, but in the form in which medical imaging information reaches it?

Keywords

artificial intelligence; radiology; medical imaging; foundation models; quantitative imaging; reference standard; clinical validation; interoperability; semantic consistency; provenance; reproducibility

1. The question has changed

As a practicing radiologist, I encounter artificial intelligence less in conference slides than in the ordinary chain of clinical work: acquisition, image review, comparison with prior examinations, measurement, reporting, communication, and follow-up. From that position, neither of the familiar extremes is convincing. Medical AI is not an empty promise, but neither is it an autonomous substitute for clinical judgement.

That makes the old yes-or-no debate increasingly unhelpful. In a growing number of defined tasks, AI can detect findings, prioritize examinations, segment anatomy, quantify abnormalities, and shorten repetitive work. In many constrained settings, the machine can identify pathology on an image. The more difficult question begins after that point.

The harder question is this: what justifies confidence that an AI output will remain reliable when it leaves the dataset on which it was developed and enters the variable world of real patients, different scanners, changing protocols, local clinical conventions, and evolving software? A high sensitivity, specificity, or area under the receiver operating characteristic curve may describe performance in a study. It does not, by itself, describe the conditions under which that performance will persist.

2. From narrow models to imaging ecosystems

The early clinical story of radiology AI was largely a story of narrow models: one system for intracranial hemorrhage, another for pulmonary embolism, another for fractures or pulmonary nodules. These tools remain useful. They simply no longer describe the leading edge of the field.

Computer vision still forms the operational core of most imaging AI. Detection, classification, and segmentation make it possible to locate findings, delineate organs, calculate dimensions and volumes, estimate disease burden, and support longitudinal comparison. Segmentation is especially important in radiotherapy, procedural planning, oncology, quantitative tissue analysis, and radiomics. It is also already an act of representation: a boundary depends on the image, the observer or algorithm, the reconstruction, the viewing conditions, and the rule used to decide what belongs inside or outside the object.

AI is also moving earlier in the imaging chain. Deep-learning reconstruction, denoising, artifact reduction, motion correction, accelerated acquisition, and dose optimization alter the image before a radiologist or downstream model interprets it. This is clinically valuable, but it expands the trust problem. An error need not arise only when a model classifies a finding; it may be introduced while the image itself is being formed or transformed.

Quantitative imaging and opportunistic screening represent another important shift. A CT examination performed for one indication may also contain measurable information about bone attenuation, skeletal muscle, visceral fat, hepatic steatosis, vascular calcification,

cardiometabolic risk, and biological ageing. Pickhardt and colleagues describe this as an opportunity to extract clinically valuable information that historically remained unused because manual measurement was laborious and some visual assessments were subjective [8]. The image is no longer merely something to be read. It is also a reservoir of potentially reusable observations.

Foundation models extend this logic. Rather than training a separate model for every task, large models are pretrained on extensive image or multimodal corpora and then adapted to detection, segmentation, retrieval, classification, or report-related tasks. A 2025 review in Radiology set out both their potential and their risks, including domain shift, bias, hallucination, automation bias, and the need for standardized evaluation [1]. The open Ark+ chest radiography model demonstrated how heterogeneous expert labels and multiple datasets can be combined into a reusable foundation model, while also illustrating how strongly such systems remain shaped by the labels and populations on which they learn [2].

Multimodal systems add clinical indications, reports, prior studies, electronic health record data, or other model outputs to the image. Their potential advantage is obvious: radiologists do not interpret images in isolation, and neither should advanced computational systems. But every additional input also creates a provenance question. Which information influenced the output? Was it current, correctly linked to the patient and examination, technically compatible, and clinically relevant? The need for the 2026 FLAIR reporting framework arose in part because studies of radiology large language models were inconsistent in their description of data, prompting, evaluation, risks, and implementation [3].

Agentic AI is the next conceptual step. An agent may retrieve prior studies, navigate clinical software, call a segmentation tool, request a measurement, query a knowledge source, and assemble a preliminary report. A 2026 review in Radiology: Artificial Intelligence describes this move from passive model output toward goal-directed systems with memory, tool use, and multi-agent coordination [4]. Once models begin to act through other systems, trust can no longer be attached only to the final sentence. It must extend to the sequence of intermediate observations and actions from which that sentence emerged.

3. What is already working — and what the evidence actually shows

A useful critique of medical AI has to begin with what it already does well. The field has moved beyond proof-of-concept demonstrations, and several recent studies report meaningful performance in real-world or multicenter settings.

In a 2026 study of 32,501 CT pulmonary angiography examinations, an FDA-cleared pulmonary embolism tool operated within a large clinical network. Concordance between AI and AI-informed radiologists was high, but expert adjudication of discordant cases supported the radiologist substantially more often than the algorithm. Performance also varied by

embolus type and location, with greater difficulty in chronic and subsegmental disease [5]. The pattern is more informative than either a success or failure label: the tool was valuable in a defined workflow, yet its performance remained uneven across the clinical spectrum of a single diagnosis.

A multicenter study of 2,980 CT angiography examinations evaluated autonomous quantification of intracranial aneurysm and parent-artery morphology. The system reduced morphology extraction from approximately 24 minutes of physician work to about 90 seconds and produced many aggregate measurements without statistically significant differences from expert-derived values [6]. The gain is not merely speed. It shows how AI can add value by producing reproducible intermediate observations, even before one asks it to issue a diagnostic conclusion.

In BreastScreen Norway, an AI model applied to more than 30,000 digital breast tomosynthesis examinations achieved an AUC of 0.93 and performance similar to independent double reading; prior examinations produced only limited additional improvement [7]. The result is clinically important, but it belongs to a specific model, population, screening programme, reference standard, and thresholding strategy. It cannot be generalized automatically to every mammography environment.

Taken together, these studies support a measured conclusion. AI can improve triage, accelerate quantification, standardize repetitive work, and detect defined abnormalities. What they do not establish is universal reliability of medical AI as a category. Every claim remains tied to a population, acquisition process, reference standard, workflow, threshold, and method of evaluation.

4. What major imaging manufacturers are building

Manufacturer materials are not clinical trials. They describe product capabilities, intended use, and strategic direction. Read with that limitation in mind, they are still useful because they show where the imaging industry is investing and what kinds of information it expects future workflows to carry.

Siemens Healthineers has moved well beyond a single abnormality detector. AI-Rad Companion Chest CT can identify and segment pulmonary lesions, calculate dimensions and volumes, measure the thoracic aorta, evaluate vertebral height, report bone attenuation in Hounsfield units, and quantify pulmonary density findings. Siemens documentation also describes writing selected measurements into DICOM Structured Reports and sending reviewed or automatically confirmed results to PACS [24]. The important point is not that Siemens has solved the whole information problem. It is that a major manufacturer already treats automated quantitative observations as objects that should pass from image analysis into the clinical information environment.

GE HealthCare follows a related but more application-specific path. Thoracic VCAR combines automated lung, lobe, and airway segmentation with attenuation-range analysis, airway-wall measurements, and relative perfusion maps based on iodine concentration [25]. These are anatomical, geometric, and physical observations, not merely binary disease predictions. At the same time, their organization within specialized applications illustrates the prevailing fragmentation of the market: powerful measurements remain attached to particular organs, modalities, protocols, and product environments.

Philips is developing both richer acquisition data and orchestration. Detector-based spectral CT preserves spectral information during routine scanning and allows retrospective analysis within the reading environment. Philips AI Manager integrates multiple algorithms with RIS and PACS and returns AI-generated results to the radiologist's primary workspace; the Vestre Viken implementation shows how this can standardize workflows across institutions [26]. Yet orchestration of applications and semantic integration of their outputs are not the same achievement. A platform may route an examination correctly and still receive results whose methods, units, anatomical identities, uncertainty, and comparability differ.

The lesson I take from these examples is not that the major manufacturers are approaching the problem incorrectly. On the contrary, they validate the direction: imaging is becoming more quantitative, more structured, more integrated, and more computational. The unresolved opportunity lies between successful specialist applications. How can observations created by different tools retain their meaning when they move across systems, vendors, institutions, and time?

5. Medical AI learns representations, not disease in the abstract

The deepest limitation of contemporary medical AI is not a shortage of computing power. It is the character of the knowledge from which the system learns.

A model does not encounter disease as a philosophical or physically complete object. It encounters images linked to information that people and institutions have produced or selected: contours, measurements, diagnostic categories, reports, pathology, clinical outcomes, registries, or expert consensus. The evolution of image annotation described by Flanders and colleagues shows how large language models can extract and normalize labels from radiology reports at scale [9]. This can transform dataset construction, but the underlying source remains a human clinical document with its own omissions, conventions, and uncertainty.

Self-supervised learning changes, but does not eliminate, this dependence. A foundation model can learn statistical structure from millions of unlabelled images. It may discover patterns no radiologist explicitly annotated. Yet those patterns become medically meaningful only when they are related to clinical concepts, outcomes, interventions, or expert

interpretation. The bridge from correlation to clinical meaning still passes through human medicine.

Medical AI does not learn disease in the abstract. It learns from the representations of disease that people, institutions, devices, measurement methods, and datasets make available to it.

That observation does not diminish the importance of human knowledge; without it, clinical AI would have no clinical meaning. The difficulty is that variability, omissions, local conventions, and implicit assumptions enter the system as well, while the underlying observation is not always preserved in a form that allows those influences to be examined.

6. One finding, several legitimate representations

A pulmonary nodule offers a familiar example. One radiologist may select an axial image, another a multiplanar reconstruction. One may report the maximum diameter, another the mean of two orthogonal diameters. A contour may include or exclude an attached vessel. The result may change with lung window settings, slice thickness, reconstruction kernel, partial-volume effects, motion, or the chosen software tool.

Two experienced radiologists can recognize the same lesion yet produce different boundaries, diameters, volumes, morphology descriptors, risk categories, or conclusions about interval change. Such variation is not always simple error. Sometimes the image does not contain enough information to support a single unambiguous boundary.

A 2025 study in which six clinicians segmented 232 pulmonary nodules found that inter-observer variation affected radiomic features, although most features remained relatively stable; the authors still concluded that improving segmentation consistency reduces variability [10]. ReaderAdaptNet, developed in work involving GE HealthCare researchers, takes the argument further. Instead of treating inter-reader variation in breast density and background parenchymal enhancement as noise to be averaged away, it models reader-specific patterns directly [11].

AI systems inherit the same structure. A model trained on one institution's contours, reports, or thresholds may not represent an object in the same way as a model trained elsewhere. One system returns a diameter, another a volume, another a malignancy score, another a risk category, and another a free-text statement. Each output may be clinically useful. They are not automatically equivalent.

The distinction matters: a lesion is not identical to its contour, measurement, probability, risk category, or sentence in a report. Each is an informational view of the same underlying phenomenon, produced by a particular method for a particular purpose.

7. Variability is also engineered, institutional, and regulatory

Observer variability is only one source of instability. The image itself is generated through different manufacturers, detector technologies, scanner models, protocols, doses, reconstruction methods, slice thicknesses, post-processing pipelines, and software versions. These choices can affect conspicuity, attenuation, segmentation, texture features, quantitative biomarkers, and model performance.

DICOM remains the indispensable foundation for image exchange, but interoperability at the file level does not erase manufacturer-specific encoding. A 2025 Scientific Data project assembled validated CT and MR DICOM datasets specifically to show how acquisition details are stored differently across manufacturers and software versions, including the use of private tags, proprietary structures, and inconsistent terminology [22]. The paper also documents how third-party PACS systems may alter tags or compression, adding another layer between acquisition and later analysis.

Clinical meaning is further shaped by local classification systems, thresholds, reporting customs, population characteristics, and legal frameworks. The European Union places certain medical-device AI systems within the high-risk provisions of the AI Act alongside the existing medical-device framework, while the United States has developed lifecycle and predetermined-change-control guidance through the FDA [27,28]. International guidance such as the 2025 IMDRF good machine-learning-practice principles seeks convergence, but it does not make the legal and operational environments identical [29].

An apparently simple number therefore carries a history. A measurement is not only a value; it is a value linked to a source image, anatomical target, acquisition context, method, software version, quality state, and set of conventions.

8. Performance is relational, not intrinsic

One of the most persistent findings in medical AI research is performance loss on external data. In a systematic review of 83 studies reporting 86 algorithms, 81% showed some decrease in external performance; 49% showed a decrease of at least 0.05 on the unit scale and 24% a decrease of at least 0.10 [12].

A 2026 systematic review and meta-analysis of externally tested models for lung-nodule malignancy classification found pooled sensitivity of 88% and specificity of 75%, alongside very high heterogeneity, inconsistent reporting, and important risk of bias [13]. Those figures do not make the models clinically useless. They show that performance depends on where, on whom, and under which conditions a model is used.

Disease prevalence, referral patterns, patient demographics, scanner mix, acquisition protocols, case selection, image quality, and reference-standard construction all alter the

relationship between a model and its environment. Even fairness properties may change when a model moves between institutions or populations.

Accuracy is therefore better understood as a relationship among a model, a task, a population, an acquisition process, a threshold, and a reference standard. When one of those elements changes, the meaning of the reported metric may change with it.

9. The reference standard also contains uncertainty

In clinical conversation, the phrase ground truth is convenient. In medical imaging, it can also be misleading. A reference standard may be a single report, consensus reading, pathology, longitudinal stability, a registry entry, an automatically extracted label, or a combination of sources. These are not interchangeable forms of truth.

The rapid use of large language models to create labels from reports makes this problem newly visible. Chavoshi and colleagues showed through simulation that small errors in LLM-generated reference labels can produce large and prevalence-dependent distortions in the apparent performance of a diagnostic model [14]. A subsequent commentary situated the finding within the established framework of imperfect reference-standard bias [15].

Scale does not cure a systematic error; it can amplify it. A million automatically generated labels may be more statistically powerful than ten thousand manually curated labels, but the same bias can spread through the larger dataset. More data are not always more knowledge. Sometimes they are more examples of the same representation.

A trustworthy dataset therefore requires more than volume. It requires an account of who or what produced the label, which evidence supported it, how disagreement was resolved, what uncertainty remains, and whether the label corresponds to the clinical question the model is intended to answer.

10. Human oversight is not a simple failsafe

One familiar reassurance is that AI will propose and the radiologist will verify. Human review is essential, but it is not a complete safety model. Verification is itself influenced by time pressure, workload, interface design, display timing, confidence scores, and the location and salience of prompts.

A 2026 eye-tracking study of screening mammography showed that incorrect AI prompts changed both diagnostic accuracy and visual-search behaviour. False-negative suggestions produced the greatest negative effect, and visible prompts also altered reading time and fixation patterns [19]. The study matters because it makes a broader point: human and machine errors are not independent once they share an interface.

The 2026 Radiology primer on interacting with AI results emphasizes that clinical value depends on how results are generated, displayed, accepted, rejected, stored, and integrated into workflow [18]. A good model presented at the wrong time, in an opaque form, or outside the reporting environment may deliver little value. A flawed result presented with excessive authority may create harm.

A generic 'human in the loop' label is therefore not enough. Safety and trust have to be designed into the interaction itself.

11. Trust is becoming a lifecycle property

Contemporary guidance increasingly rejects the idea that trust can be established by one validation study or one regulatory authorization. FUTURE-AI defines six principles — fairness, universality, traceability, usability, robustness, and explainability — and applies them across design, development, validation, deployment, and monitoring [16].

The ACR's Assess-AI registry is an operational response to the same problem. It is designed to receive de-identified AI output, report text, and DICOM metadata, compare algorithm output with surrogate labels, monitor longitudinal concordance, and support local review of discordant cases [17]. The ACR–SIIM Practice Parameter for Imaging AI, approved in 2026, similarly emphasizes governance, local acceptance testing, version tracking, ongoing monitoring, and defined responses to safety concerns [23].

These initiatives reflect a practical reality: a model may remain unchanged while the world around it changes. A scanner may be replaced, a protocol revised, a patient population shifted, a dependent software component updated, or reporting practice altered. Static authorization cannot substitute for observation of the deployed system.

Clinical trust is not granted once. It has to be maintained as the relationship between a model and its environment evolves.

12. Interoperability is not semantic equivalence

Radiology does not lack standards. DICOM supports the storage and exchange of images and related objects. DICOM Structured Reporting and DICOM Segmentation can represent measurements and masks. HL7 FHIR exchanges health information. RadLex, SNOMED CT, LOINC, and UCUM provide terminology, codes, and units.

The 2026 Radiology Reimagined report demonstrates how DICOM, FHIR, and IHE profiles can connect AI across ordering, protocoling, acquisition, interpretation, reporting, communication, follow-up, and billing. More than 20 vendor partners have participated in annual demonstrations, and the scenarios have become progressively more complex [20]. This is substantial progress.

But the ability to transmit a result does not guarantee that two systems mean the same thing by it. Suppose two applications report an 8-mm nodule. Semantic equivalence requires more than the number and unit. It requires identity of the lesion, source examination and series, anatomical location, plane and method of measurement, boundary rule, windowing or preprocessing conditions, software and model version, human correction status, quality assessment, and uncertainty.

The 2025 FHIR body-composition study described interoperable exchange of AI-derived quantitative CT observations as a necessary step toward clinical integration and linked measurements to anatomy and source imaging using standardized terminologies [21]. Its importance lies partly in the work it makes visible: every observation type needs clear definitions, provenance, and relations before it can be compared or reused safely.

Interface integration answers the question, 'Can the result move?' Information integration answers a more difficult question: 'Does the result retain its meaning after it moves?'

13. Many systems, many representations — and an underdeveloped foundation

Modern radiology is accumulating classifiers, segmentations, biomarker pipelines, radiomic signatures, risk scores, structured reports, foundation models, and agents. A single lesion may simultaneously exist as voxels, a segmentation mask, a diameter, a volume, an attenuation value, a feature vector, a probability, a risk category, a sentence, and a recommendation.

These are not synonyms. They are different informational objects, created by different methods and carrying different uncertainty. The clinical report remains indispensable because it integrates findings into a patient-specific judgement. But a report is also a deliberate compression. It is not designed to preserve every measurable physical, geometric, technical, and temporal property of the source examination.

It would be inaccurate to claim that medical imaging has no semantics or no structured data. DICOM SR and SEG, common data elements, quantitative imaging biomarkers, FHIR observations, and structured reporting already provide important components. Major manufacturers generate structured and quantitative outputs. Research groups are building increasingly sophisticated exchange frameworks.

What remains underdeveloped is a shared cross-system informational foundation. Such a foundation would preserve an observation together with its anatomical, physical, geometric, technical, temporal, methodological, and provenance context, without binding it permanently to one radiologist, one algorithm, one classification, or one vendor environment.

It would not abolish disagreement, nor should it. A more realistic purpose is to make disagreement interpretable: to show whether two values are genuinely comparable, why they differ, and which parts of their derivation can be reproduced.

14. The questions that remain

In only a few years, radiology AI has moved from narrow detection systems toward foundation models, multimodal reasoning, opportunistic quantification, automated reporting, and agentic workflows. At the same time, major manufacturers have begun embedding measurements, segmentation, spectral information, orchestration, and structured outputs into their imaging ecosystems.

The basis of clinical trust has moved more slowly. We still tend to judge a model by its final metric, although the output depends on human representations, scanner and protocol characteristics, reference standards, local populations, software versions, interface design, and post-deployment monitoring.

This leaves a set of questions that cannot be answered by a larger model alone:

> *What should be preserved between the source image and the final clinical interpretation?*
>
> *Can an observation remain comparable when it moves across scanners, institutions, algorithms, software versions, and time?*
>
> *How should we distinguish a reproducible physical or geometric property of an image from a category that depends on a particular observer, guideline, or classifier?*
>
> *Can a future AI system understand the provenance of a result rather than merely receive a number or sentence?*
>
> *And what if part of the limitation of modern medical AI lies not in the model itself, but in the form in which medical imaging information reaches it?*

This article deliberately stops at those questions. The second article in the series, We Invest Too Much in the Model — and Too Little in the Input, will examine a possible direction of answer at a public and non-proprietary level: whether a complementary layer of reproducible, physically grounded, structured, and traceable image-derived observations could improve how radiologists, clinical systems, and future AI models work together.

Conclusion

Artificial intelligence in radiology already delivers real clinical value. It can detect defined abnormalities, accelerate quantitative work, support triage, and recover information that once remained unused. That achievement deserves recognition. Clinical usefulness, however, is not the same as universal reliability.

Every model output is an account of medical reality, not medical reality itself. That account is shaped by people, devices, protocols, classifications, institutions, and law. Different radiologists and different AI systems may therefore produce different, partly legitimate descriptions of the same object.

The field is beginning to respond. Evaluation is moving beyond isolated accuracy metrics toward external testing, traceability, usability, governance, and lifecycle monitoring. Manufacturers are moving from isolated algorithms toward integrated platforms and structured quantitative output. Interoperability standards are connecting AI more closely to the clinical environment.

The next step may demand more than stronger models and tighter application integration. It may require a clearer account of what information is preserved when an image becomes a computational observation — and whether that observation can retain its meaning beyond the system that created it.

About the author

Vladimir A. Novikov, MD, is a practicing radiologist and Head of a Radiology Department with more than 15 years of experience in computed tomography. His work focuses on clinical imaging, quantitative methods, and the clinical validation of medical AI. He is a Co-founder of ResetRay.

Disclosure

The author is a practicing radiologist, Head of a Radiology Department, and Co-founder of ResetRay, a medical imaging technology initiative concerned with structured, reproducible, and traceable quantitative information derived from existing imaging data.

This article is a clinical and scientific perspective, not a systematic review, clinical guideline, comparative product assessment, or claim of superiority for any technology. Commercial product descriptions are based on publicly available manufacturer materials and are used as evidence of strategic direction, not as independent evidence of clinical effectiveness.

Source and evidence note

Peer-reviewed publications provide the principal evidence for foundation models, clinical implementation, observer variability, external validation, label noise, opportunistic screening, human factors, interoperability, and post-deployment monitoring. Professional and regulatory documents support the discussion of governance and lifecycle principles. Manufacturer sources are used only to describe publicly documented product capabilities and strategic direction.

References

1. Paschali M, Chen Z, Blankemeier L, et al. Foundation Models in Radiology: What, How, Why, and Why Not. Radiology. 2025;314(2):e240597. doi:10.1148/radiol.240597
2. Ma D, Pang J, Gotway MB, et al. A Fully Open AI Foundation Model Applied to Chest Radiography. Nature. 2025;643:488-498. doi:10.1038/s41586-025-09079-8
3. Kottlors J, Iuga AI, Bluethgen C, et al. Guidelines for Reporting Studies on Large Language Models in Radiology: An International Delphi Expert Survey. Radiology. 2026;318(2):e250913. doi:10.1148/radiol.250913
4. Khosravi B, Rouzrokh P, Akinci D'Antonoli T, et al. Agentic AI in Radiology: Evolution from Large Language Models to Future Clinical Integration. Radiology: Artificial Intelligence. 2026;8(2):e250651. doi:10.1148/ryai.250651
5. Goldberg-Stein S, Gandomi A, Barish MA, et al. Clinical Implementation of AI for Pulmonary Embolism Detection in over 30,000 CT Pulmonary Angiography Examinations. Radiology: Artificial Intelligence. 2026;8(4):e250017. doi:10.1148/ryai.250017
6. Pettersson SD, Filo J, Skrzypkowska P, et al. End-to-end Autonomous Quantification of Brain Aneurysm and Parent-artery Morphology on CT Angiography. Radiology: Artificial Intelligence. Published online May 20, 2026. doi:10.1148/ryai.251093
7. Moshina N, Larsen M, Holen AS, et al. Artificial Intelligence for Digital Breast Tomosynthesis Screening with and without Prior Examinations in BreastScreen Norway. Radiology: Artificial Intelligence. 2026;8(4):e250988. doi:10.1148/ryai.250988
8. Pickhardt PJ, Lee MH, Warner JD, Summers RM, Garrett JW. CT-based Opportunistic Screening for Adding Clinical Value: How I Do It. Radiology. 2026;319(1):e252106. doi:10.1148/radiol.252106
9. Flanders AE, Wang X, Wu CC, et al. The Evolution of Radiology Image Annotation in the Era of Large Language Models. Radiology: Artificial Intelligence. 2025;7(4):e240631. doi:10.1148/ryai.240631
10. Zhu W, Xu F, Lou K, et al. The Impact of Inter-observation Variation on Radiomic Features of Pulmonary Nodules. Frontiers in Oncology. 2025;15:1567028. doi:10.3389/fonc.2025.1567028
11. Ripaud E, Jailin C, Milioni de Carvalho P, Vancamberg L, Bloch I. ReaderAdaptNet: Modeling Reader Variability in Breast Imaging with Reader-specific Embeddings. Physics in Medicine & Biology. 2026;71(9):095032. doi:10.1088/1361-6560/ae6227
12. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiology: Artificial Intelligence. 2022;4(3):e210064. doi:10.1148/ryai.210064
13. Asmara OD, Steenhuis EGM, de Jong K, et al. Externally Tested AI Models for Malignancy Classification of Lung Nodules at Chest CT: A Systematic Review and Meta-analysis. Radiology: Artificial Intelligence. Published online June 3, 2026. doi:10.1148/ryai.250331
14. Chavoshi M, Trivedi H, Mansuri A, et al. Impact of Label Noise from Large Language Model-generated Annotations on Evaluation of Diagnostic Model Performance. Radiology: Artificial Intelligence. 2026;8(2):e250477. doi:10.1148/ryai.250477
15. Lee JH, Shin J. LLM Label Noise and the Established Framework of Imperfect Reference Standard Bias. Radiology: Artificial Intelligence. 2026;8(3):e260153. doi:10.1148/ryai.260153
16. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare. BMJ. 2025;388:e081554. doi:10.1136/bmj-2024-081554
17. Coombs LP, Brink L, Kim W, et al. ACR's Assess-AI: A Registry for Real-World Performance Monitoring of Clinical Imaging Artificial Intelligence. Journal of the American College of Radiology. Published online April 29, 2026. doi:10.1016/j.jacr.2026.04.024
18. Tejani AS, Kohli M, Rauschecker AM, et al. AI for Radiology: A Primer Part II. Interacting with AI Results. Radiology. 2026;320(1):e252702. doi:10.1148/radiol.252702
19. Taib AG, Partridge GJW, Phillips P, et al. Automation Bias in Action: Eye Tracking of Humans Reading Screening Mammograms with and without AI Prompts. Radiology. 2026;320(1):e252590. doi:10.1148/radiol.252590

20. Jiao A, Hussain M, Sippel Schmidt T, et al. Radiology Reimagined: Interoperability and Lessons Learned from the Imaging AI in Practice Demonstration. Radiology. 2026;319(3):e252225. doi:10.1148/radiol.252225

21. Wen Y, Choo VY, Eil JH, et al. Exchange of Quantitative Computed Tomography Assessed Body Composition Data Using Fast Healthcare Interoperability Resources as a Necessary Step Toward Interoperable Integration of Opportunistic Screening Into Clinical Practice. Journal of Medical Internet Research. 2025;27:e68750. doi:10.2196/68750

22. Rorden C, Beranger B, Cheng H, et al. DICOM Datasets for Reproducible Neuroimaging Research across Manufacturers and Software Versions. Scientific Data. 2025;12:1168. doi:10.1038/s41597-025-05503-w

23. American College of Radiology and Society for Imaging Informatics in Medicine. ACR-SIIM Practice Parameter for Imaging Artificial Intelligence. Approved May 2026. Official ACR source

24. Siemens Healthineers. AI-Rad Companion and AI-Rad Companion Chest CT: official product and workflow documentation. Accessed July 2026. Official Siemens source

25. GE HealthCare. Thoracic VCAR: official product documentation. Accessed July 2026. Official GE HealthCare source

26. Philips. Detector-based Spectral CT; AI Manager; Vestre Viken AI-powered radiology workflow. Official product and implementation materials. Accessed July 2026. Spectral CT · Vestre Viken workflow

27. European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act); MDCG 2025-6 FAQ on the interplay between the MDR/IVDR and the AI Act. AI Act · MDCG 2025-6

28. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations; Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. 2025. Lifecycle guidance · PCCP guidance

29. International Medical Device Regulators Forum. Good Machine Learning Practice for Medical Device Development: Guiding Principles. IMDRF/AIML WG/N88 FINAL:2025. Official IMDRF source