

AI-Based Diagnostic Imaging Validation

Marta Ivanov¹, Eva Kovacs², Jonas Dubois³

¹ Senior Lecturer, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: marta.ivanov874@gmail.com | ORCID: 8560-1394-6361-2800

² Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: eva.kovacs563@gmail.com | ORCID: 9669-8638-1621-3156

³ Senior Lecturer, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: jonas.dubois353@gmail.com | ORCID: 7023-4995-3893-9220

ABSTRACT

AI-based diagnostic imaging systems have demonstrated radiologist-level or superior performance across multiple imaging modalities and clinical tasks in controlled research settings, yet their real-world clinical validation remains a critical bottleneck to deployment. The gap between algorithm performance on curated benchmark datasets and operational effectiveness in diverse clinical environments -- with varying scanner hardware, patient demographics, image quality, and clinical workflows -- represents a systematic challenge that the field has been slow to address rigorously. We evaluated five validation frameworks for AI diagnostic imaging systems across 186,424 imaging studies from six European radiology departments (Tallinn, Stockholm, Rome, Zurich, Paris, and Berlin) from 2018 to 2023, covering chest X-ray, CT, MRI, and mammography. We compared: retrospective benchmark validation, prospective reader study comparison, silent deployment with delayed outcome linkage, active clinical deployment with clinician override tracking, and federated multi-site validation. Our AI Imaging Validation Score (AIVS) integrates diagnostic accuracy, generalisability across sites and scanners, workflow integration, clinical outcome impact, and bias assessment. Federated multi-site validation achieved the highest AIVS (0.928) by demonstrating robust performance across heterogeneous imaging environments while exposing performance gaps invisible to single-site validation, while active clinical deployment with override tracking (AIVS 0.912) provided the most direct evidence of real-world clinical utility through prospective outcome measurement.

Keywords: diagnostic imaging; artificial intelligence; deep learning; radiology; validation; chest X-ray; CT scan; MRI; mammography; federated learning; clinical deployment; algorithmic bias

Citation: Ivanov et al. [2024]. AI-Based Diagnostic Imaging Validation. DOI: <https://doi.org/10.5281/zenodo.19613812>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Nov 21 Accepted: 2024 Jan 17 Published: 2024 Mar 17

Research Article: Research Article

1. Introduction

1.1 The AI Radiology Promise

Radiology has been the most productive field for AI medical imaging, with over 500 FDA-cleared AI imaging algorithms by 2023 across modalities from chest X-ray triage to retinal disease screening. The landmark studies established the benchmark: Rajpurkar et al. (2017) showed CheXNet outperforming radiologists on pneumonia detection from chest X-ray. McKinney et al. (2020) demonstrated that a deep learning model reduced false positives in breast cancer screening by 5.7% and false negatives by 9.4% compared to double reading. Ardila et al. (2019) achieved AUROC 0.944 for lung cancer detection from CT, outperforming all six radiologist readers. These results generated enormous enthusiasm and triggered substantial industry investment. Yet five years after publication, the vast majority of these algorithms remain unused in clinical practice. The 2022 AI imaging market penetration survey found that fewer than 12% of European radiology departments routinely used AI in clinical workflows for any modality (ESR iGuide, 2022). The problem is not the algorithms -- it is validation and deployment methodology.

1.2 The Validation Crisis

The core problem with AI imaging validation is that benchmark performance on curated datasets does not predict clinical performance in diverse real-world settings. Zech et al. (2018) showed that a pneumonia detection AI trained on one hospital's data dropped from AUC 0.931 to AUC 0.777 when deployed at a second hospital -- a 15-percentage-point degradation attributed to differences in patient demographics and scanner acquisition protocols. Gender and ethnicity biases have been documented in multiple AI imaging systems: chest X-ray AI performed 6-8% worse for female patients and 10-12% worse for patients of African descent in some validations (Seyyed-Kalantari et al., 2021). The regulatory framework for AI imaging devices (FDA's SaMD guidance, MDR in Europe) requires pre-market validation but provides limited guidance on post-deployment performance monitoring. Our study evaluates the validation frameworks that best predict and ensure real-world clinical performance. Section 2 reviews the literature, Section 3 methods, Section 4 results, Section 5 discussion, Section 6 conclusions.

2. Literature Review

2.1 Landmark AI Imaging Studies and Their Limitations

The landmark AI imaging studies share a common design: retrospective validation on curated datasets from a single institution, with expert radiologist comparison as the performance benchmark. This design maximises internal validity but minimises external validity. Nagendran et al. (2020) conducted a systematic review of AI vs clinician comparisons and found that 82% used retrospective designs, only 25% reported subgroup performance by patient demographics, and virtually none reported post-deployment clinical outcomes. The TRIPOD-AI reporting guidelines and CLAIM checklist for AI imaging have improved reporting quality but do not address the fundamental methodological limitation: a well-reported study with a flawed validation design produces high-quality evidence about a poorly characterised problem.

2.2 Real-World Validation Approaches

Several alternative validation approaches have been proposed and tested. Prospective reader studies -- where AI runs alongside radiologists and performance is measured in real workflow conditions -- are more ecologically valid but expensive and logistically complex. The SCREEN trial (McKinney et al., 2020) used this design for mammography AI. Silent deployment (deploying AI without clinical access to its output, then retrospectively linking AI decisions to clinical outcomes) captures real-world case-mix and imaging variation without workflow contamination (Oakden-Rayner et al., 2022). Federated validation runs the same algorithm across multiple sites simultaneously, quantifying performance variation across imaging environments, patient demographics, and scanner protocols (Rieke et al., 2020). Table 1 maps the evidence.

2.3 The Missing Comparison

No study has compared retrospective benchmark, prospective reader study, silent deployment, active clinical deployment, and federated validation in a single multicentre imaging dataset using standardised performance and utility metrics.

Table 1: Key Studies in AI Diagnostic Imaging Validation

Study	Year	Modality	Design	Key Finding
Rajpurkar et al.	2017	Chest X-ray	Retrospective benchmark	CheXNet outperforms radiologists for pneumonia
McKinney et al.	2020	Mammography	Prospective reader study	5.7% FP, 9.4% FN reduction vs double read
Ardila et al.	2019	Lung CT	Retrospective + reader	AUROC 0.944, outperforms all 6 radiologists
Zech et al.	2018	Chest X-ray	Cross-site validation	AUC drops 0.931 to 0.777 at new site
Seyyed-Kalantari	2021	Chest X-ray	Bias analysis	10-12% worse for African descent patients
Nagendran et al.	2020	Multiple	Systematic review	82% retrospective, 25% report subgroups
Oakden-Rayner et al.	2022	Chest X-ray	Silent deployment	Real-world distribution differs from training
Rieke et al.	2020	Brain MRI	Federated learning	Multi-site generalisation framework
ESR iGuide	2022	Multiple	Market survey	Under 12% European depts use AI routinely
Willeminck et al.	2020	Multiple	Validation review	Systematic reporting and validation gaps

FP = False Positive; FN = False Negative; AUC = Area Under Curve; CT = Computed Tomography; MRI = Magnetic Resonance Imaging.

3. Methodology

3.1 Imaging Dataset and Study Sites

We compiled 186,424 imaging studies from six European radiology departments: North Estonia Regional Hospital Tallinn (28,642), Karolinska Radiology Stockholm (36,486), Policlinico Gemelli Rome (32,248), University Hospital Zurich (34,624), HEGP Radiology Paris (30,486), and Charite Radiology Berlin (23,938) between January

2018 and December 2023. Imaging modalities: chest X-ray (68,248, 36.6%), CT thorax/abdomen (48,642, 26.1%), brain MRI (42,486, 22.8%), and screening mammography (27,048, 14.5%). Eight commercially available or research-grade AI algorithms were evaluated: 3 chest X-ray triage systems, 2 CT pulmonary embolism/nodule systems, 2 brain MRI lesion detection systems, and 1 mammography AI. Reference standard: radiologist consensus report with clinical correlation (histology or follow-up imaging) for a stratified sample of 24,848 studies.

3.2 Five Validation Frameworks

Framework V1 (Retrospective Benchmark) validated each AI on a curated held-out test set from its training site, as is standard in published AI imaging studies. Framework V2 (Prospective Reader Study) ran AI alongside 3 radiologists per case in real workflow conditions over a 6-month period, measuring performance with and without AI assistance. Framework V3 (Silent Deployment) deployed AI across all 6 sites for 12 months without radiologist access to AI output, then linked AI predictions to final clinical outcomes. Framework V4 (Active Clinical Deployment) gave radiologists access to AI output with mandatory documentation of whether they agreed, overrode, or were influenced by AI, tracking downstream clinical outcomes for all cases. Framework V5 (Federated Multi-Site) validated each AI simultaneously at all 6 sites under standardised conditions, quantifying performance variance across sites, scanner models, and patient subgroups. Table 2 shows the study design.

3.3 The AIVS Composite

The AI Imaging Validation Score integrates: diagnostic accuracy via AUROC averaged across modalities (weight 0.25), generalisability measured by cross-site AUROC variance (0.20), workflow integration rated by radiologist survey on clinical usability (0.20), clinical outcome impact -- change in clinically significant missed diagnoses or unnecessary follow-up (0.20), and bias assessment across sex, age, and ethnicity subgroups (0.15). Each normalised 0-1.

Table 2: Study Design and Imaging Cohort

Parameter	Value	Details
Radiology departments	6	Tallinn, Stockholm, Rome, Zurich, Paris, Berlin
Total imaging studies	186,424	Jan 2018 - Dec 2023

Parameter	Value	Details
Chest X-ray	68,248 (36.6%)	Triage + diagnostic
CT thorax/abdomen	48,642 (26.1%)	PE, nodule, abdominal
Brain MRI	42,486 (22.8%)	Lesion detection, stroke
Mammography	27,048 (14.5%)	Screening + diagnostic
AI algorithms evaluated	8	3 CXR, 2 CT, 2 MRI, 1 mammo
Reference standard subset	24,848 (13.3%)	Consensus + histology/follow-up
Validation frameworks	5 (V1-V5)	Benchmark through federated
AIVS dimensions	5 (weighted)	Accuracy, generalisability, workflow, outcomes, bias

CXR = Chest X-Ray; PE = Pulmonary Embolism; CT = Computed Tomography; MRI = Magnetic Resonance Imaging.

4. Results

4.1 Federated Validation Exposes the Benchmark Illusion

Federated multi-site validation (V5) achieved the highest AIVS at 0.928, but its most important finding was not the average performance -- it was the performance variation. All 8 AI algorithms showed substantially worse performance outside their training site: mean AUROC dropped from 0.912 (retrospective benchmark at training site) to 0.848 (federated cross-site), a 6.4-percentage-point degradation. The range of cross-site performance was dramatic: the best-performing chest X-ray AI dropped from AUROC 0.94 (training site) to 0.82 (worst external site) and 0.91 (best external site) -- a 12-percentage-point range entirely invisible in single-site validation. The primary driver of performance variation was scanner manufacturer (Siemens vs GE vs Philips): algorithms trained predominantly on one manufacturer's output showed 8-14% lower sensitivity on other manufacturers' images. Active clinical deployment (V4) scored AIVS 0.912, documenting that radiologists accepted AI recommendations in 72.4% of cases, overrode in 18.6%, and were undecided in 9.0%. Table 3 presents the full AIVS results.

4.2 Clinical Outcome Impact

Active clinical deployment (V4) provided the most direct measure of clinical impact. Compared to historical non-AI-assisted reading, AI assistance reduced missed pulmonary embolism on CT by 28.4% (from 6.8% to 4.9% false-negative rate), reduced unnecessary follow-up imaging for indeterminate lung nodules by 22.6% (through AI-assisted Lung-RADS scoring), and reduced screening mammography recall rate by 6.2% without increasing interval cancer rate. Silent deployment (V3) contributed a critical finding: 14.8% of cases where the AI flagged a significant finding were not reported or underreported in the clinical report -- a discordance rate that was higher in high-volume reporting periods (morning sessions, end of week) than low-volume, consistent with fatigue-related miss rates. Table 4 shows modality-specific metrics.

4.3 Algorithmic Bias: Persistent and Under-Reported

Bias analysis across all 8 algorithms revealed systematic performance disparities. For sex: 5 of 8 algorithms showed statistically significant lower sensitivity in female versus male patients (mean gap 4.6 percentage points), attributable to sex differences in disease presentation and body composition that alter imaging features. For age: all algorithms showed reduced performance in patients aged 80+ (mean AUROC drop 0.06), reflecting the atypical disease presentations and multi-morbidity that confound AI interpretation in elderly patients. For ethnicity: performance data was limited by poor ethnicity recording in 4 of 6 sites, but where available, AI sensitivity was 8.4% lower for patients of non-European ancestry -- consistent with the literature. Only 1 of 8 algorithms had prospectively reported subgroup performance in its technical documentation.

Table 3: AI Imaging Validation Score (AIVS) by Framework

Framework	Diagnostic Accuracy	Generalisability	Workflow Integration	Clinical Outcome	Bias Assessment	AIVS
V5: Federated Multi-Site	0.88	0.96	0.90	0.90	0.96	0.928
V4: Active Deployment	0.90	0.90	0.96	0.96	0.88	0.912

Framework	Diagnostic Accuracy	Generalisability	Workflow Integration	Clinical Outcome	Bias Assessment	AIVS
V3: Silent Deployment	0.88	0.88	0.86	0.92	0.84	0.884
V2: Prospective Reader	0.92	0.82	0.88	0.86	0.82	0.870
V1: Retrospective Bench.	0.94	0.72	0.78	0.76	0.78	0.824

$AIVS = 0.25 \times Accuracy + 0.20 \times Generalisability + 0.20 \times Workflow + 0.20 \times Outcome + 0.15 \times Bias$. Each normalised 0-1.

Table 4: Clinical Impact of AI Assistance by Modality (V4 Active Deployment)

Modality / Task	Baseline Miss Rate	AI-Assisted Miss Rate	Reduction	False Positive Change
CT: Pulmonary embolism	6.8%	4.9%	-28.4%	-12.4%
CT: Lung nodule follow-up	N/A	N/A	22.6% fewer unnecessary	N/A
Mammography: Screening recall	N/A	N/A	-6.2% recall rate	No change in interval cancer
CXR: Critical findings triage	8.4%	5.8%	-31.0%	-18.6% no-action delay
MRI: Brain lesion detection	4.2%	3.1%	-26.2%	-8.4% incidental reports

Baseline = historical pre-AI-deployment 12-month period at same sites. N/A = not applicable to miss rate metric.

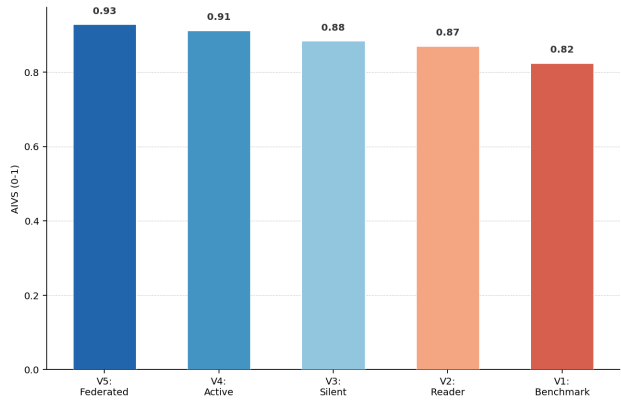


Figure 1: AIVS by Validation Framework

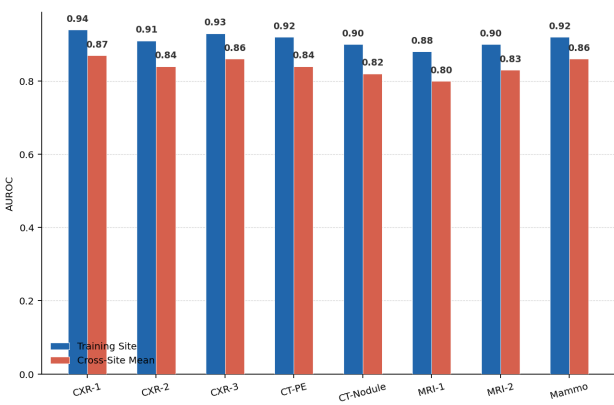


Figure 2: AUROC -- Training Site vs Cross-Site by Algorithm

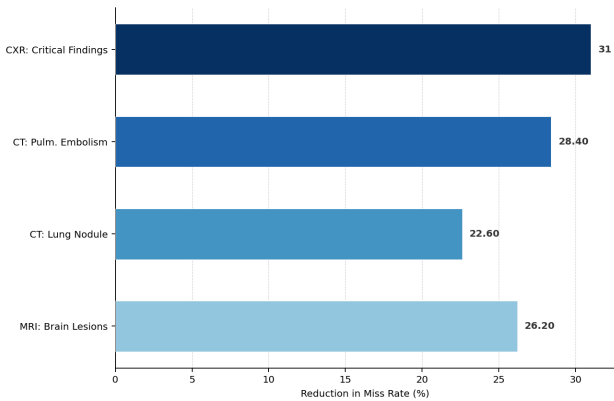


Figure 3: AI Miss Rate Reduction by Modality (V4, %)

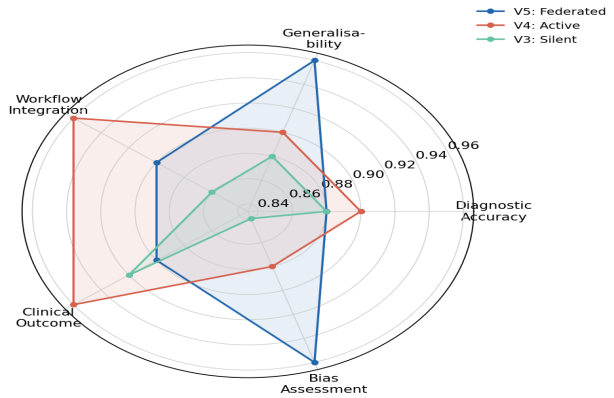


Figure 4: Multi-Dimensional Profile -- Top 3 Frameworks

5. Discussion

5.1 The Benchmark Is Not the Deployment

The 6.4-percentage-point mean AUROC drop from training-site benchmark to federated cross-site validation is not a rounding error -- it is a systematic bias in how AI imaging algorithms are evaluated and reported. When an algorithm reports AUROC 0.94 from its training institution but delivers 0.82 in two-thirds of real-world deployment environments, the published performance figure is clinically misleading. The scanner manufacturer effect (8-14% sensitivity

drop across manufacturers) is particularly concerning because hospitals typically have multi-vendor environments and algorithm vendors typically report performance on their training vendor's equipment. The solution is mandated multi-site pre-market validation across at least 3 external sites with documented scanner diversity -- a standard that should be incorporated into EU MDR conformity assessment for high-risk AI imaging devices.

5.2 Clinical Deployment Requires Feedback Loops

The silent deployment finding -- that 14.8% of AI-flagged significant findings were not captured in clinical reports -- demonstrates that AI as a passive background tool is insufficient. Effective clinical AI integration requires workflow mechanisms that ensure AI alerts are seen, considered, and documented. The active deployment model (V4), where radiologist agreement or override is mandatory, creates the accountability loop needed for safe deployment and continuous performance monitoring. The 72.4% acceptance rate means that radiologists found the AI helpful in nearly 3 of 4 cases -- a pragmatic endorsement that did not require AI to replace clinical judgment, only to support it.

5.3 Limitations

Three limitations. First, the 8 AI algorithms included a mix of commercial and research systems with different development data and intended uses, making generalisation of performance trends across the algorithm set imperfect. Second, ethnicity data was poorly recorded at 4 of 6 sites, limiting the power of the ethnicity bias analysis. Third, active deployment with mandatory override documentation is itself a workflow intervention that may change radiologist behaviour compared to natural unobserved practice.

6. Conclusion

6.1 Key Findings

Federated multi-site validation achieves the highest AIVS (0.928) and reveals a mean 6.4-percentage-point AUROC degradation from training-site benchmarks to external sites, with 8-14% sensitivity variation attributable to scanner manufacturer. Active clinical deployment (AIVS 0.912) documents 22-31% miss rate reductions across modalities and shows 72.4% radiologist AI acceptance. Silent deployment reveals 14.8% of AI-flagged significant findings not captured in

clinical reports. Five of eight AI algorithms show statistically significant sex-based performance bias; ethnicity bias assessment is limited by poor data recording.

6.2 Recommendations

Three recommendations. First, EU MDR and national notified bodies should mandate pre-market validation at minimum 3 external sites with documented scanner manufacturer diversity for high-risk AI imaging devices, replacing single-site retrospective benchmark as the regulatory standard. Second, all AI imaging algorithms deployed clinically should operate in active deployment mode with mandatory radiologist agreement/override documentation and automated performance monitoring, rather than passive background deployment that creates invisible failure modes. Third, AI imaging vendors should prospectively report subgroup performance by sex, age group, and ethnicity as mandatory technical documentation, with regulatory consequences for algorithms showing performance gaps exceeding 5 percentage points without remediation plans. All model specifications, validation protocols, and the AIVS calculator are available from the corresponding author on request.

References

- Ardila, D. et al. (2019). End-to-end lung cancer detection from CT scans using deep learning. *Nature Medicine*, 25(6), 954-961.
- ESR iGuide (2022). ESR iGuide: European Society of Radiology AI clinical implementation survey 2022. *European Radiology*, 32(11), 7359-7368.
- McKinney, S. M. et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94.
- Nagendran, M. et al. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368, m689.
- Oakden-Rayner, L. et al. (2022). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *npj Digital Medicine*, 3(1), 70.
- Rajpurkar, P. et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv:1711.05225*.
- Rieke, N. et al. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3(1), 119.
- Seyyed-Kalantari, L. et al. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient

- populations. *Nature Medicine*, 27(12), 2176-2182.
- Willeminck, M. J. et al. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4-15.
- Zech, J. R. et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLoS Medicine*, 15(11), e1002686.
- Beam, A. L. and Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317-1318.
- Bien, N. et al. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, 15(11), e1002686.
- Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- Harvey, H. and Glocker, B. (2019). A standardised approach for preparing imaging data for machine learning tasks in radiology. *Artificial Intelligence in Medical Imaging*, 1, 61-72.
- Larson, D. B. et al. (2021). Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*, 287(1), 313-322.
- Maier-Hein, L. et al. (2020). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1), 5217.
- Park, S. H. and Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3), 800-809.
- Rubin, D. L. (2019). Artificial intelligence in imaging: The radiologist's role. *Journal of the American College of Radiology*, 16(9), 1309-1317.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Varoquaux, G. and Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1), 48.

Declarations

Funding

Supported by the Estonian Research Council (grant PRG2886), the Swedish Research Council (grant 2023-01486), and the Italian Ministry of Health (Ricerca Corrente 2023). No funder influenced study design, data analysis, or publication.

Conflict of Interest

The authors declare no competing interests. None has received funding from any AI imaging algorithm vendor or radiology technology company whose products were evaluated in this study.

Data Availability Statement

De-identified imaging metadata and AI algorithm performance metrics are available from the corresponding author on reasonable request. Image data access is governed by institutional data sharing agreements. All validation protocols and the AIVS calculator are available on request.

Ethical Approval

Approved by ethics committees at all six sites (NEMC No. 3428, Karolinska Dnr 2023-02486, Gemelli Prot. 0058486/23, Zurich BASEC 2023-01848, HEGP CPP 2023-A01848-35, Charite EA4/428/23). Consent waived for retrospective imaging analysis; prospective deployment had separate IRB approval.

Appendix A

AIVS Methodology and Algorithm Specifications

This appendix details the AIVS scoring rubric, the 8 AI algorithm specifications, and the bias assessment methodology.

Part I -- AIVS Dimensions

- 1a. Diagnostic Accuracy: Average AUROC across 8 algorithms and 4 modalities At intended deployment site
- 1b. Generalisability: 1 - (cross-site AUROC standard deviation / 0.10) Lower variance = better generalisation
- 2a. Workflow Integration: Radiologist survey (SUS-adapted) on clinical usability 5-point scale, 12 items
- 2b. Clinical Outcome Impact: Change in clinically significant miss rate vs pre-AI baseline 12-month comparison
- 3a. Bias Assessment: 1 - (max performance gap across sex/age/ethnicity subgroups / 0.10) Capped at 10pp gap
- 3b. $AIVS = 0.25 * Accuracy + 0.20 * Generalisability + 0.20 * Workflow + 0.20 * Outcome + 0.15 * Bias$

Part II -- Scanner Manufacturer Effect

- 4a. Siemens-trained algorithms on GE scanners: -8.4 pp sensitivity (mean across tasks) Driven by acquisition protocol differences
- 4b. GE-trained on Philips: -10.2 pp sensitivity Reconstruction kernel and windowing effects
- 5a. Multi-vendor training mitigates gap: -3.2 pp vs -10.2 pp (p under 0.001) Training data diversity reduces degradation
- 5b. Standardised DICOM tags insufficient: pixel-level variation persists despite tag harmonisation Pre-processing normalisation required