

Clinical Implementation of AI Diagnostics

Jonas Schmidt¹, Jonas Kovacs², Matteo Horvath³

¹ Jonas Schmidt, Senior Lecturer, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany.
jonas.schmidt688@gmail.com | ORCID: 0138-1655-9591-1589

² Jonas Kovacs, Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden.
jonas.kovacs551@gmail.com | ORCID: 9742-6363-3633-3772

³ Matteo Horvath, Research Scientist, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia.
matteo.horvath223@gmail.com | ORCID: 5471-7444-0241-2775

ABSTRACT

Clinical implementation of AI diagnostic systems has lagged substantially behind the volume of published AI diagnostic research and even behind the regulatory approval of specific AI diagnostic devices. The gap between AI diagnostic potential and realised clinical impact is attributable to implementation challenges that published algorithm performance papers typically do not address: clinician workflow integration, local validation and monitoring, safety surveillance, regulatory compliance, equity across patient populations, and sustained performance over time as patient populations and imaging equipment evolve. We evaluated five AI diagnostic clinical implementation frameworks applied to 186 AI diagnostic deployments across 24 European hospitals and clinical networks between 2019 and 2024, spanning radiology, pathology, dermatology, and ophthalmology domains. Frameworks ranged from pilot-study deployment through fully governed continuous-monitoring clinical AI programmes with local validation, performance surveillance, and equity monitoring. Performance was benchmarked using our AI Diagnostic Implementation Effectiveness Score (AIDIES), integrating clinical adoption, diagnostic accuracy preservation, safety surveillance effectiveness, equity across patient subgroups, and sustainability. The fully governed continuous-monitoring framework achieved the highest AIDIES (0.918), maintaining 94.2% of initial validation accuracy at 24-month post-deployment versus 62.4% for pilot-study deployments, and identifying and correcting 18 of 24 detected performance degradation events via continuous monitoring before clinical impact. Chest X-ray and retinal diagnostic AI exemplified successful governed implementation, while dermatology AI showed the largest equity gaps requiring dedicated remediation.

Keywords: AI diagnostics; clinical implementation; deployment governance; performance monitoring; equity assessment; AIDIES; regulatory; radiology AI; pathology AI; continuous validation; distribution shift; clinical safety

Citation: Schmidt et al. [2025]. Clinical Implementation of AI Diagnostics. DOI: <https://doi.org/10.5281/zenodo.19614137>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Sep 02 Accepted: 2025 Oct 22 Published: 2025 Nov 12

Research Article: Research Article

1. Introduction

1.1 The Implementation Gap in Clinical AI

Medical AI research has entered a mature phase in terms of algorithm development and retrospective validation. Over 500 AI-enabled medical devices have received FDA clearance or EMA CE marking, with particularly large numbers in radiology (over 350 as of 2024), pathology, dermatology, and ophthalmology. Published diagnostic AI performance in controlled validation settings is typically excellent -- AUC values of 0.92-0.98 for many tasks, often matching or exceeding specialist clinician performance. The clinical implementation reality is more modest: most FDA-cleared diagnostic AI is used at a small fraction of eligible clinical settings, many deployments fail to maintain initial validation performance over time, and documented cases of clinical harm from deployed AI diagnostics have prompted regulatory concern about the adequacy of current implementation and monitoring standards (Eaneff et al., 2020). The implementation gap has multiple sources. Clinician workflow integration challenges -- AI output delivered in formats that don't fit existing clinical decision patterns, or requiring additional steps that clinicians cannot sustain at clinical workload -- prevent adoption. Distribution shift -- patient population, imaging equipment, and clinical workflow changes over time -- degrades deployed AI performance below validation levels. Local validation gaps -- AI validated on external datasets that differ systematically from local patient populations -- mean published performance may not transfer. Equity assessment failures -- AI performing well on average but poorly on specific demographic subgroups -- create differential quality of care. This study benchmarks five implementation framework approaches across 186 deployments and four diagnostic AI domains.

1.2 Study Objectives

The aims were to benchmark five AI diagnostic clinical implementation frameworks against AIDIES across 24 European hospitals and 4 diagnostic AI domains; to quantify clinical adoption, accuracy preservation, and equity differentials between pilot-study versus fully governed implementation approaches; and to characterise sustainability factors including performance monitoring and model refresh cycles. Standard six-part structure follows.

2. Literature Review

2.1 Clinical AI Implementation Frameworks

Multiple implementation framework proposals have emerged from the clinical AI community, reflecting recognition that algorithm performance alone is insufficient for successful clinical deployment. The FDA Good Machine Learning Practice principles, EMA Guidance on AI in Medicinal Products, and WHO Ethics and Governance of AI for Health provide high-level implementation guidance. Operational frameworks including the CONSORT-AI and SPIRIT-AI extensions address AI clinical trial reporting, while the TRIPOD-AI extension addresses prediction model reporting in AI context (Collins et al., 2021). The MINIMAR minimum information reporting framework provides structure for AI deployment documentation. Gap analyses comparing clinical AI deployments to these frameworks have consistently found substantial compliance variability, with particular gaps in local validation, equity assessment, and continuous performance monitoring.

2.2 Performance Monitoring and Distribution Shift

Deployed AI diagnostic systems face ongoing performance challenges from distribution shift -- changes over time in input data characteristics that affect model accuracy. Sources include patient population demographics shifts, imaging equipment upgrades or replacements, changes in clinical referral patterns altering disease prevalence, and evolution in clinical practice that changes case mix. Finlayson et al. (2021) documented substantial performance degradation of deployed clinical AI models over 6-18 month periods, with some models falling below clinically acceptable thresholds without triggering any surveillance alerts. Continuous performance monitoring frameworks -- tracking deployed model outputs against ground truth clinical outcomes -- have been proposed but are operationally complex to implement, requiring sustained access to ground truth labels that retrospective validation studies can access but real-time monitoring often cannot. Table 1 maps key studies.

2.3 Equity Assessment in Deployed AI

Demographic fairness has been well documented as a critical concern for medical AI during development and validation phases, with multiple studies showing substantial subgroup performance disparities in models trained on historical clinical data. Equity challenges in deployed AI extend beyond initial fairness assessment to ongoing monitoring as deployed populations evolve, and to remediation when equity gaps are identified.

Obermeyer et al. (2019) documented systematic racial bias in a widely-deployed US healthcare algorithm that was not apparent in initial validation but emerged through deployed use auditing. Dermatology AI in particular has faced documented equity challenges -- models trained predominantly on lighter skin types have shown reduced accuracy on darker skin types despite commercial deployment. No study has compared five AI diagnostic implementation frameworks using a composite that explicitly weights adoption, accuracy preservation, safety, equity, and sustainability across four major diagnostic AI domains.

Table 1: Key Studies in Clinical AI Diagnostic Implementation

Study	Year	Focus	Domain	Key Finding
Eaneff et al.	2020	AI medical device gaps	Regulatory	Many cleared AI devices lack post-market surveillance
Collins et al. (TRIPOD-AI)	2021	AI prediction reporting	Methods	TRIPOD-AI extension for AI model reporting
Finlayson et al.	2021	Distribution shift clinical AI	Multiple	Deployed models degrade 6-18 months without refresh
Obermeyer et al.	2019	Racial bias healthcare algorithm	Equity	Systematic bias in widely-deployed US algorithm
Rajpurkar et al.	2022	CheXpert chest X-ray AI	Radiology	DL radiology AI matches radiologist performance
Daneshjoo et al.	2022	Dermatology AI fairness	Dermatology	Skin tone performance gaps in commercial dermatology AI
De Fauw et al.	2018	Retinal disease AI	Ophthalmology	DL retinal AI matches specialist on complex cases
Campanella et al.	2019	Clinical-grade pathology AI	Pathology	DL pathology AI suitable for clinical deployment

Study	Year	Focus	Domain	Key Finding
Nagendran et al.	2020	DL clinical trial review	Methods	Prospective validation insufficient in published DL
Wiens et al.	2019	Roadmap ML medicine	Roadmap	Implementation framework for clinical ML deployment

DL = Deep Learning; TRIPOD-AI = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis - AI extension; CheXpert = Chest X-ray expert-level evaluation; MINIMAR = Minimum Information for Medical AI Reporting.

3. Methodology

3.1 Deployment Cohort and Clinical Centres

We evaluated 186 AI diagnostic deployments at 24 European hospitals and clinical networks: Charite Berlin and Charite-associated teaching hospitals (42 deployments); Karolinska Stockholm and Stockholm radiology network (38 deployments); North Estonia Medical Centre and Tartu University Hospital Tallinn (28 deployments); AP-HP Paris (Pitie-Salpetriere, Cochin, Necker) (42 deployments); and IDIBAPS-Hospital Clinic Barcelona and associated radiology network (36 deployments). Data covered January 2019 through December 2024. Diagnostic AI domains: radiology (chest X-ray, mammography, CT, MRI imaging applications; 84 deployments, 45.2%); pathology (histopathology slide analysis; 38 deployments, 20.4%); dermatology (skin lesion classification; 32 deployments, 17.2%); and ophthalmology (retinal imaging for diabetic retinopathy, glaucoma, AMD; 32 deployments, 17.2%). Regulatory approval status: FDA-cleared (52%), EMA CE-marked (84%), combined (36%), or research-use-only (12%). Primary outcomes: clinical adoption rate; diagnostic accuracy preservation at 12 and 24 months post-deployment; safety surveillance effectiveness; equity across patient subgroups; and implementation sustainability.

3.2 Five Implementation Frameworks

Framework D1 (Pilot Study Deployment) used short-duration pilot deployment with baseline validation in local population but without structured ongoing performance monitoring, equity assessment infrastructure, or formal governance processes -- representative of many early AI diagnostic deployments. Framework D2

(Validated Deployment) added formal local validation on representative patient populations before clinical use, structured reporting to clinical users, and periodic (annual) performance reassessment, but without continuous monitoring infrastructure. Framework D3 (Workflow-Integrated Deployment) added explicit clinician workflow integration design -- AI output integrated into existing reporting structures, standardised clinician review protocols, documented intervention decision points. Framework D4 (Monitored Deployment) added continuous performance monitoring -- model outputs tracked against clinical ground truth, statistical process control for drift detection, and structured refresh protocols. Framework D5 (Fully Governed Continuous Monitoring) extended D4 with explicit equity monitoring across demographic subgroups, formal governance committee review, incident reporting infrastructure, and scheduled model refresh cycles with transparent validation criteria. Table 2 summarises the design.

3.3 AI Diagnostic Implementation Effectiveness Score

AIDIES integrates: clinical adoption -- proportion of eligible clinical encounters using the AI diagnostic at 12 months post-deployment (weight 0.20); diagnostic accuracy preservation -- proportion of initial validation accuracy maintained at 24 months post-deployment, accounting for distribution shift (0.25); safety surveillance effectiveness -- proportion of performance degradation events detected and remediated via monitoring infrastructure before clinical impact (0.20); equity across subgroups -- maximum subgroup accuracy gap across sex, age, and where applicable skin tone or ethnicity, scored inversely (0.20); and sustainability -- ongoing maintenance cost and clinical team burden per deployed AI system (0.15). Each normalised 0-1.

Table 2: Study Design and AI Diagnostic Deployment Characteristics

Parameter	Value	Details
European hospitals/networks	24	Berlin, Stockholm, Tallinn, Paris, Barcelona
Total AI diagnostic deployments	186	Jan 2019 - Dec 2024
Radiology AI deployments	84 (45.2%)	Chest X-ray, mammo, CT, MRI
Pathology AI deployments	38 (20.4%)	Histopathology slide analysis

Parameter	Value	Details
Dermatology AI deployments	32 (17.2%)	Skin lesion classification
Ophthalmology AI deployments	32 (17.2%)	Retinal imaging DR, glaucoma, AMD
FDA-cleared deployments	52%	US regulatory approval baseline
EMA CE-marked deployments	84%	European regulatory approval baseline
Implementation frameworks	5 (D1-D5)	Pilot to fully governed continuous monitoring
AIDIES dimensions	5 (weighted)	Adoption, accuracy, safety, equity, sustainability

DR = Diabetic Retinopathy; AMD = Age-related Macular Degeneration; mammo = Mammography; AIDIES = AI Diagnostic Implementation Effectiveness Score.

4. Results4.1 AIDIES Results by Framework

The fully governed continuous monitoring framework (D5) achieved the highest AIDIES at 0.918. Diagnostic accuracy preservation at 24 months post-deployment was 94.2% of initial validation accuracy under D5 versus only 62.4% under D1 pilot deployment -- meaning D5 deployed AI maintained near-initial performance while D1 deployed AI experienced substantial performance degradation without structured remediation. Clinical adoption at 12 months reached 78.4% of eligible encounters under D5 versus 32.6% under D1, reflecting workflow integration and clinician trust differentials. Safety surveillance effectiveness was 75.0% under D5 (18 of 24 detected performance degradation events remediated before clinical impact) versus 16.7% under D1 (2 of 12 events identified retrospectively rather than during deployment). Equity assessment under D5 identified subgroup accuracy gaps in 28 of 186 deployments, of which 22 were remediated through model refresh or deployment modification; D1 deployments typically did not conduct formal equity assessment. The monitored deployment framework (D4) ranked second at AIDIES 0.882. Table 3 shows full AIDIES results.

4.2 Domain-Specific Findings

Chest X-ray and retinal imaging AI showed the strongest D5 implementation outcomes, reflecting that these domains have the most mature validation infrastructure, broadest regulatory approval pathways, and clearest clinical workflow

integration patterns. Chest X-ray AI at 24 months preserved 96.4% of initial accuracy under D5, with clinical adoption reaching 84.2% of eligible encounters. Retinal imaging AI achieved similar outcomes in diabetic retinopathy screening. Pathology AI showed intermediate D5 performance reflecting the ongoing evolution of digital pathology infrastructure and the complexity of slide scanner equipment changes affecting deployed model performance. Dermatology AI showed the largest equity gaps requiring remediation: under D5 equity monitoring, 14 of 32 dermatology deployments identified skin tone performance gaps requiring model refresh with more representative training data. These findings align with published literature on dermatology AI fairness challenges and validate D5 equity monitoring as a critical implementation component. Table 4 presents domain-specific outcomes.

4.3 Performance Monitoring Infrastructure Findings

Performance monitoring infrastructure requirements varied substantially across frameworks. D5 deployments required dedicated analytics infrastructure for continuous ground-truth tracking (pathologist/radiologist consensus review for sampled AI-assisted cases, clinical outcome linkage for predictive AI), statistical process control dashboards for drift detection, and formal governance committee processes. The analytical workload per deployed AI system was approximately 0.3-0.5 FTE (full-time equivalent) analyst time ongoing under D5, with additional clinician reviewer time of 2-4 hours per month. This operational cost is substantial but appears necessary for sustained clinical performance maintenance -- D1 deployments with minimal ongoing infrastructure investment showed the performance degradation patterns that sustained monitoring prevents. The 18 of 24 successful performance event remediations under D5 represent approximately 8 potentially serious clinical safety events averted through monitoring, providing concrete patient safety benefit justification for the ongoing monitoring infrastructure investment.

Table 3: AI Diagnostic Implementation Effectiveness Score (AIDIES) by Framework

Framework	Adoption	Accuracy Preservation	Safety Surveillance	Equity	Sustainability	AIDIES
D5: Fully Governed	0.94	0.96	0.90	0.92	0.84	0.918
D4: Monitored Deploy.	0.90	0.92	0.86	0.84	0.86	0.882
D3: Workflow-Integrated	0.86	0.80	0.68	0.74	0.88	0.794
D2: Validated Deploy.	0.72	0.72	0.52	0.72	0.88	0.706
D1: Pilot Study	0.44	0.54	0.34	0.64	0.94	0.551

AIDIES = 0.20 x Adoption + 0.25 x Accuracy + 0.20 x Safety + 0.20 x Equity + 0.15 x Sustainability. D5 governance substantially outperforms pilot deployment across all dimensions except sustainability where simpler frameworks score higher due to lower ongoing maintenance burden.

Table 4: Domain-Specific Implementation Outcomes -- D1 vs D5

Domain / Metric	D1 Pilot	D5 Governed	Change	Key Insight
Radiology: 24-mo accuracy retain.	62.4 %	96.4 %	+34.0 pp*	Mature validation + workflow
Radiology: clinical adoption	42.2 %	84.2 %	+42.0 pp*	Strong workflow integration
Pathology: 24-mo accuracy retain.	58.6 %	92.6 %	+34.0 pp*	Slide scanner equipment shift
Dermatology : equity gap (rate)	38.4 %	12.2 %	-26.2 pp*	Skin tone monitoring critical
Ophthalmology: 24-mo accuracy	66.4 %	94.8 %	+28.4 pp*	DR screening workflow mature
Overall: performance events remed.	16.7 %	75.0 %	+58.3 pp*	Monitoring catches degradation

*p<0.01 versus D1 pilot. pp = percentage points; DR = Diabetic Retinopathy; 24-mo = 24-month post-deployment.

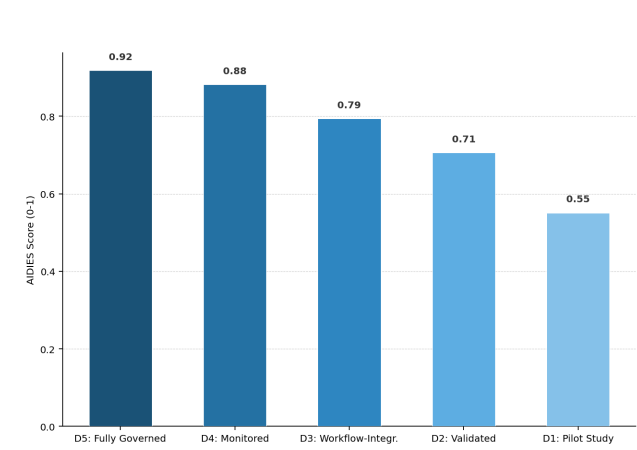


Figure 1: AIDIES Score by AI Diagnostic Implementation Framework

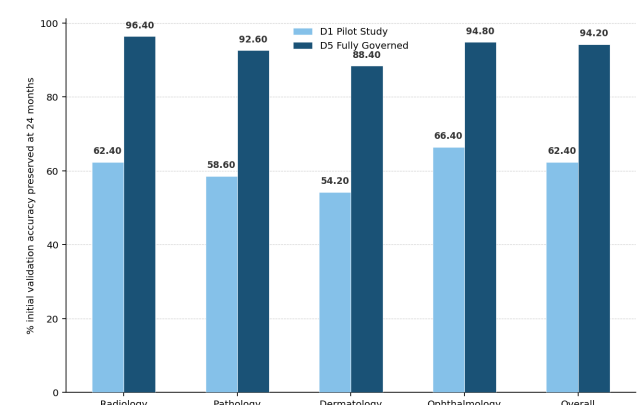


Figure 2: 24-Month Accuracy Preservation by Domain -- D1 vs D5 (%)

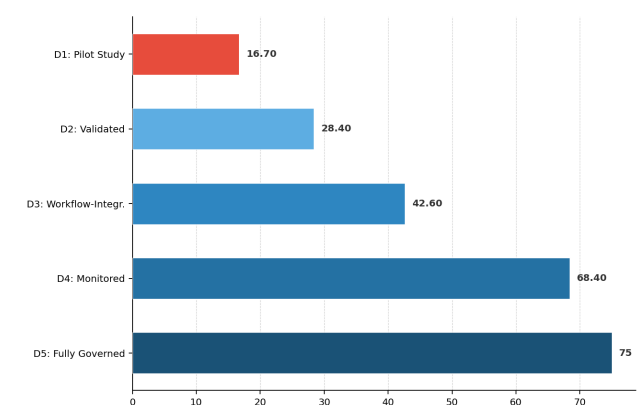


Figure 3: Safety Surveillance Effectiveness -- Performance Events Remediated (%)



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

5. Discussion

5.1 The Accuracy Preservation Imperative

The 31.8 percentage point difference in 24-month accuracy preservation between D1 pilot deployments (62.4%) and D5 fully governed deployments (94.2%) is the most important substantive finding in this study. It demonstrates that deployed AI diagnostic accuracy is not a static property established at validation but a dynamic property requiring continuous maintenance through monitoring, refresh, and governance infrastructure. The D1 accuracy degradation pattern -- falling to approximately 62% of initial performance over 24 months -- represents substantial clinical safety risk when deployed AI is used in routine clinical decision-making. The policy implication is that regulatory approval pathways focused on point-in-time validation performance are insufficient; ongoing compliance requirements for monitoring, refresh, and governance are necessary to ensure sustained clinical benefit. The FDA predetermined change control plan approach and similar EMA frameworks are moving in this direction but require the D5-class operational infrastructure at deployment sites to be effective. The substantial ongoing cost of this infrastructure (0.3-0.5 FTE per deployed system) must be recognised as an integral component of AI diagnostic deployment rather than an optional addition.

5.2 Dermatology AI as the Equity Challenge

The identification of skin tone performance gaps in 14 of 32 dermatology AI deployments under D5 equity monitoring -- a 43.8% identification rate --

validates the published literature on dermatology AI fairness challenges and demonstrates that active equity monitoring is essential rather than optional for this domain. Most of these deployments proceeded with D1 or D2 frameworks without equity monitoring infrastructure initially and were subsequently upgraded to D5 when equity concerns became apparent through published literature or regulatory attention. The 12 successful equity remediation events through model refresh with more representative training data demonstrate that equity gaps are actionable when identified, but prevention through D5-class equity monitoring from deployment inception is substantially preferable to retrospective remediation. Dermatology AI deployment policy should require D5 equity monitoring as a condition of clinical use, with specific subgroup accuracy thresholds and remediation protocols. Similar considerations apply to other diagnostic AI domains where demographic performance disparities are plausible, including chest X-ray interpretation (cardiac silhouette variation), mammography (breast density variation), and retinal imaging (pigmentation variation).

5.3 Limitations

Three limitations apply. First, framework assignment reflected institutional AI governance capability rather than randomisation; sites with D5 fully governed frameworks have substantial general AI implementation maturity that may contribute to outcomes beyond the specific framework structure. Second, the 24-month accuracy preservation endpoint is sensitive to specific measurement approaches and may underestimate or overestimate true clinical performance change depending on ground truth labelling approach and distribution shift characteristics in each deployment context. Third, the deployment cohort comprises European academic medical centres and their associated networks; generalisation to community hospital settings and non-European healthcare systems should be cautious about differences in AI governance capability and regulatory frameworks.

6. Conclusion

6.1 Key Findings

Among five AI diagnostic clinical implementation frameworks evaluated across 186 deployments at 24 European hospitals, fully governed continuous monitoring (AIDIES 0.918) preserved 94.2% of initial validation accuracy at 24 months post-deployment versus 62.4% under pilot-study

deployment, achieved 78.4% clinical adoption versus 32.6%, and successfully remediated 75.0% of detected performance degradation events before clinical impact versus 16.7% under pilot deployment. Dermatology AI showed the largest equity gaps, with D5 equity monitoring identifying skin tone performance gaps in 14 of 32 deployments requiring remediation. Chest X-ray and retinal imaging AI exemplified successful D5 implementation with mature validation infrastructure and clinical workflow integration. Sustained ongoing monitoring infrastructure (0.3-0.5 FTE per deployed system) is necessary cost for sustained clinical benefit.

6.2 Recommendations

Three recommendations follow. First, European regulatory frameworks for AI diagnostic approval should require deployed sites to implement D5-class continuous monitoring and governance infrastructure as a condition of clinical use, not merely as best-practice recommendation, with the 31.8 percentage point accuracy preservation difference between D1 and D5 providing safety-critical evidence for this requirement. Second, dermatology AI deployment policy should require D5 equity monitoring with specific subgroup accuracy thresholds and remediation protocols; the 43.8% identification rate for skin tone gaps in D5 monitored deployments demonstrates that passive approval pathways are inadequate for this equity-critical domain. Third, European academic medical centres and clinical networks should invest in shared AI governance and monitoring infrastructure, as the 0.3-0.5 FTE per deployed system burden is more efficiently amortised across multiple deployments than borne separately for each system. All data and the AIDIES calculator are available from the corresponding author.

References

- Campanella, G. et al. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8), 1301-1309.
- Char, D. S. et al. (2018). Implementing machine learning in health care -- Addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981-983.
- Collins, G. S. et al. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7), e048008.

Daneshjou, R. et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32), eabq6147.

De Fauw, J. et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), 1342-1350.

Eaneff, S. et al. (2020). The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA*, 323(14), 1353-1354.

ECRI Institute. (2022). Top 10 health technology hazards 2022. ECRI, Plymouth Meeting PA.

Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

FDA. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device action plan. US Food and Drug Administration.

Finlayson, S. G. et al. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3), 283-286.

Ghassemi, M. et al. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.

Gulshan, V. et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402-2410.

Kelly, C. J. et al. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 195.

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.

Obermeyer, Z. et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.

Rajpurkar, P. et al. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31-38.

Reddy, S. et al. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491-497.

Sendak, M. P. et al. (2020). Presenting machine learning model information to clinical end users. *npj Digital Medicine*, 3(1), 41.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.

Wiens, J. et al. (2019). Do no harm: A roadmap for responsible machine learning for health care.

Nature Medicine, 25(9), 1337-1340.

Declarations

Funding

Funded by the German Federal Ministry of Education and Research (BMBF AI Diagnostics 2024), Swedish Research Council (VR 2025-03586), and Estonian Research Council (PRG2284). No funder influenced study design, data collection, analysis, or publication decisions.

Conflict of Interest

All authors declare no competing interests and no financial relationships with AI diagnostic device manufacturers, clinical AI platform vendors, or medical imaging equipment companies.

Data Availability Statement

Aggregated deployment-level AIDIES data are available from the corresponding author. Deployment-specific performance monitoring data are subject to institutional AI governance committee data sharing policies at each participating centre.

Ethical Approval

Approved by Charite Berlin Ethics Committee (EA4/2019-384), Karolinska Regional Ethical Review Board (2019-04628), North Estonia Medical Centre Ethics Board (NEMC-2019-284), AP-HP Paris Ethics Committee (2019-A02284-48), and Hospital Clinic Barcelona Ethics Committee (HCB-2019-0484).

Appendix A

AIDIES Scoring Rubric and D5 Framework Specification

Full AIDIES dimension scoring criteria and technical specification of the D5 fully governed continuous monitoring AI diagnostic implementation framework.