Assuring Trust in Clinical AI: Monitoring Real-World Performance of Diagnostic AI Tools

Dr Jesus Perdomo Lampignano[1], Prof. David J Lowe[2], Dr Mark Hall[3]

[1]NHS Clinical AI Fellow and ST3 Radiology Registrar, NHS Greater Glasgow & Clyde, [2]Clinical Director, Innovation , University of Glasgow , Emergency Consultant at Queen Elizabeth University Hospital, NHS Greater Glasgow and Clyde.
[3]Clinical Radiology Consultant at Queen Elizabeth University Hospital, NHS Greater Glasgow & Clyde.

NHS Education for Scotland

NHS Greater Glasgow and Clyde

DHVL DIGITAL HEALTH VALIDATION LAB

NHS

Fellowship in Clinical AI

NHS Digital Academy Programme

Why Monitoring AI Matters

AI is entering clinical practice, but who makes sure it remains safe and effective after deployment?

Radiology departments are under immense pressure, with increasing scan volumes and a global shortage of radiologists(1). AI tools are now being deployed to help – from flagging lung cancer on chest X-rays to detecting brain bleeds on CT. But real-world performance can change. AI tools may drift, misclassify, or perform inconsistently across patient populations or scanner types. Without ongoing monitoring, these risks go undetected, affecting diagnostic safety and fairness(2).

AIVAL and FAMOS

A solution: platforms for continuous and responsive monitoring.

During my Clinical AI Fellowship, I worked on two separate projects that explore how healthcare providers can monitor the safety and effectiveness of AI tools in radiology after deployment.

AIVAL Analysis Lab

AIVAL is a local monitoring platform piloted across two health boards to retrospectively assess the performance of AI models used for chest X-rays (lung cancer triage) and CT brain scans (intracranial haemorrhage). The project focused on performance benchmarking, stratified analyses to detect potential algorithmic bias, and the generation of audit-friendly outputs to support clinical governance.

FAMOS (Federated AI Monitoring Service)

A multi-site initiative piloted across to test federated performance monitoring without centralising patient data. The platform enables longitudinal tracking of AI accuracy by assessing concordance with radiologists, automation bias detection, and tracks input data quality as a marker for performance drift.

While each project evaluated a different platform and technical approach, both reflect the NHS's growing need to independently verify how AI tools behave in live clinical environments.

Progress and Milestones

Although data analysis is ongoing, several key findings and learning points have started to emerge from each project:

Bias detection in real-world settings

- Stratified performance analyses revealed **variation in AI performance** across patient groups, such as age and sex, as well as across scanner manufacturers, which may be an early indicator of **domain shift**.
- AI vendors exhibited different patterns of variation, highlighting the importance of **site-specific monitoring** rather than relying solely on vendor-reported benchmarks without **institutional partnership.**

Automation bias and radiologist interaction

- Using longitudinal concordance tracking between radiologists and AI outputs, we observed signs that more junior radiologists may be more prone to automation bias.

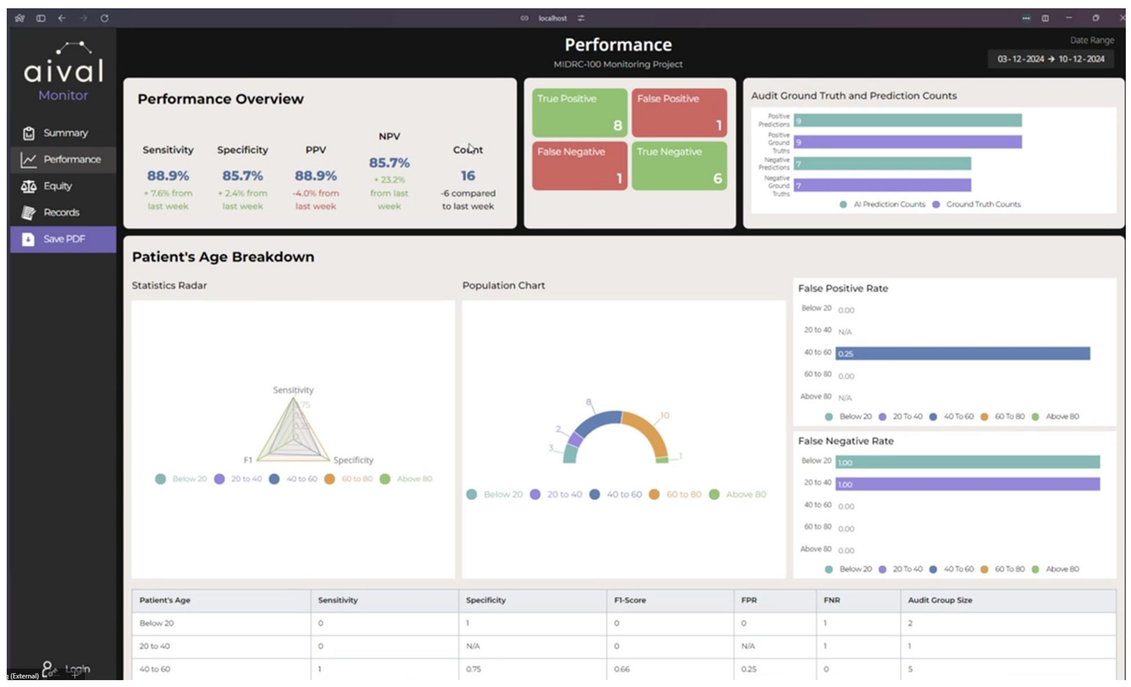
Platform feasibility and technical milestones

- We successfully deployed both platforms in simulated monitoring environments, including integration with RIS/PACS gateway systems to closely replicate real-time clinical workflows.
- For the FAMOS project, we implemented a federated monitoring architecture that preserved patient privacy while enabling cross-site performance comparisons across two major NHS trusts demonstrating the feasibility of scalable, privacy-preserving AI oversight.

References

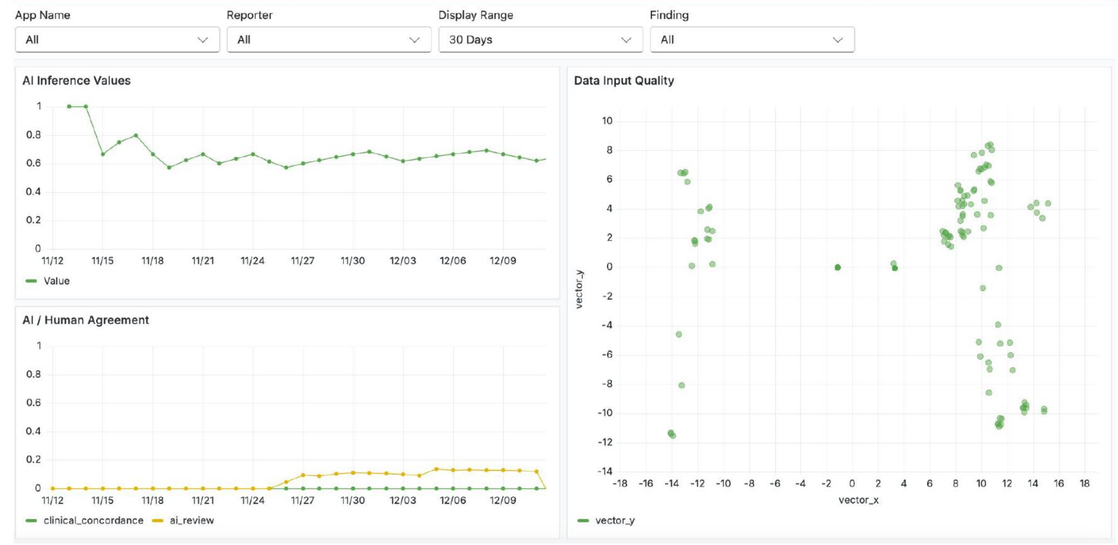
1. Clinical Radiology Workforce Census 2023 The Royal College of Radiologists.
2. Hasanzadeh, F., et al. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. npj Digit. Med. 8, 154 (2025).

Figure 1 – AIVAL Monitoring Dashboard



AI performance overview, stratified analysis, and audit-friendly reporting

Figure 2 – FAMOS Dashboard



AI inference, AI-Radiologist concordance, and data quality plots to assess performance longitudinally.

My Contributions & Next Steps

During the fellowship, I played a key role in aligning governance, ethics, and technical implementation workflows across both projects. I designed and curated calibration and monitoring datasets, contributed to protocol development for performance auditing and bias detection, and regularly engaged in multidisciplinary collaboration with radiologists, vendors, academics, and governance teams. I also synthesised emerging findings into actionable insights to support deployment decisions and academic dissemination.

These projects lay the foundation for safer, more transparent AI integration in radiology. Looking ahead, I aim to:

- Embed monitoring into routine deployment via SOPs and governance frameworks.
- Scale monitoring practices across NHS sites.
- Develop dashboard tools for radiologists and oversight boards.
- Contribute to national policy on post-market AI surveillance.

Broader Fellowship Contributions

- Presented at RCR Global AI Conference on lessons from real-world deployment.
- Authored an editorial on autonomous AI in Rad Magazine and a narrative review (under submission) on post-market surveillance.
- Joined the WoS AI strategy group, contributing to local governance policy.
- Currently designing a dataset repository to support reproducible AI evaluation.
- Working on service-wide implementation of a chest X-ray AI tool in the regional lung cancer pathway, including local SOP development.
- Currently developing AI education resources for radiology registrars and medical students at the University of Glasgow.

I plan to remain actively involved in this space beyond the fellowship, working at the intersection of clinical radiology, AI safety, and implementation science.