Post-deployment monitoring and safety reporting of AI medical imaging devices in clinical practice

March 2026

The Royal College of Radiologists

Contents

01 Scope

This guidance contains recommended monitoring activities for clinical providers after they have deployed imaging-based artificial intelligence (AI) solutions within clinical radiology and/or oncology departments. Non-imaging-based AI productivity tools used within these clinical departments are outside the scope of this document. Post-market surveillance (PMS) responsibilities of vendors are also outside the scope of this guidance.

Departmental AI medical device monitoring should be performed in partnership with multidisciplinary stakeholders (including medical physics, clinical engineering, IT and information governance teams) and comply with local governance procedures. Communication with external stakeholders (such as vendors and governmental regulatory bodies) should be maintained to ensure clinical safety for all patients.

This guidance was developed by an expert panel and incorporates feedback following a public consultation. The expert panel acknowledges the AI field is constantly evolving. The content of this document details the standards to monitor AI tools in radiology and oncology at the time of publication, but departments should consider how their monitoring activities need to evolve as AI medical imaging devices and their usage change.

The Royal College of Radiologists (RCR) recognises that appropriate infrastructure, expertise and funding are essential for post-deployment monitoring within local institutions, imaging networks and health boards.

02 Introduction

The use of AI in radiology and oncology departments is growing rapidly, with AI medical devices being deployed to support image interpretation and streamline patient pathways. While these technologies have the potential to enhance efficiency and patient care, their real-world performance must be routinely monitored to ensure they remain safe, effective and fit for purpose.

Performance of AI tools can change over time due to factors such as software updates, shifts in patient demographics and variations in imaging equipment and protocols across different departments. Ongoing performance monitoring is therefore essential for assuring that AI medical devices are delivering reliable results in clinical practice.

This document provides practical, step-by-step guidance on how to undertake post-deployment monitoring and safety reporting effectively, ensuring that AI medical devices in NHS radiology and oncology departments continue to support the delivery of safe and high-quality patient care.

This guidance primarily focuses on AI medical devices that are directly involved in diagnostic or clinical decision-making workflows. While broader AI applications such as large language model tools are not included within the scope of this guidance, awareness of such technologies and their potential integration into clinical workflows should be reflected in training and capacity-building activities.

Guidance that focuses on the transition from regulatory approval to routine clinical deployment is available in AI deployment fundamentals for medical imaging.[1] The Guidance on auto-contouring in radiotherapy includes recommendations on commissioning and post-implementation monitoring.[2]

03 Understanding post-deployment monitoring

Post-deployment monitoring is the ongoing process, undertaken by deploying institutions, of monitoring an AI medical device and the working environment in which it has been deployed in clinical practice. It ensures that the AI continues to work safely, effectively and reliably in real-world settings, beyond the controlled conditions of regulatory approval. The frequency and degree of monitoring should be tailored for each AI medical device and may be dependent on the use case.

Key aspects of post-deployment monitoring include:

- Tracking AI medical device performance (diagnostic accuracy, consistency and reliability)
- Identifying safety concerns (errors, biases, unexpected results)[3]
- Evaluating AI's impact on patient care and clinical workflows, including its intended purpose, such as workflow efficiency or diagnostic aid, and how this affects staff roles and decision-making
- Ensuring compliance with regulatory requirements (eg Medicines and Healthcare products Regulatory Agency (MHRA)).

Think of post-deployment monitoring as the 'MOT check' for your AI medical device – just because it worked when first approved does not mean it will continue performing well indefinitely.

04 How to do post-deployment monitoring

Post-deployment monitoring should be undertaken through a multidisciplinary approach, involving clinical teams, medical physics/engineering, digital safety leads and informatics support where appropriate.

- Post-deployment monitoring activities should be embedded within the organisation's clinical governance framework to ensure appropriate oversight, accountability and integration with existing quality and safety processes.
- It is important to designate an individual responsible for post-deployment monitoring and establish a reporting mechanism to provide updates to the clinical director on a regular basis.
- Departments are encouraged to notify the RCR when new AI medical devices are deployed so that these can be included in the RCR AI registry, supporting national visibility and shared learning and enabling identification of emerging trends.[4]
- Departments are encouraged to share their audit process for specific use cases by submitting audit templates to be published within AuditLive (Clinical radiology) or Audit Library (Clinical oncology).[5] [6]
- Departments should confirm that suppliers are meeting their post-deployment responsibilities. Please refer to Appendix A (supplier questions checklist) for suggested discussion points.

Each department or trust should establish local protocols that designate which individuals and departments hold local responsibility for post-deployment monitoring, how the activities will be performed and how outcomes of audits and investigations are communicated.

Step 1: Create an inventory of AI medical devices being used in the department

The Ionising Radiation (Medical Exposure) (Amendment) Regulations 2024 (IR(ME)R), which came into force on 1 October 2024, require all providers to maintain a registry of software that assists in interpretation of medical imaging that uses ionising radiation.[8] The inventory must contain the following information:

- Name of software company
- Brand name
- Current software version
- Year of original installation
- Year of current software version installation in clinical use.

While IR(ME)R explicitly states that the inventory is needed for AI medical devices that are associated with ionising radiation, we suggest that departments also include other types of AI medical devices being used.

04

Step 2: Define what to monitor

Monitor AI in three key areas (easily recalled as 'the three Ps'):[9]

- **PEOPLE**: Human–AI interaction – is the AI being used as clinically intended by all users?
- **PROCESS**: Workflow integration – is the AI processing all relevant imaging and delivering results promptly?
- **PRODUCT**: Algorithm performance – is the AI detecting findings accurately?

The specific aspects to be monitored may evolve over time as technology and clinical practice develop. Additional monitoring areas may also be identified through national or international regulatory guidance, or by other governmental or professional organisations.[10] [11] [12]

Step 3: Set up data collection

- Regularly audit AI outputs against a reference standard, where possible. Ideally this will be pathology results or final patient outcome but may be the radiologist report or follow-up imaging.
- Track AI–human disagreement rates and assess clinical impact.
- Monitor AI performance on different image-processing software and across different scanners to ensure consistency in output (see case study in Appendix B).
- Monitor for population drift (eg demographic shifts, referral pathway changes, disease prevalence), which may impact AI performance even when the AI itself has not changed.
- Review error patterns (eg is the AI consistently missing a certain type of pathology?) and ensure this is highlighted during staff training.
- Ensure vendors provide performance reports on operational metrics such as processing times and failure rates.
- Monitor the impact of AI on downstream clinical actions (eg follow-up scans, referrals), and track changes in actionable findings over time (see case study in Appendix C).

Step 4: Take action when issues arise

In addition to existing hospital incident reporting processes (eg Datix) and statutory MHRA adverse event reporting, report any adverse incidents to the MHRA using the yellow card scheme, Datix or other incident reporting tool, independent of vendor involvement.[7]
An adverse incident is defined as an event that caused (or almost caused) an injury to someone or affected the treatment or diagnosis they could receive. Use the same hospital reporting system to flag adverse events.

- If AI performance changes, liaise with the vendor to discuss the reasons (Appendix A). They may need to recalibrate, retrain or upgrade the model.
- Use existing clinical risk management processes (eg DCB0160) when issues arise.[13]
- Discuss AI errors in radiology/radiotherapy events and learning meetings (REALM) and clinical governance meetings.[14]
- Share audit findings across NHS networks to spot trends early. The RCR AI registry can be used to locate trusts using the same AI tools.
- Be particularly vigilant after any changes in the AI model, system updates or introduction of new imaging hardware (eg scanners) or image-processing software, as these can affect performance unexpectedly. Please refer to Appendix D: Impact of AI software or scanner update.
- Share risk-mitigation strategies with peers as part of post-incident learning.

05 Training

It is essential that all staff receive appropriate training in the safe and effective use of AI before clinical deployment. Training should also cover AI-specific considerations in healthcare and begin early in the acquisition and planning stages of implementation. Further recommendations on staff training are provided in the RCR guidance AI deployment fundamentals for medical imaging, and additional learning resources are available via the RCR e-learning hub.[1] [15]

- Vendor training should include:
  - Training materials, including case-based examples, and offline training where appropriate.
  - Information on which patient groups and imaging modalities the algorithm is validated and approved for use with.
  - What clinical findings or abnormalities the algorithm is intended to detect, and what are the known limitations.
  - What areas the algorithm is prone to making errors in (eg inability to detect nodules adjacent to the mediastinum).[16] [17]
- To ensure this information is available easily, vendors should provide an AI model card (as promoted by Coalition for Health AI) that lists the important criteria and specifications of the AI tool.[18]
- Training should be provided on how to interpret AI output in the appropriate clinical context, including recognising when to question or override the algorithm. Ideally, this would be via a simulation-based or offline training environment that allow users to practise interpreting AI outputs and overriding incorrect results in a controlled, non-patient-facing setting.
- Training programmes should be updated to reflect the increasing use of emerging AI technologies such as large language and multimodal models, so that staff remain informed about tools that may in future support clinical decision-making, even if such systems are not currently within the scope of this guidance.[19]

06 Conclusion

Post-market surveillance of AI medical devices is a legal requirement for manufacturers and is essential to ensuring patient safety. However, safe clinical use also relies on post-deployment monitoring within trusts to detect any performance issues and escalate to the relevant stakeholders.

Ongoing training and education for healthcare professionals, including induction of new staff and updates when there are significant changes to a tool, are an integral part of safe use. Regular engagement with users, through surveys or feedback mechanisms, can also provide early warning of issues not detected through formal audit.

Departments are encouraged to share findings with the RCR to enable early identification of emerging trends.[4] Local monitoring activities not only safeguard performance in clinical practice but also contribute to national oversight and learning.

By prioritising these foundational steps, clinical teams can help ensure the safe and effective integration of AI into patient care.

07 References

1. The Royal College of Radiologists. AI deployment fundamentals for medical imaging. London: The Royal College of Radiologists, 2024. www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/ai-deployment-fundamentals-for-medical-imaging

2. The Royal College of Radiologists. Guidance on auto-contouring in radiotherapy. London: The Royal College of Radiologists, 2024. www.rcr.ac.uk/our-services/all-our-publications/clinical-oncology-publications/auto-contouring-in-radiotherapy

3. International Medical Device Regulators Forum. Machine learning-enabled medical devices: key terms and definitions. IMDRF AIMD WG Final Document N67/2021, May 2022; Section 6.1. www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf

4. The Royal College of Radiologists. AI Registry. www.rcr.ac.uk/our-services/artificial-intelligence-ai/ai-registry

5. The Royal College of Radiologists. AuditLive (Radiology) templates. www.rcr.ac.uk/career-development/audit-quality-improvement/auditlive-radiology-templates

6. The Royal College of Radiologists. Audit Library (Clinical oncology) templates. www.rcr.ac.uk/career-development/audit-quality-improvement/audit-library-oncology-templates

7. Medicines and Healthcare products Regulatory Agency (MHRA). Yellow card: report a problem with a medicine or medical device. https://yellowcard.mhra.gov.uk

8. The Ionising Radiation (Medical Exposure) (Amendment) Regulations 2024. SI 2024/896. www.legislation.gov.uk/uksi/2024/896/pdfs/uksi_20240896_en.pdf

9. Dean G, Montañà E, Kyriazi S et al. Real-world monitoring of artificial intelligence in radiology: challenges and best practices. Korea J Radiol 2025; doi:10.3348/kjr.2025.

10. National Institute for Health and Care Excellence (NICE). Evidence standards framework for digital health technologies. NICE, updated 2025. www.nice.org.uk/what-nice-does/digital-health/evidence-standards-framework-esf-for-digital-health-technologies

11. NICE, Medicines and Healthcare products Regulatory Agency (MHRA), Care Quality Commission (CQC) and Health Research Authority (HRA). AI and digital regulations service. https://www.nice.org.uk/what-nice-does/digital-health/artificial-intelligence-ai-and-digital-regulations-service

12. National Institute for Health and Care Excellence (NICE). National Institute for Health and Care Excellence (NICE). NICE real-world evidence framework. www.nice.org.uk/corporate/ecd9

13. NHS England, NHS Digital. DCB0160: Clinical risk management: its application in the deployment and use of health IT systems. https://digital.nhs.uk/data-and-information/information-standards/governance/latest-activity/standards-and-collections/dcb0160-clinical-risk-management-its-application-in-the-deployment-and-use-of-health-it-systems

14. The Royal College of Radiologists. Standards for radiology events and learning meetings. London: The Royal College of Radiologists, 2021. www.rcr.ac.uk/media/4vaobs4a/bfcr201-standards-for-radiology-events-and-learning-meetings.pdf

15. The Royal College of Radiologists. RCR Learn. www.rcr.ac.uk/cpd-and-events/rcr-learn

16. Scaringi JA, McTaggart RA, Alvin MD et al. Implementing an AI algorithm in the clinical setting: a case study for the accuracy paradox. Eur Radiol 2025 Jul; 35(7): 4347–4353. doi.org/10.1007/s00330-024-11332-z.

17. Bernstein MH, Sheppard B, Bruno MA, Lay PS, Baird GL. Just because you're paranoid doesn't mean they won't side with the plaintiff: examining perceptions of liability about AI in radiology. medRxiv 2024. https://doi.org/10.1101/2024.07.30.24311234

18. Coalition for Health AI. The CHAI applied model card. https://chai.org/draft-chai-applied-model-card

19. Park SH, Dean G, Ortiz EM, Choi J I. Overview of South Korean guidelines for approval of large language or multimodal models as medical devices: key features and areas for improvement. Korean J Radiol 2025 Jun; 26(6): 519–523. doi:10.3348/kjr.2025.0257.

08 Acknowledgements

Acknowledgements

Guideline development group

Chair – Prof Fiona Gilbert, Department of Radiology, Lead AI Advisor of The Royal College of Radiologists, Professor of Radiology, University of Cambridge

Chair – Dr Katharine Halliday, GIRFT lead for diagnostic capacity, Consultant Paediatric Radiologist, Nottingham University Hospitals

Dr Sarim Ather, Consultant Radiologist, Co-Director, Oxford Clinical AI Research Group

Dr Rhidian Bramley, Consultant Radiologist, The Christie NHS Foundation Trust

Dr Geraldine Dean, Consultant Radiologist, Epsom and St Helier University Hospitals NHS Trust, AI lead UK TMC/ Unilabs

Dr Rajesh Jena, Academic Radiation Oncologist, Cambridge University Hospitals NHS Foundation Trust

Prof Amaka Offiah, Vice-President of The Royal College of Radiologists, Division of Clinical Medicine, University of Sheffield, Honorary Consultant Paediatric Radiologist, Sheffield Children's NHS Foundation Trust

Dr Susan Shelmerdine, Consultant Academic Paediatric Radiologist, Great Ormond Street Hospital for Children NHS Foundation Trust

Editorial and project support

Ms Kwun-Ye Chu, AI Delivery Partner, The Royal College of Radiologists

Ms Stephanie James, Professional Standards Manager, The Royal College of Radiologists

Ms Jemma Venables, AI Programme Manager, The Royal College of Radiologists

Consultation

This guidance was circulated for public consultation in August–September 2025. A total of 129 comments were received from a range of stakeholders, including clinical professionals, professional and regulatory bodies, academic institutions, and industry, and were all carefully reviewed and considered by the guideline development group in finalising the guidance.

A1 Appendix A: Questions to ask your AI supplier about post-market surveillance

Manufacturers remain legally responsible for post-market surveillance (PMS) under the UK medical devices regulations. Radiology and oncology departments are responsible for local post-deployment monitoring and clinical governance. Effective collaboration and data-sharing between suppliers and healthcare providers are essential to enable both parties to meet their respective obligations.

From 1 October 2024, the Medical Devices (Post-Market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024 introduced strengthened obligations for manufacturers to maintain a PMS system proportionate to the device risk, including performance monitoring and feedback analysis throughout its lifetime. NHS providers should ensure suppliers are meeting these requirements and collaborate in incident reporting and corrective actions.[1]

Note: Some aspects of PMS (eg performance dashboards, audit trails, incident logging) may also be supported by the platform supplier rather than the AI model vendor. It is important to clarify the division of responsibility and ensure coordination between platform and model suppliers.

The following questions are designed to help radiology and oncology departments assure themselves that suppliers have appropriate PMS, safety and quality systems in place. These questions do not replace the legal responsibilities of manufacturers or regulators but support local due diligence under NHS clinical governance and DCB0160 requirements.

Performance and validation

- What post-deployment validation has been performed on the AI in UK clinical settings and on which datasets?
- What is the accuracy, consistency and reliability of the tool?
- How do you monitor for performance degradation or data drift?
- Do you provide support for local audit and validation activities, including revalidation following software or model updates?

Regulatory compliance

- Can you provide a copy of your PMS plan, or a summary of its key elements, as required under the MHRA PMS regulations? This should outline how performance monitoring, feedback and corrective actions are managed. NHS providers have a responsibility under DCB0160 to assure themselves that an appropriate PMS system is in place.
- What is your process for updating documentation and risk assessments as required under the MHRA PMS regulations?
- Can you provide evidence of compliance with DCB0129 (Clinical risk management: manufacture of health IT systems), including a clinical-safety case report or hazard-log summary describing identified risks and mitigations?

A1

- Is the device subject to a NICE evidence generation plan (EGP) or early value assessment (EVA), and how are the associated data-collection requirements supported at deployment sites?

Monitoring and reporting

- What metrics do you provide to customers (eg processing time, failure rate, flagged abnormality rate)?
- Do you provide performance dashboards or reports to users?
- How do you ensure users can report errors or discrepancies easily?
- How do you investigate false positives or false negatives raised by customers?
- Are equity and fairness metrics reported or available (eg algorithm performance across sex, age and ethnicity groups)?

Incident management

- What is your process for reporting and responding to adverse incidents?
- Do you assist customers in reporting incidents to the MHRA (eg by supplying required technical details such as software version, device identifier and error logs, and by submitting follow-up reports through your regulatory reporting channels such as the Manufacturer Online Reporting Environment (MORE) system)?
- Do you collate adverse incidents from all your sites and report to customers?

Updates and change control

- How are software updates managed and communicated?
- Is performance revalidated after updates or when new scanners or image-processing updates are introduced?

Bias and equity

- Have you evaluated algorithm performance across different patient demographics (sex, age, ethnicities, deprivation categories)?
- What steps have you taken to mitigate potential biases? [2]

Note: It is recognised that complete demographic data (eg ethnicity or deprivation indices) may not always be available to suppliers due to privacy or data-access restrictions. Providers and suppliers should work collaboratively, within information governance frameworks, to evaluate algorithm performance across available demographic groups and share insights that support fair and equitable AI deployment.

Post-market clinical follow-up (PMCF) activities, such as large-scale analyses of bias or performance variation across demographic groups, are normally undertaken by manufacturers as part of their PMS obligations. Clinical departments are encouraged to participate in or contribute to such initiatives, where feasible, through national or multicentre collaborations rather than performing these independently.

Training and education

- What training do you provide to ensure safe and effective clinical use?
- Is there ongoing training for new users?
- How do you communicate known limitations or potential pitfalls of the algorithm?

A1

Supplementary questions for AI platform suppliers

- Does the platform (eg picture archiving and communications system or AI orchestration system) provide integrated tools to support PMS activities (eg performance dashboards, AI usage logs, incident reporting)?
- How is information shared between the platform and AI model supplier to support regulatory compliance and safety monitoring?
- Who is responsible for generating PMS data in a multi-vendor environment?

References

1. Medical Devices (Post-Market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024. Statutory Instrument 2024 No. 1368. www.legislation.gov.uk/uksi/2024/1368

2. International Medical Device Regulators Forum. Machine learning-enabled medical devices: key terms and definitions. IMDRF AIMD WG Final Document N67/2021, May 2022; Section 6.1. www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf

A2 Appendix B: AI 'outside intended use' – misdiagnosis in paediatric patients

This case study is a hypothetical illustration intended to highlight potential risks and monitoring considerations associated with AI deployment, particularly for specific patient groups. It is for illustrative purposes only to demonstrate how post-deployment monitoring may identify safety risks and unintended consequences of AI deployment. It does not represent an RCR-endorsed clinical or operational use of AI, nor does it imply acceptance of delayed reporting or prioritisation decisions based solely on AI outputs.

Caution in paediatric use of AI tools

AI tools trained primarily on adult datasets may not generalise safely to paediatric populations. An open letter from the American College of Radiologists (ACR) to the US Food and Drug Administration (FDA) in 2021 highlighted many risks that inappropriately trained AI models pose to the paediatric population and asked for greater transparency on validation in this population.[1] [2] In particular, the ACR discussed the risk of using AI triage algorithms not specifically validated for children. The ACR published a separate document on its website in 2022 elaborating on this example. It warned the public about an AI tool that failed to detect intracranial haemorrhage in a child, illustrating the potential for serious harm when AI systems are applied beyond their intended use population.

Currently there are very few AI medical devices that are regulated for paediatric use. Information on the intended use and target population of the AI tool is not routinely made public by regulatory authorities, and published studies often do not specify this either (appropriate use of AI tools for paediatric populations remains unclear in 41% of devices).[3] Therefore, the responsibility lies with the deploying institution to obtain explicit written confirmation from the vendor regarding which target population and uses the AI model has been trained and validated for.

This case underscores the critical importance of understanding the population on which an AI model was trained and validated. Departments should ensure that tools deployed |in paediatric settings have undergone appropriate testing in similar patient groups. Where tools have not been explicitly validated for children, use should be avoided or subjected to rigorous local oversight and audit.

An anonymised clinical example of misclassification in a paediatric patient has been included here to illustrate the potential risks and inform future post-market surveillance (PMS) efforts.

Case study: paediatric misclassification

In a busy general hospital emergency department, an AI triage tool was implemented to prioritise computed tomography (CT) head scans with abnormalities to the top of the on-call radiologist's reporting list. The objective was to ensure acute brain anomalies were reported and flagged to emergency staff promptly and authorised by the radiologist in a timely manner. During a busy shift, a child with an intracranial haemorrhage underwent a CT scan. Unfortunately, their CT brain scan was not prioritised for urgent reporting because the AI triage tool was not regulated for paediatric cases. Consequently, the authorised report and alert for the abnormality on the CT brain scan was delayed by several hours.

A2

The issue was identified when the on-call radiologist discovered the abnormal paediatric CT head scan later in the shift, several hours after the initial scan. Consequently:

- Despite requiring an urgent review, the case had not been flagged as a priority.
- The AI tool had not triaged the case appropriately as it was not designed to evaluate paediatric examinations.
- Adult and paediatric cases were sent to the same radiology reporting list, making the distinction between adult and paediatric cases difficult.
- While the radiographer noted the abnormality on the CT head scan at the time of the scan, they did not flag the case as they assumed the AI tool would triage the urgent study.
- Staff lacked awareness of the AI tool's specifications and limitations.

The adverse event was reported internally and at local governance meetings to ensure staff awareness of the incident and the AI tool's limitations. Rather than removing the beneficial AI tool for adult cases, the department modified its reporting system to automatically move all paediatric CT head cases (regardless of abnormality) to the top of the review list, ensuring child safety when no suitable paediatric-specific AI alternative existed.

Lessons learned

- It is important to clarify which patient groups the AI tool is validated for before implementation.
- Ensure clear communication and training for all staff who interact with the tool.
- Anticipate unintended consequences of implementation.
- Consider potential errors of omission and the complexity of the care pathway.
- Establish redundant safety mechanisms, such as having radiographers flag critical findings directly to radiologists or verify the AI triage system has appropriately prioritised cases.

References

1. Kruskal JB, Allen B, Browning J et al. Evaluation of artificial intelligence triage algorithms for medical imaging: a call for regulatory oversight to ensure patient safety. JACR 2023; 20(12): 1560–1564. www.jacr.org/article/S1546-1440(23)00409-X/fulltext

2. American College of Radiology. Triage algorithms for medical imaging pose a safety risk to children. Memorandum 9/8/2022. https://edge.sitecorecloud.io/americancoldf5f-acrorgf92a-productioncb02-3650/media/ACR/Files/Advocacy/Regulatory/Triage-Algorithms-for-Medical-Imaging-Pose-a-Safety-Risk-to-Children.pdf

3. Kelly BS, Lee J, Antram E, Arthurs O, Shelmerdine SC. Safe for kids? AI medical devices in radiology overlook paediatric suitability. Eur Radiol 2025; Oct 8. doi: 10.1007/s00330-025-11975-6. Epub ahead of print. PMID: 41060417.

A3 Appendix C: Case study – post-market surveillance audit for pneumothorax detection

Evaluation, auditing and monitoring are complementary but distinct activities. Evaluation refers to the assessment of an AI medical device before or at the point of deployment to confirm expected performance. Auditing involves structured, periodic review of outputs against reference standards to verify ongoing accuracy and compliance. Monitoring is the continuous or routine observation of operational performance and safety indicators to detect change over time. Together, these activities form the basis of effective clinical safety assurance.

Audit frequency, sample size and escalation thresholds should be locally defined through governance meetings such as REALM and proportionate to the clinical risk and workload. For example, a rise in the AI–human disagreement rate or a reduction in sensitivity may trigger enhanced monitoring, additional sampling or temporary suspension pending review. Proxy indicators such as turnaround time or abnormality flag rate can also be used to detect drift between full audits.

This case study outlines how a post-market surveillance (PMS) audit can be used to evaluate the real-world performance of an AI tool for detecting pneumothorax on chest radiographs.

The audit can be conducted in two phases: initially in shadow mode (where AI results are not visible to clinicians), and subsequently during live clinical use. Radiology reports are reviewed against AI findings, with discrepancies manually examined to establish the ground truth. During the shadow phase, the enhanced detection rate – the proportion of missed cases that could have been flagged by AI – can be calculated.

Once the AI is live in clinical use, the audit can be repeated to assess detection rates in both radiology and emergency department (ED) documentation, enabling evaluation of whether AI improves early recognition and whether there is any drift in performance over time.

This type of audit can be repeated periodically to monitor for changes or drift in AI or human performance and can be scaled up using natural language processing (NLP) to extract and analyse findings from radiology reports and clinical notes.

Illustrative example: If the AI–human disagreement rate for pneumothorax cases increases from 3% to 8% over two consecutive audit cycles, this could prompt a targeted review of the affected cases and temporary enhanced monitoring of all pneumothorax detections until the cause of the variation is understood.

Example protocol for PMS audit: pneumothorax detection

Objective

To assess the performance of an AI tool in detecting pneumothorax on chest radiographs and evaluate its impact on detection by radiology and ED clinicians, both in shadow mode and after clinical deployment.

A3

Audit approaches

1. Simple audit (manual review)

- Identify all AI-positive chest radiographs over a defined period.
- Review the original radiology report and image to determine whether the AI flag was correct (calculate positive predictive value – PPV).
- Check whether ED documentation (prior to the report) recognised the pneumothorax in true positive cases.
- Track the number of pneumothorax cases detected in shadow mode versus live deployment.

2. In-depth audit (NLP)

Use NLP or keyword search to identify chest X-ray reports that mention pneumothorax, excluding negative clauses (eg 'no pneumothorax').

- Match AI findings to report content.
- Review any discrepancies (AI says positive, report says negative, or vice versa) to establish the ground truth, ideally through expert radiologist adjudication.

Each case is then classified as:

- **True positive (TP)**: AI and report agree on presence of pneumothorax
- **True negative (TN)**: AI and report agree on absence
- **False positive (FP)**: AI flags pneumothorax not confirmed in report
- **False negative (FN)**: AI misses a pneumothorax present in the report.

Metrics to calculate

- **Sensitivity** = TP / (TP + FN)
- **Specificity** = TN / (TN + FP)
- **Positive predictive value** (PPV) = TP / (TP + FP)
- **Negative predictive value** (NPV) = TN / (TN + FN)

Enhanced detection rate (EDR)

EDR assesses whether AI could improve detection by identifying cases missed by humans.

- **Step 1**: In shadow mode, identify cases where pneumothorax was missed in the original report but flagged by AI.
- **Step 2**: Confirm the ground truth by expert review.
- **Step 3**: Calculate EDR as the number of additional true positives detected by AI divided by the total number of pneumothorax cases in the dataset:

EDR = Additional TPs detected by AI / Total confirmed pneumothorax cases

Comparative evaluation

- Compare detection rates between:
  - Radiologists (pre- and post-AI deployment)
  - ED clinicians (pre- and post-AI deployment – before the radiology report is available).
- Assess whether AI implementation improves early recognition, increases overall detection or changes detection patterns over time.
- Additional measures such as reporting turnaround times, workflow impact or cost-related

A3

indicators (eg reduction in follow-up imaging or admissions) may also be evaluated where data are available. These provide useful insights into efficiency and health economic outcomes and complement diagnostic accuracy measures.

Notes

- Access to suitable digital infrastructure, such as data-integration platforms or NLP capability, can enable large-scale post-deployment monitoring and automation of audit tasks. These tools are not yet available in all NHS trusts. Where such infrastructure is lacking, simplified audit methods such as targeted manual review can provide equivalent assurance on a smaller scale. Regional or network-level platforms may support wider adoption of scalable approaches in future.
- This protocol can be adapted for other AI-detected conditions.
- Audits can be repeated periodically to assess drift in AI or human performance.
- Use of NLP can significantly scale the audit and reduce manual workload.
- Although NLP may not be 100% accurate in identifying conditions from free-text reports, it is still useful for screening and trend analysis.
- Manual review is recommended to establish ground truth when calculating diagnostic performance metrics (eg sensitivity, specificity).
- However, for a priori analysis aimed at detecting change or drift over time (eg comparing AI agreement rates or detection rates pre- and post-deployment), full adjudication of ground truth may not be necessary, provided the same method is applied consistently.
- The depth and frequency of post-deployment monitoring audits should be proportionate to the clinical risk and impact of each AI use case. Departments may prioritise higher-risk or higher-impact applications (eg acute findings such as pneumothorax) for regular review, while applying lighter-touch monitoring to lower-risk use cases. This ensures PMS remains efficient and does not negate productivity gains achieved through AI deployment.

A4 Appendix D: Impact of AI software or scanner update

Monitoring performance of AI tools after a software update or change of scanner hardware or software is critical to ensure continued diagnostic accuracy and patient safety in radiology. Such updates can alter image quality, reconstruction algorithms or post-processing parameters, potentially affecting the appearance of pathology or the performance of AI tools.

Case study 1 – Impact of mammography software upgrade on AI recall rates[1]

A commercially available AI tool for breast cancer screening was implemented at a site using digital mammography. Following a routine software upgrade to the mammography system, the clinical team observed that the AI system's recall rate tripled compared with the pre-upgrade baseline, flagging nearly 50% of studies for recall. This was far above acceptable clinical thresholds and risked overwhelming the clinical workflow with false positives.

Analysis revealed that the AI's performance was highly sensitive to image characteristics introduced by the software update, including contrast and processing changes.
A new threshold calibration, specific to the updated software version, was performed. Once recalibrated, the AI system's recall rate aligned with clinical expectations, while maintaining high sensitivity for screen-detected and interval cancers.

Lessons learned

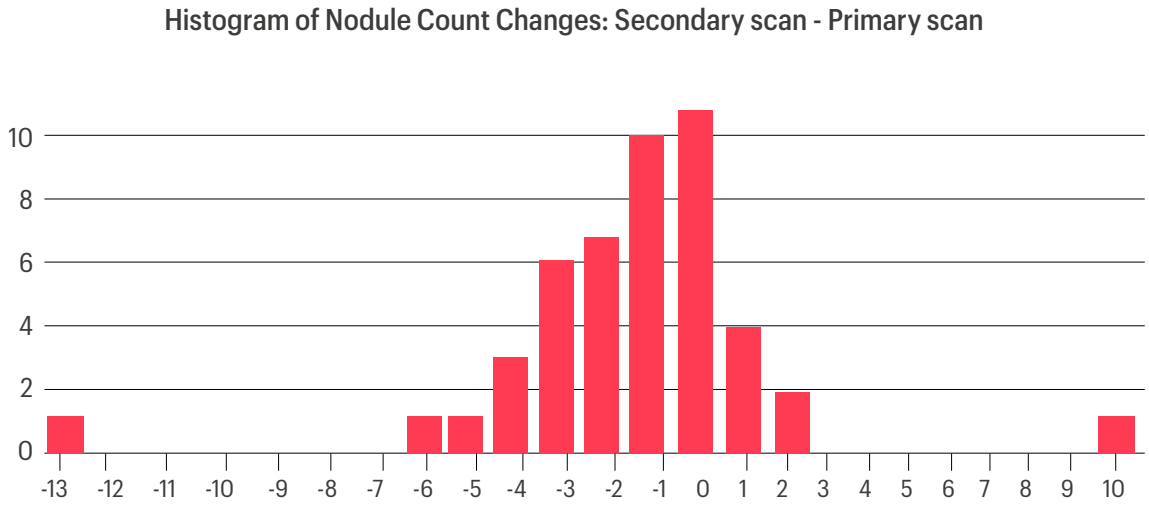
- AI tools can be **software-version dependent**; performance may degrade after imaging system updates.
- **Default AI thresholds are not generalisable** and may require version-specific calibration.
- **Post-upgrade validation** and **ongoing performance monitoring** are essential to ensure safe deployment in clinical practice.

Case study 2 – Impact of AI software update on lung nodule measurement

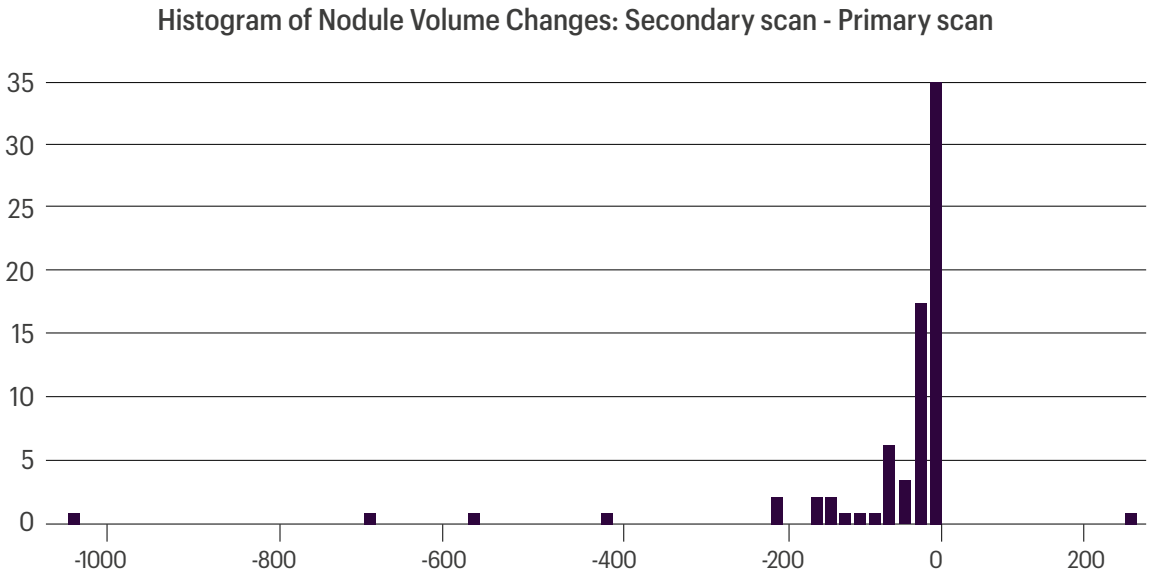
A software update was applied to a lung nodule volumetry tool used in chest CT analysis. Shortly after the update, discrepancies were observed in the number and size of nodules reported by the new version compared with the previous one.

An internal audit was conducted on scans that had been analysed by both software versions. The review revealed that in 80% of cases the two versions disagreed on the number of nodules detected.

A4



Even when the same nodules were identified, there were **significant differences in size measurements**. In approximately **20% of patients** the variation in nodule size between versions was **large enough to alter follow-up management decisions**.



Lessons learned

Software updates can significantly alter AI outputs

- Both case studies demonstrate that even routine software updates, whether to imaging systems or AI tools, can lead to substantial changes in diagnostic outputs. These include differences in detection rates, quantitative measurements and recall thresholds.

Version-specific calibration is essential

- AI tools often rely on specific image characteristics and tuning parameters that may not generalise across software versions. Thresholds or volumetric calculations calibrated on one version may become invalid after an update. Recalibration after updates is critical to restore safe and effective performance.

A4

Post-update performance may affect clinical decision-making

- Discrepancies in measurements or detection rates introduced by software updates can directly impact patient management, including follow-up intervals, referral decisions and further investigations. These changes are not always obvious without targeted monitoring.

Routine validation and auditing are crucial

- Both cases highlight the importance of internal audits or performance validation exercises following any upgrade. Comparing pre- and post-update outputs can help identify performance drift, restore confidence and prevent harm.

Ongoing post-deployment monitoring is non-negotiable

- These examples underscore the broader need for formalised, ongoing post-deployment monitoring frameworks. Imaging software, AI tools and hardware systems must be continuously monitored to detect unintended performance variation over time and across system changes.

References

1. de Vries CF, Colosimo SJ, Staff RT et al. Impact of different mammography systems on artificial intelligence performance in breast cancer screening. Radiol Artif Intell 2023 Mar 22; **5**(3): e220146. doi: 10.1148/ryai.220146. PMID: 37293340; PMCID: PMC10245180.

The Royal College of Radiologists
63 Lincoln's Inn Fields
London, WC2A 3JW, UK

The Royal College of Radiologists
is a Charity registered with the
Charity Commission No 211540.

+44 020 7405 1282
enquiries@rcr.ac.uk
rcr.ac.uk

@RCRadiologists

The Royal College of Radiologists

The Royal College of Radiologists.
Post-deployment monitoring and safety reporting of AI medical imaging devices in clinical practice.
London: The Royal College of Radiologists, 2026.

The Royal College of Radiologists is a Charity registered with the Charity Commission No. 211540