Practical guidance to Post-market surveillance of Artificial Intelligence tools for Radiology and Oncology departments

Introduction

The use of Artificial Intelligence (AI) in radiology and oncology departments is growing rapidly, with tools 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 continuously monitored to ensure they remain safe, effective, and fit for purpose.

Unlike traditional medical devices, 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 tools are delivering reliable results in clinical practice.

This document provides practical, step-by-step guidance on how to implement Post-market surveillance effectively, ensuring that AI tools in NHS radiology and oncology departments continue to deliver safe and high-quality patient care.

This guidance primarily focuses on AI tools considered medical devices or directly involved in diagnostic or clinical decision-making workflows. Broader applications such as scheduling or rostering tools may pose risks and are included in the appendix.

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

Section 1: Understanding Post-market surveillance

Post-market surveillance is the ongoing process of monitoring an AI tool 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.

Key aspects of Post-market surveillance include:

- Tracking AI performance (accuracy, consistency, and reliability)
- Identifying safety concerns (errors, biases, unexpected results)
- 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-market surveillance as the "MOT check" for your AI tool—just because it worked when first approved doesn't mean it will continue performing well indefinitely.

Section 2: How to do Post-market surveillance

- It is important to designate an individual responsible for Post-market surveillance and establish a reporting mechanism to provide updates to the clinical director on a regular basis.
- Departments are encouraged to notify the Royal College of Radiologists when new AI tools are deployed, supporting national visibility and shared learning.[3]
- Departments should ensure suppliers are meeting their Post-market responsibilities. Please refer to Appendix A: Supplier questions checklist for suggested discussion points.

Step 1: Create an inventory of AI Tools being used in the department[4]

The Ionising Radiation (Medical Exposure) (Amendment) Regulations 2024, which came into force on the 1 October 2024, require all providers to maintain a registry of software that assists in interpretation of medical imaging that uses ionising radiation.[5] The inventory must contain the following information:

(a) name of software company
(b) brand name
(c) current software version
(d) year of original installation
(e) year of current software version installation in clinical use.

Step 2: Define what to monitor

Monitor AI in three broad key areas (easily recalled as the three 'P's):

- **PEOPLE: Human-AI interaction** – are reporters using AI appropriately (neither over relying on nor disregarding it) to its intended clinical use?
- **PROCESS: Workflow integration** – Is the AI processing all relevant imaging and delivering results promptly?
- **PRODUCT: Algorithm performance** – Is the AI still detecting findings accurately?

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 patient demographics and across different scanners to ensure consistency in output (See case study in Appendix B.
- 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

- Report any adverse incidents to the MHRA using the yellow card scheme or Datix, independent of vendor involvement.[4] An adverse incident is defined as an event that caused (or almost caused) an injury to someone or affected the treatment or diagnosis one could receive. Use same hospital reporting system to flag adverse events,
    - Consider using departmental clinical risk management processes (eg DCB 0160) when issues arise.
    - Discuss AI errors in radiology events and learning meetings and clinical governance meetings.[6]
    - Share audit findings across NHS networks to spot trends early. The RCR registry can be used to locate Trusts using the same AI tools.
    - If AI performance changes, liaise with the vendor to discuss the reasons (Appendix A). They may need to recalibrate, retrain, or upgrade the model.
    - 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/scanner update.
    - Share risk mitigation strategies with peers as part of post-incident learning.

Training

It is important to ensure that training of all staff in the appropriate use of AI is given prior to clinical use.[1] Further information on AI training pathways is available via the RCR's training resources.

The training should include:

- Vendors should provide training materials, including case-based examples, and offer "off-line training" where appropriate.
- These should include information on
  which patient groups and imaging modalities the algorithm validated and approved for use.
    - what clinical findings or abnormalities the algorithm is intended to detect, and what are the known limitations
    - what areas is the algorithm prone to making errors (eg inability to detect nodules adjacent to the mediastinum).[7,8]
- 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 above.[9]
- Training on how to interpret AI output in the appropriate clinical context, including recognising when to question or override the algorithm should be considered. Ideally, this would be via a simulation-based or 'offline' training environment which 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 large language and multimodal models, with particular attention to tools that combine imaging and textual data to support clinical decision-making.[10]

Conclusion

Post-market surveillance of AI tools is a legal requirement for the company and essential to ensuring patient safety. It is therefore vital that a Post-market surveillance plan is established when deploying an AI tool and any performance issues are detected and highlighted to the relevant stakeholders.

Equally important is the need for ongoing education and training for healthcare professionals, as well as regular surveys with users. Robust Post-market surveillance practices not only safeguard local performance but also contribute to national oversight.

Departments are encouraged to share findings with the RCR to enable early identification of emerging trends.[3] These efforts provide valuable insights into how AI tools are being used in practice and can help identify potential issues that formal monitoring systems may miss. By prioritising these foundational steps, clinical teams can build a robust framework for the safe and effective integration of AI into patient care.

References

1 Royal College of Radiologists. AI Deployment Fundamentals for Medical Imaging, 2024. Available at: https://www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/ai-deployment-fundamentals-for-medical-imaging/

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

3 Royal College of Radiologists. AI registry. Available at: https://www.rcr.ac.uk/our-services/artificial-intelligence-ai/ai-registry/

4 Medicines and healthcare products Regulatory Agency. Yellow Card: Report a problem with a medicine or medical device. Available at: https://yellowcard.mhra.gov.uk/

5 UK Government (2024) *The Ionising Radiation (Medical Exposure) (Amendment) Regulations 2024*. SI 2024/896. Available at: https://www.legislation.gov.uk/uksi/2024/896/pdfs/uksi_20240896_en.pdf

6 Royal College of Radiologists. Standards for Radiology Events and Learning Meetings, 2021. Available at: https://www.rcr.ac.uk/media/4vaobs4a/bfcr201-standards-for-radiology-events-and-learning-meetings.pdf

7 Scaringi, J.A., McTaggart, R.A., Alvin, M.D. *et al.* Implementing an AI algorithm in the clinical setting: a case study for the accuracy paradox. *Eur Radiol* (2024). https://doi.org/10.1007/s00330-024-11332-z

8 Bernstein, M.H., Sheppard, B., Bruno, M.A., Lay, P.S., & Baird, G.L. (2024). Just because you're paranoid doesn't mean they won't side with the plaintiff: Examining perceptions of liability about AI in radiology. medRxiv. https://doi.org/10.1101/2024.07.30.24311234

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

10 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 Journal of Radiology. 2025 Jun;26(6):519–523. doi:10.3348/kjr.2025.0257

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

From 1 October 2024, the Medical Devices (Post-Market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024 introduced strengthened obligations for manufacturers to maintain a Post-market surveillance 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.
[Statutory Instrument 2024 No. 1368 – Regulation 44ZE]

Note: Some aspects of Post-market surveillance (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.

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?

Regulatory compliance

- Can you provide a copy of your Post-market surveillance plan as required under the MHRA Post-market surveillance regulations?
- What is your process for updating documentation and risk assessments as required under the MHRA's Post-market surveillance regulations?

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 investigate false positives or false negatives raised by customers.

Incident management

- What is your process for reporting and responding to adverse incidents?
- Do you assist customers in reporting to MHRA via the Yellow Card scheme
- 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 re-validated after updates or when new scanners or image processing updates 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?

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?

Supplementary questions for AI platform suppliers

- Does the platform (eg picture archiving and communications system or AI orchestration system) provide integrated tools to support Post-market surveillance 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 Post-market surveillance data in a multi-vendor environment?

Appendix B: AI ‘outside intended use’ – misdiagnosis in paediatric patients

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 FDA (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 discuss the risk of using AI triage algorithms not specifically validated for children. The ACR published a separate document on their website (2022) further elaborating on this example, warning 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.
- 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 subject to rigorous local oversight and audit.
- An anonymised clinical example of misclassification in a paediatric patient has been included below to illustrate the potential risks and inform future Post-market surveillance efforts.

Case study

In a busy general hospital emergency department, an AI triage tool was implemented to prioritise CT head scans with abnormalities to the top of the on-call radiologists' 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. However, 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.

The issue was identified when the on-call radiologist discovered the abnormal paediatric CT head 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 are 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 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 their 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 from this incident include:

- the importance of clarifying which patient groups the AI tool is validated for before implementation
- ensuring clear communication and training for all staff who interact with the tool,
- anticipating unintended consequences of implementation
- considering potential errors of omission and the complexity of the care pathway
- establishing 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, J.B., Allen, B., Browning, J., Kansagra, A.P., Larson, D.B., Allen, T.A., Dickerson, J., Thrall, J.H. and Dreyer, K.J., 2023. Evaluation of Artificial Intelligence Triage Algorithms for Medical Imaging: A Call for Regulatory Oversight to Ensure Patient Safety. *Journal of the American College of Radiology*, [online] 20(12), pp.1560–1564. Available at: https://www.jacr.org/article/S1546-1440(23)00409-X/fulltext [Accessed 17 Jun. 2025].

2 American College of Radiology, n.d. Triage Algorithms for Medical Imaging Pose a Safety Risk to Children. [pdf] Available at: 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 [Accessed 17 Jun. 2025].

Appendix C: Case Study – Post-market surveillance audit for pneumothorax detection

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 X-rays.

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.

Example Protocol for PMS Audit: Pneumothorax Detection

Objective:
To assess the performance of an AI tool in detecting pneumothorax on chest X-rays and evaluate its impact on detection by radiology and emergency department (ED) clinicians, both in shadow mode and after clinical deployment.

Audit Approaches:

1. Simple Audit (Manual Review):

- Identify all AI-positive chest X-rays over a defined period.
- Review the original radiology report and image to determine whether the AI flag was correct (i.e. 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 (e.g. “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.

Notes:

- 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 (e.g. sensitivity, specificity).
- However, for *a priori* analysis aimed at detecting change or drift over time (e.g. 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.

Appendix D: Impact of AI software/scanner update

Monitoring performance of AI tools after a software update as well as 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 evaluation retrospectively 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 to 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.

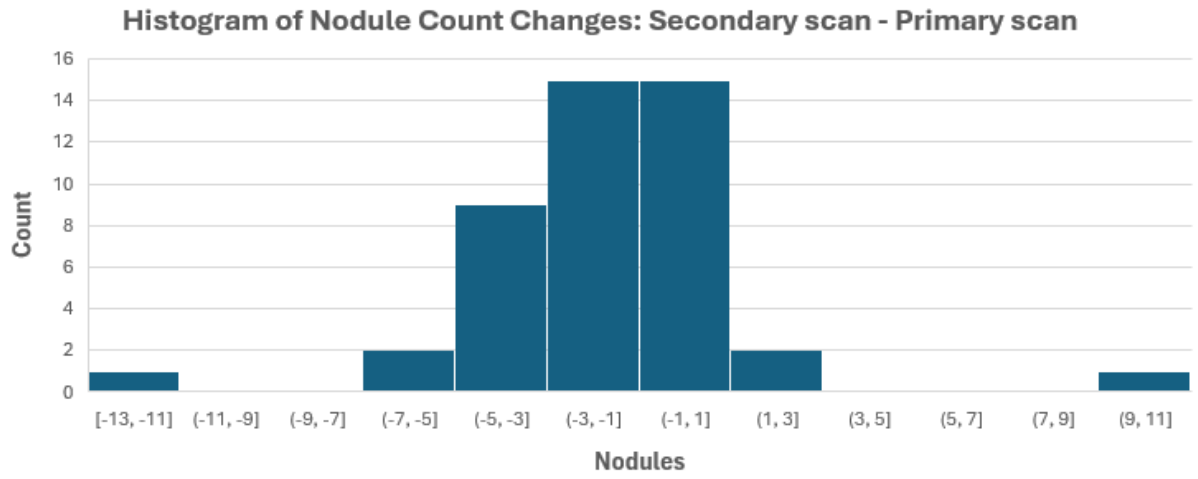
Lessons learned:

- 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.

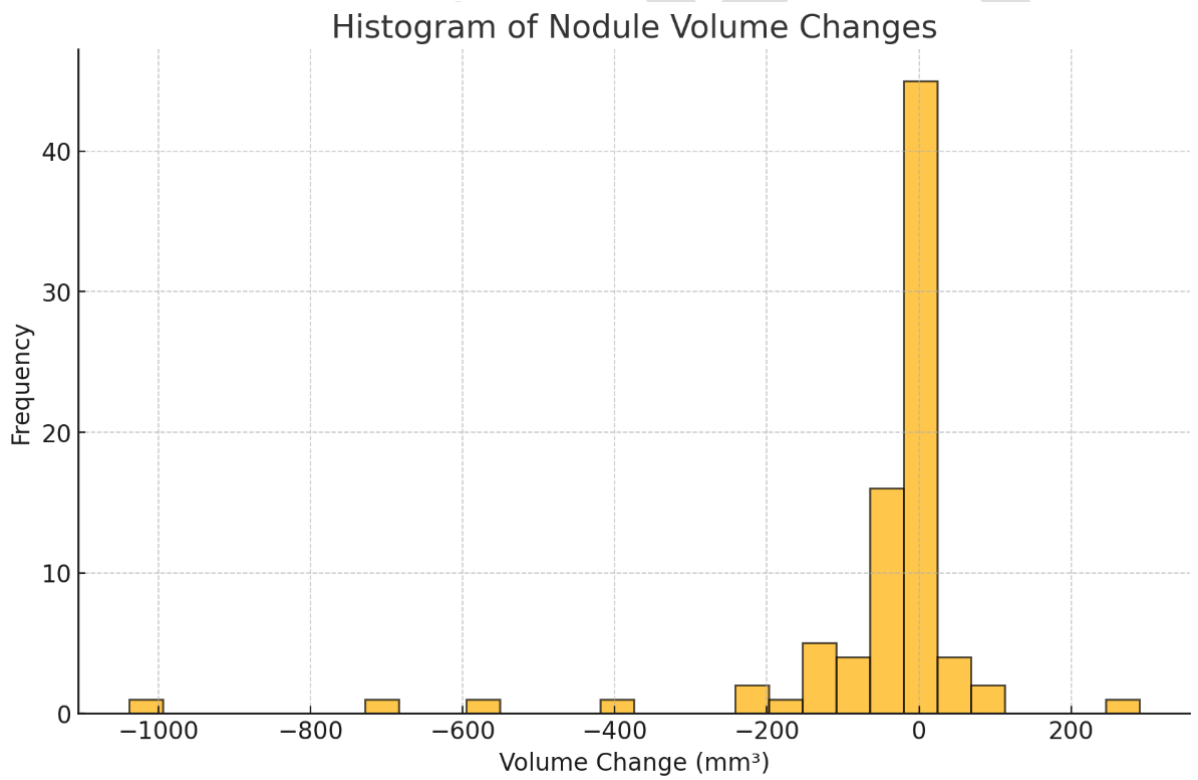
Case Study 2 – Impact of AI software update on lung nodule measurement

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 to 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.



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 post-update is critical to restore safe and effective performance.

- **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-market surveillance is non-negotiable**
  These examples underscore the broader need for formalised, ongoing post-market surveillance 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, Dymiter JA, Yearsley J, Dinneen D, Boyle M, Harrison DJ, Anderson LA, Lip G; iCAIRD Radiology Collaboration. 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.