Medicines & Healthcare products
Regulatory Agency

AI Airlock Sandbox Pilot Programme Report

Published by the Medicines and Healthcare products Regulatory Agency (MHRA)

Published October 2025

Contents

Executive Summary

The AI Airlock is the UK's first regulatory sandbox for Artificial Intelligence as a Medical Device (AIaMD), developed and delivered in partnership with the NHS and the Department of Health and Social Care (DHSC). The pilot phase was launched in May 2024 and ran through to March 2025, aiming to explore the regulatory challenges faced by real-world AI medical devices and generate actionable insights to support the development of future guidance and regulatory changes. Key insights and recommendations generated through the AI Airlock will directly inform and support the work of the National AI Commission, ensuring that its outputs are grounded in real-world evidence and regulatory experience.

Building on emerging global best practice and learnings from already well-established regulatory sandboxes in other sectors, we employed the sandbox mechanism to establish a safe space for medical device manufacturers to work directly with us. Manufacturers also worked with the programme's unique expert stakeholder network to design and execute bespoke testing plans to address current regulatory challenges associated with their AIaMD product. Regulatory sandboxes, like the AI Airlock, take regulators into uncertain areas of regulation. These represent the most contemporaneous regulatory problems that need urgent exploration, which often cannot be explored within normal business processes.

Applications and regulatory testing approach

Candidates were invited to join the AI Airlock via a public call for applications. The submissions received during this call were diverse: 31% of applications used generative AI, 23% used machine learning, 13% predictive AI, and 10% multimodal systems. The most common clinical use cases were diagnostics (25%), screening/imaging (19%), and clinical decision support (19%). The clinical conditions spanned oncology, women's health, respiratory care, and clinical workflow optimisation. Of the 40 applications received, 5 were taken forward to participate in the pilot and 4 successfully completed the full pilot.

The AI Airlock testing approach was built around a four-stage delivery model:

1. **Orientation:** Situation assessment to understand the pilot technologies and relevant regulatory context

2. **Plan:** Regulatory gap review and test plan development

3. **Test:** Testing conducted in different environments, including simulation, virtual, real-world. Concurrently, in-depth regulatory gap analysis was conducted.

4. **Review:** Synthesis of findings and recommendations.

The test plans employed 3 testing environments: Simulation, Virtual and Real-World Airlocks. Final outputs from each project included test plans, results from testing from regulatory gap analyses.

Testing environments

Simulation workshops involved experts across different disciplines: regulators, clinicians, technical experts, academics, legal representatives, and others. When testing in a virtual setting, candidates used a digital testing space built into or around an AI system, such as an AI model on a computer. Here, the AI functions were explored safely, allowing testers to interact with the AI, feeding it information or questions. They would then check how it responded and inspect how its internal processes worked. This mimicked real tasks, so teams could assess the accuracy, reliability, and reasoning within the AI itself. The real-world environment allowed testing in the setting where the AI product would be used (e.g. a hospital). The AI was used alongside doctors and real (anonymised) patient data, following the usual clinical processes, but its outputs were not used to make decisions about patients. Instead, the AI was observed in practice: how it fit into workflows, how accurate or helpful it was, and how safe and reliable it was for standard use.

Regulatory gap analysis

Our existing regulatory frameworks provide a strong foundation for ensuring the safety and effectiveness of medical technologies. However, AI introduces new characteristics such as continuous learning, complex data dependencies, and adaptive behaviour that challenge how some of these principles are applied in practice. That means parts of our framework will remain highly relevant, while others may need to evolve or be complemented by new approaches to ensure patient safety and public trust.

Regulatory gap analyses are difficult to undertake; such analyses performed in such a novel field are particularly challenging. The team enlisted experts from various sectors in an initial regulatory mapping activity. This resulted in a collation of relevant regulatory documents and a list of associated challenges. To account for limited resource and programme timelines, we prioritised 5-10 most relevant regulatory documents for in-depth review per regulatory challenge.

This foundational exercise proved extremely productive. By showing us how broad the term “regulatory gaps” could be in the context of novel AI devices, we could account for this in our definitions. We therefore used an equally broad definition to capture the gaps. This allowed us to capture wording gaps in existing guidance, context gaps in frameworks and even legislative loopholes.

Within the AI Airlock, therefore, the term ‘gap’ is used as a shorthand term for identifying any areas of improvement. Recognising that many of the existing parts of the regulatory system

must cater for the millions of hardware and software products used in the health and care system. It is the objective of the AI Airlock to identify these areas for improvement and recommend solutions.

Regulatory gap key insights and recommendations

The AI Airlock pilot identified regulatory gaps that currently challenge the safe and effective deployment of AIaMD. While the nature of these gaps varied across use cases, some recurring themes emerged across all projects, particularly in areas of risk management, validation of text-based data from large language models (LLMs), AI errors, inaccuracies and non-determinism, explainability, and post-market surveillance. These gaps and challenges were explored in collaboration with four innovators: Philips Healthcare, AutoMedica, OncoFlow and Newton's Tree.

Synthetic Data and LLMs

Current Medical Device Regulations (MDR) 2002 and associated guidance do not define if / how / where synthetic data can be used to provide evidence of compliance. In addition, there is no guidance on quality or validation requirements for text-based synthetic data, despite the growing use of this data type in AIaMD. Standards such as BS/AAMI 34971 acknowledge synthetic data as a potential input in AI model development and note the importance of managing privacy risks, though specific guidance for its use for privacy preservation is limited. During the pilot, we worked closely with the programme investigating the Synthetic data for development of AIaMD and therefore our recommendations are aligned with the findings of that report published after the AI Airlock pilot closure. The report addresses the AI Airlock recommendation for the MHRA to develop best practices for synthetic data generation, validation and assessment for fidelity and representativeness – particularly where the synthetic data are used to support regulatory submission in AIaMD development.

In collaboration with Philips Healthcare, the AI Airlock explored the generation and assessment of 4,000 synthetic radiology reports which were generated using an LLM. While evaluation methods, including automated and human reviewers, showed promise in assessing quality, they also highlighted key regulatory uncertainties. Notably, the data from using an LLM to judge outputs from another LLM raised concerns about the errors and biases being reinforced rather than independently checked. This risk of 'circularity' means the system may not be fully reliable. This highlights the need for independent validation standards and more regulatory clarity. Given the limited timeframes of the AI Airlock pilot and complexity of the issues, further investigation is needed into AI validation of AI. The MHRA should therefore continue to investigate the use of LLMs in synthetic data pipelines (and AI evaluation of AI-generated data) with a focus on building an evidence base that can inform

future guidance on quality assurance, validation methods, and the acceptability of such data in regulatory contexts.

Hallucinations, Non-Determinism and Risk Management

The risk management framework underpinning the MDR 2002, and standards such as ISO 14971, ISO 13485 and IEC 62304, offers a robust foundation for general medical device regulation. However, these frameworks were not designed to address the unpredictable and dynamic nature of AI systems. There is a lack of consensus on how to account for AI-specific risks such as hallucinations, non-determinism or model drift within existing risk management approaches. These risks could be viewed as falling within a manufacturer's performance claims; however, outputs incorporated into the medical record are also subject to the Health and Social Care Act 2008, regulation 17. This creates a tension between different regulatory requirements, where clarity is needed on how performance considerations intersect with obligations relating to the accuracy of health records.

The AI Airlock worked with AutoMedica and their product 'SmartGuideline' and explored how Retrieval Augmented Generation (RAG), could be used to ground LLM outputs in verified sources to reduce such risks. By anchoring responses in trusted clinical content (NICE guidelines), RAG added a safety layer to the system. In an experiment where 436 medical questions were asked, SmartGuideline (with RAG enabled) produced no hallucinations compared to 23 hallucinations from a baseline model. The RAG-based model also demonstrated consistent, deterministic outputs – returning highly similar results across repeated prompts. These results show how targeted AI design strategies can mitigate known risks. Through evidence from virtual testing and the simulation workshop discussions, the AI Airlock programme recommend that the MHRA should explore how such techniques could be accommodated within existing risk management frameworks. Safety measures such as retrieval augmented generation (RAG) should be encouraged in guidance, given the demonstrated ability to reduce hallucinations and support safer more explainable outputs. Guidance should also emphasise the importance of PMS, including mechanisms like the Yellow Card Scheme and Manufacturer's Online Reporting Environment (MORE) as hallucinations may be rare during pre-market testing.

Explainability and Transparency

Terms such as explainability, transparency, and interpretability are not considered in the MDR 2002 and remain under-specified in current regulatory guidance. While standards such as ISO/IEC 22989 and BS/AAMI 34971 reference transparency, they do not outline specific explainability expectations for different user groups – such as regulators, clinicians, or patients. This lack of clarity is particularly important for AIaMD, which, unlike traditional medical devices, may adapt to data or clinical use over time. Such systems carry the risk of

model drift or unanticipated behaviour in situ, making explainability a mechanism to support trust, oversight, and safe integration into clinical practice.

Together with OncoFlow, the AI Airlock explored multiple explainability approaches, interpretability and transparency. This work included plain-language explainability statements, model rationale outputs, and quantitative techniques for feature attribution. The project also tested how explainability requirements differ across stakeholder groups and use cases. While these methods showed promise when discussed with a range of stakeholders during a simulation workshop, stakeholders agreed that their application would benefit from clearer regulatory direction. This is particularly relevant for how explainability should be documented, communicated, and tailored to different audiences. The MHRA could usefully set out high-level expectations or best practices for explainability in AIaMD, supported by examples of appropriate methods and documentation formats tailored to different stakeholder needs.

Post-Market Surveillance (PMS) and Real-Time Monitoring

The PMS requirements manufacturers must meet are outlined in legislation including the Medical Devices (Post-market Surveillance Requirements) which amends the Medical Devices Regulations, 2002. Although this legislation came into force after the closure of the AI Airlock pilot, the workshop was based on the incoming requirements that are now in effect. These regulations are enacted by the MHRA via a vigilance reporting system, which includes the Manufacturer's Online Reporting Environment (MORE). Additionally, the Yellow Card Scheme is available for medical device safety reporting more orientated towards reports from end users and patients.

The sandbox confirmed that for PMS, systems such as Yellow Card and MORE offer reporting pathways and oversight of corrective action approaches. These are, however, not used enough by end users and manufacturers, and there is a clear need for AI-specific guidance on ongoing monitoring and preventative action. Risks such as model drift and automation bias are often only detected after harm has occurred, if they are detected at all. The FAMOS project demonstrated that real-time monitoring dashboards can help to surface emerging risks earlier, particularly around user behaviour. During testing, 180 reports were reviewed at two hospital sites and continuous real-time monitoring was used to detect data quality issues, model drift and automation bias. At one hospital site, the pilot identified a pattern of over-reliance on the AIaMD. If unaddressed, such over-reliance could increase the risk of harm. In this case, contributing factors included clinician fatigue, time pressure, distraction, and experience level. This reinforces the need for guidance on threshold setting, risk flagging protocols, and shared responsibilities for monitoring AI-human interaction. The MHRA should consider publishing supplementary guidance that supports proactive safety measures for AIaMD. This could include expectations on roles and responsibilities for monitoring across different settings.

Standards like BS EN 62366-1 and guidance on applying human factors to medical devices, cover usability but do not fully address AI-specific challenges such as explainability-driven over-reliance, deskilling, or automation bias. The pilot highlighted how clinician context, i.e., fatigue, time pressure, and experience, can affect safety outcomes, and that system design must account for these variables. Cross-sector collaboration is needed to ensure current guidance, training and support for users adequately address these challenges, particularly in high-risk environments.

Next steps

The AI Airlock pilot has underscored both the potential and the practical challenges of developing and deploying AIaMD. This work has culminated in recommendations for updates to policy and guidance, some of which have been outlined in this report. By taking these forwards, the MHRA can continue to bridge the gap between broad regulatory principles and the practical needs of deploying AI safely and effectively in healthcare. Clear, accessible guidance plays a crucial role in supporting innovation, safeguarding patient safety, and building public trust in AI-driven medical technologies.

Next, these findings will inform the work of the National AI Commission, ensuring that its outputs are grounded in real-world evidence and regulatory experience. In addition, the AI Airlock recommendations will inform AIaMD guidance currently in development, and future MHRA regulatory policy including those outlined in the Software and AI as a Medical Device Change Programme roadmap. Areas requiring further exploration may be taken forward by the National Commission and MHRA partners or through the next phase of the AI Airlock programme.

1. Introduction and Context

Background

The independent investigation of the NHS in England highlights the critical state of health services. It revealed significant challenges including surging waiting lists, financial constraints, and a deteriorating national health profile. Emphasising the importance of innovation in medicines and medical devices to solve some of the healthcare system's problems and highlighting the potential of software and AI as a Medical Device (AIaMD) to enhance patient care. Only if these products are of high quality, safe, reliable, and robust.

In the UK, Software as a Medical Device (SaMD) is regulated under the Medical Devices Regulations 2002, AIaMD is a subset of SaMD under IMDRF. Our reform of the medical device legislation, including the Software and AI as a Medical Device Change Programme, is driven by innovation in the life sciences sector, learnings from the Cumberlege review and a growing understanding of AIaMD products. It aims to deliver balanced regulation which ensures products meet the necessary standards for patient safety and efficacy, without hindering or blocking the path for innovation and growth. balanced regulation which ensures products meet the necessary standards for patient safety and efficacy, without hindering or blocking the path for innovation and growth.

After all, AIaMDs have the potential to help solve a broad range of challenges facing the NHS. In radiology, for example, AIaMDs could analyse x-rays or scans for abnormalities and make recommendations, saving clinician time. In surgery, AI-powered robotic systems could assist in minimally invasive procedures by analysing real-time imaging to guide surgical instruments with sub-millimetre precision, reducing operating times and complications. Meanwhile, AI-enabled wearable devices could continuously analyse physiological data, detect early signs of disease exacerbation, and provide personalised alerts to patients and clinicians, supporting chronic disease management and potentially reducing the need for frequent laboratory tests.

These innovative products also hold unique risk profiles. We, alongside AI developers and the government, must protect UK citizens from the most significant risks presented by AI. Fostering public trust in the technology and particularly considering the interests of marginalised groups. To ensure a safe, efficient, and robust route to market for these innovative products, key regulatory challenges need to be identified and addressed systematically.

The Government's AI Opportunities Action Plan sets a clear priority to ensure AI development through regulation, safety, and assurance hopes to fuel a fast, wide, and safe deployment and adoption of AI. Regulatory frameworks must support the development and implementation of these innovative solutions. One of the best ways to deliver safe and effective regulatory transformation is through regulatory sandboxes.

The regulatory sandbox model is an internationally recognised mechanism to help address novel regulatory challenges in many sectors. These include financial services, data and information and healthcare sectors. This mechanism tests innovative products, business

services and/or frameworks, often in a virtual environment separate from the market or sector they are influencing, ensuring safe and responsible product testing at pace.

Regulatory sandboxes are increasingly used in the UK. Sandboxes can be utilised by a wide range of companies, from early start up developers to large scale organisations. Some notable examples include the NHS England test beds programme and the CQC's using machine learning in diagnostic services. The Information Commissioner's Office offers a regulatory sandbox service for products that utilise personal data in innovative ways, and the Financial Conduct Authority works with firms that want to test products with real consumers in a controlled environment.

Sandbox offerings vary. Operational sandboxes can offer access to virtual controlled environments and data sets or alternative routes to market. Regulatory sandboxes bring together regulators and organisations to understand regulatory gaps and address challenges. The Data Institute published a comprehensive report analysing the use of regulatory sandboxes in the complex data space. The report highlights how sandboxes can reduce regulatory uncertainty and build capability. Further highlighting the opportunities in emerging, cross-border sandboxes, and notes the high resource requirements and difficulties to scale.

Regulatory sandboxes provide the opportunity to address challenges often experienced by innovations. Naturally, applying an existing regulatory framework to innovative products, developers with lower regulatory knowledge, regulators with lower product knowledge, and general inexperience regarding the life cycle of innovation makes regulation and product development challenging. Regulatory sandboxes are well placed to drive the development of well-designed regulation that is proportionate, agile and offers clarity to developers and users, a key priority we share with the UK government. The AI Airlock regulatory sandbox has been developed to address the regulatory challenges and promote a pro-innovation approach to learning and experimentation as outlined in the Pro-innovation regulation of technology review .

AI Airlock

In May 2024, we announced the AI Airlock to test real world products and prototypes of innovative AI medical devices, in collaboration with Team AB (a coalition of UK Approved Bodies), the NHS AI Lab and Department of Health and Social Care (DHSC).

Team AB brings expertise from Approved Bodies to increase consistency of interpretation of GB medical device regulatory requirements. It was integral to collaborate with Team AB on the AI Airlock project to uncover and navigate medical device regulatory challenges for AIaMD using our collective regulatory knowledge to inform and standardise policy positions in this rapidly evolving area.

The NHS AI Lab provides connections and expertise within our healthcare system. Many AIaMD products are deployed via NHS infrastructure making NHS England and the health services in the devolved nations crucial to regulatory discussions such as deployment and post market surveillance.

This pioneering and unique work helps us to further understand the detailed challenges of regulating AIaMD products to ensure new regulations and guidance facilitate their rapid and safe deployment into the NHS. AI Airlock helps us to support UK innovators achieve regulatory compliance.

Purpose of this report

This report provides the outcomes from the AI Airlock pilot programme that ran between April 2024 – March 2025. It will offer insights for key audiences across the UK healthcare, regulatory and digital technology sectors and international partners. Learnings from the pilot fall into 2 main categories:

a) how to implement regulatory sandboxes to investigate and improve regulation

b) suggested changes to the current regulatory framework, both for the MHRA and stakeholders across the healthcare system

2. Programme Design and Governance

The pilot had 4 key objectives for its first year:

- to work with real world products or prototypes to evidence and investigate regulatory challenges with AIaMD
- to use these case studies to generate recommendations for regulatory framework changes
- to carry out a formal programme evaluation and learn lessons from the regulatory sandbox way of working
- to secure future year funding streams

To achieve these objectives, the dedicated MHRA programme team was created with two main work streams: operational programme management [governance, stakeholder management, risk management] and regulatory / technical case management

Sandbox Design

The AI Airlock pilot programme was intentionally designed to provide opportunities to explore all regulatory challenges for AIaMD, therefore needed to be agile enough to be able to work with candidates and products at any stages of their developments.

The AI Airlock focused primarily on the GB medical devices regulatory framework. This framework of legislation and guidance is within the MHRA remit to amend and therefore it was a key priority for us to focus the sandbox on this key area where action could be taken rapidly.

The MHRA ran a call for applications to the public, industry and academia for AIaMD developers to submit proposals to join the AI Airlock, with key eligibility criteria and testing environments used to design the pilot cohort.

Eligibility Criteria

Applicants needed to be aiming for the product to be a medical device, as defined by the MDR 2002, which utilised AI or Machine Learning. They needed to be a legal entity with the rights to market their product in the UK. Most critically, there needed to be a commitment to working with us throughout the sandbox testing. All applicants needed to meet these conditions for their application to be eligible.

Further criteria were then applied to the applications to shortlist the most eligible applicants. These criteria were:

- The product has potential to deliver benefits for patients
- The product or prototype is innovative or a novel application
- The product presents a regulatory challenge
- The proposal is ready to be trialled

These criteria ensured that the candidates selected were most able to deliver results through the pilot programme rapidly, an important consideration given the condensed timelines for testing.

Sandbox testing environments

As a result of the broad testing opportunity, we established three key testing environments to be used in the sandbox.

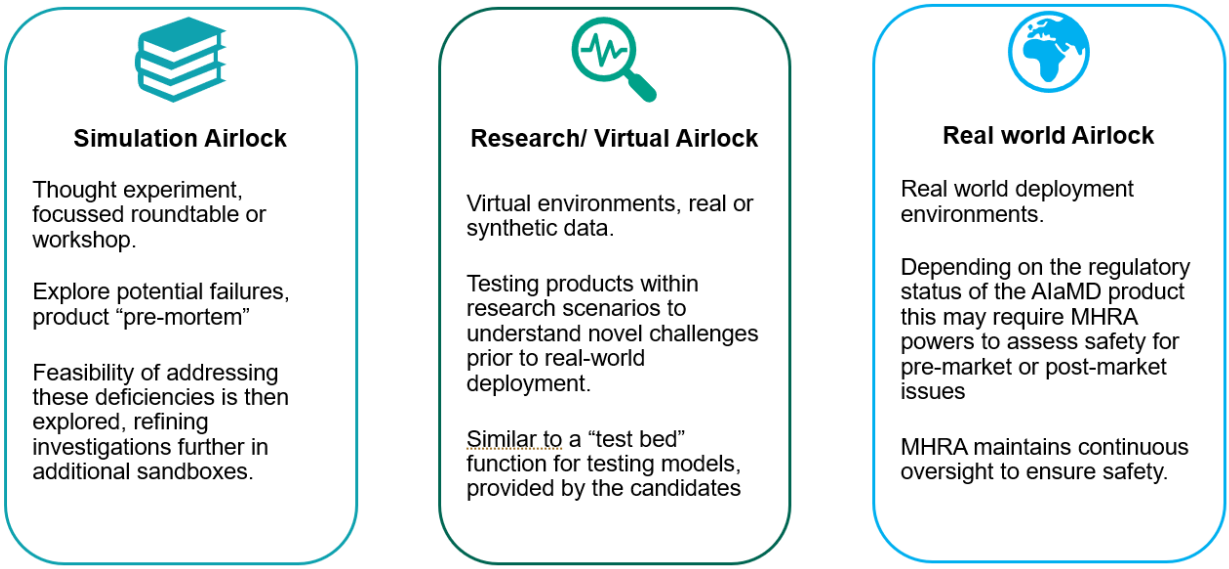

Figure 1 Overview of the three testing environments designed for the AI Airlock pilot: Simulation, Research/Virtual and Real World

Simulation Airlock is a workshop or roundtable-based activity used to address focused questions and gather various perspectives on a challenge. This type of testing can be tailored specifically to a set of key challenge areas. During the pilot the simulation workshops begun with a short presentation to set the scene and a facilitate discussion was held along with some focused breakout sessions. We looked across our broad stakeholder network and brought together a unique cast list of colleagues to bring their unique perspectives to each problem area. This type of testing environment is very flexible and can be applied to challenges experienced by products maybe at concept or early development stage or those without access or the need for data or a testing environment.

Virtual Airlock is testing a model with real or synthetic data to address research scenarios to gather insights of how the model works. This mechanism is like a “test bed” function used to validate and verify AI models and products using specific data access through the testbed. During the pilot AI Airlock these virtual environments were provided by the candidates.

Real world environment involves deploying a product in a hospital or trust without impacting patient care, that way it protects clinical pathways while gathering evidence from clinicians or hospital deployment. It is imperative that the products testing in the real-world sandbox are done so with appropriate oversight in order to protect patient safety and clinical pathways. The real world environment is not a clinical investigation.

Each of these testing environments have been tested as part of the pilot programme and will continue to be refined throughout subsequent phases of the AI Airlock.

Programme Governance

To support the AI Airlock pilot programme a new governance structure was established that included a Governance Board and Supervisory Committee. Each pilot candidate, upon onboarding, formed a project team with a dedicated case manager from the MHRA Airlock team. The project teams reported progress against the delivery plan, key technical information and regulatory challenges, via their case manager, to the Supervisory Committee and Governance Board.

- **Governance Board**: ensure that a coherent strategic approach is adopted to manage successful delivery to time, cost and sufficient quality, of the AI Airlock programme outputs. Throughout the design and testing phases of the programme the Governance Board also provide a wealth of expert input to the AI Airlock technical case studies.

- **Supervisory Committee**: oversee the operational and technical activity of the AI Airlock programme by providing an expert advisory function to the AI Airlock programme and the project teams within the cohort. Provide regulatory and scientific advice to the AI Airlock project teams in relation to specific regulatory sandbox testing plans, through active engagement at the Supervisory Committee meeting and correspondence from the AI Airlock technical team.

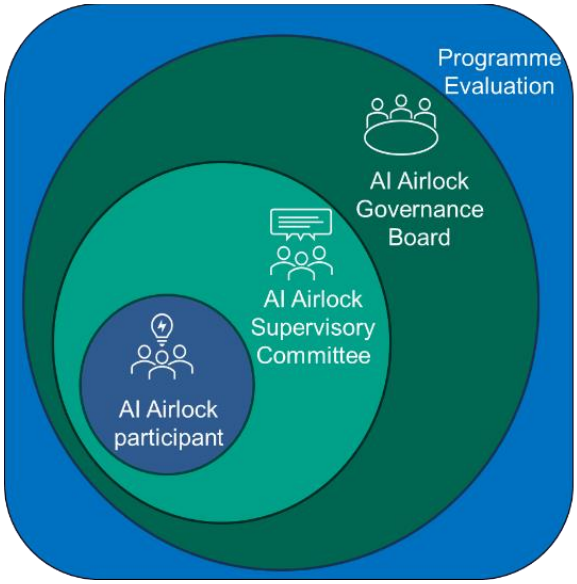

Figure 2 AI Airlock Governance Structure

All members of these groups were bound by signed confidentiality agreements and declarations of interest which increased assurance to the candidates and MHRA of the integrity of the sandbox governance. The core team, governance groups and key partners enabled multidisciplinary expertise and collaboration from across the MHRA, government departments, devolved administrations, Approved Bodies and the wider healthcare sector.

Timelines and Milestones

The pilot programme ran in the 2024-2025 financial year. It was publicly announced that the MHRA would be developing the regulatory sandbox in May 2024 and work began to rapidly design and build the sandbox in parallel to building the team.

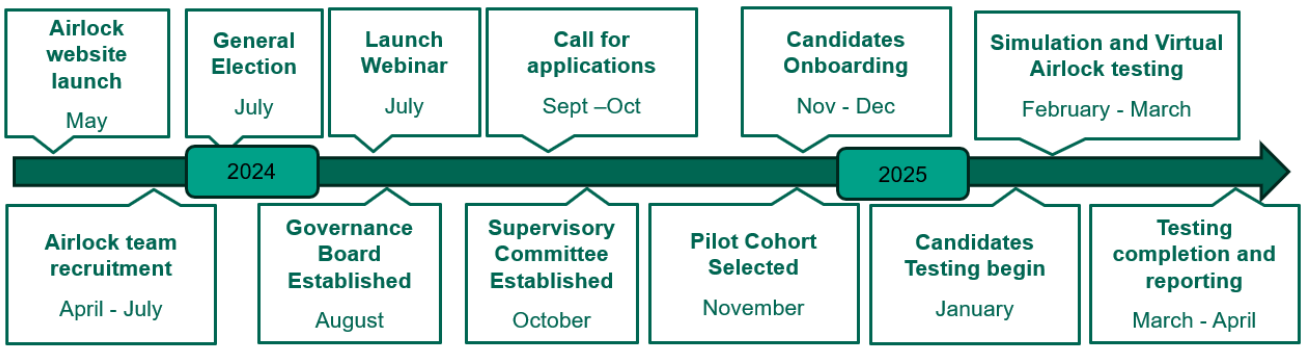

Figure 3 Timeline of the AI Airlock pilot activity

The design of the sandbox involved a comprehensive literature review of current and historic regulatory sandboxes. A network of stakeholders and experts from the AI technology and healthcare sectors alongside other government departments was established. Following the UK general election the AI Airlock held its first public webinar in July and shortly after, established its Governance Board and Supervisory Committee. Throughout Autumn 2024, we opened the call for applications for industry to join the sandbox, ran an intense and

comprehensive sifting and shortlisting mechanism and in November we designated the first cohort of candidates.

Candidates were brought onboard into the AI Airlock, after signing non-legally binding terms of engagement and dedicated case managers were assigned. The teams then worked together to design the sandbox testing plan. Throughout the winter, the project teams met at least weekly, to discuss the progression and developments of the testing plans in the Simulation, Virtual and Real-World testing environments. In March 2025 the pilot testing formally concluded, and candidates worked rapidly to produce their case study reports.

The following chapters of this report will provide additional details behind the AI Airlock, candidate selection, regulatory sandbox testing methodologies, insights from the case studies including evaluation of lessons learned.

3. Candidates and Selection

Overview of applications

The call for applications to the AI Airlock was open from 23 September to 7 October 2024. Approximately 40 applications were received, assessed and underwent a structured sifting process. Of the products applying:

- 31% used generative AI, 23% used machine learning, 13% used predictive AI
- The most common clinical uses were diagnostics (25%), screening or imaging (19%), and clinical decision support (19%).
- Clinical areas included oncology (15%), clinical workflows (12%), women's health (9%), chronic respiratory disease (9%), and health monitoring (9%).

Reflecting the complexity and evolving nature of the regulatory landscape for AI technologies, applicants identified regulatory challenges across the full product lifecycle, with most applications highlighting more than one area of uncertainty. The most cited challenge was borderline classification. Determining whether the product's intended purpose and functionality would qualify as a medical device was specifically challenging. Other areas of focus included risk classification, product training and validation, bias, explainability, reliability, post-market surveillance, and change control. Notably, most applications came from the micro and small SME sector, reflecting a strong presence of early-stage innovators, or a greater willingness within this group to engage with regulatory challenges. Fewer applications were received from large enterprises and academic institutions, perhaps due to the more complex and time-consuming internal approval processes typically required in these institutions exceeding the short application window.

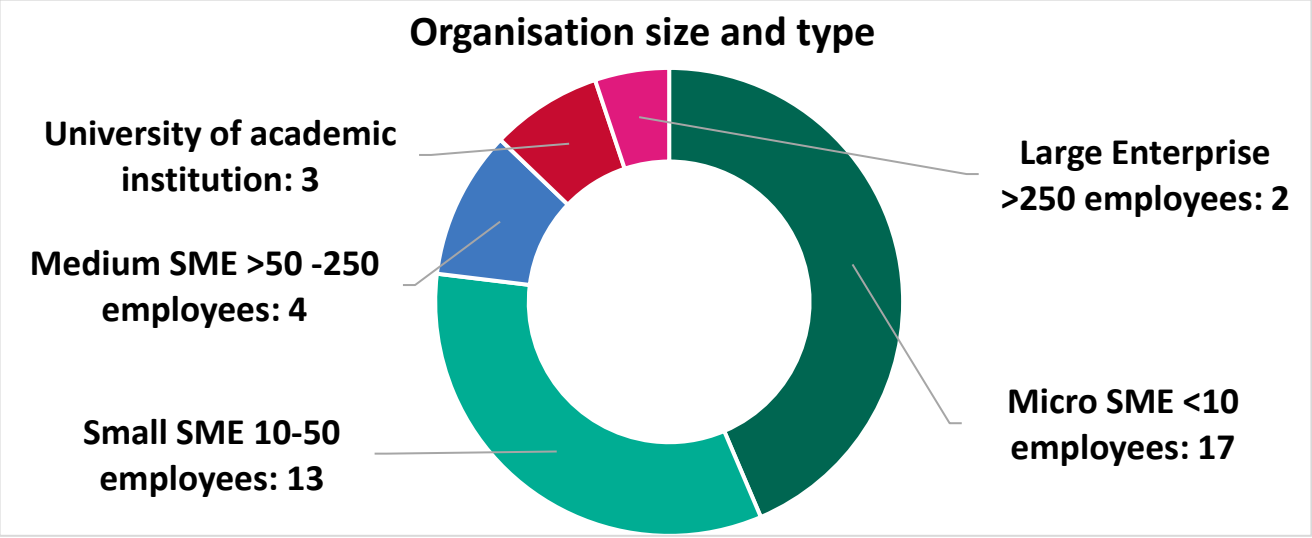

Figure 4. Applications by organisation size and type. Around 44% of applications were submitted by micro-SMEs with fewer than 10 employees, 33% by small SMEs with 10-50 employees, and the remaining 23% were split between medium and large enterprises and academia.

Application review and sifting

Each application was assigned to two reviewers from within the AI Airlock programme team: a lead reviewer and a second reviewer. The lead reviewer assessed the application against the eligibility criteria outlined in the application form, while the second reviewer conducted an independent review using the same criteria. If both reviewers agreed that an application met the required standards, it progressed to the next stage. If both reviewers agreed it did not, the application was deemed unsuccessful. Where there was disagreement between the reviewers the application was discussed in a group setting to reach a decision. This could result in the candidate being invited for a follow-up interview or being informed that their application was unsuccessful.

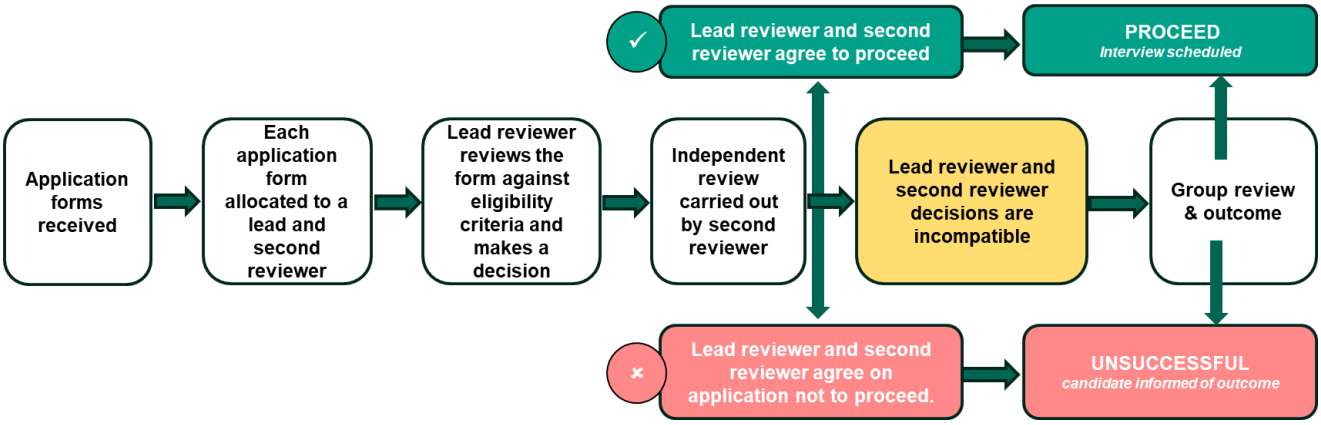

Figure 5. Application review and sifting. Each application was reviewed by members of the AI Airlock team and followed the same decision-making process which culminated in progression to interview or an unsuccessful application.

Aside from the eligibility criteria, other mandatory success criteria were measured. Applicants needed access to their own data and testing environment (if utilising a virtual or real-world Airlock), and the product had to be ready for testing. Reviewers also considered the focus of the regulatory challenge to be explored, the type of technology proposed, and the level of resource commitment available from both the applicant and the Airlock team. Through this process, a shortlist of nine candidates was selected.

A final pilot cohort of five candidates were chosen following interviews with the shortlisted candidates, considering deeper insight into their technologies, applicant readiness and resource availability. This was reduced to four candidates following the withdrawal of one participant after the organisation went into administration (an insolvency process).

The pilot cohort regulatory challenge areas spanned the product lifecycle. The Airlock team selected challenges related to data generation for training and validation, AI hallucinations, AI explainability, and post-market surveillance. While the majority (three out of four), had initially expressed interest in also exploring the borderline qualification question, time and resource constraints limited the feasibility of addressing such a broad spectrum of

challenges. The Airlock team, aware that other regulatory partners would be exploring these areas in greater depth, decided to prioritise the novel and unexplored challenges.

Table 1. Overview of candidates, regulatory challenges and key objectives to be explored during the pilot programme

Candidate & Product Details	Regulatory challenge	Key Objectives
PACS Radiology AutoImpression Philips Medical Systems Large language model Radiology	• Validation of synthetic datasets generated by an LLM • Use of LLMs to validate synthetic data	1. Understand the parameters that define good quality synthetic data 2. Challenges involved in using an LLM to generate synthetic data 3. Use of LLM to validate synthetic data
SmartGuideline by AutoMedica Large language model Clinical workflows	• Reducing hallucinations and non-determinism in LLMs	4. Evaluate LLM technology to establish how retrieval augmented generation (RAG) performs in reducing hallucinations 5. Explore how to enhance quality of recall
OncoFlow by OncoFlow Large language model Oncology	• Assessing approaches to LLM explainability and any trade-off with clinical performance	6. Compare clinical performance, and explainability of traditional NLP models vs LLMs on unstructured oncology data 7. Explore methods to make AI models explainable
Federated AI Monitoring Service (FAMOS) by Newtons Tree Virtual Environment Radiology	• Improving AIaMD safety through real-time monitoring of key metrics (model drift, data quality and automation bias)	8. Assess the effectiveness of a real-time monitoring system in enhancing the monitoring of third-party AI medical devices to improve their safety 9. Understand reporting responsibilities of different stakeholders when a trend that may lead to harm is identified

4. Regulatory Challenges and Projects

This section of the pilot programme report provides an in-depth look at the approach used to address the selected regulatory challenges within the AI Airlock. It is followed by individual case studies – Philips AutoImpression, AutoMedica SmartGuideline, OncoFlow, and Newton's Tree FAMOS. Each case study outlines the specific regulatory challenge explored, the underlying hypotheses and objectives, the testing methodologies used, and key findings from testing in virtual or real-world hospital settings.

Overview of regulatory challenges and technologies

Approach and activities overview

The AI Airlock projects followed a structured four-step delivery approach.

1. **Orientation:** Situation assessment to understand the pilot technologies and relevant regulatory context
2. **Plan:** Regulatory gap review and test plan development
3. **Test:** Testing through simulation, virtual, real-world, or hybrid environments; deeper regulatory gap analysis conducted
4. **Review:** Synthesis of findings and recommendations

In the first step, the AI Airlock team undertook a situation assessment to understand the candidate technologies. Together, the AI Airlock and candidate teams (referred to hereafter as project teams) collaboratively examined the current regulatory landscape. This review drew on regulatory documents identified by the project teams, and supervisory stakeholders, particularly Team AB (Team Approved Body). Once the relevant documents were selected, the project teams began selecting the most important ones. At the same time, testing plans were co-developed to address each project's regulatory challenges. These plans defined key hypotheses, objectives, methodologies, timelines, and anticipated risks.

Following agreement on the testing plans, projects entered the testing phase. Most were conducted through a combination of simulation, virtual and real-world Airlocks. While testing was underway, the AI Airlock team continued prioritising regulatory documents, followed by a detailed review of the prioritised guidance to identify areas requiring improvement for AIaMD regulation. In parallel, the AI Airlock team ran simulation workshops with cross-disciplinary experts. Insights from these activities were produced to inform the key findings and recommendations of the AI Airlock programme. A final meeting with all project teams marked the close of the delivery phase, followed by the completion of final reports.

Case Study: Philips Healthcare - AutoImpression and the generation, assessment and validation of synthetic data

DISCLAIMER: Methods,results of testing, verification and validation discussed in this report are for regulatory science exploration purposes only. They DO NOT imply or confer any product deficiency, nor do they imply additional requirements on Philips' processes or products.

Introduction

Radiologists and healthcare professionals face increased workloads due to staff shortages and the growing number of imaging studies. This can lead to burnout, longer wait times for patients, delayed diagnoses, and potentially missed diagnoses, compromising patient safety and outcomes. Auto Impression leverages Large Language Model (LLM) technology to automate the creation of the "Patient Impression" section in radiology reports. This automation aims to reduce errors, minimise omissions, and speed up radiology reporting, ultimately enhancing productivity for radiologists. The functionality is still at an early development stage.

AIaMDs which perform summarisation of patient health data like Auto Impression require testing against a broad spectrum of rich and representative data. Philips aimed to use an LLM to generate synthetic radiology reports to improve representation of demographics with abnormal radiology findings within the dataset. The synthetic data would then be assessed for its suitability in testing an AIaMD, aligning well with the AI Airlock objective to understand what constitutes high quality text-based synthetic data from a regulatory perspective, and what approaches can be used to assess it.

The assessment involved confirming that the data was truly synthetic rather than a direct reproduction of the original real-world or "ground truth" data. Given the large volume of data required, Philips proposed using established automated mathematical methods to carry out most of the assessment, complemented by a small sample of human-reviewed data. Once the reports were confirmed as synthetic, a "Patient Impression" would be generated by an LLM for both the synthetic and real data. These impressions would be evaluated for quality against criteria such as clarity, groundedness, accuracy, and completeness. To manage scale, an LLM would act as the primary assessor with human validation used on a small sample to cross-check the reliability of the LLM's judgement.

Regulatory challenges

Two broad categories of regulatory challenges were considered. The first relates to describing regulatory requirements for GenAI-enabled devices. The second, for those devices that do fall under regulatory oversight, involves determining the types of valid scientific evidence needed to evaluate their safety and effectiveness across the Total

Product Lifecycle (TPLC). These challenges are particularly relevant for GenAI-enabled devices that use open-ended input and output formats. These differ from the more structured formats typically found in other AIaMD and may require new evaluation approaches.

Synthetic data may create a more robust dataset and evaluation if accurate and representative. Evaluating that accuracy and representativeness is essential, considering elements including:

- the diversity of the initial clinical dataset (different geographies, modalities, users, etc.)
- the real world or ground truth data that is adequate for generation of synthetic data
- how to measure the quality of synthetic text data given the lack of standardised evaluation metrics/scores
- addressing both clinical similarity/differences and linguistic similarity/differences.

The AI Airlock pilot generated a large volume of data and scoring outputs that were difficult for humans to evaluate manually. The Philips and MHRA team worked on the assumption that these software enhancements are designed not to replace radiologists, but to improve efficiency and support the quality of their work. As such, it was important to strike the right balance between automating the analysis of this data and including human evaluation to ensure accuracy and clinical alignment. The proportion of automated versus human validation should be determined on a case-by-case basis, accounting for the specific use case, the associated risk, and the added value of human input.

Hypotheses

- *There will be no significant difference between synthetic data and real data in the quality of impressions generated in Radiology reports*
- *There will be no significant difference in the errors identified between synthetic and real data using the automated evaluation Evaluation of the performance and safety can be performed using LLMs as Judges.*
- *An LLM Judge will correctly identify errors introduced into real and generated impressions*
- *An LLM Judge will correctly ignore nonclinical counterfactual differences between reports and their generated impressions.*

Methods

Synthetic data generation

Philips used the MIMIC-IV, a large, publicly available database of de-identified health data from over 280,000 patients who stayed in critical care units at the Beth Israel Deaconess Medical Centre between 2008 and 2019. The dataset includes patient measurements, orders, diagnoses, procedures, treatments, and free-text clinical notes. MIMIC-IV is widely used for research and education, helping to reduce barriers to conducting clinical studies.

Philips used an LLM to create new, structured reports based on existing examples to support the generation of synthetic radiology reports. The goal was to maintain the structure and modality and clinical indication of the original reports while introducing new, randomised clinical content. A prompt was developed to generate a new structured report that used the format of the source (MIMIC report), but with different findings. The generated reports excluded introductory or summary text, such as impressions. The model used was configured to encourage diverse outputs. The model was chosen for its advanced language capabilities and suitability for generating clinically relevant synthetic text. The resulting synthetic reports enabled a broader range of testing scenarios while maintaining realism and structural consistency.

Generating synthetic impressions

To complement the generation of synthetic radiology reports, Philips also focused on creating the impression section, a critical component of radiology documentation that summarises the key findings and provides clinical interpretation. Care was taken to ensure that only reports without existing impressions were selected for generation of a synthetic impression. Using tailored prompts, the LLM generated impressions based on either the original findings from the MIMIC-IV dataset or the synthetic reports created earlier. The prompt used for LLM-generated impressions was designed to promote clarity, conciseness, and clinical relevance. It instructed another large language model (with a temperature to encourage more focused and deterministic outputs) to read the full body of the report and write a corresponding impression. The model was prompted to exclude normal findings and instead highlight critical abnormalities, incidental findings, and suggested follow-up actions; elements that improve the clinical utility of the output.

The full reports (including impressions) were processed in batches to ensure scalability and consistency across the dataset. These LLM-generated impressions were designed to summarise the most important content from each report. They were later assessed using another LLM acting as a “judge”, and selected samples were earmarked for further human evaluation within the sandbox environment.

Generating reports with counterfactuals

Philips manually introduced counterfactuals - small, controlled changes - to both original (ground truth) and synthetic reports to evaluate how reliably the LLM-judge distinguished genuine findings from altered or noisy content. Twenty representative samples were selected, and for each one, five counterfactual variants were generated, resulting in a balanced set of altered reports. The goal was not to maintain clinical realism in these modified reports, but to introduce noise in a structured way to test the robustness of the LLM-judge.

Seven types of counterfactuals were used, ranging from subtle character-level edits to changes in key clinical details such as organ, laterality, or gender references. Some counterfactual types relied on specific linguistic features (e.g. anatomical terms or gendered language) and could therefore not be applied equally across all samples. As a result, counterfactuals involving simple textual modifications were more frequent than those requiring specific clinical terminology. Care was taken to apply counterfactuals consistently across both synthetic and original reports despite this variation. This allowed for a fair and controlled comparison of the LLM-judge's behaviour across the two datasets.

Generating reports with errors

Philips manually applied deliberate errors to a consistent set of radiology reports to evaluate the reliability of the LLM-judge. Twenty reports were selected from both the original and synthetic datasets and five different types of clinically relevant modifications were applied, producing a total of 100 altered samples in each dataset.

Error types considered most clinically relevant were focused upon, following consultation with radiologists. These included changes that could meaningfully affect clinical interpretation such as laterality swaps, organ swaps, negations, insertions, deletions etc.

LLMs can produce slightly different responses each time they are run, even when given the same input. To manage this non-determinism and improve the reliability of the LLM-judge, two strategies were used:

Measuring consistency: For each evaluation, the same prompt was run 10 times, and the standard deviation of the results was calculated. This helped quantify how much the model's outputs varied and ensured that the results were stable enough to be trusted.

Prompt refinement: The prompt given to the LLM-judge was reviewed and adjusted to improve performance. By refining the wording and structure of the prompt, the model's responses became more focused and clinically aligned.

These steps were essential to ensure that the LLM-judge could be used reliably within the sandbox setting. More work on mitigating non-determinism can be found in the SmartGuideline case study.

Comparison of real and synthetic reports: clinical similarity

Philips used two complementary approaches to assess the clinical similarity between real and synthetic radiology findings: ICD-10 code comparison and semantic similarity analysis using BioBERT.

ICD-10 codes are internationally standardised alphanumeric codes, used to classify diseases and health conditions in clinical documentation and billing. Maintained by the World Health Organisation, each code begins with a letter representing a broad disease category (e.g. “E” for endocrine disorders), followed by numbers that specify the diagnosis (e.g. “11” for type 2 diabetes), and sometimes a decimal for additional detail (e.g. “.9” for “without complications”). To analyse how real and synthetic findings differ, Amazon Comprehend Medical Service was used to extract ICD-10 codes from both sets of findings. Philips included only codes that Comprehend Medical marked as present in the text (non-negated), and with a high confidence in the match (>0.75).

BioBERT, a deep learning model fine-tuned on biomedical texts, was used to assess clinical similarity beyond coded terms, a deep learning model fine-tuned on biomedical texts. Unlike general-purpose models, BioBERT is designed specifically for medical and life sciences language, allowing it to capture clinical nuance more effectively. It generates vector embeddings, a type of mathematical representation, of each findings section.

Comparison of real and synthetic reports: linguistic similarity & human evaluation

Linguistic similarity analysis was performed to ensure synthetic reports maintained the same linguistic style as original reports. This was evaluated using the following metrics: GLEU, ROUGE, Parts-of-Speech Overlap and Automated Readability Index (ARI) (see appendix for details). To ensure that the BioBERT, ARI and ROUGE scores effectively evaluated the similarity between synthetically generated reports and their original counterparts, a calibration step with human expert input was performed as follows:

1. Data Collection: A diverse subset of 20 original reports and their corresponding synthetically generated versions was collected. The dataset covered various topics, clinical domains, and writing styles to provide a comprehensive evaluation.
2. Initial Human Scoring: A set of 2 clinical experts from the Philips Clinical team, with experience in reading radiology reports (U.S. board certified Radiologists), blindly reviewed each original and synthetic report. They then scored the reports on a Likert scale from 1 to 5 for readability and similarity to verify that the linguistic structure, grammar, and terminologies used (focusing on clinical terms) were similar to ground truth radiology reports.

For each report, the experts scored the following points: Clinical Content Similarity, Linguistic Style Similarity, Understandability compared to the original report. In addition, an independent analysis for each report was carried out. The experts scored the following points: Language, Structure, Completeness, Clinical Soundness (see appendix for details)

Results

ICD-10 Code Comparison

Figure 6 shows the distribution of ICD-10 codes across all real and all synthetic findings. The distributions differ noticeably, with several ICD-10 codes appearing only in the synthetic dataset. This suggests that the underlying pathologies and clinical content of synthetic findings differ from real ones, which aligns with the goal of introducing variation in synthetic data generation.

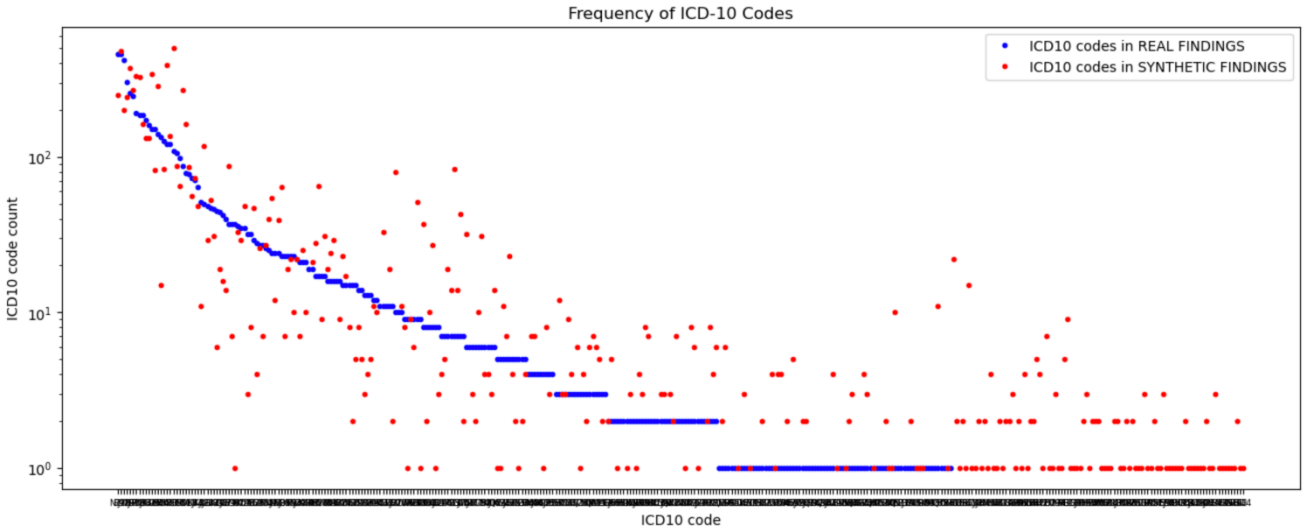

Figure 6. ICD10 Code Distribution Comparison. This chart compares the distribution of ICD-10 codes in real and synthetic findings. The blue dots represent ICD-10 codes from real radiology reports in the MIMIC dataset, while the red dots represent codes from the synthetic dataset. The x-axis shows individual ICD-10 codes (not legible due to density), and the y-axis indicates the frequency of each code across the respective datasets.

Semantic Similarity with BioBERT

Cosine similarity was used, a common measure of how similar two vectors are, to compare the embedding of each real finding with its paired synthetic version. Scores close to 1.0 indicate strong similarity, while scores near 0.0 reflect major differences. Most synthetic reports had similarity scores between 0.6 and 0.8, indicating high, but not identical, clinical similarity. This moderate distance was taken as a reflection of intentional changes introduced through prompting, such as replacing organs or conditions to increase variety. A small number of pairs scored extremely high or low. These were often due to very short reports, or errors in text parsing or generation. Further investigation of this subset could not be investigated due to resource constraints.

Human evaluation results

Once the human evaluation had been carried out, there was a feedback session. One observer mentioned that the task was notably easier when dealing with shorter reports. The same observer suggested that using a three-level scale would simplify the task, as the current five-level scale was challenging and often led them to revisit previous assessments. Additionally, one of the observers noted that they could not distinguish between synthetic and actual reports.

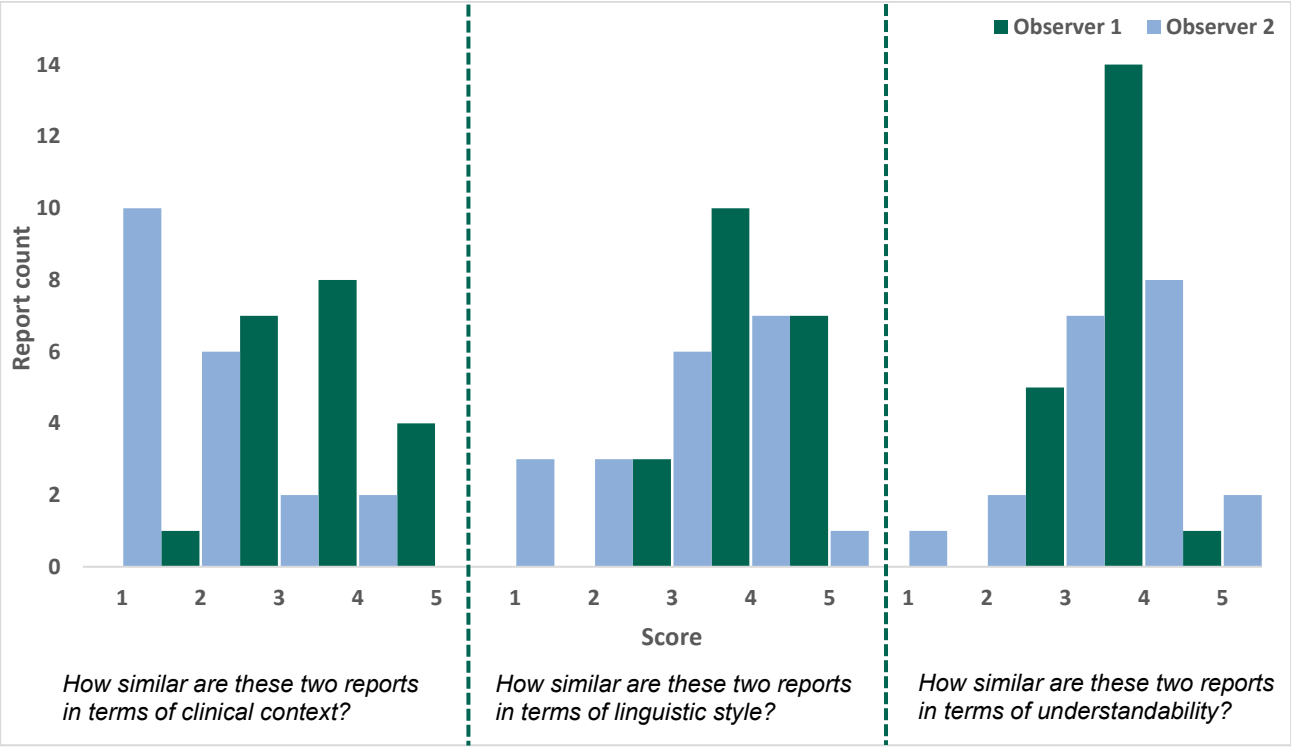

Figure 7. Inter-observer High level Analysis - Frequency distribution of scores for comparative analysis per question.

After further data processing, no significant correlation was found between the grading scores of the two observers, with one exception: a statistically significant negative correlation of -0.52 (95% CI: -0.78 to -0.11) was observed in their assessments of the professionalism of the “Language” used in the original (real-world) MIMIC report. This suggests a notable disagreement between the observers on this criterion.

There are two plausible explanations for these findings:

- underpowered multiple testing – the number of samples assessed may have been too small to detect reliable agreement across multiple dimensions.

- insufficient rater training and alignment – the observers may not have received adequate calibration or guidance on how to apply the scoring criteria consistently prior to completing the task.

The average score was used for the calibration analysis with the mathematical target metrics.

Table 2 - Correlations in between the objective metrics and observer ratings (see appendix for metric descriptions)

Question	Metric	Correlation	Count	Lower C.I.	Upper C.I.
1. How similar are these two reports are in terms of clinical content?	BIOBERT	0.390	20	-0.064	0.710
2. How similar are these two reports are in terms of linguistic Style?	ARI MIMIC Report	-0.342	20	-0.681	0.119
	ARI Synthetic Report	-0.069	20	-0.497	0.385
	Difference ARI score (Synthetic-MIMIC)	-0.315	20	-0.665	0.148
	Average sentence length MIMIC Report	**-0.481**	**20**	**-0.762**	**-0.049**
	Average sentence length Synthetic Report	-0.025	20	-0.462	0.422
3. How similar these two reports are in terms of understandability?	**ROUGE**	**0.444**	**20**	**0.002**	**0.741**
Missing value handling was conducted using pairwise exclusion, and a 95.0% confidence interval was applied. Significance was determined at the 0.05 level, with bold indicating significant correlations.					

The only statistically significant correlation was found between average sentence length in synthetic findings and similarity scores. This indicates that shorter sentences were associated with greater similarity, suggesting that simpler and more concise language may improve alignment with reference data. Question 3 showed a statistically significant correlation with ROUGE scores indicating that reports with similar phrasing have similar understandability (see appendix Table s1).

Question 1 showed a weak positive relationship with BioBERT similarity scores, suggesting a possible relationship between this metric and semantic similarity in the findings. This relationship was not statistically significant, so it should be interpreted with caution. If the relationship is underpinned by an undetected correlation, it would indicate that, when radiologists rated pairs of reports as more similar in their clinical context, BioBERT scores tended to be higher.

Question 2 showed a weak negative relationship with the Automated Readability Index (ARI). This relationship was also not statistically significant. If the relationship is underpinned by an undetected correlation, it implies that radiologists rated pairs of reports as more similar in their linguistic context as readability decreased (i.e. as text becomes more complex).

LLM as a judge results

To investigate the quality and reliability of the synthetic reports and impressions generated, Philips employed LLMs in an evaluative role, referred to as "LLM-as-a-Judge", to investigate the accuracy, clarity, groundedness and completeness of the synthetic records and impressions generated. These metrics would be used to assess the quality of the synthetic reports. The generated content was assessed using four metrics: accuracy, completeness, clarity, and groundedness. Each metric was scored on a scale from 0 to 10, accompanied by a motivation explaining the given score. Philips used an LLM model with a temperature setting of 0.6 for this evaluation. The same system prompt was used for each metric: clarity, accuracy, completeness and groundedness. The LLM judge evaluated all impressions, those generated by humans, the synthetic impressions based on real reports and synthetic impressions based on synthetic reports.

To assess the quality of impressions generated by the LLM, a human evaluation was conducted using a subset of radiology reports. A random sample was selected from two test sets: 15 original MIMIC reports with original impressions and 20 original MIMIC reports with LLM-generated impressions. Each impression was assessed using four key evaluation metrics (see appendix for details):

- Accuracy: The extent to which the impression is factually correct, including accurate values, details, and the absence of errors or misrepresentations.
- Groundedness: Whether the impression is anchored in the original findings, avoiding hallucinations and clearly distinguishing between facts and opinions.
- Completeness: Whether the impression captures all relevant information, including clinical context and necessary detail.
- Clarity: How well the impression is structured, including the use of appropriate, concise, and logically organised language.

These same metrics were applied by both the LLM Judge and certified radiologists, who independently reviewed the same set of impressions. Radiologists also evaluated the LLM Judge’s scores, offering a human perspective on the model’s performance.

The resulting scores were analysed to compare evaluations from the radiologists and the LLM Judge, identifying levels of agreement and pinpointing discrepancies. In addition to numerical scoring, qualitative comments from the radiologists were reviewed to uncover recurring strengths and weaknesses, offering practical insights into areas where LLM evaluation methods either align with or diverge from expert human judgement.

Agreement between the certified radiologist observers

Evaluators were asked to grade their agreement with the LLM judge score, for each metric, with a binary agree/disagree. All 70 pairs of Report-Impressions and LLM Judge scoring were assessed by two blinded certified radiologists.

Table 3 - Count of agreements and disagreement for the observers on the LLM judge scores per metric.

Metric	Decision	Observers count of agreement / disagreement with LLM judge		
		Obs1	**Obs2**	**Total**
Accuracy	Agree	52	53	105
	Disagree	18	17	35
Clarity	Agree	63	54	117
	Disagree	7	16	23
Groundedness	Agree	53	56	109
	Disagree	17	14	31
Completeness	Agree	46	45	91
	Disagree	24	25	49

Agreement in between Human scores and LLM judge score

Evaluators were asked to grade their assessment on a scale from 0-10 for each metric across all 70 reports-Impressions pairs. Overall, Observer 2 had higher agreements to LLM scoring than Observer 1 for all metrics, ranging from fair to good. Observer 1 had fair agreements with the LLM scoring. Weighted kappa demonstrated higher agreement (all fair) between the two Observers, suggesting that the 10-point scale was too granular, and should be reduced to 5 or 3 levels.

Feedback on Observers Recommendations

The radiologists stated that some recommendations were deemed valuable and could be incorporated by a human radiologist; however, others were inappropriate. Recommendations for surgery, for example, were seen as beyond the scope of what a radiologist would typically suggest.

The LLM Judge scored higher when the impressions included recommendations, regardless of their appropriateness. It may be necessary to fine tune the LLM Judge to avoid favouring recommendations that are deemed incorrect or irrelevant by human experts.

Conclusion

When generating synthetic radiology reports for evaluation purposes, it important to preserve linguistic similarity to real reports while introducing sufficient clinical variation. Metrics such as BioBERT similarity and ICD-10 distribution effectively captured clinical differences, while ROUGE-L, POS tag overlap, and ARI scores helped confirm stylistic alignment. Together, these metrics provided a practical framework for assessing the quality of synthetic data. Linguistic similarity is important for using synthetic data to supplement real clinical data; without it, the rationale for supplementation may be weakened.

This case study also offers insights into the evaluation of generated impressions by both human radiologists and an LLM Judge. Human observers aligned reasonably well with each other and with the LLM Judge, with Observer 2 showing stronger agreement. The LLM Judge applied less granularity in scoring, however, consistently favouring high scores (e.g., 9). Agreement metrics such as weighted kappa helped compensate for these differences and supported the overall validity of comparisons.

A key issue identified was the interpretation of “groundedness” and “completeness”. The LLM Judge tended to conflate the two, treating omissions as a failure of groundedness. whereas radiologists interpreted groundedness as the extent to which impressions are derived from report content, with omissions affecting completeness instead. This divergence likely contributed to lower agreement scores and highlights the need for clearer definitions and evaluator guidance.

While the LLM Judge shows promise in assessing radiology impressions, its interpretation and evaluation of groundedness, completeness, and recommendation relevance requires improvement, especially when distinguishing between human- and machine-generated reports. Future work should prioritise the refinement of scoring frameworks, clearer evaluation definitions, and more detailed guidance for automated evaluators to support consistency, interpretability, and alignment with clinical reasoning in the assessment of synthetic radiology data.

Case Study: AutoMedica – SmartGuideline and mitigating risks of LLM hallucination and non-determinism

Introduction

Guideline adherence improves health outcomes by promoting interventions with proven benefit, discouraging ineffective ones, and supporting greater consistency in care. Clinicians deviate from guidelines due to several factors: lack of access to guidelines at the point of care; awareness and access; clinician perception that their own experience outweighs guideline recommendations; lengthy, complex, or ambiguous guidelines.

SmartGuideline is an AI clinical decision support (CDS) tool designed to help healthcare professionals in both primary and secondary care follow best-practice recommendations. It is designed to provide real-time access to summarised clinical guidelines, answering management questions quickly and referencing the original guideline sources. The tool has potential to offer support across the full spectrum of clinical decision-making: diagnosis, drug selection, dosing, contraindications, investigations, treatments, and referral criteria. SmartGuideline aims to reduce unwarranted variation in care, improve safety, and increase clinician productivity by making guideline information more accessible and usable at the point of care.

Large language models (LLMs) are a type of AI algorithm that predict the next most likely word or part of a word based on previous context. This is achieved through creating mathematical representations of words and training the deep learning algorithms on huge volumes of textual data that pre-compute probabilities for the next most likely word. This means that AI produces plausible-sounding output but does not guarantee factual accuracy (i.e. the ground truth) which can result in plausible sounding inaccuracies (known as hallucinations or fabrications). LLMs are becoming more common in healthcare, from summarising clinical encounters and patient records to supporting clinical decisions. While this technology has great potential, the associated risks present key regulatory challenges. Even well-trained AI systems that use LLMs can confidently generate these plausible sounding inaccuracies, which in healthcare could lead to patient harm. The Medical Device Regulations 2002, which apply in Great Britain, require the benefits of the product to outweigh known and reasonably foreseeable risk. These risks should be mitigated primarily through safety by design. Retrieval-Augmented Generation (RAG) is a technique that enhances LLMs by grounding their outputs in curated external knowledge sources. By combining the conversational fluency of LLMs with the factual accuracy of retrieved content, RAG aims to address AI hallucinations. In this architecture, the LLM generates outputs using information drawn from a validated knowledge base, bridging “retrieval” and “generation.”

AutoMedica’s clinical decision support tool applies an adaptation of RAG in a healthcare context. Built using a curated knowledge graph, it applies its proprietary Tree-anchored Graph-RAG (TaG-RAG) to a structured database of National Institute for Health and Care

Excellence (NICE) resources, including clinical guidelines, knowledge summaries, patient leaflets, and a selection of medications from the British National Formulary (BNF).

Regulatory hypotheses & challenges

The team identified four primary regulatory hurdles that currently limit the safe and effective use of LLMs in regulated healthcare applications:

- Hallucination: The model may generate plausible sounding but factually incorrect or clinically unsafe content
- Non-determinism: Given the same input, LLMs can produce different outputs, which can undermine consistency and predictability
- *Training data and bias: The quality and representativeness of training data affect how well the model performs across different populations. Poor or unrepresentative data can lead to harmful inequities. More work on assessing and validating training data can be found in the Philips AutoImpression case study.*
- *Explainability: Regulatory expectations increasingly require AI systems to demonstrate how they reach a conclusion. A lack of transparency can reduce trust in AI-generated advice. More work on explainability can be found in the OncoFlow case study.*

In addition to exploring how these risks can be mitigated, the AI Airlock programme examined how SmartGuideline is positioned within the broader UK regulatory landscape. A further key objective was to assess the clinical relevance of SmartGuideline through engagement with a group of practising clinicians.

Hypothesis

The AI Airlock programme set out to test the hypothesis that key regulatory concerns surrounding the use of large language models (LLMs) in medical devices could be addressed through a safety-by-design approach by using RAG.

Methodology

To evaluate the safety and performance of SmartGuideline, AutoMedica tested the tool in the AI Airlock sandbox.

AutoMedica obtained test licences and data access from NICE, the Pharmaceutical Press, and the Medusa database. This allowed SmartGuideline to be built and tested using authoritative UK guidance, including the BNF, BNFC and NICE clinical guidelines.

Testing took place in three phases. First, SmartGuideline's responses were compared directly to the baseline model (ChatGPT-4o), to assess hallucination and omission rates. Next, reproducibility was tested by running 1,000 clinical queries through the model ten times each (n=10,000), allowing the developers to evaluate the consistency of responses. Finally, 19 GP trainees from Surrey participated in a live session where they explored the tool's usability, trustworthiness, and how it handled clinical complexity.

SmartGuideline usability

During the development of SmartGuideline, two detailed rounds of patient and public involvement (PPI) were undertaken with 30 healthcare professionals from Imperial College London and Imperial College healthcare trust. Sessions were recorded and a qualitative analysis was undertaken which highlighted several key themes: clinical utility, usability, staff compatibility and guideline comprehensiveness.

It was highlighted that to be a truly useful tool for doctors, local trust guidance must also be incorporated. Usability was addressed by redesigning the user interface during the AI Airlock.

Virtual and real-world Airlock testing

Two versions of the system were evaluated: an early proof-of-concept (PoC) version, was tested in a virtual environment to identify technical risks, and an alpha version, tested in a real-world setting Airlock. The study was designed to simulate real-world clinical scenarios and involved licensed UK medical doctors. Each doctor was randomly assigned a clinical question derived from NICE guidance or the BNF and provided an answer using either SmartGuideline or a baseline model, without being told which system they were using.

The core safety question was: how often does the model generate a clinically meaningful error?

The primary efficacy outcome was the rate of accurate query responses, and the primary safety outcome was the rate of hallucination .Hallucinations were classified as either major or minor, depending on whether they could or could not influence diagnosis or treatment decisions (as described in the CREOLA framework). Any major hallucination automatically triggered a root cause analysis, ensuring that lessons could be drawn to further improve the system.

Omissions were documented as a secondary outcome, in which the model failed to produce a response.

Non-determinism tests

The reproducibility of SmartGuideline was measured across a dataset of 1,000 queries, each submitted 10 times. This amounted to 10,000 responses, analysed using a range of natural language processing similarity metrics. These included BLEU and ROUGE (which assess word and phrase similarity for models with short textual outputs), and BERT and SBERT (which capture semantic meaning for longer outputs). The goal was to understand how variable the model was in its output, and whether such variability might pose a risk in a clinical setting. Initial results showed high internal consistency, with most variation occurring in phrasing rather than content.

Clinician workshop

Beyond technical performance, user trust and usability were also explored. In a live workshop, 19 GP trainees interacted directly with the model, using real clinical queries and assessing the tool's usefulness in real-time. Participants were asked to comment on explainability, potential biases, and overall user experience. Their feedback helped identify areas where SmartGuideline's interface and response style could be improved, for example, summarising information more succinctly or highlighting where decisions were based on national vs local guidance

While time and resource constraints limited full evaluation of all metrics, comprehensive analysis was completed for the primary outcome and overall response quality. Bias and explainability were assessed qualitatively in the GP trainee session, and reproducibility was evaluated across the full test dataset.

The study was conducted in February 2025, with statistical analysis and reporting finalised in March. The study was powered using the primary safety outcome; from pilot work, no hallucinations were expected with SmartGuideline and about 5% from chatGPT-4o. Using a power of 90% at an alpha of 0.05, a minimum of 406 responses (203) in each arm were required.

Results

PoC testing results

In 493 tests, all of which were manually evaluated (n=413 SmartGuideline vs n=80 chatGPT4o) there were no (0%) hallucinations with SmartGuideline vs 6 (7.5%) major hallucinations with chatGPT4o (P<0.0001).

In the PoC model, there was limited question rephrasing and access was restricted to NICE guidelines alone. Many of the clinical questions posed related to drug prescribing or adverse effects. As a result, once a subset of drug data from the BNF was incorporated, performance improved significantly. Moreover, to ensure the SmartGuideline model gave consistent answers, it always chose the most likely next word (a method called greedy decoding). Strong safety rules meant the model struggled to understand slightly different wording, which led to higher omission rates with SmartGuideline. Omissions were defined as failure to generate any response, or the model stating it does not have adequate information for a response when all the information needed for a response is present. Table 4 demonstrates comparative omission and hallucination rates with the SmartGuideline PoC and the baseline model. SmartGuideline omissions were significantly higher (P<0.001) whilst rates of hallucination were significantly lower (P<0.001) in the PoC testing.

Table 4 - Comparative omission and hallucination rates with the SmartGuideline PoC and chatGPT4o

	Hallucinations	Omissions
SmartGuideline	0/413 (0.0%)	35/413 (8.5%)
ChatGPT-4o	6/80 (7.5%)	0/80 (0.0%)

Alpha testing results

A total of 872 manual human evaluations were undertaken on the outputs of SmartGuideline (n=436) vs baseline GPT (n=436). There were no hallucinations and 5 omissions with SmartGuideline. There were 3 omissions (failures to generate a response) with the Chat GPT4o model, and 23 (5.3%) hallucinations. Root cause analysis divided the sources of these hallucinations as follows: n=12 faithfulness hallucinations (correct information but untrue based on underlying source material); n=7 factuality hallucinations; n=8 incorrect guideline sources. These categories were not mutually exclusive, multiple types of hallucination were present in some responses.

There were no differences in the rates of omission (P=0.725). SmartGuideline displayed statistically significant lower rates of hallucination (P<0.0001) (Table 5). Some SmartGuideline omissions were dependent on the phrasing of the question. For example, the model performed well when asking “signs that indicate need for referral for venous leg ulcer disease” but generated no response (omission) when asked about the “referral criteria for venous leg ulcer disease”. Strategies to mitigate these residual omission events are in development.

Table 5 - SmartGuideline alpha testing vs baseline GPT

	Hallucinations	Omissions
SmartGuideline	0/436 (0.0%)	5/436 (1.1%)
ChatGPT-4o	23/436 (5.3%)	3/436 (0.7%)

In all omission cases, no data were generated by the models, meaning there was no information available on which to base a clinical decision and, therefore, no associated clinical risk. The omission rates for SmartGuideline and the baseline model (1.1% and 0.7% respectively) were considered to be of low likelihood, with a likelihood score of 2 and a severity score of 1, according to ISO 14971.

The rate of hallucinations with SmartGuideline was (0), resulting in both a likelihood and severity score of 1. In contrast, hallucinations generated by ChatGPT included both faithfulness hallucinations - where information was broadly correct despite referencing non-existent sources - and factuality hallucinations, which were assessed as having a moderate severity (score = 3). Given an observed hallucination rate of 5%, the likelihood was also scored as moderate (score = 3).

When combining hallucinations and omissions, the maximum scores for likelihood and severity were used. ChatGPT's outputs resulted in an aggregate risk score of 9 (3 x 3), constituting high clinical risk. By comparison, SmartGuideline achieved an aggregate risk score of 2 (2 x 1), indicating low clinical risk for the primary outcome measure.

Table 6 – table of total risk score, using likelihood and severity as per ISO 14971.

	Consequence				
Likelihood	**Insignificant**	**Minor**	**Moderate**	**Major**	**Severe**
Almost Certain	**Medium**	**High**	**High**	**Extreme**	**Extreme**
Likely	**Medium**	**Medium**	**High**	**Extreme**	**Extreme**
Possible	**Medium**	**Medium**	**High**	**High**	**Extreme**
Unlikely	**Low**	**Medium**	**Medium**	**High**	**High**
Rare	**Low**	**Low**	**Medium**	**High**	**High**

Non-determinism

Consistency in responses is critical for AI healthcare models.

Semantic entropy was relatively high at 7, indicating low exact word match consistency. However, the BLEU score (short phrase lexical similarity) was 0.58 (high), the ROUGE score (long phrase lexical similarity) was 0.69 (very high), and the BERT and SBERT scores (which assess semantic similarity) were 94% and 97% respectively. For SmartGuideline, generating ten outputs per input was judged sufficient to capture variability in a meaningful and efficient way. The range of benchmarking scores reported here provides a promising foundation for standardising reproducibility metrics for regulatory consideration.

Secondary outcomes (Goodness of response, training data, bias and explainability)

The primary method for reducing poor-quality training data and bias within the model was the use of a proprietary information retrieval graph structure, combined with high-quality, information sources, to generate responses.

Explainability and the potential for bias were assessed during in-person user testing with 19 GP trainees. For a subset of SmartGuideline responses, participants evaluated the system using structured criteria. All (100%) of participants felt the SmartGuideline answers were moderately or strongly supported by scientific and clinical consensus; 94% felt that answers demonstrated understanding of the proposed clinical question; 100% felt that answers demonstrated correct knowledge recall; 83% felt there was moderate or strong explainability within the answers presented; 72% felt that the model answers contained unbiased information, (i.e., 28% felt there was potential for bias). Qualitative insights gathered through group discussion indicated that these concerns were less about the model itself and more about the underlying clinical guidelines, which participants felt were not always representative of diverse populations.

Importantly, this evaluation did not establish a definitive threshold for explainability required for clinician trust. The findings do suggest that perceptions of "sufficient" explainability are likely to vary between individuals, depending on their baseline clinical knowledge, familiarity with the patient in question, and comfort with digital tools. This variability underscores the importance of designing AI systems that offer layered or customisable levels of explanation to meet diverse user needs.

Conclusion

This case study highlights the value of structured, real-world testing to assess the clinical safety, reliability, and usability of AI tools designed to support clinical decision-making. It particularly demonstrates how targeted evaluation methods – such as hallucination rate assessment, deterministic decoding, and ISO 14971-based clinical risk scoring – can be applied in practice to interrogate system behaviour in a controlled environment.

The use of comparative testing against baseline models provided a clear framework for identifying specific safety gains (e.g. hallucination reduction) and trade-offs (e.g. increased omission rates). This underlines the importance of context-specific evaluation: techniques that improve safety may impact completeness or usability, and decisions on deployment should consider these trade-offs explicitly. The application of semantic similarity, BLEU/ROUGE scores, and risk-adjusted metrics enabled a multi-dimensional view of performance, which is likely to be more informative than any single indicator. These techniques could inform future guidance on performance benchmarking and validation.

Finally, qualitative testing with clinicians surfaced important insights into explainability and trust, underscoring that user perception of AI tools cannot be separated from technical performance, and may vary according to training or expectations.

Together, these techniques offer a toolkit for regulators and developers seeking to evaluate AIaMD products in a way that is robust, transparent, and proportionate to clinical risk.

Case Study: OncoFlow – OncoFlow and the trade-off between explainability and clinical performance

Introduction

OncoFlow aims to transform cancer care by tackling key inefficiencies in clinical workflows, especially those surrounding Multi-Disciplinary Teams (MDTs) and their meetings (MDTMs). Cancer treatment decisions must bring complex clinical data from multiple sources together and align it with the latest medical guidelines and best practices. Any delays in gathering this information or inconsistencies in interpreting it can negatively affect patient outcomes and delay critical care.

OncoFlow aims to improve the efficiency, accuracy, and transparency of cancer decision-making by integrating AI into cancer care workflows. Acting as an AI-powered co-pilot and decision-support tool for MDTs, OncoFlow helps clinicians by:

1. Enhancing consistency and transparency in patient care pathways.
2. Optimising the time spent discussing complex patient cases.
3. Ensuring the efficient use of clinical and diagnostic resources to support better decisions.

OncoFlow's automates key tasks that typically demand considerable clinician time and effort. The platform allows oncology teams to concentrate on critical decisions that require their expertise by providing clinicians with structured, evidence-based patient information.

OncoFlow consists of two main AI-driven components:

- Data Extraction System: This algorithm processes unstructured diagnostic and medical reports, converting raw data into clear, structured clinical summaries that can be rapidly reviewed and acted upon.
- Treatment Matching System: This algorithm aligns the extracted clinical data with current, evidence-based treatment guidelines, ensuring that MDTs have access to personalised and up-to-date treatment options for each patient.

Designed as a clinical decision-support tool that supports rather than replaces clinicians. OncoFlow maintains human oversight and accountability throughout the decision-making process. The platform seeks to enable more effective, personalised, and timely oncology care by standardising patient data and treatment pathways, ultimately improving outcomes for patients and healthcare systems alike.

Regulatory challenges

With AI-powered decision-support tools like OncoFlow operating in an evolving regulatory landscape, key regulatory considerations include compliance with the MDR 2002, adherence to Good Machine Learning Practice (GMLP) principles, and ensuring strong explainability to gain clinician trust and meet regulatory scrutiny.

A primary challenge for OncoFlow was defining and implementing explainability and transparency within its AI systems. Many AI models, particularly those based on LLMs, are often described as “black boxes.” This means it can be difficult for clinicians and regulators to understand how the AI arrives at specific decisions. Therefore, OncoFlow had to demonstrate how AI models could deliver traceable, evidence-based justifications for treatment recommendations. Whilst aligning with current regulatory expectations for safety, reliability, and clinical validity of AI medical devices. Successfully addressing these challenges was essential for building understanding of structured approaches to improving explainability, regulatory compliance, and strategies for clinical validation and trust.

Hypotheses

- *Explainability of an AI system (in this case, OncoFlow) can be verified by assessing its ability to provide clear, comprehensible, and consistent justifications for its outputs, using predefined explainability metrics and evaluation framework.*
- *Simpler models (such as traditional Natural Language Processing, (NLP), approaches) may enhance explainability, while more advanced LLM-based models may improve clinical performance.*

The primary objective of this study was to evaluate and compare the performance, explainability and clinical utility of different AI models when applied to unstructured breast cancer data, in the context of the OncoFlow use case. Breast cancer was chosen as the focus due to the availability of data, high prevalence, and burden within the NHS MDTs, and existing expertise within the OncoFlow team. Performance metrics included accuracy, precision, recall, F1 score, and computational efficiency (Graphics Processing Unit, (GPU) usage). Explainability metrics were focussed on assessing decision traceability, transparency in recommendations, and user comprehension. Clinical performance would be measured through factors such as ease of use, processing speed, and alignment with clinician needs to ensure a more seamless integration into MDT workflows.

The performance and explainability metrics were selected to ensure that the AI system meets the technical, clinical, and regulatory requirements necessary for safe and effective use in healthcare settings.

Methodology

The team explored multiple AI configurations for two key tasks:

- extracting structured clinical data and
- matching patient information to treatment guidelines,

specifically for use in breast cancer MDT meetings. Three AI approaches were compared:

- A rule-based NLP system
- A fine-tuned LLM trained on clinical data
- A hybrid method, combining rule-based NLP for straightforward data (such as dates and patient demographics) with LLMs for more complex clinical features (like lymph node status or evidence of metastasis)

Testing followed a standard train-test-validate cycle. A real-world dataset was sourced from Barts Health NHS Trust, containing 9,000 histopathology reports and 3,000 radiology reports from breast cancer patients. These reports were extracted from the hospital's Electronic Patient Record (EPR) system, anonymised by the Barts Life Sciences IT team, and clinically labelled to establish a ground truth for validation. The dataset was split 80% for training, 10% for testing, and 10% for validation. OncoFlow used current breast cancer guidelines, including the NICE Guidelines, ESMO (European Society for Medical Oncology) Guidelines, and SACT Protocols from the Clatterbridge Cancer Centre NHS Foundation Trust to guide the treatment matching process.

Performance of each AI approach was assessed using standard metrics: accuracy, precision, recall, and F1 score. This formed part of the clinical utility evaluation to determine how well each model performed in extracting relevant information and supporting decision-making.

Explainiability was tested using SHAP (SHapley Additive exPlanations), SHAP-IQ, and LIME (Local Interpretable Model-Agnostic Explanations). These methods helped to visualise which parts of the clinical text the AI model focused on when generating its output, essentially showing clinicians why the AI reached its conclusions. Outputs were reviewed by clinicians to assess whether the extracted information was meaningful, trustworthy, and fit for use in real-world practice. This testing was carried out securely on OncoFlow's AWS Cloud infrastructure, simulating the conditions of how OncoFlow would be used during actual MDT meetings. This allowed the technical and AI teams to conduct rigorous, controlled evaluations in a setting that reflected real-world clinical use.

To complement technical testing and model validation, a structured survey was conducted with a sample of clinicians to assess attitudes and expectations around explainability in AI-powered clinical decision support tools. The responses provided critical context for simulating how OncoFlow would be perceived and utilised in real-world settings, especially within MDT environments. The findings informed the fine-tuning of explainability features after initial testing of all the possible explainability approaches in the context of OncoFlow

Results

Data Extraction: NLP vs. LLMs

Both traditional NLP methods (e.g., regex-based rule extraction) and advanced LLMs were tested (Figure 2) during OncoFlow's Data Extraction System design. Traditional NLP methods worked well for pulling out structured information such as dates and numbers, but they struggled with more complex clinical reasoning, for example, linking tumour characteristics with diagnostic findings in varied report formats. As such, the NLP model performed poorly. The LLM models were better able to interpret these complex relationships, adapt to different document styles, and scale across large volumes of data. The final model configuration achieved up to 90% accuracy in extracting relevant clinical data information.

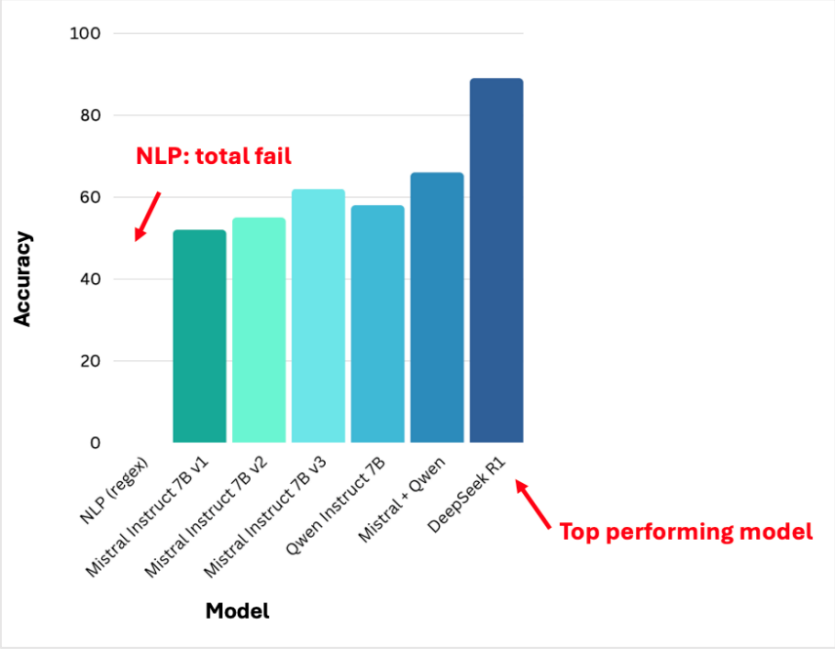

Figure 8. Comparative accuracy of NLP and LLM models for data extraction. This chart shows the performance of traditional rule-based NLP (Regex) versus multiple instruction-tuned LLMs. The NLP model underperformed significantly, while LLMs demonstrated higher accuracy, with DeepSeek R1 emerging as the top-performing model for extracting complex clinical information from unstructured texts.

A series of pre-defined clinical rules were integrated into the model's prompt design to improve the accuracy and consistency of OncoFlow's Data Extraction System. These rules were tailored to help the AI extract essential diagnostic details from unstructured medical reports, particularly in radiology and histopathology. Embedding these rules enabled the system to exceed the 80% accuracy benchmark, helping to ensure that relevant clinical information was correctly identified, interpreted, and standardised across different data sources.

Explainability tools such as SHAP (SHapley Additive exPlanations) and SHAP-IQ helped better understand how the AI evaluated inputs. These tests showed that 'distant metastasis' and 'HER2 status' were the most influential features affecting AI outputs. Other factors such as tumour size, lymphovascular invasion, and the presence of ductal carcinoma in situ (i.e., an early form of breast cancer where abnormal cells are found in the milk ducts but have not spread into surrounding breast tissue) also had a strong impact, while features like laterality (the side of the body) and biomarker presence had a more moderate effect on the system's performance.

The team observed a significant benefit from incorporating a "reasoning" feature into the model, which required the AI to explain the steps taken in generating outputs, thus providing insight into the logic used. This approach improved transparency, making it easier to trace potential sources of error.

Treatment Matching: LLM + RAG Architecture

A traditional NLP approach was initially tested but proved ineffective (Figure 9). An LLM integrated with Retrieval-Augmented Generation (RAG) was trialled. This approach was necessary to handle the complexity and frequent updates associated with clinical guidelines such as ESMO, NICE, and the SACT guidelines used in this case study. A standalone LLM (without RAG) posed a higher risk of hallucinations and reduced explainability. The RAG configuration ensured that treatment recommendations were explicitly grounded in retrieved sections of the guidelines. This improved both the traceability of outputs and alignment with regulatory expectations around safety and transparency. Retrieval efficiency was strongly influenced by the embedding model and chunking strategy used, both of which affected the accuracy and relevance of the content retrieved. The best-performing retrieval method combined PubMedBERT embeddings and a vector database (Qdrant) for optimised query expansion. The most effective model achieved 92% accuracy in retrieving relevant ESMO and NICE guidelines, and 100% accuracy in matching Systemic Anti-Cancer Therapy (SACT) protocols. Challenges were observed when guidelines included overlapping recommendations or lacked clear criteria for patient sub-groups.

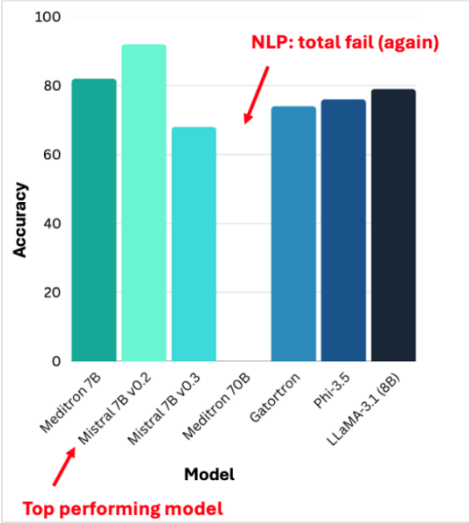

Figure 9. Accuracy comparison of language models for Treatment Matching. This chart compares the performance of various LLMs and NLP in retrieving and aligning cancer treatment recommendations with clinical guidelines. Traditional NLP again failed to handle complex guideline reasoning, while Mistral 7B v0.2 achieved the highest accuracy, outperforming other models in the Treatment Matching task.

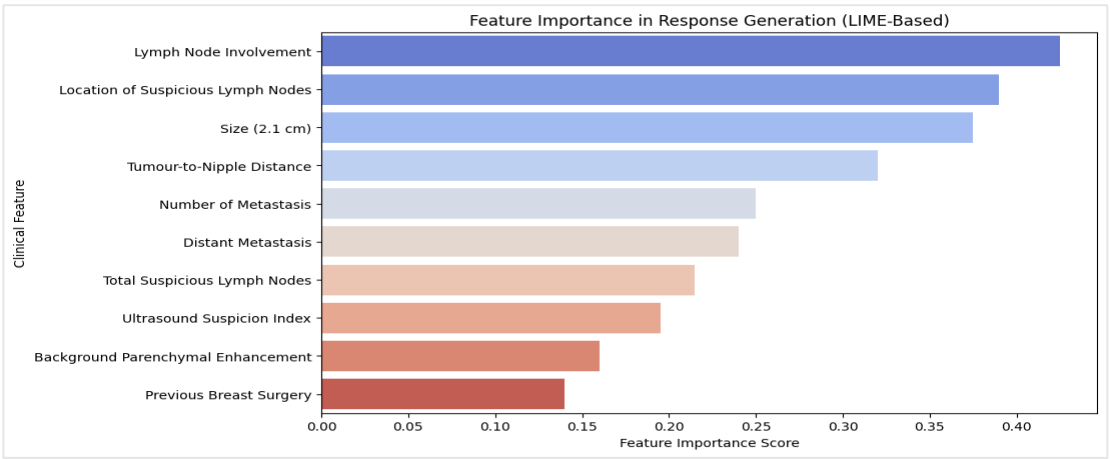

Figure 10. LIME Feature Importance in Response Generation (NICE/ESMO). This chart shows how different clinical features influenced AI-generated treatment suggestions, as measured by LIME. Lymph node involvement, location of suspicious lymph nodes, and tumour size had the highest impact on decision-making.

Despite maintaining a low hallucination rate (3.4%), testing identified overgeneralisation risks, particularly when the AI encountered gaps in guideline data. This underscores the necessity for clinician validation prior to any final, definitive decisions being made. More work on mitigating hallucinations can be found in the SmartGuideline case study.

Clinician testing results

To complement model testing, a structured survey was conducted with 16 clinicians to explore expectations around explainability in AI clinical decision support tools, particularly within MDT environments. The survey provided critical context for refining the explainability features of OncoFlow based on real-world preferences.

There was a strong consensus on the importance of explainability: 75% of respondents rated understanding AI reasoning as “extremely important”, with the remainder identifying it as “somewhat important” or “neutral”. Clinicians preferred transparency around recommendation generation, with 44% wanting a full breakdown of data inputs and algorithmic logic, 38% preferring step-by-step reasoning, and only 19% satisfied with brief summaries. Presentation style was also consistent – all respondents preferred a combination of text and visuals (e.g., graphs or decision trees) to enhance comprehension. Additionally, 94% valued understanding how clinical features were weighted in the model, validating the relevance of tools like SHAP and LIME.

Uncertainty communication was considered essential: 56% preferred detailed explanations of uncertainty or limitations, and 44% supported flagging uncertain outputs. None supported omitting uncertainty information, indicating that clinicians expect AI to reflect the inherent limitations of medical decision-making. Only one respondent was comfortable accepting recommendations without supporting reasoning. Most preferred either detailed or partial reasoning to maintain transparency and clinical trust.

Clinicians also favoured a flexible, context-sensitive approach to explainability. Higher levels of detail were expected for high-risk or acute scenarios, with lower-stakes settings permitting simpler outputs. This supports the use of explainability tiers aligned with risk classification (e.g., medical device class I–III). These insights informed the design and refinement of explainability features within OncoFlow, including the use of visual aids, confidence indicators, and transparent reasoning outputs. They also shaped how explainability was simulated in the testing environment to better reflect real-world clinical expectations.

Conclusion

In the treatment matching task, traditional NLP tools unable to handle the complexity of clinical guidelines. An LLM combined with a RAG approach proved more effective. This allowed the AI to find and quote relevant sections of guidelines when making treatment suggestions, improving both traceability and alignment with safety expectations. The model achieved high accuracy in identifying relevant guidance, although some challenges remained in cases where guidelines were ambiguous or overlapping.

Throughout the study a strong emphasis was placed on explainability. Tools such as SHAP and LIME helped to clarify which clinical factors most influenced the AI’s decisions. The addition of a “reasoning” feature gave users a clearer view of how the system arrived at its outputs. This was especially important in building confidence among clinicians.

Feedback from a survey of healthcare professionals reinforced this point. Clinicians placed high value on transparency, with most wanting detailed explanations of how AI recommendations were made. There was broad support for using a mix of text and visuals to

communicate this information, and for adapting the level of explanation based on the clinical context. The findings also highlighted the importance of acknowledging uncertainty, particularly in higher-risk decisions.

From a regulatory perspective, this case study highlights several promising techniques for improving the safety, reliability, and explainability of AI systems used in clinical settings. It also underlines the value of early engagement with users, to ensure that new technologies align with the expectations and needs of those who will use them in practice. While the results are encouraging, further work is needed to ensure these methods are robust across a broader range of clinical scenarios. Continued testing, guidance development, and cross-disciplinary collaboration will be key to supporting the safe and effective use of AI in healthcare.

Case Study: Newton's Tree – FAMOS and continuous, real time post-market surveillance monitoring

Introduction

Over time, the performance of an AI system can decline. This may happen due to shifts in the environment, such as the introduction of new types of patients, the use of different medical scanners, or other unforeseen changes in the clinical environment. This is called drift and presents a significant barrier to AI safety, and therefore its uptake.

A common mitigation strategy for model drift is to involve humans in the loop, for example by having clinicians verify AI outputs. Clinicians are, however, susceptible to automation bias, which leads them to agreeing with the AI output without thought and makes human-in-the-loop an ineffective mitigation strategy. To maintain safety, trustworthiness and regulatory compliance, continuous monitoring of deployed AI systems is essential. Continuous monitoring can detect issues that develop too slowly or too quickly to be captured through scheduled reviews and periodic audits.

The FAMOS Dashboard was designed to support this continuous oversight. It presents three real-time metrics to users: each selected through a clinical risk management approach. Potential hazards associated with AI use are identified and data are selected to most effectively signal those hazards. The three metrics align with ISO guidance on risk management for medical devices incorporating machine learning. The three metrics are:

- **Data Input Quality**: This metric tracks how the data fed into the AI system changes over time, identifying any outliers or anomalies. Ensuring high-quality data input is crucial because poor data can lead to inaccurate AI outputs.

- **AI Inference Values**: This metric monitors the AI's output over time to detect any drift. Detecting drift early helps recalibrate the AI system to maintain its accuracy.

- **AI/Human Agreement**: This metric measures how often NHS staff members agree with the AI's output. High agreement rates might indicate automation bias, where clinicians might accept AI recommendations without critical evaluation. Monitoring this helps identify and mitigate this bias.

Trends in these metrics can be monitored so preventative action can be taken before a safety incident occurs. The dashboard can be viewed by three types of organisations:

- The AI manufacturer can view the anonymised data trends in the dashboard, ensuring privacy is maintained. This can be viewed for their products across all sites they are deployed.

- The NHS organisation can view trends across all AI products deployed at their site. Additionally, the dashboard provides the ability to drill down into specific data, allowing for detailed investigation of particular trends.
- Oversight bodies, such as the MHRA or NHS England, can view combined trends across all AI products and sites. However, like manufacturers, they cannot access individual-level data, ensuring that privacy is maintained

The AI Airlock programme aimed to assess how real-time monitoring can enhance a product's risk management strategy by enabling earlier identification of performance or safety issues. An unresolved challenge was the lack of clear methods for defining thresholds for these monitoring metrics. Without robust, evidence-based thresholds, it is difficult to determine when deviations in performance should trigger interventions or regulatory actions. Establishing methods for setting these thresholds is crucial to making continuous monitoring actionable and compliant and remains a key area for further regulatory and research development.

Regulatory challenge

Annex 2, clause 3.1 of the Medical Device Directive (1993) placed an obligation on manufacturers to conduct post market surveillance (PMS), a key part of the risk management strategy. This obligation continues and grows broader as medical device regulation is updated with The Medical Devices (Post-market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024.

Following the existing methods for PMS, which relies on NHS data sharing and reporting, which may not provide enough data. One potential solution is a path that automates the sharing of information whilst preserving privacy. Further, for AI as a Medical Device, such a path could enable manufacturers to meet the 10th Good Machine Learning Practice (GMLP) principle, produced by the International Medical Device Regulators Forum (IMDRF): "*deployed models are monitored for performance*", and the 7th GMLP principle "*Focus is placed on the performance of the human-AI team*".

FAMOS offers automated continuous monitoring of AI and its (and other such platforms) use may offer a means for meeting these regulatory requirements.

Hypotheses

A real-time monitoring system can improve product safety by providing trends that can be fed into an organisation's Risk Management System. By identifying variations in data quality, AI performance and human agreement, risk can be mitigated enabling regulatory compliance.

- *Trends can reliably be presented in a real-time monitoring dashboard.*
- *Trends exist that indicate risk.*
- *There is a place for trends in a risk management system.*
- *There is a way to address the risks identified by trends.*

Methodology

The study involved a retrospective observational trial employing federated evaluation of AI models. FAMOS was implemented at two NHS hospitals: Leeds Teaching Hospitals NHS Trust and Greater Glasgow and Clyde Health Board. The AI technologies evaluated included chest X-ray algorithms from Qure.ai and Annalise.

A stratified random sampling method was required to ensure a representative sample of the hospital's patient population while meeting the study's data requirements. The sample size was predicted to be close to 3,000 patients across both sites. For the statistical analysis a two-sided p-value of less than 0.05 will be considered statistically significant.

For the primary objective:

Descriptive statistics summarised the number and percentage of chest X-rays successfully monitored by FAMOS at each site to evaluate the effectiveness of FAMOS in monitoring AI performance across different hospitals. Metrics including data capture rate (proportion of eligible chest X-rays monitored), timeliness (mean and median time delays in data capture), and monitoring accuracy (comparison of FAMOS-monitored data with source data) were calculated.

Comparative analyses between sites were conducted using chi-squared tests for proportions and t-tests or Mann-Whitney U tests for continuous variables, depending on data distribution. Visualisation tools such as flowcharts and bar charts will illustrate data flow and data capture rates. The effectiveness of FAMOS was assessed based on its ability to provide effective and timely monitoring and any discrepancies or delays in data capture.

Results

At the time of writing, the final analysis was still ongoing; however, an interim analysis had produced some early findings.

Data attributes

At NHS Greater Glasgow and Clyde (GGC) and Leeds Teaching Hospitals (LTH), chest X-rays were processed. The data attributes included comparable age ranges across the two sites, similar male-to-female ratios, and similar proportions of winter versus non-winter patient data; however, the Glasgow site had twice as many image artefacts in X-rays compared with Leeds. There was minimal time delay between image processing and display on the monitoring dashboard, supporting the first AI Airlock hypothesis: that trends can be reliably presented through a real-time monitoring dashboard.

An initial review identified that dataset outliers were primarily associated with:

- image artefacts (e.g. jewellery worn by the patient)
- variations in patient positioning
- differences in image orientation (e.g. posteroanterior vs lateral view)

At NHS Greater Glasgow and Clyde, four distinct data clusters were observed. This likely reflects the fact that GGC receives data from four separate hospital sites. These clusters highlight heterogeneity in the data, which could influence AI performance and suggests that further principal component analysis is needed to confirm the source and characteristics of each cluster.

Human-AI agreement

To ensure robust comparisons, the study design included a matched case mix across sites: there was equal distribution of normal and abnormal findings, and a balanced demographic profile (gender and age, above and below 60). Each site employed four readers, two per target disease, and used multiple AI vendors. Most major identified confounders were controlled for, such that the only meaningful differences between sites were the reader skill mix and the conditions under which readings were conducted.

Results showed that reader agreement with AI suggestions at NHS Greater Glasgow and Clyde increased over time, while remaining relatively stable at Leeds Teaching Hospital. The increase in positive findings over time observed in Glasgow was statistically significant compared to Leeds and may suggest a susceptibility to automation bias. This may be attributable to differences in reader experience. At Leeds Teaching Hospital, Newton’s Tree

recruited three reporting radiographers (each with over 10 years of experience) and one radiology trainee (ST3). However, at NHS Greater Glasgow and Clyde, Newton's Tree recruited four radiology trainees. Furthermore, Leeds reading sessions took place during standard working hours (9am–5pm), whereas Glasgow sessions were spread across irregular hours, including night shifts and early mornings - conditions more prone to fatigue, distraction, and time pressure. These are known contributors to automation bias; however, aggregating results by site or by disease may mask individual effects. It is plausible, for example, that a single reader could be disproportionately influencing overall trends, reinforcing the importance of scrutinising individual-level data for aberrant patterns. These early results support the hypotheses that trends exist that indicate risk and that trends can reliably be presented in a real-time monitoring dashboard.

Conclusion

This case study explored the use of real-time monitoring techniques to assess the performance of AI in medical imaging across two NHS sites. The study aimed to test the hypothesis that useful insights about AI performance and risk can be surfaced by monitoring data trends in real-time.

Initial findings suggest this is both possible and practical. The study was able to monitor AI performance across hospital sites with different infrastructure and staff profiles, capturing trends in data quality and human-AI agreement. It showed how differences in staffing, experience levels, and working conditions can shape how AI is used and interpreted – factors often missed in controlled testing environments. These insights underscore the value of techniques like federated evaluation, structured sampling, and human-AI agreement monitoring in post-deployment settings. The study also highlighted the need to look beyond headline metrics and examine underlying trends, including at the level of individual users, to identify potential risks or changes in behaviour over time.

While the analysis is ongoing, early results suggest that real-time monitoring methods can surface subtle patterns that may inform both safety surveillance and product improvement. Further work is needed to explore how these methods might be used in practice, including how to respond to emerging risks and what infrastructure would be required to scale these approaches more broadly.

5. Simulations and regulatory gap analysis

In parallel with testing conducted by candidates in virtual and real-world hospital settings, the Airlock team undertook a regulatory mapping and gap analysis exercise. This was complemented by a series of simulation workshops, which brought together stakeholders from across sectors and disciplines – including clinicians, regulators, academics, technical specialists, and innovators – *to explore key regulatory challenges identified through the AI Airlock. This section outlines the methodology used and summarises the key findings from both the regulatory analysis and simulation activities.*

Regulatory mapping and gap analysis

Regulatory gap analyses are difficult; such analyses performed in such a novel field are doubly difficult. The team enlisted experts from various sectors in an initial regulatory mapping activity. This resulted in a collation of regulatory documents and associated challenges. To account for limited resource and programme timelines, we prioritised 5-10 most relevant regulatory documents for in-depth review per regulatory challenge.

This foundational exercise proved extremely productive. By showing us how broad the term "regulatory gaps" could be in the context of novel AI devices, we could account for this in our definitions. We therefore used an equally broad definition to capture the gaps. This allowed us to capture wording gaps in existing guidance, context gaps in frameworks and even legislative loopholes.

Within AI Airlock, therefore, the term 'gap' is used as a shorthand term for identifying any areas of improvement, recognising that many of the existing parts of the regulatory system must cater for the millions of hardware and software products used in the health and care system. It is the objective of the AI Airlock to identify these areas for improvement and recommend solutions.

The AI Airlock pilot identified regulatory gaps that currently challenge the safe and effective deployment of AIaMD. While the nature of these gaps varied across use cases, some recurring themes emerged across all projects, particularly in areas of validation of text-based data from large language models (LLMs), risk management, non-determinism, explainability, and post-market surveillance (PMS). These gaps and challenges were explored in collaboration with four innovators: Philips Healthcare, AutoMedica, OncoFlow and Newton's Tree.

Methodology

The scope of the review was primarily GB-focused, referencing the MDR 2002 and supplementary guidance either authored in the UK or developed through UK collaboration with international partners. Relevant international standards were also considered.

The process began with regulatory mapping. Each challenge was broken down into subcomponents / questions for consideration. Next, a broad range of relevant documents were identified specific to each regulatory question explored in each project. This list was expanded in collaboration with stakeholders, including members of the supervisory committee, Team AB, and the candidate organisations. This initial mapping resulteds in approx. 30 to 40 documents mapped per project. Where regulatory information was scarce, such as text-based synthetic data, the broader topic (synthetic data) was considered.

Following this initial mapping, the documents were reviewed at a high level and a subset of 5 to 10 documents was selected for in-depth analysis. Prioritisation was based on relevance to the specific regulatory challenges being addressed, as well as practical considerations such as availability and scope. Once selected, the documents were reviewed in detail, with relevant clauses and sections extracted and evaluated against the needs and risk profiles of AIaMD.

This deeper review focused on identifying areas where existing guidance may need to improve. These included where there were missing provisions related to challenges, insufficient detail, and unclear expectations for managing risk in adaptive, continuously learning systems. The analysis also considered whether existing guidance provided clear, enforceable requirements for real-world performance monitoring and the safe deployment of AI in dynamic clinical environments.

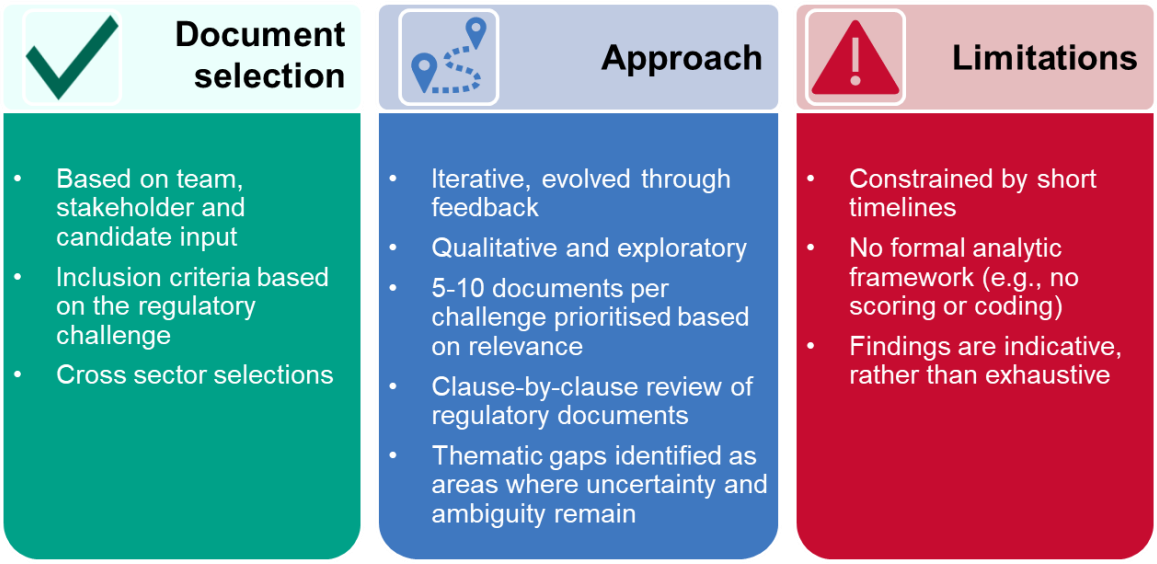

Figure 13. Regulatory mapping and gap analysis process

Where relevant, the identified gaps were thematically grouped and considered in terms of their likely regulatory impact. Particular attention was given to instances where regulations do not adequately address emerging risks or are impractical to implement. In doing so, the analysis sought not only to highlight gaps, but to build a clearer picture of where current regulation may require strengthening to ensure that AIaMD can be developed and deployed safely and effectively.

The analysis compared the coverage provided by existing documents against the level of detail needed to manage AI-specific risks in clinical settings for each challenge. Where gaps were identified, either through absence of guidance or insufficient detail, these were flagged for further consideration.

Findings from the MHRA and candidate teams were discussed collaboratively over a series of virtual meetings before being consolidated into final project reports. This process helped highlight shared gaps across projects and areas where future regulatory development or clarification may be required to ensure AI medical devices can be safely and effectively managed in real-world environments.

Results: Identification of key themes and gaps

The regulatory gaps identified in this pilot vary depending on the challenge being explored. Most gaps apply across different candidates, given the overlaps in technology types and challenges.

For the synthetic data case study, the primary focus was on unstructured text data. Synthetic data and the use of LLMs are not adequately addressed within current regulatory frameworks, including PCCP guidance and existing risk management standards, While documents, such as BS/AAMI 34971 acknowledge the potential of synthetic data to address privacy concerns and provide some guidance on its use and quality evaluation, there remains no clear direction on when the synthetic data should be generated, what prerequisites must be met beforehand, or how such data should be validated. In addition, there is ambiguity surrounding how “ground truth” should be determined, what metrics are appropriate for evaluating synthetic data, and when human validation is required. These issues are particularly relevant when synthetic data are used in place of real data for testing or training purposes. Additionally, the boundaries between synthetic and real-world data need clearer definition to determine whether synthetic datasets are exempt from GDPR and to ensure they cannot be linked back to individuals. For validating unstructured text data, there is a need for regulatory clarity on the use of LLMs in generating and validating such data. This includes the quality criteria that must be met for such data to support regulatory and clinical decisions.

Thiry-one documents were identified in the regulatory gap analysis exploring challenges posed by AIaMDs that are prone to errors such as hallucinations. Of these, six were prioritised for in-depth review based on relevance and access. These included the MDR 2002, ISO13485, ISO14971, AAMI/BS 34971:2003, IEC62304, IEC62366-1. While these documents provide a foundational framework, the analysis identified notable gaps relating to hallucinations, non-determinism and explainability, indicating that current regulations can be bolstered by additional guidance to support the safe and effective deployment of AIaMD. Several standards, including ISO14971 and ISO13485 were not developed with the dynamic

and opaque nature of AI in mind. Guidance around specific approaches for detecting and managing unpredictable outputs and validation of non-deterministic behaviour are needed. Including AI specific guidance, companion standards or best practice frameworks to bridge the gap between high level regulatory expectations and demands of implementing AI safely in healthcare. In addition, risks such as hallucinations could be viewed as falling within a manufacturer's performance claims; however, outputs incorporated into the medical record are also subject to the Health and Social Care Act 2008, regulation 17. This creates a tension between different regulatory requirements, where clarity is needed on how performance considerations intersect with obligations relating to the accuracy of health records.

Testing and simulation workshops consistently highlighted the critical importance of explainability in AIaMD, while also revealing persistent gaps in the current regulatory guidance. It was noted that many medical devices currently in use, including some in high-risk settings, are not fully explainable yet are accepted within existing frameworks. This underscores that explainability requirements cannot be considered in isolation, but rather in relation to device type, risk classification, and intended use. During the AI Airlock pilot, explainability was therefore explored more holistically, including different stakeholder expectations and the absence of clear criteria for selecting appropriate metrics. While some relevant direction exists in standards such as ISO 13485, ISO/IEC 22989, and BS/AAMI 34971, these do not define explainability requirements for AIaMD. Questions around the role of human involvement, the risks posed by non-explainable outputs, implications for usability, and appropriate approaches to transparency remain largely unresolved. Importantly, these standards do not cover the tailored approach required for different stakeholders such as clinicians, regulators, developers, patients, and others. Similarly, methods for implementing traceability and transparency to ensure usability need to be covered. Some of these regulatory gaps were addressed through virtual testing. OncoFlow, for example, introduced confidence scoring, traffic light indicators, and human-in-the-loop validation to help clinicians understand and validate outputs. In addition, traceability was enabled by logging AI recommendations, interactions, and source documents, each with unique identifiers. Simulation testing also demonstrated the varying explainability needs of different users, reinforcing the importance of intended use in designing explainability approaches.

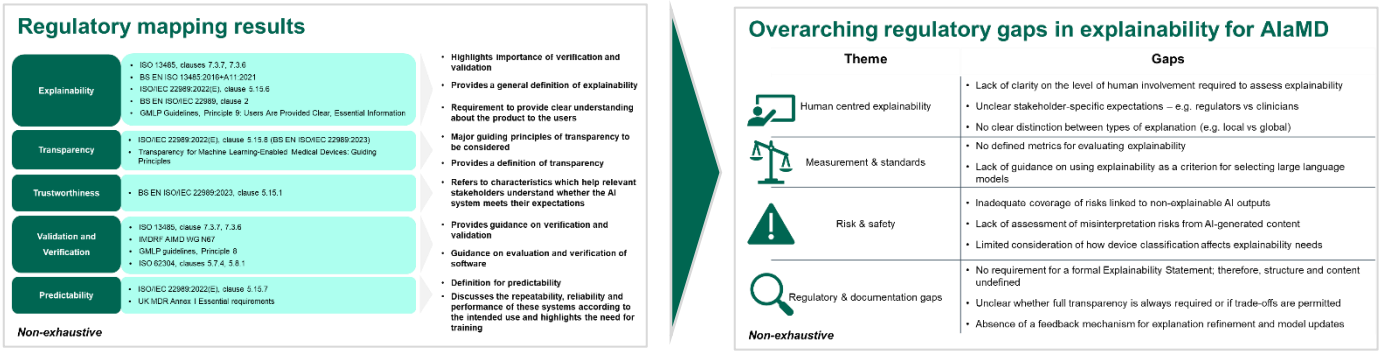

Figure 14. Example regulatory mapping and gap analysis results

Collaboration between the MHRA and OncoFlow led to the development of an explainability statement outlining the product's functionality, risks, and underlying AI. Guidance on content and structure for such documentation is still needed for innovators and products for which such statements are relevant. The approaches implemented as part of AI Airlock can be used by other AIaMD manufacturers, depending on the context of their product, its risks, intended use and intended population.

A regulatory gap analysis was conducted to identify areas needing additional support for AIaMD PMS. The reviewed documents included the MDR 2002 PMS SI 2024 (since updated to The Medical Devices (Post-market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024 amends the Medical Devices Regulations (MDR) 2002)), Yellow Card system, MORE system, ISO14971, ISO13485, AAMI/BS34971, BS EN62366-1, ISO27001. The analysis highlighted the need to additional AI specific guidance addressing changes in AI model performance over time and potential for preventative action to be taken before patient harm has occurred. In addition, there is a need to address shifts in human performance and reliance on AI when interacting with AIaMD. While the existing frameworks provide a good foundation, and the new PMS legislation addresses many of the gaps surrounding trend detection and the potential for AI model drift, specific guidance would help strengthen current regulations. When reviewing risk management frameworks such as ISO14971 and AAMI/BS34971, detailed requirements and threshold setting guidance were lacking. BS EN 62366-1 does not fully address the usability challenges introduced by AI systems such as explainability and how that could feed into over-reliance or automation bias.

An overarching challenge is the focus on corrective, reactive actions across regulatory documents. There is limited support for trend analysis which leads to early intervention and prevents an incident occurring. During testing in the AI Airlock pilot, continuous real-time monitoring was used to detect trends in data quality, model drift and automation bias. At one hospital site, the pilot identified a pattern of over-reliance on AIaMD. Given that AIaMD can produce errors, such over-reliance could lead to patient harm in a real-world setting. If trends are detected early, by contrast, there is potential to investigate and take preventative or corrective action, such as patient recall. Further analysis revealed that in this case study example, contributing factors to automation bias included clinician fatigue, distraction, time pressure, and level of experience. These findings highlight the need for AI-specific supplementary guidance on proactive safety measures, including best practices for threshold-setting, monitoring AI-human behaviour, and assigning reporting responsibilities.

Simulations methodology

Simulation workshops gathered cross-functional input on the regulatory challenges associated with AIaMD. These sessions provided a forum for discussion and collaboration, allowing participants to explore different perspectives and work towards consensus on complex issues. The workshops centred on three primary themes:

1) evaluating current risk management frameworks regarding LLM-specific risks e.g. hallucinations

2) evaluating the explainability needs of diverse stakeholders and negative consequences of limited explainability

3) evaluating the responsibilities surrounding continuous monitoring of AIaMD in real-world settings.

Figure 15. Simulations methodology illustration

Participants were selected based on their subject-matter expertise. The invitee list included technical experts from the NHS AI team, academic researchers, clinicians, representatives from approved bodies, and regulators from organisations such as the MHRA and DHSC. It also included clinical safety officers, data specialists from the CPRD, representatives from NICE with expertise in health technology assessment, and data governance experts from the ICO.

Each workshop began with a set of introductory presentations to ensure that all attendees shared a common understanding of the AI topic under discussion and the relevant regulatory landscape. These presentations outlined the current regulatory frameworks, including applicable laws, standards and guidance. A structured discussion was then facilitated to address specific regulatory challenges and known gaps. The aim was to explore these challenges from multiple angles and to identify areas where further regulatory clarity, development or alignment might be needed.

The workshops were not recorded to encourage open and candid dialogue. MHRA team members took detailed notes throughout the discussions. These notes formed the basis for the simulation reports and informed the final synthesis of findings from the AI Airlock programme. These workshops have been published as an output from AI Airlock and the recommendations incorporated with those from virtual and real-world testing during the pilot programme. The simulation reports, including an overview of the key discussion points and insights have been published separately.

Simulations summary results

Considerations for hallucination evaluation

The group explored hallucinations as a key safety risk in medical settings. Severity depends on both the nature of the error and its clinical context, with high-impact settings like

diagnosis or emergency care requiring particular scrutiny. To strengthen risk management in the deployment of AI, some members of the group suggested a three-tiered approach to categorising its use.

At the highest level, "No-Go Areas" were identified as contexts deemed too high-risk for autonomous use LLMs. These could include use cases where the consequences of errors could be severe. In such cases, the group advised that AI should not operate without human oversight. In the middle tier, "Caution Zones," the group highlighted use cases such as clinical documentation and summarisation as areas where LLMs could be used under human supervision. These use cases carry moderate risk and require close human monitoring to ensure safety and reliability. Finally, the group defined "Safe Zones" as applications where the risk is relatively low, and the benefits of AI adoption are potentially high. These include administrative support, assistance with clinical research, and patient engagement, where LLMs can provide meaningful efficiencies with minimal risk to patient safety. This risk-based approach would help prioritise regulatory attention and clarify where autonomous vs. supervised AI use is appropriate based on the risks associated.

This could help guide regulatory focus and clarify the need for human supervision. The group emphasised that AI should support, not replace, clinical judgement, with explainability features (e.g. uncertainty scores, data citations) seen as important safeguards.

Detecting hallucinations is challenging; even clinicians may miss subtle errors. A hybrid approach combining explainability tools with human review was recommended, alongside improved awareness of reporting systems like Yellow Card and MORE. Participants also raised the issue of "faithfulness hallucinations", factually correct outputs outside the AI's intended use. Faithfulness hallucinations are difficult to identify and can still influence clinical decisions or documentation, even when the AI system is not qualifying as a medical device. This revealed a potential regulatory blind spot with the group calling for clearer guidance on "function creep" and how users can identify the boundaries of appropriate use.

Current standards (e.g. ISO 14971) provide a base safety framework, but tailored metrics for hallucinations are needed. These should consider not just frequency, but detectability and clinical impact. The group also highlighted a key trade-off: reducing hallucinations too aggressively could increase omission risk. Model cards summarising known failure modes were suggested to support informed use.

Explainability requirements for different stakeholders

The group agreed that different stakeholders need different types of explanations: developers need to debug systems, regulators need evidence of safety, clinicians need to judge whether to trust a recommendation, and patients need reassurance in plain language. The group supported tailoring explainability to each user's needs, distinguishing between

global explanations (for understanding overall model behaviour) and local ones (for individual decisions). Too much or too little detail can both be unhelpful – interactive, layered explanations may offer a balanced solution.

Current standards such as BS/AAMI 34971 and GMLP offer a foundation, but the group called for clearer criteria specific to different roles. They also highlighted challenges: complex models are hard to explain, some explainability methods add confusion, and simpler models may be both interpretable and effective.

Explainability supports safety, accountability, and early error detection, especially critical where clinicians rely on AI. However, explainability must be human-interpretable. The group warned against relying on AI to check other AI, especially where models are too complex for meaningful human oversight.

Participants explored techniques to improve understanding of how AI tools generate outputs, both for clinicians and regulators. Three core approaches were discussed: explainability statements, model reasoning, and quantitative methods.

Explainability statements aim to clarify how a tool works in plain language. Participants noted that many current examples lack detail on how users interpret or act on outputs. A shared template, tailored by audience (e.g. patient, regulator), could improve consistency. Model reasoning, such as showing the steps a model takes from input to output, was seen as valuable for understanding edge cases and building clinician trust. It must be aligned with intended use and supported by post-market surveillance. Interfaces should offer layered access to detail depending on user needs. Quantitative approaches like SHAP and LIME can highlight which inputs most influenced an output. However, they require users to interpret both the model, and the method used. Participants emphasised that in most cases, these tools should supplement – not replace – clinical judgement. Familiar visual formats and summaries could support usability and trust.

Across all methods, participants stressed the importance of surfacing missing context (e.g. patient preferences or lifestyle) and maintaining transparency about model limitations. Finally, the group stressed that insufficient explainability can lead to blind trust, automation bias, or unsafe decisions. Tools like model cards and confidence scores could help users understand what the model considered - and what it didn’t. Upskilling users and designing for varied levels of AI literacy were seen as essential to safe deployment.

Post-market surveillance roles and responsibilities

Effective post-market surveillance (PMS) of AI-enabled medical devices depends on strong collaboration between manufacturers and clinical teams. Understanding how these systems perform over time, or how they might be affected by changes in their environment, requires input from both those who design the technology and those who use it day-to-day in clinical

settings. While there is broad support for this kind of collaboration in theory, in practice it often falls short. Manufacturers and healthcare providers both report frustration about a lack of engagement from the other side.

One area where this becomes especially important is in automating parts of the risk management process. AI dashboards can flag potential issues, but uncertainty around roles and responsibilities means it's not always clear who should act on these signals. This creates a risk that problems may go unaddressed – not because they aren't seen, but because it's unclear who is accountable for responding.

As data volumes grow, these systems could provide even greater insight into performance and safety. But making good use of that data requires skilled analysts and carefully calibrated thresholds to help distinguish important warning signs from background noise. Without this, there is a real possibility that existing safety reporting systems could be overwhelmed with low-value alerts, making it harder to identify real issues or take meaningful action. Crucially, monitoring alone isn't enough. When concerns are flagged, there must be clear and timely follow-up. There must be agreed protocols in place so that when performance dips or safety risks arise, there is a shared understanding of what steps to take – both to prevent harm and to ensure regulatory reporting requirements are met.

Finally, there remains uncertainty around what kinds of issues involving AI tools should be formally reported. Manufacturers are required to report incorrect diagnostic outputs. Healthcare professionals are encouraged to flag concerns through systems like the Yellow Card Scheme. Some are concerned that disagreements between clinicians and AI systems, especially if they become frequent, could flood these reporting systems – making it more difficult to focus on the most important risks.

6. Limitations and challenges

Throughout the AI Airlock programme, a range of limitations and challenges were identified by both the candidates and the Airlock team. This section focuses on those relating specifically to the regulatory gap analysis, simulation workshops, and technical testing. Broader programme-level challenges are addressed in the evaluation section that follows.

Gap analysis

The regulatory gap analysis followed a pragmatic rather than predefined approach, with documents selected based on team and stakeholder input. While not systematic in the formal sense, the method was informed by expert judgement and evolved iteratively as the most relevant materials and techniques became clearer. Philips, AutoMedica, OncoFlow, Newton's Tree and the MHRA reviewed different sets of documents, so some divergence in findings was anticipated. Prioritising a subset of the regulatory mapping documents is a limitation, as it may have led to the omission of relevant information in the gap analysis.

Simulation workshops

The simulation workshops generated useful insights, but due to the absence of a structured analytic framework, findings should be treated as exploratory. Time constraints prevented non-determinism simulation or exploration, despite the able assembled stakeholders.

Virtual testing

The Airlock team did not review candidate technical files. In several cases, these were not yet available due to the early development stage of the products. Given the exploratory nature of the Airlock, this was considered consistent with the programme's objectives.

The Airlock team did not access the datasets used for testing and therefore could not verify whether they met the definition of appropriate clinical data under the MDR 2002. As the purpose of the Airlock was to surface regulatory challenges and generate insights rather than to assess conformity, this approach was regarded as appropriate.

Philips AutoImpression

The regulatory mapping revealed significant gaps, as no specific guidance, standards, or acceptance criteria were found for generative AI, synthetic data, or large language models (LLMs). A notable limitation was the overlap in data use: the MIMIC-IV dataset served both as training data for the Claude Sonnet 3.5 LLM and as the basis for testing synthetic reports and impressions, contravening best practices that require strict separation of training and validation datasets. The clinical accuracy of the "Auto Impression" feature was not assessed

due to time constraints and the volume of data, restricting insights into its safety and real-world performance.

The LLM judge demonstrated challenges when evaluating low-quality human-written text. Resulting in inconsistent scoring, and its assessment methods needing refinement to better identify clinically significant errors. While the LLM judge showed promise, the findings must be validated by human experts to ensure reliability, particularly in comparing synthetic and real data and detecting errors. Finally, the evaluation primarily focused on the quality of summarisation rather than clinical correctness, meaning the potential for harmful errors in generated impressions remains insufficiently explored**.**

SmartGuideline

The SmartGuideline model supported only limited simultaneous users, and clinicians stressed that speed is essential for decision support tools designed for point-of-care use.

The short timeframe and limited resources of the AI Airlock programme meant the team were unable to test economic evaluation, whether this approach works across different AI models, or compare the tool with manual clinical workflows. Planned semantic similarity testing using GPT-4o was also deferred due to cost but is intended for future work.

The clinician sample was skewed towards trainees, which may limit generalisability. More experienced practitioners may interact with AI outputs differently, affecting results and revealing different patterns of trust or bias.

OncoFlow

Explainability testing of the OncoFlow AI models revealed several important challenges. The data extraction model, which pulls information from different clinical documents, sometimes produced unclear outputs when those documents contained conflicting or incomplete data. For example, if a radiology report suggested metastasis but the pathology report did not confirm it, the AI might struggle to interpret the case. When key information like tumour grade was missing, the system chose not to infer values, instead marking them as “Not Reported.” While this avoided incorrect assumptions, it also reduced the completeness of summaries.

The treatment matching model information retrieval from clinical guidelines was one of the main challenges. If documents were poorly structured, the AI could miss key recommendations. In cases where retrieval failed or data was sparse – such as rare tumour types or experimental treatments – the model risked generating confident but unsupported suggestions. To reduce this risk, uncertain outputs were flagged for human review, rather than presented as clinical recommendations. Stakeholders also suggested allowing the

system to "learn" from difficult or unusual cases discussed in multidisciplinary team (MDT) meetings.

Several mitigation strategies were proposed. These included traffic-lighting low-confidence outputs, improving how guidelines are broken into searchable sections, and using explainability tools like SHAP and LIME to help users understand how AI decisions were made. Clinicians remained in full control, with the AI acting only as a support tool. All decisions made with AI assistance were logged and traceable, ensuring a clear audit trail.

Understanding which explainability approach would be optimally aligned with end-user needs, particularly in the live, fast-paced MDT settings remains an ongoing challenge. OncoFlow addressed this as part of a formal feedback discussion and simulation session with end-users to further inform regulatory frameworks for AI-assisted decision support systems.

FAMOS

The sample size may not provide sufficient power to detect small effects as a feasibility study. Results may not be generalisable beyond the participating sites due to potential site-specific factors. Limitations inherent to retrospective studies, such as selection bias and missing data, may affect the findings.

.

7. Evaluation Approach and Key Insights

Evaluation Approach

The University of Nottingham were commissioned to independently evaluate the AI Airlock pilot programme, working closely with the MHRA to understand the practices, strategies and challenges of conducting the pilot. The evaluation was led by Professor Carl Macrae. The evaluation was designed to be constructive and continuous, enabling periodic feedback to the AI Airlock team throughout the pilot. This chapter summarises key insights and findings from the independent evaluation, reflecting the MHRA's commitment to ongoing learning and improvement from the AI Airlock pilot.

Multiple sources of data were used throughout the evaluation including observations and attendance at governance board meetings, supervisory committee meetings, simulations and other events. It includes analysis of AI Airlock documentation and event feedback surveys. Evaluation evidence also includes the notes and feedback generated from an in-person evaluation workshop during the Airlock Connect event.

The data also includes detailed set of interviews conducted with many of the participants involved in the AI Airlock. These interviews allowed participants to feedback and report on experiences constructively and anonymously. 29 interview sessions were conducted throughout the process, engaging with a total of 25 individuals who were involved in the AI Airlock in different roles. These interviews were primarily conducted in two phases: towards the end of the testing design work, and towards the end of the AI Airlock pilot. Three focus groups were conducted with the MHRA AI Airlock team, supplemented by individual interviews with each team member. Interviews were conducted with each candidate organisation at two time points and included two members of each AI Airlock candidate organisation. Wider stakeholders were interviewed including representatives from Team AB, the governance board, the supervisory committee, and attendees at the simulation workshops.

A key commitment of the formative evaluation is to create a safe and protected space to gather feedback and insight from all of those involved in the process of the AI Airlock. All feedback has therefore been suitably anonymised during the evaluation process.

Insights have been themed into four broad areas;

- Motivations, Expectations and Experiences,
- Engagement, Exploration and Collaboration,
- Outputs, Impact and Value and
- Management of the Sandbox.

Evaluation Findings and Key Insights

Motivations, Expectations and Experiences

Expectations going into the novel pilot AI Airlock varied. Candidates had different assumptions but were all motivated by the aim of deepening their understanding of regulation, regulatory challenges whilst engaging with regulators and technical experts directly in a meaningful way. Many candidates shared a commitment to safety and risk management with a strong commitment to build safe technologies and systems whilst utilising the opportunity to engage with the regulator on shaping future regulation in this sector. Motivations were often linked to organisation and strategic goals, including building regulatory knowledge, developing robust assurance processes, and building organisational legitimacy. Candidates indicated that the regulatory sandbox model allowed them to engage in a unique process of regulatory science, by rigorously and scientifically generating new knowledge in a challenging and novel area with the regulator.

The application process was seen as fair, as was the emphasis in the selection process on 'Airlock readiness' which included access to data sets, virtual environments and real-world deployments, as required for the testing plan execution. This emphasis was driven by the need to identify candidates that could participate effectively with the time constrained pilot. It was also evident that these candidates were already well developed in their thinking and capability to address a regulatory challenge and, if there had been more time available for the pilot testing, would have welcomed more ambitious projects.

Some candidates found the development of testing plans more time-consuming than originally expected. All understood the need to focus on clearly defined regulatory challenges given the compressed timeline. Clearer guidance on the expectations of the important stages of candidate onboarding would enhance the experience for candidates.

> "Yes, it's difficult and time consuming, but I suppose the second part is it's something we need to do anyway and it's sort of gives us a focus on where we spend that time"
>
> (candidate interview)

Time and resource demands were challenging and at times constrained candidates' ambition for the depth of testing. They were broadly considered manageable and proportionate to the benefits of insight, learning, and influence on future regulatory thinking. It was highlighted that the testing and development activities carried out in the AI Airlock were well aligned with activities the candidate organisations would have expected to conduct at some point, and the programme accelerated and prioritised this work with associated accountability, resource and time commitments. The AI Airlock regulatory sandbox mechanism can both accelerate important regulatory work and provide structure to maintain focus and attention on otherwise deprioritised issues.

Engagement, Exploration and Collaboration

Candidates and stakeholders consistently reported that the AI Airlock created a constructive environment for exploring and understanding the key regulatory challenges associated with AI medical devices. The AI Airlock process was generally experienced as open, exploratory, and highly collaborative. Candidates were actively engaged, proactive and aligned with the programme objectives throughout. Sandbox activities were experienced as genuinely collaborative, and this approach which was viewed as essential to delivering a successful AI Airlock pilot.

Candidates particularly valued the range of expertise and stakeholder input facilitated by the AI Airlock, though many would have welcomed closer and deeper interaction with the broader expert network, particularly through in-person or direct conversations. The diversity of perspectives was widely recognised as one of the key strengths of the AI Airlock. Candidates felt that the short timescales limited the depth of exploration, and several candidates felt the involvement had prepared them for more in-depth work in future but had not provided the length of time or exploration that was truly needed to develop their analyses and address the regulatory challenge fully.

Simulation events were understood as a learning opportunity given the high stakeholder engagement involved. Candidates and stakeholders felt they were the centre piece of the testing noting these events were especially useful because they enabled candidates to present detailed work and explore practical cases with diverse stakeholders and experts.

Features such as real-world examples, small-group discussions, and relevant expertise were viewed as being particularly supportive of learning. The importance of balancing the breadth and depth of the simulation workshops was highlighted. As was the need for carefully curating the knowledge, expertise and representation for each activity, ensuring there is space for everyone to have an effective voice and avoid discussion tangents.

> “With some workshops… it seemed to be a very open invitation. There were a lot of people, but not all people were contributing. So again, being clear about — do you want people who are interested, or just curious about it? Is that the right thing for them to be invited to these workshops? Or is it that you want people who are representing different groups? Because if you have too many, then maybe some people had relevant things to say, but maybe they didn't” (Stakeholder interview)

There was a general preference for in-person, or single modality, meetings to support learning and relationship-building, though it was appreciated that the compressed timelines made this challenging.

Outputs, Impact and Value

The regulatory challenges addressed within the AI Airlock were viewed by candidates and stakeholders as important and relevant, and represented priority areas for regulatory attention. It was understood that the scope of these problems was necessarily constrained by the AI Airlock's tight timelines, meaning there was a focus on feasible aspects of these problems. Participants recognised this as a pragmatic trade-off. Some stakeholders believed that it would be valuable to define a key focus area or agenda for future Airlock phases to clarify a set of priorities from the outset and to inform the cohort design. Future Airlocks could be streamlined by supporting the 'economies of attention' that can result from all candidates focusing on a particular technology, application or set of closely related regulatory challenges. It was also noted that there may be benefit in bringing other regulators across the system into discussions in a more detailed and structured way, noting that the challenges often span multiple regulators.

There was widespread agreement that the key outputs of the AI Airlock should be regulatory guidance that is practical, public, and informative. Outputs should take a range of engaging formats and both articulate future regulatory principles and share the regulatory sandboxing methods and insights so that these can be learnt from and built on by others. Stakeholders indicated that the outputs needed to balance being general and widely applicable, while also being detailed enough to move the field forward and contribute to a clearer understanding of the regulatory problems being addressed. Interest was shown in the sharing of the key regulatory challenge areas that the MHRA may look to prioritise and transparency of the current regulatory gaps that the MHRA would be addressing in future Airlocks or change programmes.

> "…the goal was to expose the challenges and needs, and it did so" (Candidate interview).

Candidates reported a range of broader benefits of their involvement in the AI Airlock, including increased knowledge and deeper understanding around regulatory gaps, strengthened approaches to assurance and validation, and secondary gains such as raised public profile and enhanced legitimacy. Candidates noted that engaging in the sandbox had helped their organisations expand and develop their thinking and practices, particularly regarding the developing systematic approaches to safety testing and evaluation.. The AI Airlock testing allowed them to learn more clearly about which problems they should be approaching and how clarifying these questions and the areas of uncertainty, with a wide range of experts and stakeholders, was viewed as beneficial.

<u>Management of the Sandbox</u>

Managing and organising the AI Airlock required a broad skillset and at times involved managing an intensive workload. The AI Airlock team was able to effectively embed information sharing mechanisms, ongoing cycles of learning and iterative improvements throughout the process. This iteratively improved processes and created a collaborative

environment that facilitated mutual learning among all stakeholders. Several stakeholders and candidates highlighted the AI Airlock's openness to learning and improvement and emphasised the constructive sense of experimentation that was perceived to exist throughout the process.

> "often what regulators do is, they know the law and then they will paraphrase law in all their guidance… and that's not that helpful, it doesn't really fill in the grey areas, and it's a very safe playing field for the regulator, and what's nice about this is that we can actually venture into the grey, and think about, ok, what is a good method?" (Candidate interview)

The AI Airlock team needed to carefully manage and facilitate relationships between stakeholders and candidates, and there was a focus on ongoing two-way learning between the Airlock team and candidates, with support and input from expert stakeholders. Significant effort was put into building and maintaining relationships with key stakeholders such as Team AB. As with many regulatory sandbox initiatives, there were inherent tensions in balancing working on specific cases and organisations, with the wider aim of generating broadly applicable regulatory insights. Managing these tensions of critical independence and constructive guidance is key to an effective regulatory sandbox.

> "We have an Airlock firewall between us and the companies in the form of the committee, in the form of the Airlock group themselves, which I think has been important that's allowed us to maintain independence" (Stakeholder interview)

These findings illustrate the importance of the MHRA's AI Airlock management approach and the styles of interaction that supported and encouraged learning, that should be carried forward into future Airlocks: building a space for learning within a regulatory sandbox requires regulators themselves to acknowledge and engage in learning processes and acknowledge and accept uncertainties.

8. Policy Implications and Recommendations

Approach to recommendations

The following recommendations draw on findings from the AI Airlock pilot and are intended to inform the MHRA's development of targeted regulatory guidance for AIaMD, as well as contribute evidence to the National AI Commission. They reflect recurring challenges observed across multiple projects and highlight areas where existing regulation could benefit from greater clarity.

Each recommendation was developed through a structured process that involved:

- Assessing the underlying evidence and the level of confidence in that evidence – for example, insights supported by candidate testing data from virtual or real-world settings, and by simulation workshop discussions had high confidence.
- Determining whether the recommendation related to updates in existing policy or identified an area for future research or collaboration with AI Airlock partners.

The recommendations cover key regulatory challenges, including:

- AI generation and validation of synthetic unstructured text data
- Managing risks associated with AI inaccuracies such as hallucinations and omissions
- Reducing misuse and mistrust through explainability
- Mitigating patient harm through proactive monitoring and intervention

Addressing these areas will support the MHRA in advancing a regulatory framework that promotes the safe, effective, and trustworthy development and deployment of AIaMD.

Product development: Data validation

- Develop guidance on expectations for good quality synthetic data and how it should be assessed during development and regulatory evaluation. Guidance is also needed on the responsible use of synthetic data in the validation of AIaMD. This should include outlining appropriate use cases, minimum quality requirements for synthetic data in terms of fidelity, utility, and relevance of data features. These validation techniques may also be applicable to verifying reliability of text-based LLM outputs.

- Introduce guidance for the use of LLMs in training and testing AIaMD. Consider covering best practices for using open-source LLMs and the quality of prompts used during development. Guidance should outline approaches for prompt safety, for example having pass/fail criteria for prompts in case significant clinical errors have been generated by the LLM (rather than only scoring the quality of the outputs).

- Form a position on approaches to evaluating the quality of text outputs, including considerations when outputs statistically significantly differ from human written reports but have been scored as better than human written reports (mathematical or human).

Risk management: hallucinations and omissions

- Provide clear guidance on evaluating and mitigating LLM risks such as hallucinations and omissions in AIaMD. This should include support for end users with clear information on identifying and reporting safety events. Guidance should also emphasise the importance of PMS, including mechanisms like the Yellow Card scheme and MORE as hallucinations may be rare during pre-market testing.

- Encourage, in guidance, safety measures such as retrieval augmented generation (RAG) given the potential to reduce hallucinations and support safer, more explainable outputs. As part of this guidance, it is important to highlight the importance of holistic safety testing after any model adjustments in case other errors, e.g., omissions, increase.

Risk management: Explainability

- Emphasise in regulatory guidance the importance of developers adopting context-aware explainability approaches that account for the clinical question, associated risk and time constraints on the end user. Different stakeholder groups may require different types of explanations: global explanations may be most useful for regulators and policymakers while local case-specific explanations may be more important for clinicians and patients.

- Strengthen in guidance transparency and training requirements for AIaMD and warn against the risks of insufficient user training. Structured outputs such as confidence indicators or other visual aids should be encouraged. Context-specific clarity around the human-AI teaming expectations should be promoted to improve shared decision-making. Training should be implemented with clear accessible user-facing communication about AI-specific risks.

- Emphasise, in guidance, the importance of explainability in conjunction with good machine learning principles and other safety and risk management approaches. Where appropriate, AI models should communicate both what has informed a decision and what has been excluded to reduce the risk of automation bias, defensive medicine, and risks to patient safety. Developers can implement various approaches such as model reasoning and model cards to make these factors transparent to the end user.

Post-market surveillance

- Promote greater awareness and use of reporting tools such as MORE and Yellow Card among clinicians, patients, and the obligations on manufacturers. This should be supported by educational initiatives and stronger collaboration between manufacturers and healthcare organisations to improve reporting and ensure PMS data are complete and meaningful.

- Provide guidance which supports the development and implementation of continuous monitoring for AIaMD. These mechanisms should track key metrics to detect issues such as automation bias. Proactive monitoring enables early detection of safety concerns and should be accompanied by clear follow-up protocols for investigation and action.

- Develop guidance which provides examples of approaches for threshold setting for continuous monitoring. Manufacturers should determine thresholds through a socio-technical approach, combining clinical and technical expertise which balance early warnings with minimising false alarms. A tiered system may be useful for triggering additional monitoring before requiring intervention.

- Emphasise the importance of follow-up protocols for threshold breaches in the post-market surveillance process. Monitoring alone may be insufficient to ensure safety; breaches should be investigated and where appropriate acted upon. Guidance needs to be clear on regulatory requirements for reporting preventative and corrective action taken.

9. Conclusion and Next Steps

The AI Airlock pilot explored the potential and the practical challenges of developing and deploying AIaMD. Our synthetic radiology case study demonstrated that large language models can generate diverse, clinically relevant text data, provided that rigorous automated and human evaluation methods are employed to confirm data fidelity, clarity, accuracy and completeness. The standardisation of these evaluation methods needs to be outlined in guidance.

Our data extraction and treatment-matching studies demonstrated that retrieval-augmented architectures and rule-based prompting can be effective while maintaining high accuracy on complex tasks. Holistic testing helped to minimise omissions and other safety risks. Explainability mechanisms enhanced user trust and understanding, though these must be complemented by ongoing user education and continuous monitoring. Our real-time surveillance case study demonstrated that a monitoring dashboard can be effective at identifying trends in data quality, model drift and automation bias, enabling early intervention before patient harm could occur. Our regulatory gap analyses and simulation workshops identified key areas where AI-specific supplementary guidance is needed to strengthen existing frameworks.

Through the AI Airlock pilot, we have identified a series of key challenges that illustrate the novel regulatory considerations posed by AI in healthcare. This work has culminated in 40 recommendations, covering both areas for immediate implementation and those requiring further exploration. These findings will inform the AIaMD guidance currently in development, as well as future MHRA policy.

The AI Airlock directly supports the work of the National AI Commission in establishing a credible, international regulatory framework for the safe and effective use of AI in health. The pilot will provide the practical evidence base and operational insights needed to address complex regulatory questions.

Building on this foundation, the next phase of the Airlock will generate further evidence and test solutions to these challenges. Its findings will be fed directly into the Commission’s work, ensuring that the Commission’s recommendations are informed by robust, real-world evidence and implementation experience.

Plans for the next phase

The government confirmed in its Regulatory Action Plan on 17 March 2025 that the AI Airlock will continue into Phase 2 for the FY 25/26.

Our strategy for Phase 2 will have two main missions: Unlock and Expand.

Our Unlock mission includes the sharing of the learnings from the pilot phase, both regulatory and the use of sandboxes. We are also developing a programme of activity to boost the impact of the insights in the sector alongside policy changes.

Our Expand mission involves our work with a new cohort and looking to identify areas across the MHRA, healthcare academic and international regulatory sectors for collaboration opportunities to expand the scope of regulatory challenges the AI Airlock can investigate.

Phase 2 will run until end of March 2026. We will work with a new cohort of candidates and test a new set of regulatory challenges for these innovative devices to improve regulations, guidance and, ultimately, patient outcomes.

Acknowledgements

The MHRA would like to acknowledge the hard work of the pilot programme candidates at Philips Healthcare, AutoMedica, OncoFlow and Newton's Tree. Each organisation dedicated their time and motivation to the pilot testing phase and their experience will not only shape the next phase of the AI Airlock sandbox but also inform future guidance and policy within the MHRA.

We would also like to thank our expert stakeholder network, starting with the members from the Governance Board and Supervisory Committee. Representatives within these groups have included throughout the pilot, Royal Derby Hospital, Department for Health and Social Care, NHS AI Lab, Brunel University, University of Birmingham, Health Research Authority, Information Commissioner's Office, Team AB, Scottish Government, Health Technology Wales, Department of Health Northern Ireland, National Institute for HealthCare Excellence (NICE), University of Nottingham, University of Oxford, and across the MHRA.

Thank you to our partners within the Team AB Software and AI working group who have been open to a new way of working throughout the AI Airlock regulatory sandbox, responsive and engaged when asked to support throughout the testing and regulatory mapping activities. We look forward to continuing to work with these partners in the future phase.

Thank you to our pilot phase sponsors at the NHS AI Lab team and the Department of Health and Social Care.

10. Appendices

AI Airlock Candidates

Philips Healthcare: Results

Linguistic similarity

To make sure that the synthetic reports maintained the same linguistic style as original reports, Philips evaluated linguistic similarity using the following metrics: GLEU, ROUGE, Parts-of-Speech Overlap and Automated Readability Index (ARI).

Table s1. Linguistic similarity metrics

Metric	What it measures	What it means for comparing radiology reports	Limitations
GLEU	Balanced n-gram precision & recall	Could be used to measure sentence-level similarity of real and synthetic reports	Sentence-level metric, not informative for texts with multiple sentences and no direct sentence-to-sentence mapping
ROUGE-L	Longest common sequence (LCS) overlap	High ROUGE score indicates that reports have similar phrasing but slight variations	Doesn't capture deep semantics
POS Tag Overlap	Part-of-speech usage	Ensuring similar grammatical structure	Doesn't capture semantics
ARI	Automated Readability Index (ARI)** estimates understandability of English texts based on evaluating chars/words and words/sentences ratios.	Similar ARI scores indicate that real and synthetic findings have similar length of words and sentences => have similar writing style.	Non-robust with respect to slight variations in punctuation and numbers

For each report, the experts scored the following points: Clinical Content Similarity, Linguistic Style Similarity, Understandability compared to the original report.

Evaluation dimensions of the Impression section using LLM as a judge

The Impression section was evaluated against four metrics: accuracy (faithfulness to the content of the findings), completeness (coverage of important findings), clarity (readability and interpretability of the test), and groundedness (ensuring statements are justified by underlying findings).

Accuracy was measured as alignment between “Findings” and “Impression,” quantified on a 0-10 scale, and justified with structured text explanations.

- **Definition of accuracy:** whether the IMPRESSION section correctly reflects the key findings and conclusions from the clinical report.
- Scoring method: LLM judge was asked to give a numerical score from 0 to 10.
    - A score of 0 is required if there is no IMPRESSION section at all.
    - Other scores (1–10) depend on how well the IMPRESSION matches the findings.
- Justification: the LLM judge must provide two kinds of written reasoning:
    - "motivation" ➔ general reasoning.
    - "scoremotivation" ➔ reasoning tied to why that specific score was given.
- Format: output must be structured JSON with the fields "motivation", "scoremotivation", and "score", following strict formatting rules.

Completeness was measured as coverage of all relevant findings in the IMPRESSION, expressed on a 0-10 scale, with structured justifications.

- Definition of completeness: whether the IMPRESSION section contains all relevant information from the findings and does not leave out any critical details.
- Scoring method: the LLM judge must assign a numerical score between 0 and 10.
    - A score of 0 must be given if there is no IMPRESSION section at all.
    - Scores 1-10 reflect the extent to which the IMPRESSION covers all key details.
- Justification: the LLM judge must provide written reasoning in two parts:
    - "motivation" ➔ overall explanation.
    - "scoremotivation" ➔ specific justification for the chosen score.
- Format: output must be structured JSON with the fields "motivation", "scoremotivation", and "score", following strict formatting rules.

Clarity was measured as how understandable and well-written the IMPRESSION is, expressed on a 0–10 scale, with structured justifications.

- Definition of clarity: whether the IMPRESSION section is clearly written and easy to understand.

- Scoring method: the LLM judge must assign a numerical score from 0 to 10.
    - A score of 0 is required if no IMPRESSION section is present.
    - Scores 1-10 indicate the degree to which the text is understandable and clearly expressed.
- Justification: the LLM judge must provide explanations in two fields:
    - "motivation" ➔ general explanation of the assessment.
    - "scoremotivation" ➔ specific reasoning for the numerical score chosen.
- Format: output must be structured JSON with the fields "motivation", "scoremotivation", and "score", following strict formatting rules.

Groundedness was measured as the extent to which the IMPRESSION is supported by evidence in the findings, expressed on a 0-10 scale with structured justifications.

- Definition of groundedness: whether the IMPRESSION section is based on, and supported by, the evidence and findings in the clinical report.
- Scoring method: the LLM judge must assign a numerical score between 0 and 10.
    - A score of 0 is mandatory if there is no IMPRESSION section.
    - Scores 1-10 reflect how well the IMPRESSION is tied to the underlying findings (e.g., not introducing unsupported statements).
- Justification: the LLM judge must explain their assessment in two fields:
    - "motivation" ➔ general explanation of their reasoning.
    - "scoremotivation" ➔ specific justification tied to the numerical score.
- Format constraint: assessment must be returned in JSON with the required fields, following strict formatting rules.

Table s2. Example of scoring system for comparative analysis

	Not similar at all	Slightly similar	Moderately similar	Very similar	Extremely similar
	1	2	3	4	5
Clinical Content Similarity					
Linguistic Style Similarity					
Understandability					

In addition, an independent analysis for each report was carried out. The experts scored the following points: Language, Structure, Completeness, Clinical Soundness
The scale for the reports was as follows:

0: Not Professional / Not Expected in Radiology Report
1: Minimally Professional / Below Clinical Standards
2: Somewhat Professional / Needs Improvement
3: Moderately Professional / Meets Basic Clinical Standards
4: Professional / Above Clinical Standards
5: Highly Professional / Fully Expected in Radiology Report