Medicines & Healthcare products Regulatory Agency

AI Airlock Sandbox Phase 2 Programme Report

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

Published June 2026

Contents

Executive Summary

Artificial intelligence is transforming healthcare at pace, creating significant opportunities for patients, clinicians and the National Health Service (NHS). It is also raising complex regulatory questions that existing frameworks, designed for traditional medical technologies, were not built to answer. The MHRA AI Airlock regulatory sandbox was established to create a safe, structured environment for MHRA to work directly with AI medical device developers to understand how current regulation performs in practice, and where there is opportunity to change.

This report presents the findings, insights and recommendations from Phase 2 of the AI Airlock, which ran from April 2025 to March 2026. Phase 2 built directly on the pilot programme (April 2024 to March 2025), deepening the programme's investigation through a structured cohort of seven candidates, three regulatory challenge areas and a programme of expert simulation workshops. Supported by the Department of Health and Social Care (DHSC) and the Regulatory Innovation Office (RIO), Phase 2 received 51 applications from industry and selected candidates representing a broad range of AI technologies, clinical areas and development stages.

Phase 2 was structured around three regulatory challenge areas that the pilot identified as critical and unresolved:

- **Scope of intended purpose and validation**: how to define, maintain and enforce the intended purpose of AI systems that can evolve in functionality and clinical influence over time

- **Predetermined change control plans (PCCPs) and post-market surveillance (PMS):** how to manage iterative updates to AI systems safely and proportionately across the device lifecycle

- **Performance evaluation for AI-powered in vitro diagnostics (IVDs)**: how to assess and evidence the performance of AI-powered IVDs

Key insights from phase 2

1. AI medical devices benefit from a lifecycle approach, where post-market monitoring complements pre-market evidence. Real-world performance is not sufficiently replicated or represented by controlled testing alone. Therefore, it is important for pre-market evidence generation to be designed with realistic deployment conditions in mind, and post-market monitoring should be designed and specified from the outset.

2. Adequate human oversight may look different over a device's lifecycle and may be inadequate to treat as a "fixed" safety control. When AI systems demonstrate consistent accuracy and users become familiar with using them, users may trend toward applying less scrutiny to device outputs. This natural shift in user behaviour can reduce the effectiveness of human-in-the-loop arrangements over time. When developing and designing human oversight mechanisms as a risk mitigation, considering how human interactions with the device may change over time is an important facet of ongoing monitoring.

3. Ensuring a device continues to operate within its intended purpose is important across a product's lifecycle. For generative AI and large language model (LLM)-based products, significant shifts in device behaviour can occur through model behaviour alone. The Airlock found that when LLM-based systems lack active guardrails, devices may begin to exhibit functions beyond their intended purpose.

4. It is important for performance thresholds for AI to be clinically meaningful and statistically detectable. Across all three challenge areas, the Airlock consistently found that statistically significant change and clinically important change are not the same thing. PMS plans and PCCP acceptance criteria and thresholds should be grounded in clinical relevance, including to set risk-proportionate and pragmatic expectations.

5. Products that do not qualify as medical devices are outside MHRA's regulatory remit. However, the Airlock surfaced a broader ecosystem consideration regarding shared interest across developers, deployers, commissioners and professional bodies in ensuring adequate safety and accountability mechanisms exist regardless of device status. Manufacturers are responsible for remaining diligent in maintaining their product's intended purpose, as products may change over time and will need to remain compliant with applicable regulations if the changed product has a medical intended purpose.

Phase 2 programme recommendations

The recommendations follow directly from the insights. Each recommendation is grounded in evidence generated through the Phase 2 case studies, simulation workshops, and working sessions with ACE. The National Commission into the regulation of AI in Healthcare draws on learning from the AI Airlock and other strategy initiatives as part of its broader consideration on how the UK's regulatory framework should evolve as AI becomes more widely used. The recommendations in this report will therefore inform the policy landscape in the UK.

1. **Develop PCCP guidance.** Providing principles-based clarity on the limits of changes that could be included in a PCCP, expectations on the content of a PCCP for AI medical devices, and AI-specific guidance on when changes to a device are significant and

aligned to the broader medical device framework would better enable manufacturers to make responsible use of this change management mechanism.

2. **Update software qualification and intended purpose guidance.** It is important for these foundational pieces of guidance to clearly address the changing device landscape, including AI-specific considerations and adaptive systems, and for the guidance to make clear how they should be applied throughout the device lifecycle.

3. **Develop guidance on performance metrics for AI-powered IVDs.** These products face unique challenges and opportunities that would benefit from specific clarity in guidance. Practical worked examples addressing how and when additional metrics add value, and how analytical performance concepts apply to AI systems that do not rely on physical analyte measurement, would help advance designing successful validation paradigms for AI-powered IVDs.

4. **Consider ecosystem-wide approaches to ensuring responsible deployment and accountability for products with low or no device regulatory requirements.** Risk-proportionate and consistent governance approaches for products that may not meet the definition of a device but may be used in healthcare will help to ensure users and deploying entities have adequate visibility and clear accountability for these products. Manufacturers have the responsibility to provide transparency about their products and users and deployers need to understand how they can use appropriate products safely, regardless of regulatory status.

Programme impact and next steps

Phase 2 was independently evaluated by the research agency Woodnewton, whose findings confirmed strong stakeholder satisfaction with the programme's sandbox methodology, facilitation quality and creation of a trusted space for open regulatory dialogue. The most consistent feedback was a call for the programme's insights to be translated into clearer, system-level regulatory guidance, a priority the MHRA is currently addressing and committed to continue in the next phase. The AI Airlock has demonstrated that collaborative, evidence-based regulatory sandboxing can provide MHRA with the practical insight needed to keep pace with a rapidly evolving technology landscape.

Phase 3 of the AI Airlock, funded by DHSC for three years from April 2026, will build on the foundations of the programme to date with a focus on refining processes to ensure sustainable delivery of long term programme benefits and delivering tangible outputs mapped to specific MHRA actions and policy priorities as the next step in interrogating challenges in the AI medical device regulation. Its priorities will be shaped by the recommendations of the National Commission into the Regulation of AI in Healthcare, expected in 2026, alongside the evidence generated through this report and building upon learnings across prior phases of the Airlock.

1. Introduction and Context

Over the past year the UK Government has strongly reaffirmed its ambition for AI in healthcare, aiming for the NHS to be the most AI-enabled care system in the world.

The Regulation Action Plan (2025) aims to overhaul the regulatory system to enable growth and reduce investment barriers to the UK market. Key actions include tackling complexity and burden of regulation, reducing uncertainty, and challenging risk aversion. Within the Life Sciences Sector Plan the Government commits that by 2030, the UK will be one of the top three fastest places in Europe for patient access to medicines and MedTech. The 10 Year Health Plan "*Fit for the Future"* outlines 3 key strategic shifts for modernising the NHS, ensuring it is sustainable and capable of meeting future health challenges. These shifts move healthcare, from hospital to community, from analogue to digital and from sickness to prevention.

Further, the UK Government's AI Opportunities Action Plan is a national strategy to accelerate AI adoption across the UK to boost economic growth, improve public services, create future jobs and support everyday life. Published in January 2025, it includes recommendations focused on strengthening AI infrastructure, data access, skills, regulation, public-sector adoption and sovereign AI capability. The MHRA is committed to the UK Government's AI Opportunities Action Plan by supporting safe, evidence-based AI adoption in a high-impact public service area: healthcare.

To meet these ambitions and continue to ensure UK patients can safely access the most innovative products on the market, without unnecessary barriers, the MHRA has prioritised the AI Airlock regulatory sandbox programme. Backed by the Department of Health and Social Care (DHSC) and the Regulatory Innovation Office (RIO) in its second phase, the Airlock enables the MHRA to examine the existing framework through collaborative and flexible ways of working that can keep pace with emerging technologies like AI.

AI Airlock regulatory sandbox in context

The AI Airlock was established as a pilot in April 2024 and is designed to investigate and address the challenges experienced with regulating Artificial Intelligence (AI) as a Medical Device (AIaMD). When AI is intended to be used for a medical purpose, it will qualify as a medical device and needs to be regulated under the UK Medical Device Regulation 2002. The AI Airlock is a regulatory sandbox, intended as a safe, controlled space for industry and regulators to explore new, cutting-edge solutions to challenges experienced by innovations that push the limits of the current regulatory framework. Regulatory sandboxes are becoming well established across sectors and the AI Airlock continues to learn from well-established UK

sandboxes (for example the Information Commissioner's Office) and internationally through initiatives including the DataSphere Initiative Global Sandbox Forum.

Following a successful pilot programme between April 2024 to March 2025, the Department for Health and Social Care (DHSC) funded a second phase. The strategy for the second phase of the AI Airlock focused on two goals: 1) to unlock the findings from the pilot by further sharing our recommendations for regulation of AIaMD. These publications will ensure we meet this mission and we are taking forward ways we can increase our reach and accessibility, and 2) to increase the portfolio of the Airlock regulatory sandbox. We worked with a new, bigger cohort of AIaMD candidates and engaged new partners and experts across the sector.

Building Regulatory Science Partnerships for AI Innovation

In 2025 the MHRA established the Centres of Excellence for Regulatory Science and Innovation (CERSIs) to foster collaboration between academia, industry and regulators and to accelerate the delivery of safe innovation in human health. Two CERSIs in Digital Health and AI (CERSI-AI and RADIANT-CERSI) aim to maximise the potential of AI-powered healthcare solutions, including diagnostics and digital health technologies have been working closely with the AI Airlock programme in this second phase, sharing learnings and expertise across these complex regulatory themes.

The National Commission into the Regulation of AI in Healthcare launched by MHRA in September 2025 is advising MHRA on a new regulatory framework for AI in healthcare, with recommendations due in 2026 that will be informed by UK-wide public, clinical, patient and industry engagement. The AI Airlock has engaged with the Commission during phase 2, sharing practical insights from its regulatory sandbox activity. In turn, the Commission's recommendations will influence the design of AI Airlock Phase 3, including its priorities, focus areas and role in supporting implementation of the emerging framework.

In addition, the AI Airlock works closely with the AI Airlock Committee of Experts (ACE), which includes representatives from UK Approved Bodies (ABs), NHS, the National Institute for Health and Care Excellence (NICE), the Health Research Authority (HRA), the Information Commissioner's Office (ICO), the Care Quality Commission (CQC), academia and technical AI experts. Governance oversight is provided by the AI Airlock Governance Board, an expert board comprising clinicians, MHRA experts, including from the Clinical Practice Research Datalink (CPRD), government representatives such as DHSC, Department for Science, Innovation and Technology (DSIT), NHSE, academics and industry representatives.

Purpose of this report

This report covers each of the 7 case studies that were carried out in phase 2 of the AI Airlock, across 3 key regulatory challenge areas, and includes relevant insights and recommendations. The sandbox methodology can be found in Appendix 1 and governance and timelines in Appendix 2. The audience for the report is intended to be broad, with relevance for UK stakeholders, the public, international regulators and partners, and AI medical device developers.

2. Regulatory challenges

Medical devices in the United Kingdom are regulated under a combination of retained EU directives in Great Britain and EU regulations in Northern Ireland. While this framework provides a well-established foundation for assessing safety, performance, and clinical effectiveness, it was not developed primarily for rapidly developing software and AIaMDs and modern, adaptive software and AI systems. As a result, applying the existing framework to AIaMDs introduces several challenges. Many regulatory concepts — such as defining intended purpose, performance evaluation, and change management — were largely developed with devices in mind that are static and functionally deterministic. In contrast, AI systems can be adaptive, probabilistic, and highly dependent on data quality and deployment context.

This creates uncertainty for manufacturers and regulators in interpreting how existing requirements should be applied in practice. For example:

- Defining and maintaining intended purpose for systems that are meant to evolve over time
- Evidencing device performance where pre-market testing may not replicate real-world conditions
- Managing iterative updates without clear thresholds for changes of regulatory significance

Ongoing UK regulatory reform aims to modernise the framework to support innovation while maintaining patient safety. However, at the time of phase 2 of the AI Airlock, there remain implementation challenges between the structure of existing regulation and the practical realities of AI development and deployment. The AI Airlock programme provides a mechanism to test how current regulatory approaches perform in practice, and to identify where interpretation, clarification or adaptation is needed to ensure the framework remains effective for AI technologies.

Phase 2 candidates

Phase 2 of the AI Airlock was structured around 3 key regulatory challenge areas. The portfolio approach to selection allowed the programme to design a thematic cohort, with a tiered approach to a specific challenge, allowing investigation of the same challenge from multiple product perspectives.

Table 1 Overview of the candidates, regulatory challenges and key objectives explored during phase 2 of the AI Airlock

Candidate & Product	Challenge	Candidate Type	Key Objectives
TORTUS by Tortus AI, Ambient Voice Technology	Scope of Intended purpose	Multi-environment	1. Define when TORTUS shifts from documentation (speech-to-text) to clinical decision influence (diagnostic coding/clinical decision support) and determine regulatory trigger for reclassification. 2. Validate LLM outputs for diagnostic impact and ensure safe human-AI workflows.
Nu & Aegis by Numan, AI conversational & monitoring system, weight loss management		Simulation only	3. Borderline Qualification: When does the product move from wellness support to regulated medical use? Which features or claims trigger AIaMD qualification? 4. Evidence & Documentation: What evidence and controls are needed to keep a product in compliance or safely expand the product's scope?
Safe Summarisation by NHSE, Patient Discharge summaries	PMS & PCCPs	Multi-environment	5. How can the Federated Data Platform (FDP) be applied to performance monitoring and PMS activities of AIaMDs, specifically the Safe Summarisation tool? 6. How can FDP support development and implementation of a PCCP for AIaMD?
Eye2Gene, genetic eye disease detection tool		Simulation only	7. How should PMS plans be designed for AIaMDs in rare disease settings, where data is scarce, to ensure meaningful and proportionate performance monitoring metrics? 8. What regulatory considerations arise when developing real-world performance feedback loops for AIaMDs?
DeepX AI by DeepX Health, skin cancer lesion triage tool		Simulation only	9. How should PMS for AIaMDs be designed to monitor performance across different skin tones, ensuring that bias is detected and mitigated? 10. Can data from other jurisdictions be used to support PMS activities in the UK where there is scarce demographic data available in the UK?
PANProfiler Colorectal by Panakeia, histopathology slide review for cancer detection	AI as an IVD	Multi-environment	11. What steps can be taken to establish clear and consistent definitions and metrics for evaluating AI algorithms, considering the variable definitions of positive and negative events that affect sensitivity and specificity results? 12. How can we harmonise assay protocols and implement quality control measures to improve concordance across different labs, given the sensitivity of concordance to ground truth variability?
Octopath, histopathology slide review for cancer detection		Simulation only	13. How can we develop a performance evaluation strategy which will be applicable across various AI algorithms? 14. How can post-market updates and model retraining be effectively managed?

Scope of intended purpose and validation

The intended purpose defines the scope of the deployed product; it is used to determine qualification and classification of the device, which in turn informs the nature and extent of the evidence required to demonstrate its safety and performance.

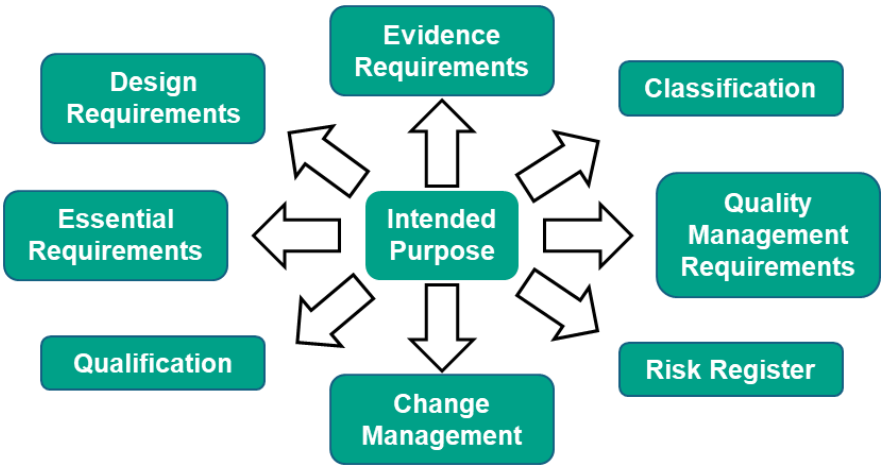


Figure 1. The intended purpose sits at the centre of the regulatory framework, informing qualification, classification, clinical evidence requirements, and post-market obligations.

The existing MDR 2002 outlines intended purpose in broad terms, placing it at the centre of most regulatory requirements illustrated in Figure 1. This is further reinforced by the MHRA SaMD guidance for crafting an intended purpose in the context of software as a medical device (SaMD) and MHRA Software flowchart which clarifies its role as the foundation for risk classification, clinical evidence generation, and post-market activitiesFigure 1. While the guidance recognises software-specific challenges and provides a structured approach to defining intended purpose, it largely assumes a static, well-bounded, and functionally deterministic system.

In practice, this assumption becomes increasingly challenging for modern AI technologies. Despite being considered within the scope of SaMD, MHRA experience indicates that manufacturers can face challenges articulating and supporting a clear intended purpose with evidence and ensuring that devices operate within the limits of that purpose over time. It is important that verification and validation activities account for how AIaMD, particularly LLM-based products, behave in real clinical contexts, where outputs may vary substantially from controlled, pre-market performance depending on data, user interaction, and deployment environment. This places particular emphasis on post-market processes, where surveillance and follow-up activities will need to be able to detect performance drift or other emerging risks over time. At the same time, ongoing change management presents an additional complexity, as updates such as prompt modifications, model retraining, and iterative improvements need to be managed in a way that maintains alignment with the intended purpose and regulatory requirements.

Performance evaluation for AI-powered in vitro diagnostic (IVD) devices

At the time of writing, the regulatory framework for in vitro diagnostic (IVD) medical devices is under reform, and this is expected to lead to future updates to the guidance on the regulation of IVD medical devices in Great Britain. Alongside these regulatory changes, there is increasing emphasis on strengthening how performance is defined, evidenced and assessed, including growing focus on the selection and justification of performance metrics as part of performance evaluation. While these developments aim to improve consistency, transparency, and robustness in regulatory assessment, they are largely based on assumptions that device performance can be characterised using stable, well-defined metrics applied to bounded and controlled systems.

In practice, AI-powered IVDs introduce complexities that this assumption does not fully account for. Performance is often influenced by variability in data inputs, laboratory processes and deployment environments, and may not be fully captured through a fixed set of traditional metrics. This creates uncertainty for manufacturers in selecting appropriate metrics, and for regulators in interpreting whether performance has been adequately demonstrated. As a result, it is not always clear whether existing or emerging approaches to performance metrics are sufficient for AI systems, or how they should be adapted to reflect the context-dependent nature of AI performance. A central regulatory challenge is determining how performance for AI-powered IVDs should be assessed in a way that captures their evolving and context-dependent behaviour. However, it is important to simultaneously build on established metrics such as accuracy, precision, sensitivity, and specificity. This creates uncertainty around what constitutes sufficient evidence of performance over time, including how to interpret reproducibility, variability, and sustained performance in real-world use.

Predetermined change control plans and post-market surveillance

Predetermined change control plans

There is growing international interest in approaches to managing iterative changes to AIaMD, including the emerging regulatory concept of predetermined change control plans (PCCPs). PCCPs aim to provide a structured framework for defining and managing anticipated modifications to software systems over time in a safe and effective manner. The concept involves engaging Approved Bodies on predefined changes during initial conformity assessment, reducing the need for repeated reassessment and streamlining the process for iterative updates. PCCPs also enhance transparency and predictability for manufacturers by clearly defining the scope and boundaries of changes, supporting ongoing safety and

performance across the device lifecycle. This approach supports innovation by enabling more efficient deployment of updates, particularly for AIaMD that may require frequent optimisation.

However, PCCPs are not yet formally embedded within the UK regulatory framework, and their role and application in practice remain unclear. Nonetheless, guidance has been developed internationally and within the UK, including MHRA's guiding principles on PCCPs and through collaboration with international forums such as IMDRF.

Current regulatory approaches to change management are based on whether a modification is considered significant (or substantial) and therefore requires notification or regulatory reassessment. In the UK context, this principle is established in the conformity assessment annexes of the directives, and supporting guidance does exist, most notably the older NBOG guidance on substantial change. However, in practice this framework can be difficult to apply consistently, particularly for software and AI-enabled devices, where the nature of modifications is highly variable and understanding of what constitutes a significant change is still evolving. This is compounded by the fact that many manufacturers, especially those more familiar with EU MDR processes, are more likely to look to newer MDCG guidance than to older directive-based guidance that remains relevant in Great Britain. In addition, Approved Bodies themselves are continuing to interpret how these principles should be applied in the context of emerging AI technologies, which further contributes to the reliance on case-by-case engagement. This creates uncertainty not because the principle itself is absent, but because the applicable framework is fragmented, dated, and challenging to interpret consistently in practice. It also raises broader questions about which types of updates could be managed through a more structured mechanism such as a PCCP, and how regulators can ensure that devices remain within the scope of their original approval while continuing to be implemented safely.

Post market surveillance

Regulatory expectations for post-market surveillance (PMS) have recently been updated through The Medical Devices (Post-market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024, which amend the Medical Devices Regulations 2002 (MDR 2002) by introducing a new Part 4A for PMS requirements. These requirements apply to medical devices, including in vitro diagnostic (IVD) devices and active implantable medical devices (AIMD), within Great Britain. The updated framework places greater emphasis on continuous monitoring of device performance and the use of real-world evidence to identify risks across the product lifecycle. More broadly, manufacturers are encouraged to take a more proactive and systematic approach to monitoring device performance beyond initial approval.

However, as currently structured, PMS requirements assume that changes in device performance are discrete and detectable, with regulatory action typically triggered by identifiable incidents or adverse events such as within guidance for manufacturers on reporting adverse incidents involving software as a medical device under the vigilance system. In practice, some AI systems may change continuously and in ways that are not easily observable, making it difficult to determine when regulatory intervention is required. There remains a lack of clarity regarding which aspects of system behaviour should be monitored, the appropriate frequency of monitoring, and which signals should be regarded as indicative of meaningful changes in performance.

3. Key insights

The regulatory challenges described in Section 2 formed the investigative framework for phase 2 of the AI Airlock. The seven case studies and three simulation workshops were designed to test how these challenges manifest in practice, across real products and real deployment contexts. See Appendix 1 for full methodology. The insights that follow are drawn from that evidence, grounded in candidate data and validated through expert stakeholder engagement in simulation workshops or through technical working sessions with the Airlock Committee of Experts.

Detailed Case studies

The seven case studies generated through Phase 2 of the AI Airlock form the primary evidence base for the insights and recommendations set out in this report. Each case study represents the candidate's own exploration of the regulatory challenges identified for their product, documenting their testing approach, the evidence they generated, and the findings that emerged. They are illustrative of how manufacturers are approaching these challenges in practice.

The insights and recommendations in this report have been developed through MHRA's analysis of the candidate-level evidence, considered alongside broader stakeholder input gathered through simulation workshops and working sessions with the Airlock Committee of Experts (ACE). Where the case studies provided product-specific exploration, this report synthesises what was found consistently across the cohort and identifies the regulatory implications.

The case study reports will be published separately following the release of this programme report. Readers seeking candidate-level detail on the regulatory questions explored, the methodologies applied, and the findings that emerged for individual candidates are directed to those documents when available.

Cross-cutting Insights

Robust pre-market evidence is essential and should be complemented by post-market monitoring in real-world conditions

Robust pre-market evidence is the foundation of confidence in a medical device's safety and performance, but a key question is what level of pre-market evidence is sufficient and meaningful, given the known limitations of what controlled testing can replicate.

The TORTUS case study tested two safety mechanisms in both a controlled virtual environment and in real-world conditions through a shadow-run infrastructure. In offline testing, both performed as designed. In production, the results diverged. The Shell, designed to identify and remove hallucinated content, remained consistently effective across both environments. The Context Guardrail, designed to prevent adversarial prompt injection, significantly reduced scope drift in virtual testing but had no significant effect in real-world deployment, because real users were not attempting to manipulate the system. While the guardrail addressed some foreseeable misuse, it was not reflective of the routine clinical use presented in the case study. This illustrates a broader point raised by the Airlock Committee of Experts (ACE): manufacturers may assume that synthetic or adversarial testing is sufficient to demonstrate real-world validation, but risk optimising for conditions that rarely materialise in practice while missing the variability that does.

Candidates noted that third party foundation models can also introduce functional evolution that is difficult to map to discrete change events, meaning ensuring that device performance remains constrained to the intended purpose becomes more difficult to manage and requires ongoing monitoring. Stakeholder workshops also reinforced the need for performance metrics with clinical grounding, not merely statistical framing, and the importance of real-world testing and monitoring as a standard expectation.

Taken together, these findings suggest that post-market monitoring is most effective when designed to complement rather than follow from pre-market testing. Manufacturers should consider the appropriate balance of evidence generation, understanding that pre-market testing may not be sufficiently suited to demonstrate the long-term, actual performance of the product in practice. At the pre-market stage, manufacturers can usefully design validation studies with real-world deployment in mind, identifying which assumptions about user behaviour, data quality, and clinical context may not hold in production, and specifying from the outset how those assumptions will be monitored.

Human oversight is a lifecycle variable with regulatory significance

Human review of AI outputs featured regularly across the Airlock challenge areas as part of manufacturers' risk controls. The evidence raised a consistent question: how can manufacturers and deployers be confident that human oversight is providing the assurance expected, particularly as use patterns may change over time?

TORTUS achieved a real-world precision score of 0.989, and workshop delegates noted that users may reduce scrutiny of outputs over time in response to consistent and reliable product performance. As a product becomes more trusted, human review may become less rigorous, making residual errors less likely to be detected. Further, these user trends are also unlikely to be uniform across domains. It is therefore important to consider the dynamic and variable nature of user interactions with a product, particularly when user oversight may be intended

as a risk mitigation measure. When monitoring for changes in oversight quality, monitoring plans will be more effective if they are designed to consider the type and severity of potential errors, not only their frequency, as well as other factors that may affect the potential impact of risks, including where and how human-in-the-loop steps intersect with device function. Assessing oversight in practice benefits from multidisciplinary expertise and input from both the AI development and clinical communities.

However, it is important to consider AI performance and potential errors compared to clinician error rates in the same domain. Some ACE members noted that for automated transcription a performance better than 1 error in 100 might be desired, while clinicians acknowledged that their own transcription accuracy cannot be guaranteed to exceed 99% and, further, that error severity is often only assessable retrospectively. Therefore, the relevant comparator for AI performance should be considered in context of current practice in the same domain. A framework combining error frequency with severity was proposed as preferable to a single precision metric. Stakeholders suggested this should be tabulated consistently with DCB and ISO standards. In addition, delegates noted that performance metrics are most effective when grounded in both clinical and statistical reasoning.

During simulation workshops, DeepX AI explored the deliberate version of oversight reduction: whether a PCCP could support staged reductions in clinician confirmation of AI outputs, while Safe Summarisation raised a less structured version: review depth declining as users become accustomed to high-precision outputs without any deliberate system change. In either case, behavioural signals such as review time and edit depth were considered as potential indirect indicators of engagement quality.

Findings suggested that it is important to consider human oversight as a variable prone to change over the device lifecycle and, therefore, monitoring to detect meaningful changes in behaviour can be helpful to understand realistic device use over time. This may be particularly helpful when human-in-the-loop functions are intended as explicit risk controls and should be part of reassessment when performance or effectiveness changes just as other relevant risk controls would also be re-evaluated. It was also found that overarching performance metrics should consider both current clinical performance and statistical reasoning to develop proportionate performance expectations.

Scope expansion and validation

Intended purpose remains the reference point for regulation, but generative AI systems can make it harder to limit actual device functionality in practice

Intended purpose remains central to device qualification, but for generative AI products, it can be challenging to limit real-world use to the stated purpose. One example is identifying when a wellness product begins to provide medical device functions without being explicitly programmed to do so, or a product similarly exhibiting functionality that could move it from Class I to a higher device classification.

The TORTUS case study showed empirically that without active guardrails, out-of-scope performance was detected in approximately 39% of real-world notes, falling to 20% with guardrails active. This suggests LLM behaviour alone, without appropriate design controls, can generate behaviour outside the product's intended purpose. TORTUS reinforced that while disclaimers function as passive controls; they do not prevent generative AI systems from producing content that constitute functionality outside of the stated intended purpose.

During the simulation workshop, Numan illustrated a similar challenge from a different angle. Features that are each individually low-risk can together create a product that functions differently from the sum of its parts. The combination of clinician-reviewed biomarker data, longitudinal memory, and condition-specific personalisation within a clinical pathway shifted the overall impression of the product from wellness to medical device for participants. This is also practically significant because manufacturers develop products iteratively. Features that are low risk individually may together lead to additional risks that would not be anticipated without an impact assessment of the collective changes. As such, periodic reviews that consider feature interactions are more reliable than reviews of individual additions. However, further work is needed to provide clarity on what may constitute effective monitoring and reviews, and the expected response to findings from such monitoring, to ensure continued regulatory compliance.

ACE members also noted that current market entry processes for Class I devices do not include verification of intended purpose statements. Publication of intended purpose statements alongside registration was suggested as a practical step.

Qualification and classification depend on function, context and influence and direct diagnosis cannot be defined by language alone

Assessment of intended purpose was found to be more reliable when grounded in consideration of the whole product than when focused on wording and claims in isolation. Workshop participants treated disclaimers and specific language as meaningful qualification signals when discussing intended purpose in the abstract. However, when the same participants assessed concrete product scenarios with functionality, user experience, and clinical context made explicit, their judgements depended far more on the actual use and functional design of the product than on how it was labelled (Figure 2).

Relatedly, there was persistent uncertainty about what constitutes direct diagnosis for AI systems generating natural language outputs that emerged across candidate discussions and workshop sessions, and a further challenge discussed is whether an LLM may be considered to provide a diagnosis when it gives information that strongly and specifically influences clinical or patient behaviour, even without a definitive diagnostic label. Stating that a user appears to have iron deficiency anaemia may be more clearly diagnostic than noting a low haemoglobin value, but both may direct behaviour depending on user context, follow-up advice, and the specificity of the output. This demonstrates that understanding the presentation of natural language outputs is highly context dependent. As such, it was evident that not only is it challenging to understand an intended purpose based upon claims and language alone, interpreting intended behaviour of a product can be increasingly challenged by nuanced, natural language outputs.

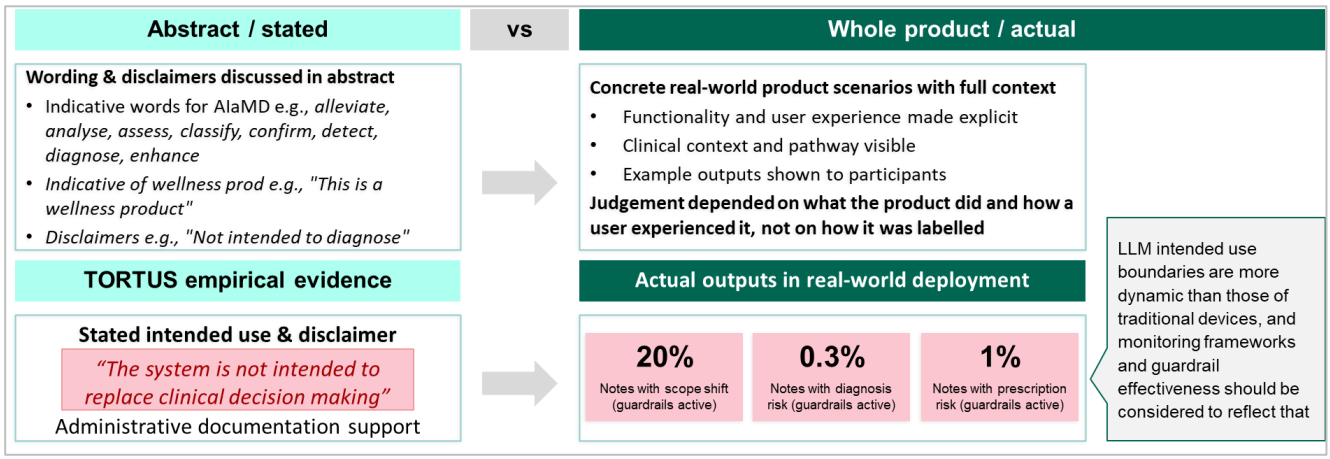


Figure 2 illustrates the shift in participant judgement between abstract and scenario-based assessment of intended purpose during the scope of intended purpose simulation workshop.

The Numan case study provides an additional example of the impact of data presentation and context. Workshop participants drew a consistent distinction between generic support and personalised interpretation of clinical data. General guidance on diet, movement, or habit formation was treated as sitting comfortably within a wellbeing boundary. Perceptions shifted when the product began to explain biomarkers in a patient-specific way. This illustrated that the same type of output may be accepted as “generic” educational information in one context but could be interpreted as potentially diagnostic in another, depending on whether it is personalised to an individual, combined with clinical data, or delivered within a healthcare pathway. Direct diagnosis was therefore not felt by participants to be a fixed property of the output itself but impacted by the intended purpose of the output as demonstrated by multiple factors.

The TORTUS findings add a dynamic dimension that compounds this challenge. A product that is positioned away from direct diagnosis at the point of design can still provide such functions through hallucination. Without active guardrails, some evidence of possible out-of-

scope outputs was detected in approximately 39% of real-world notes. Similarly, the boundary between documentation support and clinical decision support can be crossed erroneously through features of the model behaviour alone. Responses to this challenge include structured provision of clinical context to the LLM, which was associated with improved precision and reduced hallucination, and active hallucination reduction systems, which demonstrated meaningful impact in testing. Neither, however, resolves the direct diagnosis question and the need for further regulatory clarity on this interpretation, but both point toward functional design and workflow choices that can help manufacturers keep products more reliably operating within the bounds of their intended purpose.

Products that do not qualify as medical devices are outside MHRA's regulatory remit. However, it is understood that these products may evolve over time, including potential addition of medical device functionality. The Numan case study illustrated how this can happen: semantic memory could in some configurations impact a product's perceived intended purpose without any deliberate change to its intended purpose statement. More specifically, where an AI system accumulates a longitudinal record of symptoms, biomarker trends, or health behaviours, the distinction between personalisation and monitoring becomes less clear. These two functions may be architecturally identical while potentially being regulatorily distinct, and the shift in practical use and understanding may be invisible to regulators until a product is already widely deployed.

The Airlock surfaced a broader ecosystem consideration regarding shared interest across developers, deployers, commissioners, and professional bodies in ensuring adequate safety and accountability mechanisms exist regardless of device status. The concept of creating mechanisms to help drive consistent good practice and visibility was supported by ACE members. Nonetheless, manufacturers are responsible for remaining diligent in maintaining their product's intended purpose, as products may change over time and will need to remain compliant with applicable regulations if the changed product has a medical intended purpose.

AI-powered IVDs

AI-powered In vitro diagnostics (IVDs) are a rapidly growing category of devices that apply machine learning to diagnostic data, most commonly histopathology. IVDs have distinct regulatory framework and IVD conformity assessment is primary centred around performance evaluation – including analytical performance, clinical performance and scientific validity. Currently, performance expectations are influenced by the assumption that devices remain static and performance will remain stable once deployed. However, AI IVD outputs can be more sensitive to factors such as data quality, laboratory workflows, scanner settings, staining variation, population characteristics, model updates and preprocessing pipelines compared to other IVDs. The AI Airlock programme explored this challenge through the selected candidates for phase 2, Panakeia's PANProfilerColorectal and Octopath, providing the programme with case studies.

Current performance expectations for IVDs remain applicable for AI, though discretion is needed to determine if additional metrics may help provide a more comprehensive profile

The Airlock found that existing performance expectations which include demonstrating adequate sensitivity, specificity, analytical performance, clinical validity, and scientific validity remain important and applicable to AI-powered IVDs. However, there remains an open question on how to apply them appropriately to AI products. AI IVD performance may benefit from additional metrics, as these alternatives could demonstrate more suitable data compared to traditional measures. A device that classifies a biomarker from a digital pathology image may need to demonstrate, among other things, how often it produces a definitive result, how it handles indeterminate cases, how performance varies across scanning and staining environments, how it performs in borderline cases, and how stable performance is over time. The PANProfiler case study showed that the choice between diagnostic accuracy metrics and agreement metrics is not only a technical consideration but also a regulatory one as it depends on whether a gold standard comparator exists. Agreement metrics may be more appropriate where the comparator is itself imperfect.

The Octopath case study added a further dimension to this challenge. Within the simulation workshop, Octopath was described as detecting 32% more mitoses than pathologists, with molecular staining indicating that a high proportion of these detections aligned with biological markers of mitotic activity. These findings were presented as illustrative evidence to challenge existing validation assumptions and to raise an important point for regulatory consideration: where an AI system consistently detects beyond the range of human review, and where that detection appears biologically grounded, the appropriate comparator and the metrics used to assess agreement or accuracy may themselves need to be reconsidered.

The PANProfiler case study also introduced the concept of test replacement rate: the proportion of cases for which the AI device provides a definitive result and could therefore replace additional laboratory testing. While this metric is not part of standard diagnostic performance evaluation frameworks, it is relevant to clinical workflow, operational impact and real-world value. Taken together, PANProfiler and Octopath illustrate that metric selection should be driven by what is relevant to the device's intended purpose, clinical role and user context and that for some AI IVDs, the evidence base required to demonstrate performance may need to go beyond what existing frameworks currently anticipate.

In simulation workshops, participants suggested that implementing a structured justification process for selected metrics, in the context of the specific device’s intended purpose and risks, could help to address these areas.

Real-world variability is already within the IVD regulatory framework, but AI makes its management more complex

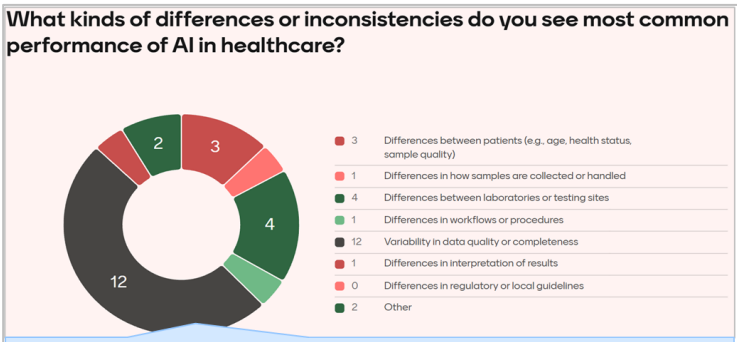
Traditional IVD regulation already recognises that performance can vary across analytical conditions, laboratories and patient populations. What the Airlock found is that AI-powered IVDs can be particularly sensitive to variability in data inputs and deployment environments, and that managing this variability requires more deliberate regulatory handling.

In digital pathology, differences in staining intensity, scanner type, image resolution or pre-processing pipeline can affect what the model detects and how confidently it performs. Workshop participants identified data quality and completeness as the most important source of performance inconsistency for AI IVDs and for AIaMD (Figure 3, left). They also observed that digital pathology does not eliminate technical variation in specimen preparation; in some respects, it exposes that variation more clearly than traditional visual assessment.

Simulation workshop contributions from UK National External Quality Assessment Service (UK NEQAS), National Pathology Imaging Co-operative (NPIC), and National Physical Laboratory (NPL) highlighted the importance of external quality assessment as a mechanism for detecting and managing image quality issues that affect AI performance, and for providing a framework for quality improvement across the digital pathology pathway. The PANProfiler case study reinforced the need for reference datasets that reflect real-world laboratory variability, including variation across NHS sites, borderline cases and demographic diversity. However, there remains open questions around who would be accountable for such datasets and their governance.

Further, datasets used for model development and training serve a fundamentally different purpose from those used for independent benchmarking. A benchmarking dataset used for regulatory assessment needs to be independent of the manufacturer's development process. Similarly, a national benchmarking dataset may not substitute for local deployment validation, and a real-world evidence dataset for PMS purposes needs to reflect the deployment environment rather than a controlled validation setting.

Based on a literature review, the PANProfiler case study proposed a seven-domain framework for a national reference dataset (Figure 3, right), encompassing intended-use alignment, ground truth derivation, real-world laboratory variability, borderline and challenging cases, demographic representativeness, statistical evaluation design, and governance. Workshop discussion added important practical dimensions: ownership, maintenance, curation, open versus restricted access, patient opt-out mechanisms, and how to ensure that datasets remain representative and independent over time.



Domain	Purpose
Intended Use Alignment	Case mix and statistical powering matched to the NHS pathway the product will operate in
Ground Truth Derivation	Transparent, reproducible reference standard with defined scoring thresholds and documented label provenance
Real-World Laboratory Variability	Multi-site inclusion covering variation in fixation, staining, scanners and workflow artefacts – reflects actual NHS practice
Borderline and Challenging Cases	Equivocal cases, low-expression tumours and artefact-containing samples to stress-test robustness at the clinically consequential edges
Demographic Representativeness	Age, sex and ethnicity documented; population distribution reflective of UK cancer demographics to support equitable deployment
Statistical Evaluation Framework	Pre-defined analysis plan with harmonised metrics, confidence intervals and subgroup and site-level analysis to enable fair cross-developer benchmarking
Governance and Oversight	Defined custodianship, multi-stakeholder oversight, transparent access policy and update mechanisms to sustain national benchmarking infrastructure

Figure 3 illustrates two related findings from the AI-powered IVDs simulation workshop. Left: when asked to identify the most significant sources of AI performance variability, delegates ranked data quality and completeness as the leading factor. Right: drawing on a structured literature review, the PANProfiler case study proposed a seven-domain framework for a national reference dataset to support consistent AI IVD validation

The risk of poorly governed datasets is significant. A dataset treated as a definitive standard may become outdated, unrepresentative, or subject to overfitting if used for both development and independent assessment. If used inappropriately, it could narrow innovation or fail to detect performance issues in underrepresented groups or difficult cases. However, there remains opportunity to develop shared infrastructure that improves the efficiency and trustworthiness of AI IVD assessment while maintaining the integrity necessary for regulatory confidence.

AI IVD updates require clearer linkage between performance evaluation, change control and PMS

While PANProfiler colorectal is a general AI-powered IVD and therefore a PCCP would not be applicable, the case study tested directly whether iterative AI-powered IVD updates can be managed effectively under current regulatory frameworks. It compared two versions of their device, v1.0 and v1.1. The update involved dataset refinement with improved quality assurance criteria and model optimisation. It improved performance, but did not change the intended purpose, clinical claim, decision logic or risk class. This created a practical regulatory question that current guidance does not answer clearly: when does a performance-improving update remain a non-significant change?

The case study concluded that v1.1 demonstrated meaningful performance improvement without a change in intended purpose or risk, illustrating a scenario where current regulation can in principle be applied, but where clearer guidance on classification and management of such changes would substantially reduce uncertainty for manufacturers and Approved Bodies. The case study also explored whether PCCPs could provide an appropriate framework for iterative AI-powered IVD updates, concluding that they offer a promising mechanism. However, without clearer expectations, manufacturers may default to full resubmissions for bounded updates, slowing beneficial improvements and increasing regulatory burden. Conversely, under-escalating change could risk clinically meaningful alterations being introduced without sufficient scrutiny.

Explainability for AI IVDs should be designed to support safe use, auditability and scientific validity

The Airlock found that explainability is important for AIaMD and AI IVDs and should complement usability and transparency guidance. Explainability needs to serve specific regulatory and clinical purposes, supporting safe clinical use, enabling regulatory assessment, demonstrating scientific validity, and providing auditability.

Different audiences require different forms of explanation, a concept which was explored in depth during the AI Airlock pilot. Regulators may need global explainability: what the model is designed to do, how it was developed, what features or biological characteristics it uses, what its limitations are, and how its behaviour is controlled and monitored. Clinicians may need local explainability: why a particular result was produced in a specific case, how confident the system is, and when they should consider challenging or overriding the output. Patients may need accessible information about whether AI is being used, what role it plays, and what safeguards are in place.

The PANProfiler case study developed a quantitative explainability framework using tissue and tumour composition metrics to provide auditable context for diagnostic outputs. This approach linked AI outputs to measurable biological features, creating a more structured and verifiable explanation than a purely visual heatmap. The Octopath case study contributed additional perspectives on real-world evidence, performance instability and performance beyond human review. Stakeholders expressed a preference for combining quantitative and qualitative approaches, using global explainability for regulatory assurance and local explainability for clinical use. Explainability requirements were noted as most effective when linked to usability, risk management and scientific validity, and when designed and validated with clinicians and patients from the outset of development rather than as a visual feature at the end of development.

Predetermined change control plans and Post-market surveillance

Post-Market Surveillance and Predetermined Change Control Plans are central lifecycle management tools for AIaMD. Three case studies: NHS Safe Summarisation, DeepX Health, and Eye2Gene generated structured evidence on how these mechanisms should be designed, linked and governed for AI systems. At the time of AI Airlock phase 2 testing and initial reporting, there were no formal PCCP provisions in UK regulations. During reporting, a draft statutory instrument (UK Medical devices (Amendment) Regulations 2026 regulation) was published and contained formal PCCP provisions which applied only to products subject to Approved Body oversight. They do not apply to Class I devices or general IVDs. The insights in this section are therefore most directly relevant to products in Class IIa or self-testing IVDs and above, and the recommendations should be read in that context.

PMS and PCCPs are most effective when appropriately bounded, evidence-linked and integrated with PMS

PCCPs were widely supported by workshop participants and candidates as a mechanism to allow planned, controlled updates to AIaMD without full regulatory reassessment for each change. This is genuinely valuable as AI systems that cannot be updated efficiently in response to drift, performance signals or new clinical knowledge may become less safe and effective over time.

A PCCP is intended to define the specific scope of modifications anticipated, the conditions under which each can occur, the evidence required before and after implementation, acceptance thresholds, version control arrangements, rollback routes and monitoring obligations. Safe Summarisation and Eye2Gene provided detailed operational examples within the Airlock. For example, Safe Summarisation's PCCP covered prompt changes, referential data changes, foundation model changes and monitoring configuration changes,

each with defined triggers, evidence requirements, RAG-status governance and human sign-off requirements.

The Airlock consistently found that PCCPs are most effective when structurally linked to PMS. PMS provides the real-world evidence that should determine whether PCCP-controlled changes can proceed, be paused, or require the PCCP itself to be revised. The Safe Summarisation monitoring framework illustrates this in practice. When a planned modification is initiated, the relevant dimensions are flagged for review and cannot proceed until defined acceptance checks are passed. Where thresholds are not met, the change is blocked. The same performance dimensions monitored continuously by PMS (hallucinations rates, groundedness, completeness) are the same criteria that determine if a PCCP change can proceed, creating a direct procedural link between ongoing surveillance and change management.

At the time of AI Airlock phase 2, the absence of formal PCCP provisions in UK regulations was identified by candidates as a significant barrier to operationalising the approaches described above, and is directly addressed in the recommendations that follow.

Significant change for AIaMD cannot be determined by the type of technical change alone

Across the PMS/PCCP workshop and all three PMS/PCCP case studies, a recurring finding is that the regulatory significance of a change cannot be determined from the type of change in isolation. The same technical modification (e.g., model retrain, prompt revision, workflow adjustment, interface change) may be significant in one context and not in another. DeepX AI illustrated this: a workflow change that reduces or removes clinician confirmation of AI outputs may involve no change to model architecture or clinical logic, but may alter the device's use and risk profile, its classification, and the assumptions underpinning conformity assessment. Workshop delegates confirmed that human factors and workflow changes are often at least as important as technical performance metrics. Safe Summarisation similarly found that the same level of technical change in a foundation model could represent a major or minor PCCP event depending on whether the model is within the same family and whether reasoning architecture differs.

The practical implication is that significance assessment will need to start with intended purpose but extend to function, clinical role, user interaction, level of autonomy, deployment environment, benefit-risk impact and human oversight. A decision framework that makes these contextual factors explicit would support more consistent assessment by manufacturers, Approved Bodies and MHRA alike.

Performance thresholds should be clinically meaningful, not only statistically detectable

PMS and PCCP governance depends on performance thresholds that can trigger meaningful review and action. The Airlock found that threshold-setting is one of the most practically difficult aspects of AI device lifecycle management, and that statistically detectable change and clinically meaningful change are not the same thing.

Eye2Gene illustrated this in a rare disease context where patient populations are small, outcome timelines may be long, and genetic confirmation is inconclusive in many inherited retinal disease cases. Where clinical ground truth is uncertain, reliable PMS triggers are difficult to define, and statistically significant drift may not correspond to clinically important performance change. Safe Summarisation illustrated the challenge in an LLM context: automated metrics including groundedness, completeness, conciseness, readability and toxicity support monitoring but cannot replace clinician calibration, and LLM-as-a-judge evaluation was confirmed to be insufficient as a sole assurance mechanism. DeepX raised the further challenge of subgroup-level thresholds where overall acceptable performance may mask material performance gaps in specific patient subgroups.

From another regulatory challenge area, Octopath, an AI-powered IVD raised the same point. Octopath presented a model with approximately 91% sensitivity but with notable week to week variation, attributed to real-world population shift. Discussions highlighted that statistical performance thresholds alone may not fully capture clinically relevant stability, prompting consideration of how performance variability should be monitored and interpreted.

Subgroup performance improvement to support equity can be enabled by PMS and PCCPs

AI devices may perform acceptably against overall performance metrics while underperforming for specific patient subgroups. These disparities may be invisible at pre-market evidence but become apparent after deployment in a broader or different population. Without clear regulatory pathways for subgroup remediation, known performance disparities may persist without a fast and agile way for manufacturers to correct them so the product can continue benefiting patients.

DeepX AI placed this issue at the centre of its PCCP and PMS investigation. It explored whether PMS-identified subgroup performance gaps, for example in underrepresented skin tones in a UK dermatology dataset, could trigger model remediation within a PCCP. Key questions that were discussed in simulation workshops included: what data sources are acceptable for subgroup remediation where UK population data is limited; how to validate

remediation to avoid improving one subgroup while degrading another; and whether synthetic or international data can be used to supplement insufficient UK evidence.

Similarly, Eye2Gene's Airlock work explored whether PCCPs could provide a controlled pathway for evidence-driven use expansion – for example, validating additional genes as they are characterised – provided the expansion follows a predefined protocol and does not alter the broader intended purpose. This points to a broader need, for knowledge-based AIaMD, regulatory frameworks should consider mechanisms that support controlled updating to remain aligned with current clinical evidence. Octopath raised a parallel question from a different angle. Its underlying AI model demonstrated strong and broadly consistent diagnostic performance, correctly identifying most relevant cases across multiple tissue types, including breast cancer, neuroendocrine tumours and melanoma. This consistency was used to support discussion around whether a model that performs reliably across several cancer types could, under a structured validation framework, have its use extended to new indications without requiring a full revalidation process from the outset. This concept of a pan-cancer or modular approach to validation was presented as a potential mechanism for proportionate evidence generation. While such approaches may be conceptually feasible within a PCCP, the specific approach and requirements to do so successfully require further exploration.

4. Recommendations

The recommendations in this section follow directly from the insights in Section 3. Each recommendation is grounded in evidence generated through the Phase 2 case studies, simulation workshops, and working sessions with ACE. For ease of reference, the key insight driving each recommendation is signposted where relevant.

Cross-Cutting Recommendations

Lifecycle considerations: pre-market evidence, post-market monitoring and human oversight

For consideration by the MHRA:

1. Provide guidance on principles/approaches for evidencing meaningful human oversight of AI systems and highlight when human-in-the-loop approaches are meaningful, how to demonstrate them, and how to evaluate when effectiveness may change over time.

2. Establish how “State of the Art” for key AI technology areas could be defined centrally.

3. Consider additional guidance on effective PMS reporting. In complement to this, ensure sufficient regulatory capacity and resource within the MHRA to meet any additional PMS needs.

4. Yellow card and incident reporting serve a complementary function to PMS monitoring. Consider how vigilance and incident reporting mechanisms can be strengthened for AIaMD.

For consideration by Manufacturers:

5. Propose and validate appropriate human oversight arrangements for specific product and clinical contexts and monitor whether that oversight remains meaningful in practice as part of PMS obligations.

6. Design validation studies with real-world deployment in mind, using real-world conditions wherever feasible. Consider post-market monitoring plans from the outset of design and development and address the performance factors that pre-market testing could not fully validate.

7. Document where training data or validation datasets underrepresent specific demographic groups or deployment contexts and specify how this will be addressed through PMS or PMCF.

For consideration by Approved Bodies:

8. Continue to assess the appropriateness of human review based on the product, its intended purpose and its clinical context, and consider how the adequacy of oversight can be evaluated and maintained across a product's lifecycle as applicable.

9. Assess the adequacy of pre-market evidence in the context of planned real-world deployment, including whether the validation dataset and testing environment are sufficiently representative of the intended deployment context.

Guidance hierarchy and collaborative development

For consideration by the MHRA:

10. Publish broad guidance that explains in practical terms how to apply the regulatory framework to AI medical devices specifically.

11. Publish more detailed application guidance on specific topics, such as PCCP design, significant change assessment and performance metric selection. Some such guidance may be developed in collaboration with or primarily by other entities as appropriate, and the scope and utility of each guidance within the broader policy landscape would benefit from further clarity.

For consideration by Team AB:

12. Develop collective guidance on AI-specific topics, published on behalf of the sector. Such guidance should be consistent with MHRA's legal framework and policy but can provide additional, practical application detail that may benefit manufacturers.

13. Publish best practice examples to facilitate consistency and clarity for manufacturers where possible, in collaboration with the MHRA where appropriate.

Scope, Qualification and Validation

Intended purpose as a lifecycle obligation

For consideration by the MHRA:

14. Clarify further that maintaining device performance consistent with intended purpose is a lifecycle obligation rather than a market-entry declaration and provide AI-specific examples of medical intended purposes. Guidance could address how manufacturers can demonstrate device operation remains aligned with stated intended purpose and what signals should prompt a manufacturer to seek regulatory input.

15. Provide additional resources such as training and education for other stakeholders in the regulatory ecosystem to support consistent interpretation and application of new or revised guidance.

16. Consider whether publication of intended purpose statements alongside device registrations would be a practical step to improve public visibility and enable comparison across products.

For consideration by Manufacturers:

17. Treat intended purpose as evidenced through product design, output constraints, monitoring metrics and human oversight arrangements, not simply by a label applied at market entry. Where a product is developed iteratively, conduct periodic scope reviews that consider the impact of feature interactions, not only individual additions.

For consideration by Approved Bodies:

18. Specify what evidence is needed to maintain ongoing conformity with stated intended purpose across the device lifecycle and publish best practice examples where possible. Consider where example outputs and user journeys should be requested to support qualification assessment.

Qualification based on function and context

For consideration by the MHRA:

19. Update software qualification guidance to address AI-specific features and the ways these functions may be used for a medical purpose. Guidance could address output decisiveness, clinical specificity, user population and the likelihood of influencing diagnosis or clinical management when describing how the nuances of a device's intended purpose may be demonstrated.

For consideration by Manufacturers:

20. Demonstrate product function and purpose through example outputs and user journeys. Qualification cases cannot effectively rely on wording choices or disclaimers alone. Where

products sit near the qualification boundary, engage with MHRA through the Innovation Office before deployment rather than after.

Ecosystem-wide approaches to ensuring responsible deployment

For consideration by all Stakeholders:

21. Consider ecosystem-wide approaches to ensuring responsible deployment and accountability for products that are not medical devices. Risk-proportionate and consistent governance approaches for products that may not meet the definition of a device but may be used in healthcare will help to ensure users and deploying entities have adequate visibility and clear accountability for these products.

For consideration by Manufacturers:

22. Consider how to provide transparency about products so that users and deployers can understand how to select and use appropriate products safely, regardless of regulatory status.

AI-Powered IVDs

Performance metric selection and validation

For consideration by the MHRA:

23. Provide AI-specific guidance on performance metrics and selection for AI IVDs, focusing on worked examples and practical approaches and addressing the circumstances in which additional metrics add value, and how analytical performance concepts can be adapted for AI systems.

For consideration by Manufacturers:

24. Select and justify performance metrics by reference to their device's design, clinical context and intended comparator. Where deviation from traditional metrics is proposed, provide explicit rationale.

25. Implement change management documentation from the outset and design explainability approaches in close collaboration with clinical users.

For consideration by Approved Bodies:

26. Encourage engagement with manufacturers early on validation design, and work to pre-agree categories of change and associated thresholds for regulatory notification. Where

representative validation data is difficult to obtain, consider providing clarity on what may be practical and sufficient.

PCCPs and PMS

For consideration by the MHRA:

27. Develop statutory PCCP provisions and supplementary guidance. To be most effective, guidance should cover PCCP structure covering modification scope, PMS linkage, trigger logic, acceptance criteria, version control and rollback arrangements.

28. Provide clarity on significant change in an AI-specific context, aligned with the broader MD significant change framework, which addresses context-based assessment covering intended purpose, clinical role, autonomy, oversight, benefit-risk and cumulative change.

For consideration by Manufacturers:

29. Design PCCPs from the outset of product development, integrated with risk management and PMS plans, rather than as a retrospective exercise. Manufacturers may want to consider explicit PMS-to-PCCP conditional logic: which monitored metrics, at which thresholds, will enable, pause or terminate planned changes.

30. Justify performance thresholds using clinical as well as statistical reasoning. Where subgroup performance is clinically relevant, subgroup monitoring should be included in PMS plans. Further, monitoring approaches may benefit from including behavioural signals of oversight quality, particularly where human oversight is intended as a risk control.

For consideration by Approved Bodies:

31. Engage early with manufacturers on PCCP design as part of conformity assessment, with pre-agreed changes and thresholds to support efficient and consistent assessment.

5. Evaluation Approach and Key insights

Evaluation Scope and Objectives

An independent evaluation partner, Woodnewton, was appointed to evaluate phase 2 of the AI Airlock programme. The use of an external partner provided independent perspectives, supported objectivity in assessing programme performance, and enabled the programme to draw on specialist evaluation expertise. The scope of the evaluation was to cover all key aspects of the programme.

Figure 4 Pictographic representation of the key areas of focus for the AI Airlock programme evaluation

This evaluation explored the experiences of participating firms, including whether the programme met their expectations and how it could be improved or scaled. It assessed stakeholder and partner engagement, as well as patient perspectives on AI in healthcare, and examined the programme's wider impact across the AI, health and MedTech sectors (Figure 4). It also captured the views of programme team members on delivery experience and feedback, supporting an overall assessment of the programme's effectiveness, maturity and future development needs. There were three core objectives of the evaluation exercise:

- Assess delivering against objectives and areas of improvement: To understand the extent to which the Airlock is meeting its objectives, progressing towards a business-as-usual model, responding to regulatory challenges, building partnerships, and strengthening team capacity and capability.

- Evaluate continual improvement and learning: To assess how phase 2 has incorporated lessons from pilot, improved engagement and processes, and generated learning that can inform sandbox approaches within MHRA and wider regulatory settings.

- Consider future scalability and development: To explore how the Airlock model, ways of working, frameworks and product testing environments could be scaled to respond to evolving needs in AI and health technology.

Evaluation Methods and Evidence Base

The evaluation used a mixed-method approach, drawing on programme documentation, delivery data, stakeholder feedback, participant insights, patient perspectives and programme team reflections (Figure 5).

Figure 5 Methods used in the AI Airlock programme evaluation approach

The evidence base for the evaluation activities comprised learning from pilot, briefing materials on the programme participants and their technologies, topic guides, questionnaires, horizon scanning or challenges and risks, and programme planning documents.

Lessons Learned and Recommendations

Key Strength in Programme Delivery

AI Airlock phase 2 was widely praised for its sandbox methodology, effective programme management, and creation of a trusted space for open, high-value engagement between regulators, industry and wider stakeholders, successfully meeting its objectives to both share pilot learning and explore new regulatory challenges. Survey and qualitative feedback show consistently strong satisfaction with delivery, communication and stakeholder inclusion, alongside clear evidence of organisational learning and capability building within MHRA.

The main area for improvement is strengthening tangible learning on potential regulatory solutions, with stakeholders expecting clearer guidance and regulatory impacts to emerge as the programme progresses into phase 3.

Focus on Simulation Workshops

The Airlock phase 2 Simulation Workshops were a core element of the programme and were widely seen as a valuable forum for open, cross-disciplinary exchange between MHRA, industry and other experts. Survey and qualitative feedback were strongly positive on the quality of pre-workshop information, including the informative baseline presentations set at the start of each workshop, clarity of purpose and the value of joint discussion, though some participants felt that the sessions could be overly theoretical, unevenly facilitated, or dominated by those with stronger regulatory expertise.

Stakeholder quote:

"It's an excellent, innovative programme of work and the team are brilliant – we are really pleased to be involved."

Stakeholders Suggestions for Improvements

Feedback clustered around five themes: improving timelines and process efficiency, reducing administrative burden and improving meeting effectiveness, providing clearer terminology and upfront onboarding, better deployment of expertise, and strengthening the translation of discussions into tangible outputs.

While participants valued the learning, engagement and clarity gained for individual use cases, the most consistent concern was the need for clearer, system-level regulatory guidance and policy change emerging from the AI Airlock. Overall, stakeholders remain supportive of the programme but emphasised that its long-term credibility will depend on demonstrable regulatory impact and clearer outputs in future phases.

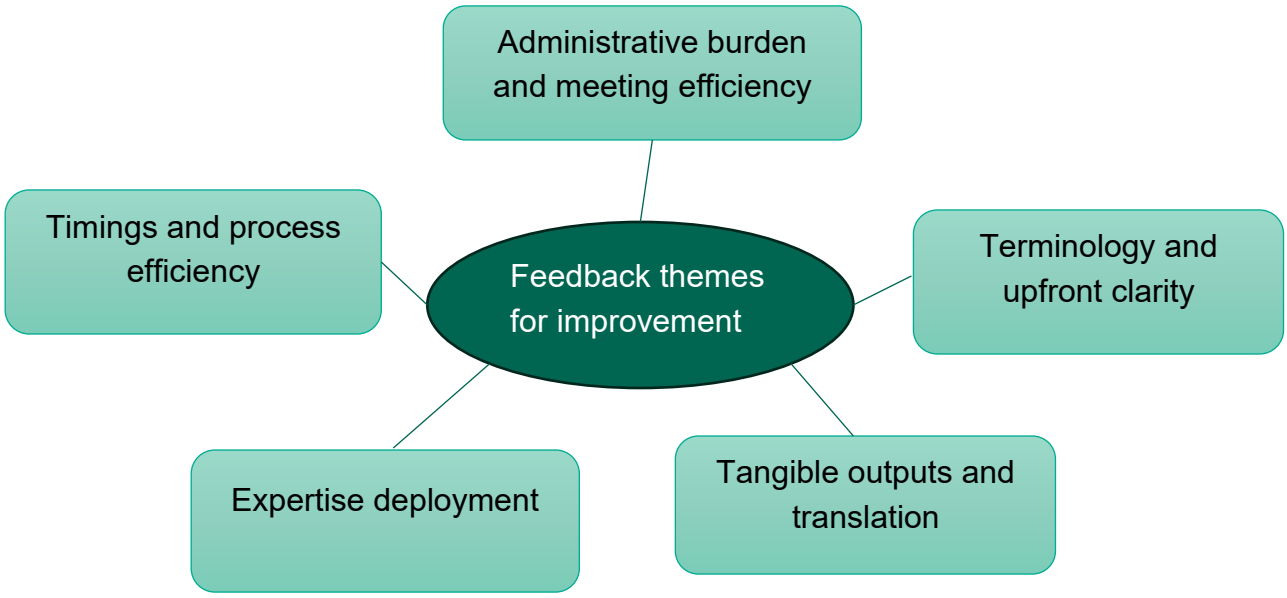


Figure 6 Diagram overview of the 5 key themes of areas for improvement from the Woodnewton programme evaluation report

The six-month phase 2 engagement period created delivery pressures due to the short, focused timing, and managing multiple workstreams in parallel. Airlock technical experts highlighted that testing timeframes between 6-12 months would be more appropriate. Stakeholders highlighted the need to reduce administrative burden and make roles and expectations on time and resource commitments clearer. They also called for clearer terminology and stronger onboarding for innovators and experts. Greater deployment of expertise was recommended, including more embedded technical AI capability, stronger senior sponsorship, and improved patient and clinician engagement. Stakeholders also highlighted the need for greater interaction with other sandbox participants and innovators. A key priority remains translating programme insights into tangible regulatory guidance, policy, and wider sector impact.

Stakeholder quote:

"Great forum for open discussion between various stakeholders. Good to dig into details and get perspectives across different sets of expertise – fantastic that manufacturers, consultants and MHRA are all in the same room to align."

Future Focus Areas

Participants recommended that phase 3 should address both emerging and persistent regulatory challenges and technology, including generative and agentic AI, post-market surveillance, lifecycle management, and a range of cross cutting issues such as evidence standards, cybersecurity and global harmonisation. Many stakeholders emphasised the value

of deeper exploration of learnings from pilot and phase 2 where this would generate new insight, alongside new topics. There was also a clear call for a more strategic, system-level examination of AIaMD regulation and integration, deployment and adoption challenges, recognising the disruptive nature of AI and the limitations of applying traditional regulatory approaches.

The three-year funding for phase 3 presents a significant opportunity to consolidate learning from earlier phases and evolve the AI Airlock model while retaining its successful sandbox approach. Woodnewton's reflections highlight several priorities for adaptation and improvement:

- Experiment with more agile, multi-model engagement approaches, or iterative testing cycles across multiple environments rather than replicating previous formats.
- Simplify and clarify terminology, governance structures and administrative processes.
- Strengthen advisory capacity through a larger, more flexible expert committee and enhanced patient and public involvement.
- Prioritise the translation of Airlock insights into tangible regulatory guidance and demonstrable impact on MHRA business-as-usual practice.

Stakeholder quote:

"The real value will come when the learnings from Airlock are translated into guidance and regulation."

6. Conclusion and Next Steps

The three regulatory challenge areas investigated in phase 2 of the Airlock: scope of intended purpose and validation, performance evaluation for AI-powered IVDs, and lifecycle change management are not isolated technical questions. They are interconnected challenges derived from the same underlying problem: maintaining regulatory confidence throughout the lifecycle of devices that have complex behaviour, outputs and risks can shift over time, intentionally or unintentionally.

The evidence generated through Phase 2 suggests that the existing framework requires targeted clarification, practical guidance and, in some areas, formal statutory development to keep pace with the realities of AI deployment. The programme has identified areas where those developments are needed and tested practical approaches with real products and real stakeholders to develop its recommendations.

In bringing together innovators, regulators, and healthcare providers, the AI Airlock has demonstrated how collaborative, iterative approaches can drive the safe regulation of AI. Its impact extends beyond individual use cases, contributing to a broader shift in how emerging technologies can be evaluated and governed within the UK healthcare system.

The recommendations here represent important input for future policy and guidance development. Translating sandbox findings into durable regulatory change requires sustained effort across the MHRA and other stakeholders, and that translation is one of the defining challenges for Phase 3. The AI Airlock is well placed to support it, with three-year funding secured for the next phase to create the conditions to do so with the depth, continuity and ambition the task requires.

Phase 3

The Department for Health and Social Care has awarded the AI Airlock funding to develop the programme further over the next 3 years. This third phase will allow the MHRA to further develop the regulatory sandbox framework, addressing more complex regulatory challenges, in greater depth, with increasingly innovative technologies and systems.

The AI Airlock signals a step change in how the UK approaches innovation in healthcare, balancing pace with safety, and ambition with accountability. As the programme matures, it will continue to shape not only individual technologies, but the wider ecosystem in which AI is developed, assessed, and adopted. Learnings from this report will influence the coming months of programme refinement and redesign. The programme will continuously embed the lessons highlighted by the stakeholder insights and candidate experiences captured by Woodnewton.

While the key insights from the regulatory analysis and innovator case studies have informed the work of the National Commission this year, equally the programme anticipates the recommendations of the Commission will influence the key priority areas that the AI Airlock may be best placed to address. Alongside the MHRA's upcoming 5-year Strategy, we are well aligned with wider Agency priorities and objectives.

Key pillars for the AI Airlock strategy in the next phase are to continue to refine the programme, become more efficient and fully effective to deliver tangible outputs to benefit the sector, healthcare providers, and patients. We will aim to do this in several ways.

Enhancement: focus on key foundational needs for regulating cutting edge technology that are impacting innovators, regulators, and the healthcare sector and have been surfaced over the prior phases of work. Explore options for maturing and targeting the Airlock's delivery by providing enhanced testing environments and considering the Airlock's positioning to address both integration and access challenges, and international regulatory challenges.

Efficiency: continuously improve the processes and methodology of the regulatory sandbox, balance resources in the most effective way and streamline the workflow while experimenting with more agile methodologies for testing. Learnings from the evaluation reports in key areas including timeframes for testing, information sharing and thematic cohesion of regulatory challenges will be implemented. Take steps to fully embed into business operations and envision the lasting and sustainable role the AI Airlock can serve within the MHRA.

Effectiveness: Maximise the impact of the insights and the outputs of the Airlock testing, addressing a key learning from this phase to prioritise translation of the evidence into policy, resource development, and regulatory change. Use the learning from the prior phases of the Airlock to explore critical challenges further and at depth with purpose.

As the UK's National AI Strategy continues to prioritise innovation, safety, and public trust, the Airlock provides a critical bridge between policy intent and innovation development and design. Its ongoing evolution will not only accelerate future regulatory advances for the UK but also responsible innovation across the NHS, ensuring that the UK remains at the forefront of high quality, AI-enabled healthcare.

Appendices

Appendix 1: Sandbox Methodology

Applications, sifting and selection

The application to join the second phase of the AI Airlock was open between June and July 2025 and the programme received 51 diverse applications from developers.

Most applications were received from micro and small enterprises (76%) with some from large enterprises (8%), universities (8%) and government entities, namely NHS. This is similar in representation to the pilot cohort and reaffirms the value of the Airlock for smaller organisations. Similarly, most applications were from products that were within operational prototype stage or pre-clinical (68%), while post market products made up 14% of the applications, either recently launched or within the first 3 years. The remaining applications were from products in concept or design verification stages.

Represented AI technologies were predominantly Generative AI, Large Language Models (LLMs), Natural Language Processing (NLP), Deep learning and predictive models. AI-driven functions across these products including decision support, risk prediction, diagnosis and information retrieval and summarisation. Most of the products described in the applications were disease agnostic, meaning they could be applied to multiple clinical areas. Other prevalent clinical areas include oncology, mental health, orthopaedic and weight management. Applicants were asked to indicate their specific regulatory challenge as part of the application form and the most commonly selected challenges, of the 22 identified were clinical or performance evaluation (12%), Human AI interactions (13%), post-market challenges (8%), Scope of intended purpose (7%), and international alignment (7%).

Applications were processed similar to the pilot programme methodology, and a longlisting process was carried out against specific eligibility criteria to select applicants to attend a selection interview. The programme team provided a review to finalise the candidates, applying a portfolio approach, and decision by the Governance Board before communicating the offer to the successful candidates.

Portfolio Structure

Figure 7 provides an overview of the tiered approach to the portfolio, working with multi-environment and simulation-only candidates within multiple regulatory challenge areas. This has enabled an increased breadth of learning, from across multiple product types, within the same challenge area, allowing the Airlock to gather a rich picture of learnings. The multi

environment candidates worked with the AI Airlock team to test the regulatory challenge in more than one of the AI Airlock's testing environments, described as follows:

Simulation environment: a focused roundtable workshop bringing together multiple stakeholder perspectives to address simulated testing scenarios

Virtual or Research environment: testing of the AI product in a controlled data environment to address specific research questions and generate evidence

Real-world environment: deployment of a product in the intended purpose environment, separate from the clinical pathways and decision-making processes, to generate evidence of the product's performance without impacting the healthcare outcomes. "Real-world testing" for purpose of the Airlock indicates use of retrospective or synthetic data and is not a clinical investigation.

The Simulation environment candidates participated in the Simulation environment only.

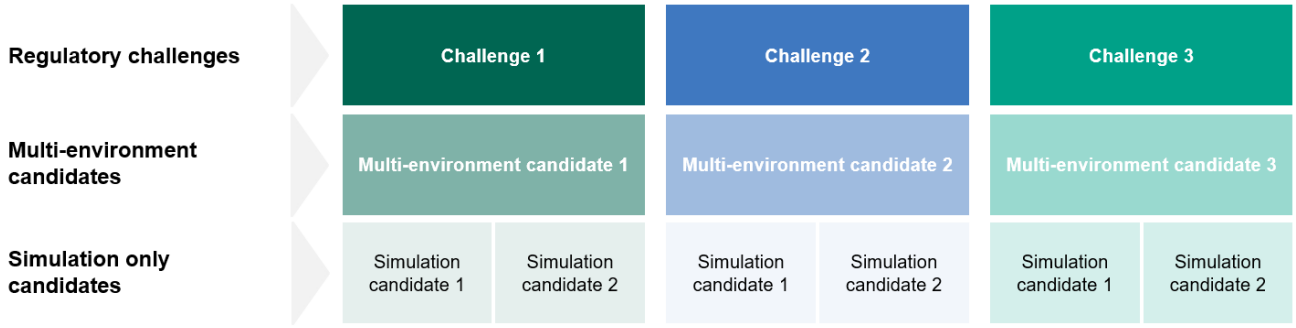


Figure 7 Overview of how the phase 2 AI Airlock portfolio has been structured across multiple regulatory challenges, with multi-environment and simulation only candidates layered throughout the portfolio

Sandbox Testing Approach

Following candidate selection, the programme team then worked to onboard the innovators. As with the pilot programme, the phase 2 projects followed a structured four-step delivery approach. The phase 2 cohort generated 7 case studies, across 3 key regulatory challenges. Three simulation workshops were held in February 2026 and each multi-environment candidate also carried out virtual and real-world testing.

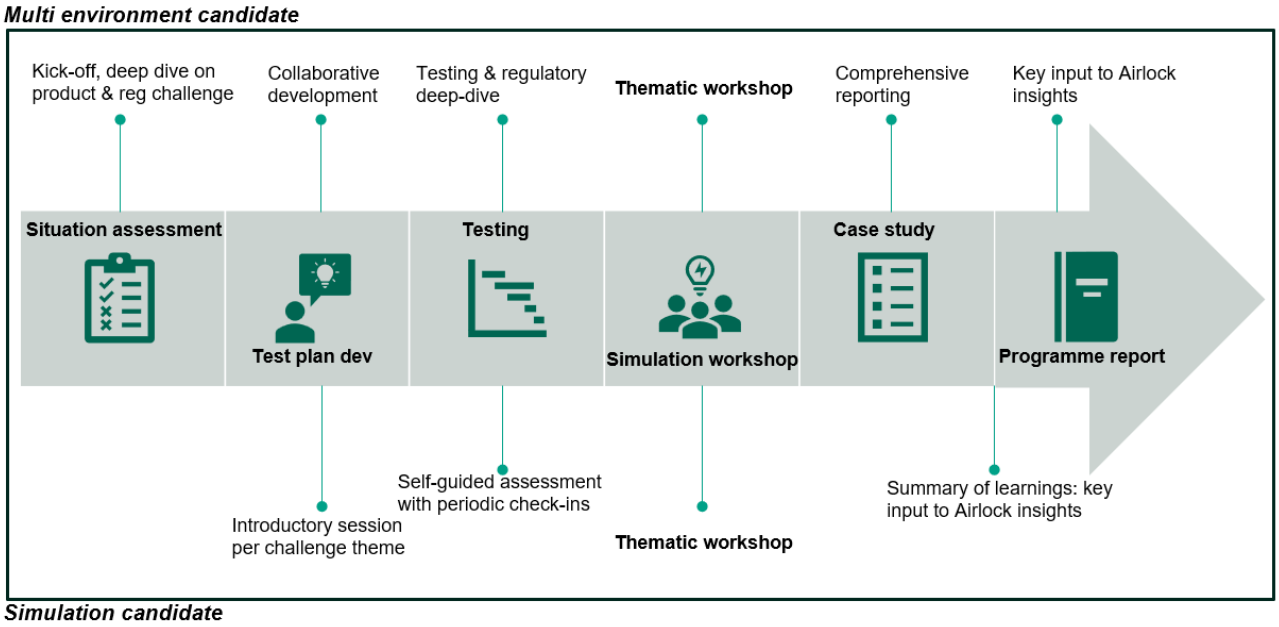


Figure 8 Overview of the workflow for the multi-environment and simulation candidates throughout the testing phase of AI Airlock phase 2.

Situation assessment and test plan development

During onboarding of the candidates, a comprehensive situation assessment was carried out. This involved working collaboratively with the candidate teams to better understand the product and the regulatory challenge to be investigated. Testing plans were co-developed, detailing bespoke testing design, across various testing environments and deployment methods.

Regulatory gap analysis

Regulatory gap analysis is the structured process of identifying areas within the existing regulatory framework where legislation, standards, or guidance are ambiguous, incomplete, outdated, or insufficiently aligned with the characteristics and risks of AIaMD. This process was led by the AI Airlock review team, with support from regulatory specialists, technical leads, and relevant stakeholders, to ensure a coordinated and expert-informed assessment.

Virtual / real world testing

Virtual and real-world testing was conducted across the three multi-environment candidates over approximately 5 months of the programme period. Following the close of testing, candidates spent approximately one month on analysis. The programme team provided support to candidate teams, including ongoing regulatory and technical input at key decision points, supporting candidates in interpreting emerging findings in the context of their regulatory challenge and refining the framing of results ahead of analysis.

Simulation workshop collaboration

The workshops were jointly facilitated by PHG Foundation, a not-for-profit health policy think-tank that works with various stakeholders to understand how new technologies could improve healthcare and health outcomes. Each workshop was designed to cover key focus areas from each candidate's testing plan.

Table 2 Overview of the Simulation workshop's key focus areas

Regulatory Challenge	Key focus areas
Scope of intended purpose	• The scope of intended purpose workshop examined how to distinguish medical from non-medical AI products, using objective criteria such as indicative language and risk-based justification. • Sessions explored real-world signals for detecting early use and function changes, and how added functionality, system integration and deployment context can impact a product's classification, alongside the effectiveness and limitations of controls for managing drift and off-label use.
PMS & PCCPs	• The PMS and PCCPs workshop examined how PCCPs can be defined, structured and operationalised in practice, including the evidence, validation criteria and assurance requirements needed to support them. • Sessions explored the integration of PMS and PCCPs across the AIaMD lifecycle, the challenge of defining "significant change", and how findings could inform the development of future regulatory guidance.
AI as an IVD	• The AI as an IVD workshop examined how pre-analytical variability, including differences in slide preparation, staining, digitisation, scanner hardware and image quality, affects AI-powered IVD performance and what infrastructure, standards and governance mechanisms are needed to address this. • Sessions also explored the limitations of the traditional IVD framework for AI and considered practical regulatory approaches for managing adaptive AI-powered IVDs, including update triggers, drift monitoring and post-market change management, with the aim of informing clearer expectations for the development, evaluation and ongoing management of these devices.

The workshops included a plenary session, for all experts to discuss focused questions in table discussions and interactive scenarios, followed by break out rooms that were each focused on a specific candidate's product and challenge. Delegates represented a broad range of expertise and organisations including MHRA, Team AB, NHS clinicians, patient and public representatives, Information Commissioners Office, Health Research Authority, Care Quality Commission, academia and Centre's of Excellence in Regulatory Science and Innovation, NICE, AI technical experts, and the innovators.

Insights from the workshops were refined and used to inform the candidate case studies. Full reports of the simulation workshops can be found on the AI Airlock website.

Insight generation

Insights were extracted from candidate case study reports, simulation workshop outputs, and the regulatory gap analysis conducted across the cohort. Thematic analysis was applied to identify convergent findings within and across the three regulatory challenge areas, with cross-cutting themes surfaced through structured comparison of evidence generated by both multi-environment and simulation-only candidates. Key findings were validated through the simulation workshops, where expert delegates reviewed and stress-tested emerging conclusions across stakeholder perspectives.

Appendix 2: Programme Design and Governance

Team and Governance Structure:

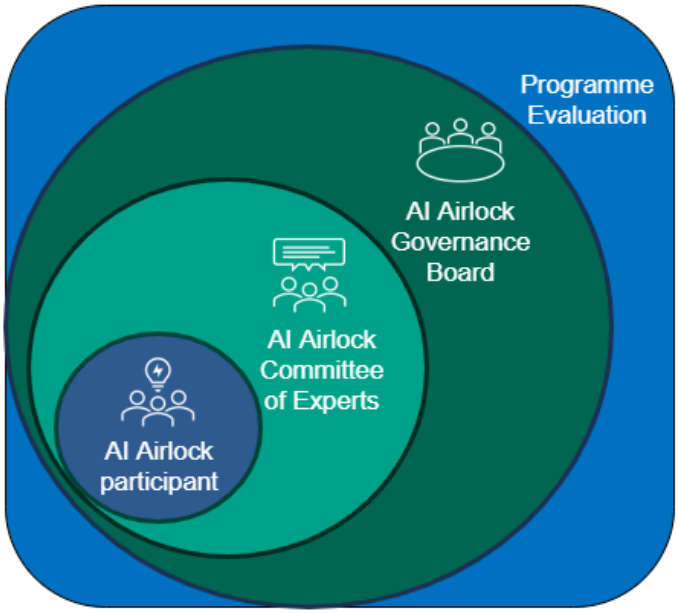


Figure 9 AI Airlock Phase 2 Governance Structure

Phase 2 adopted the same overarching governance structure as the Pilot with delivery centred on a core project group comprising the participating organisation and the AI Airlock team. The AI Airlock team includes both delivery and regulatory technical capability: the delivery team manages programme coordination, planning, governance, stakeholder engagement and delivery tracking, while the regulatory technical team provides specialist input on regulatory questions, testing approaches and the interpretation of evidence generated through the Airlock.

This core group was supported by the Airlock Committee of Experts (ACE), previously referred to as the Airlock Supervisory Committee during the Pilot and rebranded in Phase 2 to more accurately reflect its advisory role and breadth of specialist expertise. ACE brings together representatives from government, other regulators, UK Approved Bodies, clinical practice, academia and technical AI specialists, including expertise relevant to data protection, research governance, care quality, clinical safety and AI assurance. Its role is to provide targeted challenge, specialist input and cross-sector perspectives to support project teams in exploring technical, regulatory and implementation questions.

The core project group and ACE report into the AI Airlock Governance Board, a clinician-led strategic board with representation from DHSC, MHRA, academia, clinical practice and industry. The Board provides programme-level oversight, supports decision-making, and considers escalated risks, issues and complex technical or regulatory questions. In Phase 2, the governance model was also complemented by an overarching programme evaluation activity, delivered by independent partner Woodnewton, to assess how the programme is demonstrating continuous improvement and to capture learning and experience to inform the design of future Airlock rounds.

Timeline and Milestones

Phase 2 formally launched with a webinar on 19 June 2025, which announced the call for applications. The application window opened on 23 June 2025 and ran for three weeks, closing on 14 July 2025. This was followed by a structured sifting and shortlisting process across July and August. The 51 applications received during this period assessed against the published eligibility criteria. Shortlisted innovators were invited to interview and present their proposals.

Following Governance Board ratification of the proposed portfolio and cohort in September, offer letters were issued to the selected candidates. Candidates started onboarding at the beginning of October and the AI Airlock 2 Kick-off Connect event was held on 16 October 2025 to announce the cohort. The event provided an early opportunity for candidate and stakeholder engagement. Ideas for regulatory testing were also workshopped during the event. This was followed by orientation through situation assessment. Participants and the Airlock regulatory technical team subsequently co-created test plans during November and December 2025, with some activity extending into January 2026. Testing commenced on a staggered basis between

December 2025 and January 2026, supported by three in-person simulation workshops held in February 2026. Multi-environment candidates concluded their testing and analysis in between April and May 2026. In parallel, all candidates prepared case study reports to present their analysis, findings and learning from the programme.

In April, the Department of Health and Social Care awarded multi-year funding to expand AI Airlock activities. At the point of drafting this report, the programme is drawing on learning from the Pilot and Phase 2 to inform the design of Phase 3.