Medicines & Healthcare products Regulatory Agency

PHG FOUNDATION
making science work for health

Insights from the AI Airlock Phase 2 Simulation Workshop: Scope of Intended Purpose and Validation

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

This document is not an MHRA policy position but represents an overview of findings from a simulation workshop carried out by the AI Airlock.

Acknowledgements
This session brought together a diverse group of representatives from across industry, clinical practice, regulation, academia, and technology. Participants included experts from the MHRA, NICE, the NHS, and a range of partner organisations. Their insights and collaborative contributions were instrumental in shaping the discussions and recommendations summarised in this report.

This workshop was co-facilitated by the PHG Foundation, who also led on the report drafting.

Contents

Contents 2
Overview 3
Introduction 3
Workshop focus and scope 3
Multi-environment candidate: TORTUS 4
Simulation candidate: Numan (Nu & Aegis AI Conversational & Monitoring System) 4
Key regulatory concepts 4
Intended use or intended purpose 4
Medical purpose 5
Qualification and classification of medical devices 5
Direct diagnosis 5
Key insights from the workshop 6
Qualification of a product as AIaMD 6
Uncertainty around current AIaMD classification rules 7
Managing AI-specific risks 7
Monitoring AIaMDs over its product lifecycle 8
Lessons from real-word case studies 9
Numan Case Study 9
TORTUS Case Study 9
Conclusions 9
Key suggestions for the MHRA 9
Key suggestions for the manufacturers 10

Overview

Introduction

This report outlines key findings from a workshop on *Scope of intended purpose and validation* held on 5th February 2026. This simulation workshop - an interactive, focussed roundtable discussion designed to bring together multiple stakeholder perspectives across a general audience with a broad range of subject expertise, was organised by the Medicines and Healthcare products Regulatory Agency (MHRA) as part of Phase 2 of its AI Airlock programme. The recommendations from this workshop are interwoven in the discussion overview, and a list for implementation is included in the AI Airlock Phase 2 Programme Report.

Workshop focus and scope

Artificial intelligence as a medical device (AIaMD) is regulated as a subset of software as a medical device (SaMD) under the UK Medical Devices Regulations (UK MDR) 2002. A core principle of medical device regulation is intended purpose, which defines the scope and function of a device.

Unlike the stable, relatively narrow intended purpose of traditional devices with fixed designs, AI-based systems introduce new challenges because their potentially broad functionality can evolve through learning from real-world data or iterative updates. Harnessing the dynamic nature of AIaMDs creates regulatory complexities: AI systems may exhibit capabilities beyond their validated scope, increasing the risk of unintended use and potential patient harm. Furthermore, frequent updates, whether driven by manufacturers or identified through post-market surveillance (PMS), complicate the assessment and maintenance of an up-to-date understanding of the device's safety and performance.

Discussion for this workshop was restricted to large language models (LLMs), as all candidate products were LLM-based. The risks associated with these products stem from challenges such as, non-deterministic behaviour, a lack of AI explainability and transparency, and hallucinations, and that the manufacturer's intent does not always determine the LLM's output, especially without appropriate guardrails in place.

The objectives of this simulation workshop on *Scope of intended purpose and validation* were to:

- Discuss clear indicative wording to distinguish medical from non-medical AI
- Assess how intended purpose can be evidenced using objective, risk-based justification
- Identify practical real-world signals for early detection of scope expansion

- Examine how functionality, integration and real-world use drive scope and classification change, and test the effectiveness of controls to manage performance drift and off-label use

The workshop brought together a diverse group of stakeholders including two real-world candidates: TORTUS (a multi-environment candidate), and Numan (a simulation candidate). Stakeholder expertise spanned multiple sectors, including healthcare, academia, regulatory and compliance, policy and ethics, research and innovation, AI specialists, patient representatives, and NHS England.

Multi-environment candidate: TORTUS

TORTUS is an LLM-based ambient voice technology (AVT), currently self-declared as a Class I medical device. It is designed to support healthcare professionals by providing real-time, efficient and accurate clinical record-keeping through the transcription of patient consultations, extraction of clinically relevant information, and structuring of this information into clinical notes and letters, thereby reducing administrative burden and improving quality of care. The central regulatory challenge for TORTUS is defining the specific point at which its functionality transitions from documentation support into diagnostic or clinical decision-support functionality, which would likely trigger reclassification under the UK MDR.

Simulation candidate: Numan (Nu & Aegis AI Conversational & Monitoring System)

Numan is a digital healthcare provider whose LLM-powered AI health coach, Nu, is currently positioned as a wellness tool (non-medical device). Nu is supported by Aegis, an AI conversational and monitoring system. The primary regulatory challenge for Numan is defining the boundary at which the addition of planned functionality (for example, explanations of personalised blood test results) would trigger a reclassification of its product to a medical device.

Key regulatory concepts

Intended use or intended purpose

‘Intended purpose’ is defined in UK MDR; which refers to the manufacturer's official claims regarding the device’s function, target users, clinical context, and how its outputs are intended to influence clinical or patient decision-making. For traditional devices, intended purpose is typically stable across the product lifecycle, unless deliberate, planned and discrete updates are made by manufacturers. For AIaMD, functionality may evolve through software updates, learning from real world data, or expansion into new clinical applications. While some updates may be accommodated within existing change control mechanisms within the product’s quality management system (QMS), more substantive shifts such as expanded diagnostic claims, increased influence on clinical decision making, or increased

autonomy may constitute a change in intended purpose. This may trigger reclassification, renewed conformity assessment, and additional post-market requirements.

Medical purpose

Medical devices are products intended by the manufacturer to be used for a defined medical purpose. Medical purpose includes, but is not limited to, the diagnosis, prevention, monitoring, treatment, mitigation, or prediction of a disease or condition, as well as the investigation, replacement, or modification of the anatomy or a physiological process, in accordance with the Medical Devices Regulations (MDR 2002). The medical purpose of a product is determined by the manufacturer's intended purpose, as defined in its labelling, instructions for use, and promotional material.

Qualification and classification of medical devices

The regulatory process begins by determining whether a product meets the criteria to be classified as a medical device. To qualify as a medical device, a product must be intended for a medical purpose.

Once a product qualifies as a medical device, its intended purpose informs its classification. Classification determines risk class, evidence requirements, and regulatory obligations under medical device law. Devices are assigned to Class I, IIa, IIb, or III through the application of classification rules, which are designed to reflect the level of risk a device poses to patients and users. Medical device classification rules have historically focused on physical harm, which raises questions about how risk should be understood for software and AIaMD, where potential harms may be indirect, cognitive, or systemic. Challenges can arise where AI systems provide explanatory, informational, or supportive outputs that may indirectly influence clinical or patient decision making. These interpretation challenges can create uncertainty around how intended purpose can be effectively framed for AIaMDs and how changes to functionality can be assessed over time.

Direct diagnosis

A device is considered to "allow direct diagnosis" when it provides the diagnosis of a disease or condition independently, provides decisive information for making a diagnosis, or claims are made that it can perform as, or support the function of a clinician in performing diagnostic tasks. For devices intended for use by lay users, the provision of an indicative diagnosis may be sufficient to imply that the device is allowing direct diagnosis.

While the definitions of medical purpose, qualification, classification, reclassification and direct diagnosis are established in regulation and guidance, questions remain about how these concepts are interpreted and applied in practice for AIaMD.

Key insights from the workshop

Qualification of a product as AIaMD

The central challenge discussed by delegates was the regulatory ambiguity surrounding AIaMD qualification, which determines whether a product requires regulatory oversight. While there is guidance available to support the qualification determinations, there was some clarity required to better understand the terminology for AIaMD qualification.

Delegates noted that terms like "lifestyle" and "wellness" are often confusing and highly context dependent. For example, a recommendation to walk 10,000 steps is typically seen as lifestyle advice, but delegates felt that the same advice given to patients on obesity medication could be considered medical advice. Therefore, to support objective qualification decisions, delegates suggested the development of tools (such as flowcharts) and the use of well defined, objective terms in regulatory guidance to reduce ambiguities where possible.

Discussions also revealed specific linguistic indicators in a product's intended purpose statement tended to imply a medical purpose (while remaining subject to the relevant contexts), noting that singular words alone do not sufficiently capture the intended purpose of a product in a clinical context. During the workshop, these included:

- **Verbs** such as alleviate, analyse, assess, diagnose, estimate, evaluate, prescribe, reduce, and screen.
- **Language alluding to a condition** such as abnormality, disease, disorder, syndrome
- **Device outputs** such as those informing users on dosage changes, risk prediction linked to a condition, or treatment changes.

Furthermore, it was highlighted that qualification decisions will also be informed by the intended user and who makes decisions based on a product's output, rather than relying solely on terms such as *clinical* or *medical* in the intended purpose statement.

Delegates agreed that the phrasing of the LLM's output is also important for determining device qualification. The delegates did not perceive language that shifts decision making onto the user (such as "You may have this disease") to suggest a product was a medical advice. However, delegates felt that LLMs that output definitive language about new information (such as "You have this disease") did constitute function as a medical advice and are likely to qualify as AIaMD. In such cases, delegates suggested that a manufacturer's stated intent or simple disclaimers (e.g., 'Not a substitute for a doctor's advice') are insufficient and felt that qualification should ultimately be determined by the product's actual functionality and the risks it poses.

Uncertainty around current AIaMD classification rules

Discussions revealed that delegates perceived the existing device classification rules as complex, difficult to navigate, and outdated for AIaMDs, noting that they were developed for “static” devices. The delegates felt this was a significant barrier to innovation, particularly for new manufacturers who often struggle to navigate regulatory nuances. Delegates called for the use of more definitive and objective terminology in regulatory guidance, citing the International Medical Device Regulators Forum (IMDRF) framework as an example of harmonised language.

Delegates noted a lack of clarity in existing rules regarding what constitutes ‘direct diagnosis’ for LLMs. For example, current rules do not clearly delineate whether medically relevant information provided by an LLM (for example, stating a user is anaemic vs. having low iron) constitutes direct diagnosis. To address this ambiguity, delegates emphasised the need for clear guidance on "direct diagnosis" and thresholds relevant for differentiating between medical advice and diagnostic functions of an LLM. Furthermore, to address wellness products that could impact patient health but did not meet the definition of a device, delegates considered and spoke positively of mechanisms to allow additional oversight and management of non-device products used in healthcare. Products that do not qualify as medical devices are not subject to post-market surveillance obligations under MDR 2002. However, it is understood that these products may evolve over time, including potential addition of medical device functionality. It was proposed that ecosystem-wide mechanisms that could promote visibility and responsibility for all products used in healthcare, ensuring adequate transparency into how the products are meant to be used, would be beneficial. Manufacturer responsibilities would remain to ensure they are assured their product remains in regulatory compliance as its functionality is changed.

Managing AI-specific risks

Delegates also identified several critical AI-specific gaps, emphasising that current rules do not fully account for the unique safety risks these technologies can pose. For example, current regulation does not specifically consider the potential for indirect influence of software devices on clinical workflows, which could lead to outcomes like additional diagnoses (accurate or inaccurate) and therefore additional downstream impacts that may be less apparent than, for example, direct patient harms. Delegates highlighted a need for clearer guidance on evaluating the risks and impact of human–AI co-working and the definition of respective roles and responsibilities to support comprehensive governance that considers the impact of relevant products in practice.

Delegates discussed the risks associated with the evolving nature of AI outputs (e.g., what is the impact of risk based on the content being summarised, such as an error in summarising allergy information versus other potential summarisation errors). It was also noted that a low error rate, as evaluated pre-market, can still pose risk to a large number of patients when deployed at scale. Consequently, and given the black box nature of the LLMs, delegates

emphasised the need for a nuanced, risk-based approach to classification that focuses on the risks posed by the device's outputs on patient health and ensures that sole focus is not placed on analysing the LLM's underlying mechanism of action.

Monitoring AIaMDs over its product lifecycle

Delegates emphasised the need for robust, continuous oversight of AIaMDs once they are in use to ensure patient safety post-deployment. This requires manufacturers to regularly monitor the interplay between the AI models (LLMs), the specific instructions they are given (prompts), and the clinical outcomes. Such monitoring was seen as essential to prevent errors and to detect changes in performance beyond validated functionality ('model drift'). Delegates also highlighted a need for proactive monitoring, involving manufacturers analysing internal performance metrics and external signals (such as customer reviews) to detect and mitigate unintended use. It was suggested that lay users may not have the oversight of the product's validated intended purpose scope and may not be able to recognise when a product drifts or operates outside of its intended purpose. The organised health system and external clinicians were suggested as "smoke detectors".

The risk of off-label use was seen as a significant concern for AIaMDs, whether by users seeking self-diagnosis via unregulated tools or using tools beyond their intended medical purpose. To mitigate these hazards, delegates suggested that manufacturers embed technical safety guardrails against both foreseeable and unforeseeable use, as simple disclaimers alone offer insufficient protection.

However, this approach presented a conflict: if manufacturers implement strong guardrails that restrict user access to device functionality, it could result in decreased user interest. While restricting access may impact the availability of healthcare tools, it improves patient safety by preventing device use beyond its validated functionality. General purpose AI models may not impose such guardrails, allowing patients to access the information they want and potentially reducing reliance on regulated tools. Consequently, delegates strongly recommended that unregulated LLMs should cease providing medical advice to users, thereby ensuring that patients rely on regulated sources for medical information.

Delegates also stressed the need to demonstrate that devices adhere to their intended purpose throughout the product lifecycle. To encourage industry compliance, delegates suggested that the MHRA could develop clear guidelines and tools to empower the sector to conduct its own safety checks. Furthermore, it was recommended that a mechanism that helps to hold non-compliant manufacturers accountable, including a system for reporting devices used outside their intended scope would be helpful. Delegates also suggested that manufacturers could still publish an intended purpose statement as good practice for non-medical products, or a "Disqualification Statement" that could clearly declare the rationale for why a product is not a medical device and how their product guardrails against functionality creep, though these kinds of mechanisms, while potentially valuable to users, would not be enforced by the MHRA because such products are not medical devices.

Lessons from real-word case studies

Numan Case Study

The primary regulatory challenge for Numan was defining the boundary at which its planned feature to provide personalised, evidence-based explanations of blood test results, would cause it to qualify as a medical device, and at which point its AI health coach, Nu, would then be considered an AIaMD.

Delegates felt that a key factor for reclassification would be the use of personalised, contextual patient data. It was discussed that the regulatory boundary may be approached when Nu begins to use or interpret an individual's medical data (such as blood test results) to, or makes explicit claims to, manage, diagnose, or treat a medical condition.

TORTUS Case Study

The main regulatory challenge for TORTUS was defining the exact point where its functionality moves from clinical documentation support to providing diagnostic or clinical decision support, which would trigger reclassification under the UK MDR 2002. This difficulty is heightened by the lack of established regulatory standards for benchmarking AVT products. Delegates noted that TORTUS's deeper integration with electronic health records (EHRs) significantly increases the potential severity of errors requiring appropriate mitigation. Discussions also highlighted the potential for a "reliance bias paradox", where clinicians may increasingly rely on the system's demonstrated accuracy and stop manually checking notes as intended, reducing the utility of that oversight mechanism. This highlighted the changing interactions users may have with devices over time and the need to consider the impacts of human-AI interactions among other variables like changes in population or use environment.

Conclusions

The workshop highlighted regulatory uncertainty among stakeholders regarding AIaMD qualification and classification, and emphasised the need for clear, objective guidance to manage its evolving functionality. To address these areas of uncertainty, the delegates proposed the following recommendations.

Key suggestions for the MHRA

Update classification rules and risk-based frameworks: Delegates suggested that MHRA could adopt a modified, risk-based framework that focuses on the risks posed by the AI's outputs on patient health. To clarify regulatory boundaries, delegates suggested that new guidance could benefit from using objective terminology and establish specific thresholds for LLMs to resolve the current ambiguity between "direct diagnosis" and "indirect diagnosis".

To monitor wellness products that still impact health, delegates proposed a lightweight mechanism, which was envisioned as an ecosystem wide way for a regulatory body to promote visibility of all products used in healthcare, ensuring adequate transparency into how the products are meant to be used, including of non-device products, but not imposing a full QMS. Furthermore, it was suggested that the Class I device registration process should be re-evaluated to reduce the risk of self-certification being misinterpreted as a mark of quality for commercial gain. It was also suggested that a process could be introduced for manufacturers to publish “Disqualification Statements” in an effort to clearly evidence why their product does not meet the definition of a medical device and could include how they monitor and prevent function creep.

Enhance AIaMD monitoring: To support accountability, delegates suggested that additional mechanisms for reporting devices used outside their intended purpose (potentially similar to vulnerability reporting) would be helpful. To incentivise industry compliance, delegates suggested that the MHRA could publicise the benefits and successful case studies of developers who proactively engage with the Agency.

Improve accessibility of regulatory guidance: Current guidance materials could be made easier to understand and more accessible for innovators without a medical background, using conversational front ends, videos, and webinars. Delegates suggested that the MHRA launch an education campaign to clarify that device classification signifies registration status and the level of regulatory oversight, and it is not a “quality stamp” nor a guarantee of device safety.

Key suggestions for the manufacturers

Proactively monitor AIaMDs: Delegates suggested that manufacturers should continuously monitor internal performance metrics and external signals (such as customer reviews) to ensure devices adhere to their intended purpose throughout the product lifecycle and to detect off-label use. In addition to Instructions for Use (IFU) requirements, delegates suggested that manufacturers should ensure that user-facing information and warnings are kept up to date and reflect emerging risks identified through real-world use.

Mitigate off-label use: Delegates suggested that manufacturers embed strong, technical safety guardrails to address both foreseeable and unforeseeable risks to prevent unauthorised or off-label use of their products that would extend their scope into unvalidated functionality, as simple disclaimers are insufficient to guarantee patient safety. Examples of such guardrails may include restricting responses to predefined use cases, disabling diagnostic outputs for non-clinical tools, or prompting users to seek professional advice when high-risk inputs are detected.

For non-medical device products operating close to the regulatory boundary, delegates also suggested that manufacturers could consider publishing a "Disqualification Statement" that clearly articulates evidence and rationale for why their product is not a medical device.

Manage user perception: Delegates suggested that manufacturers communicate information on key model updates for LLMs, as well as any system prompts that may affect user understanding of the product's capabilities or limitations. This would support the ongoing alignment between product performance and its stated intended purpose.

Delegates also emphasised that manufacturers should carefully consider product branding and naming (for example, avoiding terms such as "Clinical Companion" for non-clinical tools), as this can imply a medical purpose that may not be substantiated. Ensuring that product claims, including implied claims through language and branding, are accurate and appropriately evidenced is critical to preventing user misunderstanding and unintended use.

© Crown copyright 2026

Open Government Licence

Produced by the Medicines and Healthcare products Regulatory Agency. www.gov.uk/mhra

You may re-use this information (excluding logos) free of charge in any format or medium, under the terms of the Open Government Licence. To view this licence, visit http://www.nationalarchives.gov.uk/doc/open-government-licence or email: psi@nationalarchives.gsi.gov.uk.

Where we have identified any third-party copyright material you will need to obtain permission from the copyright holders concerned.

The names, images and logos identifying the Medicines and Healthcare products Regulatory Agency are proprietary marks. All the agency's logos are registered trademarks and cannot be used without the agency's explicit permission.