Medicines & Healthcare products Regulatory Agency

PHG FOUNDATION
making science work for health

Insights from the AI Airlock Simulation Workshops: Post-market surveillance and continuous monitoring of AIaMD

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

This is not a formal MHRA policy position, but 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 outlined in this report.

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

Overview

Introduction

This is a report on a workshop that we organised on 4th March 2025, as part of the AI Airlock pilot programme. The AI Airlock operates as a sandbox through which real world products can be explored collaboratively. Developing a shared understanding of regulatory challenges posed by AI technologies that classify as medical devices, also known as AI as a medical device (AIaMD), and possible solutions to these.

Manufacturers of medical devices, including AIaMD, are legally required to carry out post-market surveillance (PMS) of their products. This is necessary to monitor for safety concerns, maintain an up-to-date assessment of the benefits versus the risks of the device, and meet established standards for safety and performance. The PMS requirements manufacturers must meet are outlined in legislation including the Medical Devices (Post-market Surveillance Requirements) which amends the Medical Device Regulation, 2002. Although this legislation came into force after the closure of the AI Airlock pilot, the workshop was based on the incoming requirements that are now in effect. These regulations are enacted by the MHRA via a vigilance reporting system, which includes the Manufacturer's Online Reporting Environment (MORE). Additionally, the Yellow Card Scheme is available for medical device safety reporting more orientated towards reports from end users and patients.

AIaMD can undergo performance changes after being deployed. This may be driven by changes to the device's algorithm ('model drift'), to the characteristics of the input data ('data drift'), or in the rate of agreement of human users and the device (which might suggest automation bias), among others. PMS is both more important and more challenging in AIaMD products which can evolve whilst in use. Undetected changes pose a risk to product safety, and as a result patient safety, making continuous monitoring a potentially important safeguard.

As part of the AI Airlock, the MHRA investigated how a monitoring system might improve the risk management of AIaMD by identifying performance and trends which could lead to safety issues. This workshop brought together a diverse group of stakeholders with clinical, AI technology, policy, regulatory, and legal expertise to better understand and align on the roles and responsibilities involved in monitoring and reporting on incidents involving AIaMD. The recommendations from this workshop are interwoven in the discussion overview, and a list for implementation is included in the AI Airlock Pilot Programme Report. The National AI Commission represents a key opportunity to establish a credible, international regulatory framework for the safe and effective use of AI in healthcare. The insights generated through the AI Airlock will directly inform and support the work of the Commission, ensuring that its outputs are grounded in real-world evidence and regulatory experience.

AI Airlock Case Study: Newton's Tree's Federated AI Monitoring Service (FAMOS)

FAMOS is a real-time dashboard designed to monitor the performance of AI products in healthcare settings. It provides three key metrics:

- **AI Inference Values**[1] – Tracks model outputs over time to identify trends or potential model drift. Data can be filtered by medical condition to distinguish between model-related issues and public health changes.
- **AI/Human Agreement**[2] – Measures how often healthcare professionals agree with AI outputs, helping to detect signs of automation bias during clinical use.
- **Data Input Quality** – Clusters input data to highlight anomalies that may affect AI performance.

The dashboard is available in two versions: a local version for use within healthcare sites, and a version for manufacturers and approved parties to review anonymised, multi-site data. Newton's Tree has an overview of performance data across all products and sites via the external dashboard. At the time of reporting, FAMOS is being trialled in two NHS hospitals with support from academic partners, alongside its inclusion in the AI Airlock programme.

Key insights from the workshop

Existing approaches to PMS of AIaMD

PMS activities currently undertaken by AIaMD manufacturers tend to be limited and reactive to incidents. Examples given of more proactive approaches include monitoring the success/failure rates of device executions, sometimes with consideration to variation between sub-populations of patients. Trend analysis may also be undertaken proactively, for example to monitor for changes in data. Expert panels and Radiology Events and Learning Meetings (REALMs) are also convened in clinical settings to review individual cases. REALMs primarily serve educational purposes but may inform incident reporting mechanisms.

[1] AI inference value refers to the behaviour of the algorithm, representing its routine behaviour in processing data. However, variations in this value may stem either from the algorithm itself, its environment, or from changes in the underlying population, and distinguishing between these is not always straightforward.

[2] AI human concordance is measured by comparing between AI predictions and human outputs, to assess whether there was agreement. This level of agreement is used as a proxy measure for automation bias, highlighting where human decisions may be influenced by AI outputs. The approach does not focus on determining whether results are right or wrong, but rather on identifying patterns of behaviour that may indicate hazardous situations or performance anomalies.

There was some discussion about a potential hesitancy among manufacturers with regards to engaging with MORE, the primary vigilance reporting system. It was suggested that this may be due to a perception that incident reporting could constitute an admission of wrongdoing or cause reputational harm. It was emphasised that – if this is the case – this mindset should be changed so that feedback to the MHRA is seen as a crucial component of a collaborative environment that prioritises patient safety while also supporting developers to meet their regulatory requirements. It was also suggested there may be low levels of awareness among healthcare professionals that they can report safety concerns relating to AIaMD via the Yellow Card Scheme.

As we had been receiving low volumes of incident reporting involving AIaMD, it was suggested this indicates that AIaMD currently deployed in clinical settings either meet the required standards, or there is a lack of awareness of what and how to report, and / or that effective monitoring is a challenge. Anecdotal evidence was shared of errors involving AIaMD in the NHS, but it was suggested there is a lack of wider awareness. An independent legal perspective, separate from the our legal team, suggested that manufacturers of AIaMD could be held liable for harm involving their products. It was added that this potential for liability may, in turn, encourage more proactive monitoring.

Responses to continuous monitoring

Participants evaluated the benefits and challenges of continuous monitoring as part of PMS for AIaMD, focusing on the three key metrics presented within the FAMOS dashboard: AI/human agreement, AI inference values, and data input quality. These metrics align well with existing risk mitigation strategies, as per guidance on the application of the risk management standards (ISO 14971) to machine learning in artificial intelligence (BS AAMI 34971). The simplicity of a dashboard, and visual elements along with clinically focused alerts were favourably received.

However, possible drawbacks include potential blind spots in the form of gradual but significant changes that take place ‘under the bonnet’ of the metrics presented in the case study example. For example, changes that affect specific sub-populations of patients may go unnoticed because demographic data – often incomplete at clinical sites – is not integrated. Likewise, participants highlighted the value of monitoring different types of AIaMD using a dashboard. For example, devices that do not require an expert-in-the-loop, such as chatbots or wearable devices, require predictive monitoring systems to enable interventions to happen before harm occurs.

Data and cybersecurity risks were discussed; if a dashboard only uses performance data without exporting personal information and acts as an intermediary tool, risks to patients are limited and comparable to other connected clinical devices. The possibility that such

dashboards could become critical infrastructure was noted as an important future consideration.

Setting thresholds for preventative action

The discussion on setting thresholds for preventative action by the manufacturer revealed a mix of social and technical considerations. It was clear that appropriate thresholds must consider both the technology and how users interact with it, varying by specific AIaMD products and their local contexts. Understanding how each metric is defined and assessed is also crucial. Questions arose about the difference between product drift and algorithmic drift, and whether evaluation should focus on device performance or algorithm analysis, noting these terms lack clear definitions in literature.

Benchmarking was proposed as one approach: establishing a standard, monitoring deviations, and triggering investigations for significant changes — for example, a 5% deviation could be defined as a threshold for significant change and warrant investigation by the manufacturer. It was noted that thresholds do not require perfect calibration if they effectively alert experts. Lower thresholds increase the likelihood of interventions. As such, some participants raised the prospect that thresholds which are set too low may be particularly problematic in the case of AIaMD as there is evidence that frequently updating the AI model may itself exacerbate algorithmic drift. Consequently, continuous updating of AI models may not be beneficial except perhaps for highly personalised devices. In addition, this highlights the importance of a risk-proportionate approach based on the AIaMD being monitored.

A distinction between types of thresholds was suggested: lower thresholds could be set to trigger enhanced monitoring with higher thresholds reserved for intervention. This risk-proportionate approach would enable changes to be detected more frequently for monitoring purposes, without being overly interventionist. Effective thresholds also need to reflect a real capacity to intervene and therefore must be set below the point at which harm cannot be mitigated. Here, the role of human factors in setting thresholds was emphasised. It was proposed that a clinical risk management approach could be adopted, with thresholds set locally at the site of deployment and aligned with existing policies. This would be consistent with guidelines for determining ‘acceptable risk’ outlined in ISO 14971, which advise against universal thresholds due to device variety and diverse legal and cultural considerations.

Follow-up actions and incident investigation

Participants’ views on setting thresholds were closely linked to the related question of appropriate responses when thresholds are breached. The dashboard’s purpose — to identify hazardous situations and predict the risk of a dangerous outcome — was emphasised, clarifying that it cannot determine if an error has occurred. Therefore, follow-up

actions, including further investigation and access to additional data, are crucial to assess the implications of alerts.

The discussion considered which alerts should be classified as notifiable incidents. It was suggested this should depend on harm severity and/or frequency. Manufacturers may hold the view that something is reportable when it is consistently wrong. Healthcare professionals may be advised to report any concerns, for example via the Yellow Card Scheme. It was highlighted that the new PMS regulations (44ZC “incident” (e)) stipulate that erroneous results for all diagnostic devices are considered reportable incidents. Concerns were raised that reporting all erroneous results and clinician concerns for AIaMD could overwhelm the system. Difficulties in managing the scale of reports should not be a consideration in meeting the legal requirements of reporting. Indeed, the risk of overwhelm should be considered and managed by the vigilance system accordingly. The MHRA offers the ability to craft more bespoke reporting activities for manufactures with large reporting loads that are willing to engage. On the user side, one approach suggested educating clinicians about AIaMD limitations so they can judge which deviations are serious enough to report.

The group also considered whether some threshold breaches might be better handled outside MHRA vigilance reporting. Since the dashboard’s metrics consider deployment context, certain alerts (for example, relating to human-AI agreement) could be managed locally by clinical governance or management teams. Issues related to standard of care might more appropriately be referred to the Care Quality Commission.

Roles and responsibilities in PMS of AIaMD

The workshop explored existing and possible roles and responsibilities for key actors in PMS of AIaMD, focusing on manufacturers, deploying sites, the MHRA, and other organisations. This underscored the necessity of improved collaboration among all parties. Best practice PMS is recognised as a shared responsibility, with each stakeholder playing a distinct yet complementary role in ensuring the safety and effectiveness of AIaMD.

Manufacturer

Manufacturers hold ultimate responsibility for PMS but face challenges managing AIaMD risks that often emerge during deployment, particularly in hospitals. Some delegates suggested that manufacturers may not be able to fully address these risks prior to deployment and advocated for a clinical risk management approach at the deployment site, requiring stronger collaboration with hospitals. Predetermined Change Control Plans (PCCPs) were noted as a regulatory tool enabling manufacturers to pre-assess which changes require MHRA reporting.

Deploying site

Participants reported relatively low levels of engagement with manufacturers post-deployment. It was emphasised that collaboration requires greater investment and capacity-building at the deployment site. Interest was expressed in a new standard that could formalise contracts in which manufacturers engage hospitals to evaluate product performance on-site, thus fostering active collaboration in PMS of AIaMD.

MHRA

The MHRA does not instruct how manufacturers monitor specific metrics such as algorithmic drift but does expect manufacturers to give appropriate consideration to risks which are relevant to the device's intended purpose and address these within risk management plans. Key priorities for the MHRA include enforcing legislative requirements, promoting awareness of existing guidance and maintaining broad awareness of data collection methods and PMS approaches employed by manufacturers.

Other organisations

NHS England sets and oversees standards for clinical risk management related to the deployment and use of health IT systems (e.g. DCB0129 and DCB0160). Compliance is mandated by law under section 250 of the Health and Social Care Act 2012 (as amended) but there are currently no specified enforcement powers, which may limit their impact. As a result, the role of other organisations in post-market surveillance remains crucial but is potentially constrained by limited legal sanction.

Conclusion

Key challenges for PMS of AIaMD

Making the first GMLP principle a reality: Effective PMS requires multidisciplinary collaboration between manufacturers' technical experts and clinical staff at deployment sites. Assessing the impact of drift, or factors in a device's external environment, throughout the total product lifecycle is likely to require insight from both the manufacturer and clinical staff. There is support in principle for this collaboration but the extent to which it happens in practice appears to be limited. Both manufacturers and deploying sites express frustration about a lack of engagement by the corresponding party.

Automating actions amid unclear roles: The dashboard can automate risk management processes that ought to be established already in clinical settings, but this may be complicated by uncertainties around existing roles and responsibilities. The workshop started to explore these issues, but further engagement is needed.

Distinguishing signal from noise: Increasing volumes of data could enhance PMS, but this will require analytical skills and competencies to be adequately resourced in the right places, and thresholds which are carefully calibrated to alert expert attention when required. Without this, it is possible that the existing vigilance reporting system could be overwhelmed with data that fails to lead to clear insights or follow-on actions.

Responses to threshold breaches requires clear protocols for follow-up actions and reporting: Monitoring, on its own, cannot determine the safety or performance of AIaMD; therefore, follow-up investigation is essential. This requires clear protocols to ensure preventative action is taken when necessary and reporting informs PMS in line with regulatory requirements.

Defining notifiable incidents involving AIaMD: There is ambiguity about what constitutes a reportable incident. Manufacturers must report erroneous diagnostic results, whilst healthcare professionals are encouraged to report any concerns via the Yellow Card Scheme. The potential volume of reporting, given expected disagreements between clinicians and AI outputs, is a concern for some.

The MHRA and its partners will continue to explore how regulation can support safe, effective use of AI medical devices in healthcare. This includes improving guidance, working with developers and clinicians, and learning from tools being tested in the field. The workshop generated practical, evidence-based insights and recommendations, captured in the AI Airlock Pilot Programme Report. These will inform the future work of the National AI Commission and shape MHRA's regulatory strategy in Great Britain, while also contributing to international regulatory discussions.