Medicines & Healthcare products Regulatory Agency

PHG FOUNDATION
making science work for health

Insights from the AI Airlock Phase 2 Simulation Workshop: Artificial intelligence (AI) in vitro diagnostics (IVDs)

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

This 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 patient representatives. Participants included experts from the MHRA, approved bodies, NEQAS, NPL, NPIC, 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 co-facilitated by the PHG Foundation, who also led on the report drafting.*

Contents

Overview

Introduction

This report outlines key findings from a workshop on performance evaluation of *Artificial intelligence (AI) powered in vitro diagnostics (IVDs)* held on 20th 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 Medical Devices Regulations (MDR) 2002. According to these regulations, medical devices are required to demonstrate sufficient evidence that they function as intended for their specified use. Performance evaluation is how manufacturers demonstrate that a device works in practice — meaning it produces accurate results, performs reliably, and is clinically relevant for the condition it is intended to assess. Performance evaluations are based on four main characteristics of reporting: i) analytical performance (such as accuracy, precision, sensitivity, and specificity); ii) clinical performance (how well the device works in a real clinical setting); iii) scientific validity (the association between the test and the clinical condition or physiological state); iv) State of the Art (SOTA) (the current best level of practice, knowledge, or performance in a field).

The central regulatory challenge for AI-powered IVDs lies in how performance evaluation should be defined, generated and maintained across the product lifecycle. While existing performance expectations and validation metrics are well established for traditional, largely static IVDs, they do not fully address the potential dynamic and adaptive characteristics of AI based systems. The objectives of this simulation workshop on performance evaluation of *Artificial intelligence (AI) powered in vitro diagnostics (IVDs)* were to:

- Explore challenges related to pre-analytical variability such as differences in slide preparation and digitisation and how these impact AI-IVD performance.
- Identify what infrastructure, standards, and governance mechanisms may be needed to support robust evaluation and patient safety.
- Define practical and proportionate regulatory approaches for managing adaptive AI- IVDs post-deployment through post-market surveillance, including update triggers, retraining, drift monitoring, and change management.

- Clarify expectations for the development, evaluation, and ongoing management of AI-IVDs.

The workshop brought together a diverse group of stakeholders including two real-world candidates: Panakeia (a multi-environment candidate), and Octopath (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: Panakeia PANProfiler

PANProfiler Colorectal (PPC) Microsatellite Instability/Mismatch Repair (MSI/MMR) is an AI-IVD that determines status of biomarkers MSI/MMR directly from routine brightfield images of confirmed colorectal cancer cells and tissues, entirely digitally. The device assists clinicians in understanding the progression of the disease and identifying appropriate treatment pathways.

The focus areas for PANProfiler included: defining appropriate performance metrics; identifying significant updates and changes to IVDs; and establishing a suitable reference dataset and framework for explainability of AI-IVDs.

Simulation candidate: Octopath

Octopath is an AI-powered platform that automatically detects, labels, and provides quantitative analysis of cellular features in digitised histopathology slides from confirmed cancer patients (malignant cancers). The product is currently for research use only and is being developed into a General IVD under MDR 2002.

The main questions for Octopath focused on: assessing the state-of-the-art in AI-IVDs; how to integrate real-world evidence (RWE) to address performance instability or drift; and expanding the scope of performance evaluation.

Key regulatory concepts

Software as an IVD medical device

IVD medical devices are defined as either reagents, kits or equipment which are intended to examine specimens from the human body such as blood or tissue that can provide information on the patient's physiological or pathological state, physical or mental health, predisposition to disease, safety of samples, treatment response and monitoring. When software is intended to analyse human specimens to provide information on a physiological or pathological state, it is considered an in vitro diagnostic medical device.

Pre-analytical variability and external quality assurance

The quality of digital inputs underpins reliable AI performance, and AI must be able to cope with multiple sources of variability throughout the analytical pathway. Therefore, understanding variation, what is an acceptable amount of variation, and the impact it has on AI use in healthcare and performance evaluation is important. External quality assessment schemes currently exist which provide essential factors in the understanding of how to independently monitor the quality of tests and their reporting, with the goal of improving patient care.

The UK National External Quality Assessment Scheme (NEQAS) digital pathology external assessment scheme is working towards consistent quality assurance which is vital as UK laboratories move towards digital 'glassless' operation and increased use of AI-assisted analysis. The scheme has surfaced a range of issues that can impact quality and consistency, including whether digital images accurately reflect what is on slides, issues with slide preparation that impact analysis, and technical factors that impact the function of imaging systems.

The National Pathology Imaging Cooperative (NPIC) is working on a national digital pathology system to support clinical implementation within the National Health Services (NHS). The ENSURE-AI project, which is evaluating multiple AI products across different sites and scanner types, is demonstrating the need for reference datasets, multi-site, multi-scanner validation and independent academic oversight. The NPIC Quality Coordination Centre (QCC) is working to identify knowledge gaps and gather real-world evidence on the landscape of variability within digital pathology. They are developing a range of tools to improve quality and consistency at all stages of digital pathology.

The National Physical Laboratory (NPL) Centre for AI Measurement is carrying out research to develop technical AI standards and establishing benchmarks and developing metrics to evaluate these technologies. One of the challenges they are investigating is around standardisation and measurement. Given the complexity of health data, there is a question around which data should be incorporated into reference datasets. In silico trials, and research using synthetic data, are methods being explored to help gather standardised information. Another question persists around units of measurement, and how differences can cause challenges in terms of transferability of technology and information between different systems.

Key insights from the workshop

What do we mean by inconsistency and what is its impact?

Inconsistency is present in many factors that impact the development and performance of AI in healthcare. These include not only the performance of the AI models themselves, but also

how they are used in real world healthcare settings and how their results are interpreted and communicated. Variability in data quality and completeness was the most common inconsistency perceived by delegates to impact the performance of AI in healthcare. Factors that impact the data quality were also highlighted as issues and examples included differences between laboratories' protocols, patient demographics, sample collection, and analysis processes.

The impacts of inconsistency discussed by delegates were wide-ranging. In terms of model performance, inconsistency creates challenges for evaluation and generalisability and can reduce health system confidence in these tools, for example if they are limited in their ability to solve clinical problems. An important consideration for users is around the associated concept of reproducibility, which delegates noted required clarification in the AI context. For AI-IVDs, reproducibility does not necessarily mean generating identical numerical outputs on every run but rather producing the same clinical interpretation or classification under equivalent conditions. Production of non-identical outputs does not necessarily mean that a model is performing inconsistently, if reproducibility can be demonstrated.

Patient representatives in attendance highlighted that inconsistency can result in poor clinical decision making, introduce errors, and undermine patient confidence and trust. These impacts can also occur through misplaced confidence in underperforming models. The discussion also highlighted that wider system impacts could lead to a loss of confidence among users in AI technologies as a healthcare tool, leading to wider dissatisfaction with these tools among healthcare professionals, and loss of investment in technology development.

What can be done about inconsistency?

Measuring and understanding inconsistency are key to developing mechanisms to manage it. Delegates challenged that consistency is not the only feature that should be considered as equivalence is also necessary to compare AI model performance between clinical sites and accuracy to the ground truth. Delegates highlighted that this is where standardisation and accreditation of laboratory processes is valuable.

The discussion emphasised that good quality AI is underpinned by good quality data, therefore foundational measures to help manage inconsistency focus on data and data infrastructure. Benchmarking datasets, containing specific and restricted sets of data, were highlighted by delegates as being needed to test AI model's safety, efficacy and validity. Existing external quality assurance organisations such as NPL or NEQAS could play a role in their development and curation.

Delegates also noted that reference datasets, containing well curated, independent and representative patient data could help developers to test AI models on real world data and support consistency between AI models. However, delegates additionally considered that

continued testing against a reference dataset might fail to detect changes over time for evolving models, and that reliance on manufacturer-controlled testing alone could raise concerns of integrity and objectivity. Independent third-party organisations were proposed by the delegates as ideally positioned to maintain reference datasets so they remained impartial, secure, comprehensive and high quality.

In parallel with reference datasets, the delegates proposed open and national shared approaches to developing data standards, which would need to consider the inherent variability of AI-IVDs, compared to more static traditional IVDs. It was highlighted that healthcare professionals will also need guidance on how to identify and document inconsistencies in the models they are using in clinical practice.

What are the real -world challenges and limitations of quality standardisation?

Delegates noted that the real-world application of quality standards into already stretched services could significantly slow implementation of AI models and strategies to manage inconsistency in healthcare if those standards are not relevant to the situation on the ground. Conversely, delegates shared that if the implementation of change is too quick it can have a negative impact on staff and workflows. However, delegates noted that a balance needs to be struck between the degree of standardisation in processes and sufficient flexibility to allow for local differences in practice. Delegates suggested that standards should outline the acceptable boundaries for model use and performance.

While adherence to good practice should be encouraged, delegates recommended that a mixture of compulsory and voluntary measures, grounded in tangible benefits and enforceable requirements may be needed. These measures could include mandates for adoption and engagement, support for clinical staff including at the local level, and organisational AI champions who engage with staff about AI and provide a point of contact for queries.

How can we communicate inconsistency and promote confidence?

Once AI models are in use in clinical practice, there are varying levels of explainability and communication that will be needed on the outputs from the tools. Delegates noted that effective explainability of a product was particularly needed to enable building confidence in test results, to communicate uncertainty and risk with the use of the AI models and the meaning behind the results to the appropriate audiences, including patients who may be under high stress awaiting a healthcare outcome or experiencing other relevant factors.

It was also highlighted that it is important for patients to be asked what they want to know, and to acknowledge the needs of different individuals will vary. More graphical forms of communication, including lived experiences, were indicated as more suitable by the patient

representatives in the workshop. Healthcare professional confidence in AI was seen as being reassuring for patients. For healthcare professionals delegates note that it is essential to ensure they have enough information to question AI outputs, recognise uncertainty, and override recommendations when needed, especially because diagnostic liability ultimately sits with the clinician. Data including health economics, statistics, outcomes and model performance were highlighted as more effective. Model cards, which include key information on an AI model, could support decision making. Broader communication principles that were discussed included ensuring consistency in terminology and whether AI and machine learning should be protected terms in a clinical context to ensure that they are used appropriately.

Lessons from real-word case studies

Metrics

During the discussions with delegates on the PANProfiler case study, there was broad agreement from delegates that performance metrics should be objectively and quantitatively measurable, and essential clinical diagnostic standards, such as sensitivity and specificity metrics, should continue to be met. Other metrics, such as those which reflect agreement or concordance, could be added where appropriate and justified. Delegates felt that providing clinicians with an overwhelming amount of all metrics to gauge AI accuracy and safety may be unhelpful, and the clinician representative's preference was to get accurate results from metrics they trust as essential for making diagnosis or referrals. However, in discussion there was some tension surrounding whether manufacturers or regulators are responsible for defining and selecting the most appropriate metrics for AI-IVD performance evaluation, and the delegates concluded that clearer regulatory and professional guidance is needed on the selection and interpretation of metrics.

When discussing the Octopath case study delegates considered super-human AI performance, use of evidence to record performance instability, and approaches to performance evaluation. It has been shown that AI models can demonstrate higher sensitivity and accuracy than clinicians, therefore for these models, using human concordance as a primary performance benchmark has limitations but was deemed acceptable by stakeholders provided that the benchmark is justified. To account for super-human performance, where no clear gold standard exists (beyond identifying the biological ground truth), transparency in benchmarking, supported by scientific evidence, clinical trial or survival and follow up data, was considered critical. It was noted, however, that evidentiary feedback loops can be slow, meaning that when AI and clinician judgement diverge, resolution may take time unless performance can be assessed against an alternative, possibly with analytical datasets. Furthermore, while the pathologist acts as a 'human in the loop', this raises concerns about confirmation bias and over-reliance on AI results, suggesting a need for monitoring metrics (for example, frequency of agreement, or attention maps) to mitigate these risks.

Significant change

While model performance instability and drift should be captured through post market surveillance, the attendees also identified a need for clear metrics and thresholds for acceptable performance and a consensus on what constitutes significant change. To maintain performance of AI models through time, increasing the diversity of training data would help to future proof systems.

The discussion highlighted that delegates felt changes in intended purpose would constitute significant change, however that a balance between breadth and specificity is needed in intended purpose definitions. Early engagement with regulators was recommended to explore how these definitions could be approached. A general principle agreed during the discussion was that a change would likely be considered significant if it alters clinical outcomes. This type of change could change the qualification and/or risk classification of the device. However, decision tools or matrices were highlighted as being needed to support manufacturers in determining when a change to an AI model is considered significant and requires resubmission to the regulator. Predetermined change control plans (PCCPs) were discussed as a mechanism that could be used to predefine anticipated changes and thresholds to support iterative updates to models.

Explainability and “interpretability” for patients

The attendees stated that both quantitative and qualitative information is needed for explainability, and this would include global information, for example, that explains overall model logic, and local information, covering individual clinical predictions.

The patient representative perspective was that transparency and clear communication about when AI is used, is essential to maintain patient confidence and trust. However, during the workshop itself there was divergent views between delegates around the level of information they would prefer to receive about the use of AI in their care pathways. While there is further work to be done around what patient information is most helpful to be provided, the discussion concluded that each patient would have individual needs and clinicians, manufacturers and regulators need to be able to provide a proportionate amount of information to meet those needs and to manage informed consent in the context of AI.

A final discussion point offered by a patient representative, who highlighted that waiting for results is one of the most stressful aspects of the cancer pathway. Therefore, it was highlighted that there is a careful balance to be struck between speed to test results and accuracy.

Adoption

Delegates noted that if an AI identifies more positives, it may increase downstream costs (e.g. more biopsies), which could hinder adoption even if detection is correct. Further, seeing AI

outputs first can bias pathologists (automation bias / over-reliance), although clinicians still retain responsibility for the final diagnosis. The delegates noted that this raises a tension: if liability for accuracy sits with clinicians, they may be less willing to trust and use the tool without clear interpretability and explainability from the models.

In discussion, concerns about whether clinicians could lose interpretive skills over time were raised, but analogies (e.g., the stethoscope) suggest that skill shift is not inherently a marker of a bad tool and should not automatically limit adoption; instead, insights could be derived from usage data and supported through awareness and training.

Conclusions

The workshop discussions identified key challenges around information and data – how inconsistency and performance is measured and managed, and what resources are needed to support this. The need for guidance, standards and regulatory advice, particularly around performance and significant change, was highlighted. The attendees developed and proposed the following recommendations for further action.

Recommendations from delegate discussions

Guidance on interpreting regulation in accessible language

The delegates discussed that MHRA should provide clearer and more detailed and consistent guidance, with definitions of key regulatory principles. Areas highlighted for guidance and regulatory clarity include:

- MHRA input into, or collaboration in the development of, standards and best practise, including around how data are collected, and the features of reference datasets
- Whether the MHRA mandates the use of these developed standards
- Guidance for documenting differences in model performance
- Terminology, and whether certain terms should be standardised
- Objectively and quantitatively measurable metrics for determining performance, and performance thresholds, in a regulatory context

As in other workshops, there was also a clear call for guidance on identifying what may be considered significant change, including the types of changes that may not affect intended purpose directly but could still be considered significant and therefore requiring reassessment or management under a PCCP as applicable.

Delegates felt that consolidated guidance would support manufacturers in interpreting regulatory requirements with greater confidence and direction. They recommended that this guidance should include case-study based examples to illustrate how regulatory principles should be applied in practice, particularly for complex scenarios such as AI model updates and performance changes.

In developing further guidance, it was recommended that the MHRA should consider the feasibility and proportionality of proposed standards, particularly taking into account the current capacity of approved bodies and the volume of foreseeable change requests.

Decision making tools to easily identify when resubmissions are required for significant changes

Delegates suggested that MHRA could develop practical tools to support decision making, such as decision trees, matrices, flowcharts and case study-led worked examples, to help manufacturers determine whether changes should be classified as significant and what regulatory pathways should be followed.

Standard templates for regulatory submissions

Delegates recommended that Approved Bodies consider introducing structured templates to support manufacturer conformity assessment, enabling more consistent documentation and providing assurance that appropriate processes have been followed.

© Crown copyright 2024
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.